Table of Contents

When viewing this document in a pdf editor, click on the page number to go directly to the page.

TABLE OF CONTENTS ............................................................................................................................. 2
INTRODUCTION ...................................................................................................................................... 29
LICENSE ................................................................................................................................................ 30
SYSTEM CONFIGURATION .................................................................................................................. 33
INSTALLATION ....................................................................................................................................... 34
ADVANCED INSTALLATION ................................................................................................................... 35
  SILENT INSTALLATION BY INSTALLSHIELD SCRIPT (WINDOWS ONLY) ............................................. 35
  LANGUAGE SELECTION .......................................................................................................................... 44
  SELECTION OF THE USER FOLDER ....................................................................................................... 45
  SERVER INSTALLATION AND IMAGE CREATION .................................................................................. 47
  REFERENCES ........................................................................................................................................... 47
THE XLSTAT APPROACH ...................................................................................................................... 48
DATA SELECTION ................................................................................................................................... 48
MESSAGES ............................................................................................................................................... 51
OPTIONS ............................................................................................................................................... 52
DATA SAMPLING .................................................................................................................................... 56
  DESCRIPTION ......................................................................................................................................... 56
  DIALOG BOX ......................................................................................................................................... 57
  REFERENCES ........................................................................................................................................... 58
DISTRIBUTION SAMPLING ...................................................................................................................... 59
  DESCRIPTION ......................................................................................................................................... 59
  DIALOG BOX ......................................................................................................................................... 67
  EXAMPLE ............................................................................................................................................... 68
  REFERENCES ........................................................................................................................................... 68
VARIABLES TRANSFORMATION ............................................................................................................. 70
  DIALOG BOX ......................................................................................................................................... 70
MISSING DATA ....................................................................................................................................... 73
  DESCRIPTION ......................................................................................................................................... 73
  DIALOG BOX ......................................................................................................................................... 74
  RESULTS ............................................................................................................................................... 76
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXAMPLE</td>
<td>168</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>168</td>
</tr>
<tr>
<td>MULTICOLLINEARITY STATISTICS</td>
<td>170</td>
</tr>
<tr>
<td>DESCRIPTION</td>
<td>170</td>
</tr>
<tr>
<td>DIALOG BOX</td>
<td>171</td>
</tr>
<tr>
<td>RESULTS</td>
<td>172</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>173</td>
</tr>
<tr>
<td>CONTINGENCY TABLES (DESCRIPTIVE STATISTICS)</td>
<td>174</td>
</tr>
<tr>
<td>DESCRIPTION</td>
<td>174</td>
</tr>
<tr>
<td>DIALOG BOX</td>
<td>175</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>178</td>
</tr>
<tr>
<td>XLSTAT-PIVOT</td>
<td>179</td>
</tr>
<tr>
<td>DESCRIPTION</td>
<td>179</td>
</tr>
<tr>
<td>DIALOG BOX</td>
<td>181</td>
</tr>
<tr>
<td>RESULTS</td>
<td>184</td>
</tr>
<tr>
<td>EXAMPLE</td>
<td>185</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>185</td>
</tr>
<tr>
<td>SCATTER PLOTS</td>
<td>186</td>
</tr>
<tr>
<td>DIALOG BOX</td>
<td>186</td>
</tr>
<tr>
<td>EXAMPLE</td>
<td>188</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>188</td>
</tr>
<tr>
<td>PARALLEL COORDINATES PLOTS</td>
<td>189</td>
</tr>
<tr>
<td>DESCRIPTION</td>
<td>189</td>
</tr>
<tr>
<td>DIALOG BOX</td>
<td>189</td>
</tr>
<tr>
<td>EXAMPLE</td>
<td>191</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>191</td>
</tr>
<tr>
<td>TERNARY DIAGRAMS</td>
<td>192</td>
</tr>
<tr>
<td>DESCRIPTION</td>
<td>192</td>
</tr>
<tr>
<td>DIALOG BOX</td>
<td>192</td>
</tr>
<tr>
<td>EXAMPLE</td>
<td>194</td>
</tr>
<tr>
<td>2D PLOTS FOR CONTINGENCY TABLES</td>
<td>195</td>
</tr>
<tr>
<td>DESCRIPTION</td>
<td>195</td>
</tr>
<tr>
<td>DIALOG BOX</td>
<td>195</td>
</tr>
<tr>
<td>EXAMPLE</td>
<td>197</td>
</tr>
<tr>
<td>ERROR BARS</td>
<td>198</td>
</tr>
<tr>
<td>DESCRIPTION</td>
<td>198</td>
</tr>
<tr>
<td>DIALOG BOX</td>
<td>198</td>
</tr>
<tr>
<td>EXAMPLE</td>
<td>199</td>
</tr>
<tr>
<td>PLOT A FUNCTION</td>
<td>200</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>RESULTS</td>
<td>541</td>
</tr>
<tr>
<td>EXAMPLES</td>
<td>545</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>546</td>
</tr>
<tr>
<td>CORRELATION TESTS</td>
<td>547</td>
</tr>
<tr>
<td>DESCRIPTION</td>
<td>547</td>
</tr>
<tr>
<td>DIALOG BOX</td>
<td>547</td>
</tr>
<tr>
<td>RESULTS</td>
<td>551</td>
</tr>
<tr>
<td>EXAMPLE</td>
<td>551</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>551</td>
</tr>
<tr>
<td>RV COEFFICIENT</td>
<td>552</td>
</tr>
<tr>
<td>DESCRIPTION</td>
<td>552</td>
</tr>
<tr>
<td>DIALOG BOX</td>
<td>553</td>
</tr>
<tr>
<td>RESULTS</td>
<td>555</td>
</tr>
<tr>
<td>EXAMPLE</td>
<td>555</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>555</td>
</tr>
<tr>
<td>TESTS ON CONTINGENCY TABLES (CHI-SQUARE, ...)</td>
<td>557</td>
</tr>
<tr>
<td>DESCRIPTION</td>
<td>557</td>
</tr>
<tr>
<td>DIALOG BOX</td>
<td>561</td>
</tr>
<tr>
<td>RESULTS</td>
<td>564</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>564</td>
</tr>
<tr>
<td>COCHRAN-ARMITAGE TREND TEST</td>
<td>566</td>
</tr>
<tr>
<td>DESCRIPTION</td>
<td>566</td>
</tr>
<tr>
<td>DIALOG BOX</td>
<td>567</td>
</tr>
<tr>
<td>RESULTS</td>
<td>569</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>569</td>
</tr>
<tr>
<td>MANTEL TEST</td>
<td>570</td>
</tr>
<tr>
<td>DESCRIPTION</td>
<td>570</td>
</tr>
<tr>
<td>DIALOG BOX</td>
<td>571</td>
</tr>
<tr>
<td>RESULTS</td>
<td>573</td>
</tr>
<tr>
<td>EXAMPLE</td>
<td>573</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>573</td>
</tr>
<tr>
<td>ONE-SAMPLE T AND Z TESTS</td>
<td>575</td>
</tr>
<tr>
<td>DESCRIPTION</td>
<td>575</td>
</tr>
<tr>
<td>DIALOG BOX</td>
<td>576</td>
</tr>
<tr>
<td>RESULTS</td>
<td>578</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>578</td>
</tr>
<tr>
<td>TWO-SAMPLE T AND Z TESTS</td>
<td>579</td>
</tr>
<tr>
<td>DESCRIPTION</td>
<td>579</td>
</tr>
<tr>
<td>DIALOG BOX</td>
<td>581</td>
</tr>
<tr>
<td>EXAMPLE</td>
<td>584</td>
</tr>
<tr>
<td>Test</td>
<td>Page</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Dixon Test</td>
<td>682</td>
</tr>
<tr>
<td>Grubbs Test</td>
<td>674</td>
</tr>
<tr>
<td>One-sample runs test</td>
<td>669</td>
</tr>
<tr>
<td>McNemar's Test</td>
<td>660</td>
</tr>
<tr>
<td>Cochran's Q test</td>
<td>655</td>
</tr>
<tr>
<td>Cochran-Mantel-Haenszel test</td>
<td>664</td>
</tr>
<tr>
<td>Example</td>
<td>688</td>
</tr>
<tr>
<td>References</td>
<td>688</td>
</tr>
<tr>
<td>Dialog box</td>
<td>652</td>
</tr>
<tr>
<td>Description</td>
<td>650</td>
</tr>
<tr>
<td>Results</td>
<td>652</td>
</tr>
<tr>
<td>Results</td>
<td>658</td>
</tr>
<tr>
<td>Results</td>
<td>667</td>
</tr>
<tr>
<td>Results</td>
<td>667</td>
</tr>
<tr>
<td>Results</td>
<td>672</td>
</tr>
<tr>
<td>Results</td>
<td>678</td>
</tr>
<tr>
<td>Results</td>
<td>680</td>
</tr>
<tr>
<td>References</td>
<td>680</td>
</tr>
</tbody>
</table>
GENERALIZED PROCRUSTES ANALYSIS (GPA)......................................................................... 772
   DESCRIPTION....................................................................................................................... 772
   DIALOG BOX....................................................................................................................... 774
   RESULTS............................................................................................................................ 778
   EXAMPLE .......................................................................................................................... 780
   REFERENCES...................................................................................................................... 780

SEMANTIC DIFFERENTIAL CHARTS...................................................................................... 782
   DESCRIPTION....................................................................................................................... 782
   DIALOG BOX....................................................................................................................... 783
   RESULTS............................................................................................................................ 784
   EXAMPLE .......................................................................................................................... 784
   REFERENCES...................................................................................................................... 784

TURF ANALYSIS.................................................................................................................. 785
   DESCRIPTION....................................................................................................................... 785
   DIALOG BOX....................................................................................................................... 787
   RESULTS............................................................................................................................ 790
   EXAMPLE .......................................................................................................................... 790
   REFERENCES...................................................................................................................... 790

DESIGN OF EXPERIMENTS FOR SENSORY DATA ANALYSIS........................................... 792
   DESCRIPTION....................................................................................................................... 792
   DIALOG BOX....................................................................................................................... 796
   RESULTS............................................................................................................................ 798
   EXAMPLE .......................................................................................................................... 799
   REFERENCES...................................................................................................................... 799

DESIGN OF EXPERIMENTS FOR SENSORY DISCRIMINATION TESTS............................. 800
   DESCRIPTION....................................................................................................................... 800
   DIALOG BOX....................................................................................................................... 801
   RESULTS............................................................................................................................ 802
   EXAMPLE .......................................................................................................................... 802
   REFERENCES...................................................................................................................... 802

SENSORY DISCRIMINATION TESTS.................................................................................... 804
   DESCRIPTION....................................................................................................................... 804
   DIALOG BOX....................................................................................................................... 807
   RESULTS............................................................................................................................ 809
   EXAMPLE .......................................................................................................................... 809
   REFERENCES...................................................................................................................... 809

DESIGN OF EXPERIMENTS FOR CONJOINT ANALYSIS................................................. 810
   DESCRIPTION....................................................................................................................... 810
   DIALOG BOX....................................................................................................................... 811
   RESULTS............................................................................................................................ 813
<table>
<thead>
<tr>
<th>Example</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>DESCRIPTION</td>
<td>814</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>814</td>
</tr>
<tr>
<td><strong>DESIGN FOR CHOICE BASED CONJOINT ANALYSIS</strong></td>
<td>815</td>
</tr>
<tr>
<td>DESCRIPTION</td>
<td>815</td>
</tr>
<tr>
<td>DIALOG BOX</td>
<td>816</td>
</tr>
<tr>
<td>RESULTS</td>
<td>819</td>
</tr>
<tr>
<td>EXAMPLE</td>
<td>819</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>819</td>
</tr>
<tr>
<td><strong>CONJOINT ANALYSIS</strong></td>
<td>820</td>
</tr>
<tr>
<td>DESCRIPTION</td>
<td>820</td>
</tr>
<tr>
<td>DIALOG BOX</td>
<td>822</td>
</tr>
<tr>
<td>RESULTS</td>
<td>825</td>
</tr>
<tr>
<td>EXAMPLE</td>
<td>830</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>830</td>
</tr>
<tr>
<td><strong>CHOICE BASED CONJOINT ANALYSIS</strong></td>
<td>831</td>
</tr>
<tr>
<td>DESCRIPTION</td>
<td>831</td>
</tr>
<tr>
<td>DIALOG BOX</td>
<td>832</td>
</tr>
<tr>
<td>RESULTS</td>
<td>836</td>
</tr>
<tr>
<td>EXAMPLE</td>
<td>837</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>837</td>
</tr>
<tr>
<td><strong>CONJOINT ANALYSIS SIMULATION TOOL</strong></td>
<td>838</td>
</tr>
<tr>
<td>DESCRIPTION</td>
<td>838</td>
</tr>
<tr>
<td>DIALOG BOX</td>
<td>840</td>
</tr>
<tr>
<td>RESULTS</td>
<td>842</td>
</tr>
<tr>
<td>EXAMPLE</td>
<td>842</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>843</td>
</tr>
<tr>
<td><strong>DESIGN FOR MAXDIFF</strong></td>
<td>844</td>
</tr>
<tr>
<td>DESCRIPTION</td>
<td>844</td>
</tr>
<tr>
<td>DIALOG BOX</td>
<td>844</td>
</tr>
<tr>
<td>RESULTS</td>
<td>846</td>
</tr>
<tr>
<td>EXAMPLE</td>
<td>846</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>846</td>
</tr>
<tr>
<td><strong>MAXDIFF ANALYSIS</strong></td>
<td>847</td>
</tr>
<tr>
<td>DESCRIPTION</td>
<td>847</td>
</tr>
<tr>
<td>DIALOG BOX</td>
<td>848</td>
</tr>
<tr>
<td>RESULTS</td>
<td>849</td>
</tr>
<tr>
<td>EXAMPLE</td>
<td>850</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>850</td>
</tr>
<tr>
<td><strong>MONOTONE REGRESSION (MONANOVA)</strong></td>
<td>851</td>
</tr>
<tr>
<td>DESCRIPTION</td>
<td>851</td>
</tr>
<tr>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>--------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>COX MODEL (XLSTAT-POWER)</td>
<td>1042</td>
</tr>
<tr>
<td>Description</td>
<td>1042</td>
</tr>
<tr>
<td>Dialog Box</td>
<td>1044</td>
</tr>
<tr>
<td>Results</td>
<td>1045</td>
</tr>
<tr>
<td>Example</td>
<td>1046</td>
</tr>
<tr>
<td>References</td>
<td></td>
</tr>
<tr>
<td>PARETO PLOTS</td>
<td>1106</td>
</tr>
<tr>
<td>Attribute Charts</td>
<td>1081</td>
</tr>
<tr>
<td>Description</td>
<td>1081</td>
</tr>
<tr>
<td>Dialog Box</td>
<td>1083</td>
</tr>
<tr>
<td>Results</td>
<td>1088</td>
</tr>
<tr>
<td>Example</td>
<td>1090</td>
</tr>
<tr>
<td>References</td>
<td>1090</td>
</tr>
<tr>
<td>INDIVIDUAL CHARTS</td>
<td>1070</td>
</tr>
<tr>
<td>Description</td>
<td>1070</td>
</tr>
<tr>
<td>Dialog Box</td>
<td>1072</td>
</tr>
<tr>
<td>Results</td>
<td>1077</td>
</tr>
<tr>
<td>Example</td>
<td>1080</td>
</tr>
<tr>
<td>References</td>
<td>1080</td>
</tr>
<tr>
<td>TIME WEIGHTED CHARTS</td>
<td>1092</td>
</tr>
<tr>
<td>Description</td>
<td>1092</td>
</tr>
<tr>
<td>Dialog Box</td>
<td>1095</td>
</tr>
<tr>
<td>Results</td>
<td>1102</td>
</tr>
<tr>
<td>Example</td>
<td>1104</td>
</tr>
<tr>
<td>References</td>
<td>1105</td>
</tr>
<tr>
<td>REFERENCE PLOTS</td>
<td>1106</td>
</tr>
<tr>
<td>Description</td>
<td>1106</td>
</tr>
<tr>
<td>Dialog Box</td>
<td>1107</td>
</tr>
<tr>
<td>Results</td>
<td>1110</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>DEMING REGRESSION</td>
<td>1265</td>
</tr>
<tr>
<td>DESCRIPTION</td>
<td>1265</td>
</tr>
<tr>
<td>DIALOG BOX</td>
<td>1266</td>
</tr>
<tr>
<td>RESULTS</td>
<td>1268</td>
</tr>
<tr>
<td>EXAMPLE</td>
<td>1269</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>1269</td>
</tr>
<tr>
<td>PASSING AND BABLOK REGRESSION</td>
<td>1261</td>
</tr>
<tr>
<td>DESCRIPTION</td>
<td>1261</td>
</tr>
<tr>
<td>DIALOG BOX</td>
<td>1262</td>
</tr>
<tr>
<td>RESULTS</td>
<td>1263</td>
</tr>
<tr>
<td>EXAMPLE</td>
<td>1264</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>1264</td>
</tr>
<tr>
<td>SENSITIVITY AND SPECIFICITY</td>
<td>1237</td>
</tr>
<tr>
<td>DESCRIPTION</td>
<td>1237</td>
</tr>
<tr>
<td>DIALOG BOX</td>
<td>1240</td>
</tr>
<tr>
<td>RESULTS</td>
<td>1242</td>
</tr>
<tr>
<td>EXAMPLE</td>
<td>1242</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>1242</td>
</tr>
<tr>
<td>METHOD COMPARISON</td>
<td>1254</td>
</tr>
<tr>
<td>DESCRIPTION</td>
<td>1254</td>
</tr>
<tr>
<td>DIALOG BOX</td>
<td>1257</td>
</tr>
<tr>
<td>RESULTS</td>
<td>1259</td>
</tr>
<tr>
<td>EXAMPLE</td>
<td>1260</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>1260</td>
</tr>
<tr>
<td>ROC CURVES</td>
<td>1244</td>
</tr>
<tr>
<td>DESCRIPTION</td>
<td>1244</td>
</tr>
<tr>
<td>DIALOG BOX</td>
<td>1248</td>
</tr>
<tr>
<td>RESULTS</td>
<td>1251</td>
</tr>
<tr>
<td>EXAMPLE</td>
<td>1252</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>1252</td>
</tr>
<tr>
<td>DIFFERENTIAL EXPRESSION</td>
<td>1270</td>
</tr>
<tr>
<td>DESCRIPTION</td>
<td>1270</td>
</tr>
<tr>
<td>DIALOG BOX</td>
<td>1273</td>
</tr>
<tr>
<td>RESULTS</td>
<td>1275</td>
</tr>
<tr>
<td>EXAMPLE</td>
<td>1276</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>1276</td>
</tr>
<tr>
<td>Analysis Type</td>
<td>Page</td>
</tr>
<tr>
<td>---------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Heat Maps</td>
<td>1277</td>
</tr>
<tr>
<td>Description</td>
<td>1277</td>
</tr>
<tr>
<td>Dialog Box</td>
<td>1278</td>
</tr>
<tr>
<td>Results</td>
<td>1280</td>
</tr>
<tr>
<td>Example</td>
<td>1281</td>
</tr>
<tr>
<td>References</td>
<td>1281</td>
</tr>
<tr>
<td>Canonical Correlation Analysis (CCORA)</td>
<td>1282</td>
</tr>
<tr>
<td>Description</td>
<td>1282</td>
</tr>
<tr>
<td>Dialog Box</td>
<td>1282</td>
</tr>
<tr>
<td>Results</td>
<td>1285</td>
</tr>
<tr>
<td>Example</td>
<td>1287</td>
</tr>
<tr>
<td>References</td>
<td>1287</td>
</tr>
<tr>
<td>Redundancy Analysis (RDA)</td>
<td>1288</td>
</tr>
<tr>
<td>Description</td>
<td>1288</td>
</tr>
<tr>
<td>Dialog Box</td>
<td>1289</td>
</tr>
<tr>
<td>Results</td>
<td>1293</td>
</tr>
<tr>
<td>Example</td>
<td>1294</td>
</tr>
<tr>
<td>References</td>
<td>1294</td>
</tr>
<tr>
<td>Canonical Correspondence Analysis (CCA)</td>
<td>1295</td>
</tr>
<tr>
<td>Description</td>
<td>1295</td>
</tr>
<tr>
<td>Dialog Box</td>
<td>1296</td>
</tr>
<tr>
<td>Results</td>
<td>1301</td>
</tr>
<tr>
<td>Example</td>
<td>1301</td>
</tr>
<tr>
<td>References</td>
<td>1301</td>
</tr>
<tr>
<td>Principal Coordinate Analysis (PCOA)</td>
<td>1303</td>
</tr>
<tr>
<td>Description</td>
<td>1303</td>
</tr>
<tr>
<td>Dialog Box</td>
<td>1305</td>
</tr>
<tr>
<td>Results</td>
<td>1307</td>
</tr>
<tr>
<td>Example</td>
<td>1307</td>
</tr>
<tr>
<td>References</td>
<td>1307</td>
</tr>
<tr>
<td>Multiple Factor Analysis (MFA)</td>
<td>1309</td>
</tr>
<tr>
<td>Description</td>
<td>1309</td>
</tr>
<tr>
<td>Dialog Box</td>
<td>1310</td>
</tr>
<tr>
<td>Results</td>
<td>1318</td>
</tr>
<tr>
<td>Example</td>
<td>1320</td>
</tr>
<tr>
<td>References</td>
<td>1320</td>
</tr>
<tr>
<td>Latent Class Clustering</td>
<td>1321</td>
</tr>
<tr>
<td>Description</td>
<td>1321</td>
</tr>
<tr>
<td>Dialog Box</td>
<td>1322</td>
</tr>
<tr>
<td>Results</td>
<td>1329</td>
</tr>
</tbody>
</table>
Introduction

XLSTAT started over ten years ago in order to make accessible to anyone a powerful, complete and user-friendly data analysis and statistical solution.

The accessibility comes from the compatibility of XLSTAT with all the Microsoft Excel versions that are used nowadays (starting from Excel 97 up to Excel 2016), from the interface that is available in seven languages (Chinese, English, French, German, Italian, Japanese, Polish, Portuguese, and Spanish) and from the permanent availability of a fully functional 30 days evaluation version on the XLSTAT website www.xlstat.com.

The power of XLSTAT comes from both the C++ programming language, and from the algorithms that are used. The algorithms are the result of many years of research of thousands of statisticians, mathematicians, computer scientists throughout the world. Each development of a new functionality in XLSTAT is preceded by an in-depth research phase that sometimes includes exchanges with the leading specialists of the methods of interest.

The completeness of XLSTAT is the fruit of over fifteen years of continuous work, and of regular exchanges with the users’ community. Users’ suggestions have helped a lot improving the software, by making it well adapted to a variety of requirements.

Last, the usability comes from the user-friendly interface, which after a few minutes of trying it out, facilitates the use of some statistical methods that might require hours of training with other software.

The software architecture has considerably evolved over the last 5 years in order to take into account the advances of Microsoft Excel and the compatibility issues between platforms. The software relies on Visual Basic Application for the interface and on C++ for the mathematical and statistical computations.

As always, the Addinsoft team and the XLSTAT distributors are available to answer any question you have, or to take into account your remarks and suggestions in order to continue improving the software.
License

XLSTAT 2015 - SOFTWARE LICENSE AGREEMENT

ADDINSOFT SARL ("ADDINSOFT") IS WILLING TO LICENSE VERSION 2015 OF ITS XLSTAT (r) SOFTWARE AND THE ACCOMPANYING DOCUMENTATION (THE "SOFTWARE") TO YOU ONLY ON THE CONDITION THAT YOU ACCEPT ALL OF THE TERMS IN THIS AGREEMENT. PLEASE READ THE TERMS CAREFULLY. BY USING THE SOFTWARE YOU ACKNOWLEDGE THAT YOU HAVE READ THIS AGREEMENT, UNDERSTAND IT AND AGREE TO BE BOUND BY ITS TERMS AND CONDITIONS. IF YOU DO NOT AGREE TO THESE TERMS, ADDINSOFT IS UNWILLING TO LICENSE THE SOFTWARE TO YOU.

1. LICENSE. Addinsoft hereby grants you a nonexclusive license to install and use the Software in machine-readable form on a single computer for use by a single individual if you are using the demo version or if you have registered your demo version to use it with no time limits. If you have ordered a multi-users license, the number of users depends directly on the terms specified on the invoice sent to your company by Addinsoft or the authorized reseller.

2. RESTRICTIONS. Addinsoft retains all right, title, and interest in and to the Software, and any rights not granted to you herein are reserved by Addinsoft. You may not reverse engineer, disassemble, decompile, or translate the Software, or otherwise attempt to derive the source code of the Software, except to the extent allowed under any applicable law. If applicable law permits such activities, any information so discovered must be promptly disclosed to Addinsoft and shall be deemed to be the confidential proprietary information of Addinsoft. Any attempt to transfer any of the rights, duties or obligations hereunder is void. You may not rent, lease, loan, or resell for profit the Software, or any part thereof. You may not reproduce or distribute the Software except as expressly permitted under Section 1, and you may not create derivative works of the Software unless with the express agreement of Addinsoft.

3. SUPPORT. Registered users of the Software are entitled to Addinsoft standard support services. Demo version users may contact Addinsoft for support but with no guarantee to benefit from Addinsoft standard support services.

4. NO WARRANTY. THE SOFTWARE IS PROVIDED "AS IS" AND WITHOUT ANY WARRANTY OR CONDITION, WHETHER EXPRESS, IMPLIED OR STATUTORY. Some
jurisdictions do not allow the disclaimer of implied warranties, so the foregoing disclaimer may not apply to you. This warranty gives you specific legal rights and you may also have other legal rights which vary from state to state, or from country to country.

5. LIMITATION OF LIABILITY. IN NO EVENT WILL ADDINSOFT OR ITS SUPPLIERS BE LIABLE FOR ANY LOST PROFITS OR OTHER CONSEQUENTIAL, INCIDENTAL OR SPECIAL DAMAGES (HOWEVER ARISING, INCLUDING NEGLIGENCE) IN CONNECTION WITH THE SOFTWARE OR THIS AGREEMENT, EVEN IF ADDINSOFT HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. In no event will Addinsoft liability in connection with the Software, regardless of the form of action, exceed the price paid for acquiring the Software. Some jurisdictions do not allow the foregoing limitations of liability, so the foregoing limitations may not apply to you.

6. TERM AND TERMINATION. This Agreement shall continue until terminated. You may terminate the Agreement at any time by deleting all copies of the Software. This license terminates automatically if you violate any terms of the Agreement. Upon termination you must promptly delete all copies of the Software.

7. CONTRACTING PARTIES. If the Software is installed on computers owned by a corporation or other legal entity, then this Agreement is formed by and between Addinsoft and such entity. The individual executing this Agreement represents and warrants to Addinsoft that they have the authority to bind such entity to the terms and conditions of this Agreement.

8. INDEMNITY. You agree to defend and indemnify Addinsoft against all claims, losses, liabilities, damages, costs and expenses, including attorney’s fees, which Addinsoft may incur in connection with your breach of this Agreement.

9. GENERAL. The Software is a "commercial item." This Agreement is governed and interpreted in accordance with the laws of the Court of Paris, France, without giving effect to its conflict of laws provisions. The United Nations Convention on Contracts for the International Sale of Goods is expressly disclaimed. Any claim arising out of or related to this Agreement must be brought exclusively in a court located in PARIS, FRANCE, and you consent to the jurisdiction of such courts. If any provision of this Agreement shall be invalid, the validity of the remaining provisions of this Agreement shall not be affected. This Agreement is the entire and exclusive agreement between Addinsoft and you with respect to the Software and supersedes all prior agreements (whether written or oral) and other communications between Addinsoft and you with respect to the Software.
System configuration

XLSTAT runs under the following operating systems: Windows XP, Windows Vista, Windows 7, Windows 8.x, and Mac OSX 10.6, 10.7, 10.8, 10.9, 10.10 and 11. 32 and 64 bits platforms are supported.

To be able to run XLSTAT required that Microsoft Excel is also installed on your computer. XLSTAT is compatible with the following Excel versions on the Windows systems: Excel 97 (8.0), Excel 2000 (9.0), Excel XP (10.0), Excel 2003 (11.0), Excel 2007 (12.0), Excel 2010 (14.0), Excel 2013 (15.0) and Excel 2016 (16.0) (32 and 64 bits). Version 2011 (14.1) with Service Pack 1 (or later) of Excel is required for the Mac OSX system.

Free patches and upgrades for Microsoft Office are available for free on the Microsoft Website. We highly recommend that you download and install these patches as some of them are critical. To check if your Excel version is up to date, please go from time to time to the following web site:

Windows: http://office.microsoft.com/officeupdate

Installation

To install XLSTAT you need to:

- Either double-click on the xlstat.exe (PC) or xlstatMac.zip (Mac) file that you downloaded from the XLSTAT website www.xlstat.com or from one of our numerous partners, or available on a CD-Rom,

- Or insert the CD-Rom you received from us or from a distributor and wait until the installation procedure starts and then follow the step by step instructions.

If your rights on your computer are restricted, you should ask someone that has administrator rights on the machine to install the software for you. Once the installation is over, the administrator must let you have read and write access to the following folder:

- The folder where the XLSTAT user files are located (typically C:\Documents and settings\User Name\Application Data\Addinsoft\XLSTAT\), including the corresponding subfolders.

This folder can be changed by the administrator, using the options dialog box of XLSTAT.
**Advanced installation**

XLSTAT is easy to deploy within organizations thanks to a variety of functionalities that assist you during the installation on a server, a farm of computers or on computers with multiple user accounts.

**Silent installation by InstallShield Script (Windows only)**

XLSTAT uses an installation program that was created with InstallShield. It is based on install script only. That means that, as with any other installation package based on InstallShield, you do a silent installation.

During the installation, XLSTAT needs that MS Excel is installed on the computer. Excel will be called once to add the XLSTAT button in the Excel main icon bar. The reverse operation is performed during the uninstall process.

Use of an InstallShield script:

You can call the installation program to run a silent installation with the following options that are described in the help of InstallShield.

/\uninst**: This option forces an uninstall of XLSTAT.

/s: The installation will be done without showing the user dialogs.

/\f1 "script file": This parameter indicates the script file that should be used with an absolute path and file name.

/\f2 "log file": This parameter indicates the log file that should be used with an absolute path and file name.

/r: This parameter activates the record mode to create a script file.

/\L: This parameter allows the selection of the language used during the installation. 10 languages are currently supported as indicated in the following table:
/servername=XLSTATLICENSESERVER: this parameter gives the name of the network on which the XLSTAT server is hosted. It is only useful in the case of an XLSTAT client server concurrent license. In that case, XLSTATLICENSESERVER should be replaced by the host name of the server where the XLSTAT concurrent license is hosted.

After the installation of XLSTAT there are two sample script files for installation and uninstall of XLSTAT in the folder silentinstall under the XLSTAT installation folder. You need also the file setup.exe of the installation package to be able to work with the scripts. You obtain these scripts by dezipping the xlstat.zip file that you can download on our website.

To work in a convenient way with scripts for a silent installation, in the following examples, we suppose that the script files and the setup.exe file are located in the same MYDir folder, which is at the same time the current working folder.

**Silent installation of XLSTAT**

A call to install XLSTAT can be as follows:

```
setup.exe /s /f1"C:\MyDir\setup.iss"
```

In this case the script file setup.iss contains the following text:

```
[InstallShield Silent]
Version=v7.00
File=Response File

[File Transfer]
OverwrittenReadOnly=NoToAll

[{68B36FA5-E276-4C03-A56C-EC25717E1668}-DlgOrder]
Dlg0={68B36FA5-E276-4C03-A56C-EC25717E1668}-SdWelcome-0
```
Count=9
Dlg1={(68B36FA5-E276-4C03-A56C-EC25717E1668)-SdLicense2Rtf-0}
Dlg2={(68B36FA5-E276-4C03-A56C-EC25717E1668)-SetupType2-0}
Dlg3={(68B36FA5-E276-4C03-A56C-EC25717E1668)-SdAskDestPath2-0}
Dlg4={(68B36FA5-E276-4C03-A56C-EC25717E1668)-SdAskDestPath2-1}
Dlg5={(68B36FA5-E276-4C03-A56C-EC25717E1668)-SdComponentTree-0}
Dlg6={(68B36FA5-E276-4C03-A56C-EC25717E1668)-SdStartCopy2-0}
Dlg7={(68B36FA5-E276-4C03-A56C-EC25717E1668)-MessageBox-0}
Dlg8={(68B36FA5-E276-4C03-A56C-EC25717E1668)-SdFinish-0}
[(68B36FA5-E276-4C03-A56C-EC25717E1668)-SdWelcome-0]
Result=1
[(68B36FA5-E276-4C03-A56C-EC25717E1668)-SdLicense2Rtf-0]
Result=1
[(68B36FA5-E276-4C03-A56C-EC25717E1668)-SetupType2-0]
Result=303
[(68B36FA5-E276-4C03-A56C-EC25717E1668)-SdAskDestPath2-0]
szDir=C:\Program Files\Addinsoft\XLSTAT
Result=1
[(68B36FA5-E276-4C03-A56C-EC25717E1668)-SdAskDestPath2-1]
szDir=C:\My documents\Addinsoft\
Result=1
[(68B36FA5-E276-4C03-A56C-EC25717E1668)-SdComponentTree-0]
szDir=C:\Program Files\Addinsoft\XLSTAT
Component-type=string
Component-count=4
Component-0=Program Files
Component-1=Help Files
In this example you may replace the path "C:\Program Files\Addinsoft\XLSTAT" by your desired installation path. You can as well change the path for the user's files "C:\My Documents\Addinsoft" to a path of your choice.

**Silent uninstall of XLSTAT**

A call to uninstall XLSTAT can be as follows:

```
setup.exe /uninstall /s /f1"C:\MyDir\setupRemove.iss"
```

In this case the script file setupRemove.iss contains the following text:

```
[InstallShield Silent]
Version=v7.00
```
Silent install of XLSTAT server when using a network concurrent license

Silent installation of XLSTAT Server.

A call to install XLSTAT can be as follows:

```
setup.exe /s /f1"C:\MyDir\setup.iss"
```

In this case the script file setup.iss contains the following text:

```
[InstallShield Silent]
Version=v7.00
```
File=Response File

[File Transfer]

OverwrittenReadOnly=NoToAll

[68B36FA5-E276-4C03-A56C-EC25717E1668]-DlgOrder]

Dlg0=(68B36FA5-E276-4C03-A56C-EC25717E1668)-SdWelcome-0

Count=8

Dlg1=(68B36FA5-E276-4C03-A56C-EC25717E1668)-SdLicense2Rtf-0

Dlg2=(68B36FA5-E276-4C03-A56C-EC25717E1668)-SetupType2-0

Dlg3=(68B36FA5-E276-4C03-A56C-EC25717E1668)-SdAskDestPath2-0

Dlg4=(68B36FA5-E276-4C03-A56C-EC25717E1668)-SdAskDestPath2-1

Dlg5=(68B36FA5-E276-4C03-A56C-EC25717E1668)-SdComponentTree-0

Dlg6=(68B36FA5-E276-4C03-A56C-EC25717E1668)-SdStartCopy2-0

Dlg7=(68B36FA5-E276-4C03-A56C-EC25717E1668)-SdFinish-0

[68B36FA5-E276-4C03-A56C-EC25717E1668]-SdWelcome-0]

Result=1

[68B36FA5-E276-4C03-A56C-EC25717E1668]-SdLicense2Rtf-0]

Result=1

[68B36FA5-E276-4C03-A56C-EC25717E1668]-SetupType2-0]

Result=303

[68B36FA5-E276-4C03-A56C-EC25717E1668]-SdAskDestPath2-0]

szDir=C:\Program Files\Addinsoft\XLSTAT

Result=1

[68B36FA5-E276-4C03-A56C-EC25717E1668]-SdAskDestPath2-1]

szDir= C:\My documents\Addinsoft

Result=1

[68B36FA5-E276-4C03-A56C-EC25717E1668]-SdComponentTree-0]

szDir=C:\Program Files\Addinsoft\XLSTAT
In this example you may replace the path "C:\Program Files\Addinsoft\XLSTAT" by your desired installation path. You can as well change the path for the user’s files "C:\My Documents\Addinsoft" to a path of your choice.

**Silent install of XLSTAT Client on the user computer when using a network concurrent license**

A call to install XLSTAT can be as follows:

```
setup.exe /s /f1"C:\MyDir\setup.iss"
```
In this case the script file setup.iss contains the following text:

[InstallShield Silent]
Version=v7.00
File=Response File
[File Transfer]
OverwrittenReadOnly=NoToAll

{[68B36FA5-E276-4C03-A56C-EC25717E1668]-DlgOrder}]
Dlg0={68B36FA5-E276-4C03-A56C-EC25717E1668}-SdWelcome-0
Count=9
Dlg1={68B36FA5-E276-4C03-A56C-EC25717E1668}-SdLicense2Rtf-0
Dlg2={68B36FA5-E276-4C03-A56C-EC25717E1668}-SetupType2-0
Dlg3={68B36FA5-E276-4C03-A56C-EC25717E1668}-SdAskDestPath2-0
Dlg4={68B36FA5-E276-4C03-A56C-EC25717E1668}-SdAskDestPath2-1
Dlg5={68B36FA5-E276-4C03-A56C-EC25717E1668}-SdComponentTree-0
Dlg6={68B36FA5-E276-4C03-A56C-EC25717E1668}-AskText-0
Dlg7={68B36FA5-E276-4C03-A56C-EC25717E1668}-SdStartCopy2-0
Dlg8={68B36FA5-E276-4C03-A56C-EC25717E1668}-SdFinish-0

{[68B36FA5-E276-4C03-A56C-EC25717E1668]-SdWelcome-0}]
Result=1

{[68B36FA5-E276-4C03-A56C-EC25717E1668]-SdLicense2Rtf-0}]
Result=1

{[68B36FA5-E276-4C03-A56C-EC25717E1668]-SetupType2-0}]
Result=303

{[68B36FA5-E276-4C03-A56C-EC25717E1668]-SdAskDestPath2-0}]
szDir=C:\Program Files\Addinsoft\XLSTAT
Result=1
szDir = C:\My documents\Addinsoft\nResult = 1

szDir = C:\Program Files\Addinsoft\XLSTAT
Component-type = string
Component-count = 5
Component-0 = Program Files
Component-1 = Help Files
Component-2 = Icons & Menu
Component-3 = Client setup
Component-4 = Single Node
Result = 1

szText = XLSTATLICENSESERVER
Result = 1

[Application]
Name = XLSTAT 2015
Version = 15.4.08.2810
Company = Addinsoft
Lang = 040c

Result = 1
bOpt1 = 0
bOpt2 = 0
In this example you may replace the path "C:\Program Files\Addinsoft\XLSTAT" by your desired installation path. You can as well change the path for the user's files "C:\Program Files\Addinsoft\" to a path of your choice.

You must enter the hostname of the server where the XLSTAT server license is installed, by replacing "XLSTATLICENSESERVER" by that hostname.

**Creating a user defined script file**

For further changes to the installation you may also record a manual installation of XLSTAT to create a script file that will be used later. Please use the option /r. A sample call for a script creation might look as follows:

```
setup.exe /r /f1"C:\MyDir\setup.iss"
```

**Language selection**

In most cases, a language selection is not necessary during a silent installation. If XLSTAT was already installed on the computer, the language selection by the installation option /L or the registry entry explained hereunder will have no effect. Each user of the computer will find the language choice he has made before. The user might change the language option at any moment using the XLSTAT Options menu. A demonstration on how a user can change the language is available at:


If XLSTAT is being installed for the first time by a user, with the InstallShield interface, then the language that has just been selected for the installation, will be chosen as the default language for XLSTAT.

If XLSTAT is being installed for the first time using a silent installation, then English will be selected as default language. There are two possibilities to change the interface language of XLSTAT before the first start of XLSTAT.

- **/L**: Use this option when calling the silent installation to set the desired language for the installation and for XLSTAT.
- **Register entry**: After the installation of XLSTAT has finished and before XLSTAT is started for the first time, you may change the value of the registry key HKEY_LOCAL_MACHINE\SOFTWARE\XLSTAT+\General\Language to one of the 7 values to set the language of XLSTAT:

<table>
<thead>
<tr>
<th>Code</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>US</td>
<td>English</td>
</tr>
<tr>
<td>FR</td>
<td>French</td>
</tr>
<tr>
<td>DE</td>
<td>German</td>
</tr>
<tr>
<td>IT</td>
<td>Italian</td>
</tr>
<tr>
<td>ES</td>
<td>Spanish</td>
</tr>
<tr>
<td>PT</td>
<td>Portuguese</td>
</tr>
<tr>
<td>PL</td>
<td>Polish</td>
</tr>
<tr>
<td>JP</td>
<td>Japanese</td>
</tr>
<tr>
<td>CN</td>
<td>traditional Chinese</td>
</tr>
<tr>
<td>CS</td>
<td>simplified Chinese</td>
</tr>
</tbody>
</table>

**Selection of the user folder**

XLSTAT gives the user the possibility to save data selections and choices made in the dialog boxes that correspond to the different functions, so that you can reuse them during a future session. Further details on how to control this feature can be found in the XLSTAT Options dialog box.

**Standard installation of XLSTAT**

The selection of the user folder during a standard installation of XLSTAT is set by InstallShield to:

%USERPROFILE%\Application data\ADDINSOFT\XLSTAT

%USERPROFILE%, which is a Windows environment variable, is replaced by its current value during the installation.

Each user has the possibility to change this default value to a user defined value using the corresponding option in the “Advanced” tab of the XLSTAT Options dialog box.

Furthermore you have the possibility to directly change the value of the following registry entry to the desired user folder. The registry entry has priority over the selection in the XLSTAT Options dialog box. The registry entry is different for each user. It has the following name:

HKEY_CURRENT_USER\Software\XLSTAT+\DATA\UserPath

The value of the registry entry may contain environment variables.
Multi-user environment

There are different types of multi-user environments. One example would be a server installation in the case of the Windows Terminal Server or in the case of a Citrix Metaframe Server. Another type of environment is a pool of computers that have all the same installation, often created using an image that has been replicated on all the computers of the pool where some users are authorized to work with XLSTAT. For such cases, please take note of the following advices regarding the choice of the user directories.

In that case, for each user, the user folder should point to a personal folder, for which the user has read and write rights.

There are basically two ways to meet these requirements:

- Use of a virtual folder;
- Use of environment variables.

Virtual folder

In this case, a virtual user folder already exists and is being used. This folder has the same name for every user, but it points to a different folder. A virtual folder is often associated to a user disc like U or X. During the login this user drive is often mounted automatically by a script. The users have normally read and write rights in this folder. For XLSTAT are no further actions necessary regarding the access rights.

If for instance the virtual user folder is U, then you can choose the following XLSTAT user folder that will contain the user data following the Microsoft naming conventions:

U:\Application Data\ADDINSOFT\XLSTAT

This folder should exist for each possible XLSTAT user before starting XLSTAT. If this is not the case, an error message informs about the non existing user folder and invites the user to select another user folder.

Environment variables

With this method the value of an environment variable is used to choose a different folder for each user. The user must have read and write rights in that folder.

For instance the environment variable %USERPROFILE% can be used to define the following folder using the Microsoft naming conventions:

%USERPROFILE%\Application Data\ADDINSOFT\XLSTAT
The use of environment variables in the dialog boxes of InstallShield is not possible. You may use environment variables in a script file or directly in registry entries.

**Server installation and image creation**

Server installation and image creation should be possible without any problem. Please notice that Microsoft Excel must have been installed on the machine including all options for VBA (Visual Basic for Applications), Microsoft Forms and graphical filters. During a server installation under Windows Terminal Server, Microsoft Excel version 2003 or later is a preferable choice.

During the installation of XLSTAT, read and write rights are necessary for the folder where the Excel.exe file is located.

If you have specific questions regarding the server installation, do not hesitate to contact the XLSTAT Support.

**References**

The XLSTAT approach

The XLSTAT interface totally relies on Microsoft Excel, whether for inputting the data or for displaying the results. The computations, however, are completely independent of Excel and the corresponding programs have been developed with the C++ programming language.

In order to guarantee accurate results, the XLSTAT software has been intensively tested and it has been validated by specialists of the statistical methods of interest.

Addinsoft has always been concerned about permanently improving the XLSTAT software suite, and welcomes any remarks and improvements you might want to suggest. To contact Addinsoft, write to support@xlstat.com.

Data selection

As with all XLSTAT modules, the selecting of data needs to be done directly on an Excel sheet, preferably with the mouse. Statistical programs usually require that you first build a list of variables, then define their type, and at last select the variables of interest for the method you want to apply to them. The XLSTAT approach is completely different as you only need to select the data directly on one or more Excel sheets.

Three selection modes are available:

- **Selection by range**: you select with the mouse on the Excel sheet all the cells of the table that corresponds to the selection field of the dialog box.

- **Selection by columns**: this mode is faster but requires that your data set starts on the first row of the Excel sheet. If this requirement is fulfilled you may select data by clicking on the name (A, B, ...) of the first column of your data set on the Excel sheet, and then by selecting the next columns by leaving the mouse button pressed and dragging the mouse cursor over the columns to select.

- **Selection by rows**: this mode is the reciprocal of the "selection by rows" model. It requires that your data set starts on the first column (A) of the Excel sheet. If this requirement is fulfilled you may select data by clicking on the name (1, 2, ...) of the first row of your data set on the Excel sheet, and then by selecting the next rows by leaving the mouse button pressed and dragging the mouse cursor over the rows to select.

Notes:

- Doing multiple selections is possible: if your variables go from column B to column G, and if you do not want to include column E in the selection, you should first select columns B to D with the mouse, then press the Ctrl key, and then select columns F to G still pressing Ctrl. You may also select columns B to G, then press Ctrl, then select column E.
• Multiple selections with selection by rows cannot be used if the transposition option is not activated (button)

• Multiple selections with selection by columns cannot be used if the transposition is activated (button).

• When selecting a variable or a group of variables (for example the quantitative explanatory variables) you cannot mix the selection mode. However you may use different modes for different selections within a dialog box.

• If you selected the name of the variables within the data selection, you should make sure the “Columns labels” or “Labels included” option activated.

• You can use keyboard shortcuts to quickly select data. Notice this is possible only you installed the latest patches for Microsoft Excel. Here is a list of the most useful selection shortcuts:

  • **Ctrl A**: Selects the whole spreadsheet
  • **Ctrl Space**: Selects the whole column corresponding to the already selected cells
  • **Shift Space**: Selects the whole row corresponding to the already selected cells
  • When one or more cells are selected:
    • **Shift Down**: Selects the currently selected cells and the cells on the row below on one row
    • **Shift Up**: Selects the currently selected and the cells on the row below on one row
    • **Shift Left**: Selects the currently selected and the cells to the left on one column
    • **Shift Right**: Selects the currently selected and the cells to the right on one column
    • **Ctrl Shift Down**: Selects all the adjacent non empty cells below the currently selected cells
    • **Ctrl Shift Up**: Selects all the adjacent non empty cells above the currently selected cells
    • **Ctrl Shift Left**: Selects all the adjacent non empty cells to the left of the currently selected cells
    • **Ctrl Shift Right**: Selects all the adjacent non empty cells to the right of the currently selected cells
  • When one or more columns are selected:
    • **Shift Left**: Selects one more column to the left of the currently selected columns
• **Shift Right**: Selects one more column to the right of the currently selected columns

• **Ctrl Shift Left**: Selects all the adjacent non empty columns to the left of the currently selected columns

• **Ctrl Shift Right**: Selects all the adjacent non empty columns to the right of the currently selected columns

• When one or more rows are selected:

• **Shift Down**: Selects one more row to the left of the currently selected rows

• **Shift Up**: Selects one more row to the right of the currently selected rows

• **Ctrl Shift Down**: Selects all the adjacent non empty rows below the currently selected rows

• **Ctrl Shift Up**: Selects all the adjacent non empty rows above the currently selected rows

See also:

[http://www.xlstat.com/demo-select.htm](http://www.xlstat.com/demo-select.htm)
Messages

XLSTAT uses an innovative message system to give information to the user and to report problems.

The dialog box below is an example of what happens when an active selection field (here the Dependent variables) has been activated but left empty. The software detects the problem and displays the message box.

The information displayed in red (or in blue depending on the severity) indicates which object/option/selection is responsible for the message. If you click on OK, the dialog box of the method that had just been activated is displayed again and the field corresponding to the Quantitative variable(s) is activated.

This message should be explicit enough to help you solve the problem by yourself. If a tutorial is available, the hyperlink "http://www.xlstat.com" links to a tutorial on the subject related to the problem. Sometimes an email address is displayed below the hyperlink to allow you send an email to Addinsoft using your usual email software, with the content of the XLSTAT message being automatically displayed in the email message.
Options

XLSTAT offers several options in order to allow you to customize and optimize the use of the software.

To display the options dialog box of XLSTAT, click on “Options” in the menu or on the button of the XLSTAT toolbar.

- **Save**: Click this button to save the changes you have made.
- **Close**: Click this button to close the dialog box. If you haven’t previously saved the options, the changes you have made will not be kept.
- **Help**: Click this button to display the help.
- **Reload**: Click this button to reload the default options.

**General** tab:

**Language**: Use this option to change the language of the interface of XLSTAT.

**Dialog box entries**:

- **Memorize during one session**: Activate this option if you want XLSTAT to memorize during one session (from opening until closing of XLSTAT) the entries and options of the dialog boxes.
  - **Including data selections**: Activate this option so that XLSTAT records the data selections during one session.

- **Memorize from one session to the next**: Activate this option if you want XLSTAT to memorize the entries and options of the dialog boxes from one session to the next.
  - **Including data selections**: Activate this option so that XLSTAT records the data selections from one session to the next. This option is useful and saves your time if you work on spreadsheets that always have the same layout.

**Ask for selections confirmation**: Activate this option so that XLSTAT prompts you to confirm the data selections once you clicked on the OK button. If you activate this option, you will be able to verify the number of rows and columns of all the active selections.

**Notify me before license or access to upgrades expires**: Activate this option so that XLSTAT notifies you two weeks before your license or your free access to upgrades expires.
**Display information messages:** Activate this option if you want to see the news released by Addinsoft. This is the best way to be informed of the availability of free upgrades.

**Show only the active functions in menus and toolbars:** Activate this option if you want that only the active functions corresponding to registered modules are displayed in the XLSTAT menu and in the toolbars.

**Missing data** tab:

**Consider empty cells as missing data:** this is the default option for XLSTAT and it cannot be changed. Empty cells are considered by all tools as missing data.

**Consider also the following values as missing data:** when a cell contains a value that is in the list, below this option, it will be considered as a missing data, whether the corresponding selection is for numerical or categorical data.

**Consider all text values as missing data:** when this option is activated, any text value found in a table that should contain only numerical values, will be converted and considered by XLSTAT as a missing data. This option should be activated if you are sure that text values can not correspond to numerical values converted to text by mistake.

**Outputs** tab:

**Position of new sheets:** If you choose the “Sheet” option in the dialog boxes of the XLSTAT functions, use this option to modify the position if the results sheets in the Excel workbook.

**Number of decimals:** Choose the number of decimals to display for the numerical results. Notice that you always have the possibility to view a different number of decimals afterwards, by using the Excel formatting options.

**Minimum p-value:** Enter the minimum p-value below which the p-values are replaced by “< p” where p is the minimum p-value.

**Color tabs:** Activate this option if you want to highlight the tabs produced by XLSTAT using a specific color.

**Display titles in bold:** Activate this option so that XLSTAT displays the titles of the results tables in bold.

**Empty rows after titles:** Choose the number of empty rows that must be inserted after titles. The number of empty rows after tables and charts corresponds to this same number +1.

**Number of decimals:** Choose the number of decimals to display for the numerical results. Notice that you always have the possibility to view a different number of decimals afterwards, by using the Excel formatting options.
Display table headers in bold: Activate this option to display the headers of the results tables in bold.

Display the results list in the report header: Activate this option so that XLSTAT displays the results list at the bottom of the report header.

Display the project name in the report header: Activate this option to display the name of your project in the report header. Then enter the name of your project in the corresponding field.

Enlarge the first column of the report by a factor of X: Enter the value of the factor that is used to automatically enlarge the width of the first column of the XLSTAT report. Default value is 1. When the factor is 1 the width is left unchanged.

Charts tab:

Display charts on separate sheets: Activate this option if you want that the charts are displayed on separate chart sheets. Note: when the charts are displayed on a spreadsheet you can still transform them into a chart sheet, by clicking the right button of the mouse, and then selecting "location" and then "As new sheet".

Charts size:

- Automatic: Choose this option if you want XLSTAT to automatically determine the size of the charts using as a starting value the width and height defined below.
- User defined: Activate this option if you want XLSTAT to display charts with dimensions as defined by the following values:
  - Width: Enter the value in points of the chart’s width;
  - Height: Enter the value in points of the chart’s height.

Display charts with aspect ratio equal to one: Activate this option to ensure that there is no distortion of distances due to different scales of the horizontal and vertical axes that could lead to misinterpretations.

Advanced tab:

Random numbers:

Fix the seed to: Activate this option if want to make sure that the computations involving random numbers always give the same result. Then enter the seed value.
**Maximum number of processors:** XLSTAT can run calculations on multiple processors to reduce the computing time. Choose the maximum number of processors that XLSTAT can use.

**Use NVIDIA GPUs:** GPUs stands for Graphical Processing Units. Those units are now an integral part of many devices and allow the fast computation of high quality graphics and rendering in many applications. Alternately, they can also be used as General-Purpose computing GPUs (GPGPUs) in computational intensive algorithms to do what they do best: handle massive computations at an incredible speed.

XLSTAT chose NVIDIA, the manufacturer of the most widespread and powerful GPUs, to implement on GPUs a growing number of algorithms and offer both better performances and power savings to its users. Methods with an available GPU implementation are marked with the line "GPU accelerated" in their description in the XLSTAT help.

If your device is equipped with NVIDIA GPUs and if you are using the 64 bits version of Excel, you can activate this option to enable GPUs acceleration on the supported algorithms. You should then experience significant speedups on your usual methods.

**Show the advanced buttons in the dialog boxes:** Activate this option if you want to display the buttons that allow to save or load dialog box settings, or generate VBA code to automate XLSTAT runs.

**Path for the user's files:** This path can be modified if and only if you have administrator rights on the machine. You can then modify the folder where the user’s files are saved by clicking the [...] button that will display a box where you can select the appropriate folder. User’s files include the general options as well as the options and selections of the dialog boxes of the various XLSTAT functions. The folder where the user’s files are stored must be accessible for reading and writing to all types of users.
Data sampling

Use this tool to generate a subsample of observations from a set of univariate or multivariate data.

Description

Sampling is one of the fundamental data analysis and statistical techniques. Samples are generated to:

- Test a hypothesis on one sample, and then test it on another;
- Obtain very small tables which have the properties of the original table.

To meet these different situations, several methods have been proposed. XLSTAT offers the following methods for generating a sample of N observations from a table of M rows:

**N first rows**: The sample obtained is taken from the first N rows of the initial table. This method is only used if it is certain that the values have not been sorted according to a particular criterion which could introduce bias into the analysis;

**N last rows**: The sample obtained is taken from the last N rows of the initial table. This method is only used if it is certain that the values have not been sorted according to a particular criterion which could introduce bias into the analysis;

**N every s starting at k**: The sample is built extracting N rows, every s rows, starting at row k;

**Random without replacement**: Observations are chosen at random and may occur only once in the sample;

**Random with replacement**: Observations are chosen at random and may occur several times in the sample;

**Systematic from random start**: From the j'th observation in the initial table, an observation is extracted every k observations to be used in the sample. j is chosen at random from among a number of possibilities depending on the size of the initial table and the size of the final sample. k is determined such that the observations extracted are as spaced out as possible;

**Systematic centered**: Observations are chosen systematically in the centers of N sequences of observations of length k;

**Random stratified (1)**: Rows are chosen at random within N sequences of observations of equal length, where N is determined by dividing the number of observations by the requested sample size;
Random stratified (2): Rows are chosen at random within N strata defined by the user. In each stratum, the number of sampled observations is proportional to the relative frequency of the stratum.

Random stratified (3): Rows are chosen at random within N strata defined by the user. In each stratum, the number of sampled observations is proportional to a relative frequency supplied by the user.

User defined: A variable indicates the frequency of each observation within the output sample.

Dialog box

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- ![Reload](reload.png): Click this button to reload the default options.
- ![Delete](delete.png): Click this button to delete the data selections.

**Data**: Select the data in the Excel worksheet.

**Sampling**: Choose the sampling method (see the description section for more details).

**Sample size**: Enter the size of the sample to be generated.

**Strata**: This option is only available for the random stratified sampling (2) and (3). Select in that field a column that tell to which stratum each observation belongs.

**Weight of each stratum**: This option is only available for the random stratified sampling (3). Select a table with two columns, the first containing the strata ID, and the second the weight of the stratum in the final sample. Whatever the weight unit (size, frequency, percentage), XLSTAT standardizes the weight so that the sum is equal to the requested sample size.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.
**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Variable labels**: Activate this option if the first row of the data selections (data and observations labels) includes a header.

**Observation labels**: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

**Display the report header**: Deactivate this option if you want the sampled table to start from the first row of the Excel worksheet (situation after output to a worksheet or workbook) and not after the report header. You can thus select the variables of this table by columns.

**Shuffle**: Activate this option if you want to randomly permute the output data. If this option is not activated, the sampled data respect the order of the input data.

**References**


Distribution sampling

Use this tool to generate a data sample from a continuous or discrete theoretical distribution or from an existing sample.

Description

Where a sample has been generated from a theoretical distribution, you must choose the distribution and, if necessary any parameters required for this distribution.

Distributions

XLSTAT provides the following distributions:

- **Arcsine** \((\alpha)\): the density function of this distribution (which is a simplified version of the Beta type I distribution) is given by:

\[
f(x) = \frac{\sin(\pi \alpha)}{\pi x} \left(\frac{x}{1-x}\right)^\alpha, \quad \text{with } 0 < \alpha < 1, \ x \in [0,1]
\]

We have \(E(X) = \alpha\) and \(V(X) = \alpha(1-\alpha)/2\)

- **Bernoulli** \((p)\): the density function of this distribution is given by:

\[
P(X = 1) = p, \ P(X = 0) = 1-p \quad \text{with } p \in [0,1]
\]

We have \(E(X) = p\) and \(V(X) = p(1-p)\)

The Bernoulli, named after the Swiss mathematician Jacob Bernoulli (1654-1705), allows to describe binary phenomena where only events can occur with respective probabilities of \(p\) and \(1-p\).

- **Beta** \((\alpha, \beta)\): the density function of this distribution (also called Beta type I) is given by:

\[
f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}, \quad \text{with } \alpha, \beta > 0, \ x \in [0,1] \ \text{and} \ B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}
\]

We have \(E(X) = \alpha/(\alpha+\beta)\) and \(V(X) = \alpha\beta/[(\alpha+\beta+1)(\alpha+\beta)^2]\)

- **Beta4** \((\alpha, \beta, c, d)\): the density function of this distribution is given by:
\[
f(x) = \frac{1}{B(\alpha, \beta)} (x-c)^{\alpha-1}(d-x)^{\beta-1}, \quad \text{with } \alpha, \beta > 0, \quad x \in [c, d] \\
c, d \in \mathbb{R}, \quad \text{and} \quad B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}
\]

We have \(E(X) = c+(c-d)a/(\alpha+\beta)\) and \(V(X) = (c-d)^2\alpha\beta/[(\alpha+\beta+1)(\alpha+\beta)^2]\)

Pour the type I beta distribution, \(X\) takes values in the \([0, 1]\) range. The beta4 distribution is obtained by a variable transformation such that the distribution is on a \([c, d]\) interval where \(c\) and \(d\) can take any value.

- **Beta (a, b):** the density function of this distribution (also called Beta type I) is given by:

  \[
f(x) = \frac{1}{B(a, b)} x^{a-1} (1-x)^{b-1}, \quad \text{with } a,b > 0, \quad x \in [0,1] \quad \text{and} \quad B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}
\]

  \(E(X) = a/(a+b)\) and \(V(X) = ab/[(a+b+1)(a+b)^2]\)

- **Binomial (n, p):** the density function of this distribution is given by:

  \[
P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad \text{with } x \in \mathbb{N}, \quad n \in \mathbb{N}^*, \quad p \in [0, 1]
\]

  \(E(X) = np\) and \(V(X) = np(1-p)\)

  \(n\) is the number of trials, and \(p\) the probability of success. The binomial distribution is the distribution of the number of successes for \(n\) trials, given that the probability of success is \(p\).

- **Negative binomial type I (n, p):** the density function of this distribution is given by:

  \[
P(X = x) = \binom{n-1}{x-1} p^n (1-p)^x, \quad \text{with } x \in \mathbb{N}, \quad n \in \mathbb{N}^*, \quad p \in [0, 1]
\]

  \(E(X) = n(1-p)/p\) and \(V(X) = n(1-p)/p^2\)

  \(n\) is the number of successes, and \(p\) the probability of success. The negative binomial type I distribution is the distribution of the number \(x\) of unsuccessful trials necessary before obtaining \(n\) successes.

- **Negative binomial type II (k, p):** the density function of this distribution is given by:

  \[
P(X = x) = \frac{\Gamma(k+x) p^x}{x!\Gamma(k)(1+p)^{k+x}}, \quad \text{with } x \in \mathbb{N}, \quad k, p > 0
\]

  \(E(X) = kp\) and \(V(X) = kp(p+1)\)
The negative binomial type II distribution is used to represent discrete and highly heterogeneous phenomena. As \( k \) tends to infinity, the negative binomial type II distribution tends towards a Poisson distribution with \( \lambda = kp \).

- Chi-square (\( df \)): the density function of this distribution is given by:

\[
 f(x) = \frac{(1/2)^{df/2}}{\Gamma(df/2)} x^{df/2-1} e^{-x/2}, \quad \text{with } x > 0, \ df \in \mathbb{N}^*
\]

\( E(X) = df \) and \( V(X) = 2df \)

The Chi-square distribution corresponds to the distribution of the sum of \( df \) squared standard normal distributions. It is often used for testing hypotheses.

- Erlang (\( k, \lambda \)): the density function of this distribution is given by:

\[
 f(x) = \lambda^k x^{k-1} e^{-\lambda x} \frac{1}{(k-1)!}, \quad \text{with } x \geq 0 \text{ and } k, \lambda > 0 \text{ and } k \in \mathbb{N}
\]

\( E(X) = k/\lambda \) and \( V(X) = k/\lambda^2 \)

\( k \) is the shape parameter and \( \lambda \) is the rate parameter.

This distribution, developed by the Danish scientist A. K. Erlang (1878-1929) when studying the telephone traffic, is more generally used in the study of queuing problems.

Note: When \( k=1 \), this distribution is equivalent to the exponential distribution. The Gamma distribution with two parameters is a generalization of the Erlang distribution to the case where \( k \) is a real and not an integer (for the Gamma distribution the scale parameter \( \beta = 1/\lambda \) is used).

- Exponential (\( \lambda \)): the density function of this distribution is given by:

\[
 f(x) = \lambda \exp(-\lambda x), \quad \text{with } x > 0 \text{ and } \lambda > 0
\]

\( E(X) = 1/\lambda \) and \( V(X) = 1/\lambda^2 \)

The exponential distribution is often used for studying lifetime in quality control.

- Fisher (\( df_1, df_2 \)): the density function of this distribution is given by:

\[
 f(x) = \frac{1}{xB(df_1/2,df_2/2)} \left( \frac{df_1 x}{df_1 x + df_2} \right)^{df_1/2} \left( 1 - \frac{df_1 x}{df_1 x + df_2} \right)^{df_2/2},
\]

\( \text{with } x \geq 0 \text{ and } df_1, df_2 \in \mathbb{N}^* \)

\( E(X) = df_2/(df_2 - 2) \) if \( df_2 > 0 \), and \( V(X) = 2df_2^2(df_1+df_2 - 2)/[(df_1(df_2-2)^2)(df_2 - 4)] \)
Fisher’s distribution, from the name of the biologist, geneticist and statistician Ronald Aylmer Fisher (1890-1962), corresponds to the ratio of two Chi-square distributions. It is often used for testing hypotheses.

- Fisher-Tippett ($\beta, \mu$): the density function of this distribution is given by:
  \[
  f(x) = \frac{1}{\beta} \exp\left( -\frac{x-\mu}{\beta} - \exp\left( -\frac{x-\mu}{\beta} \right) \right), \quad \text{with} \ \beta > 0
  \]
  \[
  E(X) = \mu + \beta \gamma \quad \text{and} \quad V(X) = (\pi \beta)^2 / 6 \quad \text{where} \ \gamma \text{is the Euler-Mascheroni constant.}
  \]
  The Fisher-Tippett distribution, also called the Log-Weibull or extreme value distribution, is used in the study of extreme phenomena. The Gumbel distribution is a special case of the Fisher-Tippett distribution where $\beta=1$ and $\mu=0$.

- Gamma ($k, \beta, \mu$): the density of this distribution is given by:
  \[
  f(x) = (x-\mu)^{k-1} \frac{e^{-(x-\mu)/\beta}}{\beta^k \Gamma(k)}, \quad \text{with} \ x > \mu \ \text{and} \ k, \beta > 0
  \]
  \[
  E(X) = \mu + k \beta \quad \text{and} \quad V(X) = k \beta^2
  \]
  $k$ is the shape parameter of the distribution and $\beta$ the scale parameter.

- GEV ($\beta, k, \mu$): the density function of this distribution is given by:
  \[
  f(x) = \frac{1}{\beta} \left( 1 - k \frac{x-\mu}{\beta} \right)^{1/k-1} \exp\left( -\left( 1 - k \frac{x-\mu}{\beta} \right)^{1/k} \right), \quad \text{with} \ \beta > 0
  \]
  We have \( E(X) = \mu + \frac{\beta}{k} \Gamma(1+k) \) and \( V(X) = \left( \frac{\beta}{k} \right)^2 \left( \Gamma(1+2k) - \Gamma^2(1+k) \right) \)
  The GEV (Generalized Extreme Values) distribution is much used in hydrology for modeling flood phenomena. $k$ lies typically between -0.6 and 0.6.

- Gumbel: the density function of this distribution is given by:
  \[
  f(x) = \exp\left( -x - \exp\left( -x \right) \right)
  \]
  \[
  E(X) = \gamma \quad \text{and} \quad V(X) = \pi^2 / 6 \quad \text{where} \ \gamma \text{is the Euler-Mascheroni constant (0.5772156649…).}
  \]
  The Gumbel distribution, named after Emil Julius Gumbel (1891-1966), is a special case of the Fisher-Tippett distribution with $\beta=1$ and $\mu=0$. It is used in the study of extreme phenomena such as precipitations, flooding and earthquakes.

- Logistic ($\mu, s$): the density function of this distribution is given by:
\[ f(x) = \frac{e^{-\frac{(x-\mu)}{s}}}{s \left(1 + e^{-\frac{(x-\mu)}{s}}\right)^2}, \quad \text{with } \mu \in R, \text{ and } s > 0 \]

We have \( E(X) = \mu \) and \( V(X) = (\pi s^3)/3 \)

- Lognormal \((\mu, \sigma)\): the density function of this distribution is given by:
  \[ f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x) - \mu)^2}{2\sigma^2}}, \quad \text{with } x, \sigma > 0 \]
  \( E(X) = \exp(\mu + \sigma^2/2) \) and \( V(X) = [\exp(\sigma^2)-1]\exp(2\mu + \sigma^2) \)

- Lognormal2 \((m, s)\): the density function of this distribution is given by:
  \[ f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x) - \mu)^2}{2\sigma^2}}, \quad \text{with } x, \sigma > 0 \]
  \[ \mu = \ln(m) - \ln(1+s^2/m^2)/2 \] and \( \sigma^2 = \ln(1+s^2/m^2) \)
  \( E(X) = m \) and \( V(X) = s^2 \)

  This distribution is just a reparametrization of the Lognormal distribution.

- Normal \((\mu, \sigma)\): the density function of this distribution is given by:
  \[ f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad \text{with } \sigma > 0 \]
  \( E(X) = \mu \) and \( V(X) = \sigma^2 \)

- Standard normal: the density function of this distribution is given by:
  \[ f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \]
  \( E(X) = 0 \) and \( V(X) = 1 \)

  This distribution is a special case of the normal distribution with \( \mu=0 \) and \( \sigma=1 \).

- Pareto \((a, b)\): the density function of this distribution is given by:
  \[ f(x) = \frac{ab^a}{x^{a+1}}, \quad \text{with } a, b > 0 \text{ and } x \geq b \]
  \( E(X) = abl/(a-1) \) and \( V(X) = ab^2/[(a-1)^2(a-2)] \)
The Pareto distribution, named after the Italian economist Vilfredo Pareto (1848-1923), is also known as the Bradford distribution. This distribution was initially used to represent the distribution of wealth in society, with Pareto's principle that 80% of the wealth was owned by 20% of the population.

- **PERT** (a, m, b): the density function of this distribution is given by:

  \[ f(x) = \frac{1}{B(\alpha, \beta)} \frac{(x-a)^{\alpha-1} (b-x)^{\beta-1}}{(b-a)^{\alpha+\beta-1}}, \quad \text{with } \alpha, \beta > 0, \ x \in [a, b] \]

  \[ a, \ b \in \mathbb{R}, \ \text{and} \ B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} \]

  \[ \alpha = \frac{4m + b - 5a}{b - a} \]

  \[ \beta = \frac{5b - a - 4m}{b - a} \]

  We have \( E(X) = \frac{(b-a)\alpha}{(\alpha+\beta)} \) and \( V(X) = \frac{(b-a)^2\alpha\beta}{[(\alpha+\beta+1)(\alpha+\beta)]^2} \)

  The PERT distribution is a special case of the beta4 distribution. It is defined by its definition interval \([a, b]\) and \(m\) the most likely value (the mode). PERT is an acronym for **Program Evaluation and Review Technique**, a project management and planning methodology. The PERT methodology and distribution were developed during the project held by the US Navy and Lockheed between 1956 and 1960 to develop the Polaris missiles launched from submarines. The PERT distribution is useful to model the time that is likely to be spent by a team to finish a project. The simpler triangular distribution is similar to the PERT distribution in that it is also defined by an interval and a most likely value.

- **Poisson** (\(\lambda\)): the density function of this distribution is given by:

  \[ P(X = x) = \frac{\exp(-\lambda)\lambda^x}{x!}, \quad \text{with } x \in \mathbb{N} \text{ and } \lambda > 0 \]

  \( E(X) = \lambda \) and \( V(X) = \lambda \)

  Poisson's distribution, discovered by the mathematician and astronomer Siméon-Denis Poisson (1781-1840), pupil of Laplace, Lagrange and Legendre, is often used to study queuing phenomena.

- **Student** (\(df\)): the density function of this distribution is given by:

  \[ f(x) = \frac{\Gamma\left(\left(\frac{df}{2}\right)\right)}{\sqrt{\pi df} \Gamma\left(\frac{df}{2}\right)} \left(1 + \frac{x^2}{df}\right)^{-\left(df+1\right)/2}, \quad \text{with } df > 0 \]
E(X) = 0 if df>1 and V(X) = df/(df - 2) if df>2

The English chemist and statistician William Sealy Gosset (1876-1937), used the nickname Student to publish his work, in order to preserve his anonymity (the Guinness brewery forbade its employees to publish following the publication of confidential information by another researcher). The Student's t distribution is the distribution of the mean of df variables standard normal variables. When df=1, Student's distribution is a Cauchy distribution with the particularity of having neither expectation nor variance.

- Trapezoidal (a, b, c, d): the density function of this distribution is given by:

\[
\begin{align*}
  f(x) &= \begin{cases} 
    \frac{2(x-a)}{(d+c-b-a)(b-a)}, & x \in [a,b] \\
    \frac{2}{(d+c-b-a)}, & x \in [b,c] \\
    \frac{2(d-x)}{(d+c-b-a)(d-c)}, & x \in [a,b] \\
    0, & x < a, x > d
  \end{cases}
\end{align*}
\]

with \(a < m < b\)

We have \(E(X) = (d^2+c^2-b^2-a^2+cd-ab)/(3(d+c-b-a))\)

and \(V(X) = [(c+d)(c^2+d^2)-(a+b)(a^2+b^2)]/(6(d+c-b-a))-E^2(X)\)

This distribution is useful to represent a phenomenon for which we know that it can take values between two extreme values (a and d), but that it is more likely to take values between two values (b and c) within that interval.

- Triangular (a, m, b): the density function of this distribution is given by:

\[
\begin{align*}
  f(x) &= \begin{cases} 
    \frac{2(x-a)}{(b-a)(m-a)}, & x \in [a,m] \\
    \frac{2(b-x)}{(b-a)(b-m)}, & x \in [m,b] \\
    0, & x < a, x > b
  \end{cases}
\end{align*}
\]

with \(a < m < b\)

We have \(E(X) = (a+m+b)/3\) and \(V(X) = (a^2+m^2+b^2-ab-am-bm)/18\)

- TriangularQ (q1, m, q2, p1, p2): the density function of this distribution is a reparametrization of the Triangular distribution. A first step requires estimating the a and b parameters of the triangular distribution, from the q1 and q2 quantiles to which
percentages p1 and p2 correspond. Once this is done, the distribution functions can be computed using the triangular distribution functions.

- **Uniform (a, b):** the density function of this distribution is given by:
  \[
  f(x) = \frac{1}{b-a}, \text{ with } b > a \text{ and } x \in [a, b]
  \]
  \[
  E(X) = \frac{(a+b)}{2} \text{ and } V(X) = \frac{(b-a)^2}{12}
  \]
  The uniform (0,1) distribution is much used for simulations. As the cumulative distribution function of all the distributions is between 0 and 1, a sample taken in a Uniform (0,1) distribution is used to obtain random samples in all the distributions for which the inverse can be calculated.

- **Uniform discrete (a, b):** the density function of this distribution is given by:
  \[
  f(x) = \frac{1}{b-a+1}, \text{ with } b > a, (a,b) \in N, x \in N, x \in [a, b]
  \]
  We have \( E(X) = \frac{(a+b)}{2} \) and \( V(X) = \frac{[(b-a+1)^2 -1]}{12} \)
  The uniform discrete distribution corresponds to the case where the uniform distribution is restricted to integers.

- **Weibull \((\beta)\):** the density function of this distribution is given by:
  \[
  f(x) = \beta x^{\beta - 1} \exp\left(-x^\beta\right), \text{ with } x > 0 \text{ and } \beta > 0
  \]
  We have \( E(X) = \Gamma\left(\frac{1}{\beta} + 1\right) \) and \( V(X) = \Gamma\left(\frac{2}{\beta} + 1\right) - \Gamma^2\left(\frac{1}{\beta} + 1\right) \)
  \( \beta \) is the shape parameter for the Weibull distribution.

- **Weibull \((\beta, \gamma)\):** the density function of this distribution is given by:
  \[
  f(x) = \frac{\beta}{\gamma} \left(\frac{x}{\gamma}\right)^{\beta - 1} e^{-\left(x/\gamma\right)^\beta}, \text{ with } x > 0, \text{ and } \beta, \gamma > 0
  \]
  We have \( E(X) = \gamma \Gamma\left(\frac{1}{\beta} + 1\right) \) and \( V(X) = \gamma^2 \left[\Gamma\left(\frac{2}{\beta} + 1\right) - \Gamma^2\left(\frac{1}{\beta} + 1\right)\right] \)
  \( \beta \) is the shape parameter of the distribution and \( \gamma \) the scale parameter. When \( \beta=1 \), the Weibull distribution is an exponential distribution with parameter \( 1/\gamma \).

- **Weibull \((\beta, \gamma, \mu)\):** the density function of this distribution is given by:
\[ f(x) = \beta \left( \frac{x - \mu}{\gamma} \right)^{\beta-1} e^{-\left(\frac{x - \mu}{\gamma}\right)^{\beta}}, \text{ with } x > \mu, \text{ and } \beta, \gamma > 0 \]

We have \( E(X) = \mu + \gamma \Gamma\left(\frac{1}{\beta} + 1\right) \) and \( V(X) = \gamma^2 \left[ \Gamma\left(\frac{2}{\beta} + 1\right) - \Gamma^2\left(\frac{1}{\beta} + 1\right) \right] \)

The Weibull distribution, named after the Swede Ernst Hjalmar Waloddi Weibull (1887-1979), is much used in quality control and survival analysis. \( \beta \) is the shape parameter of the distribution and \( \gamma \) the scale parameter. When \( \beta=1 \) and \( \mu=0 \), the Weibull distribution is an exponential distribution with parameter \( 1/\gamma \).

**Dialog box**

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- : Click this button to reload the default options.
- : Click this button to delete the data selections.

**Theoretical distribution**: Activate this option to sample data in a theoretical distribution. Then choose the distribution and enter any parameters required by the distribution.

**Empirical Distribution**: Activate this option to sample data in an empirical distribution. Then select the data required to build the empirical distribution.

- **Column labels**: Activate this option if the first row of the selected data (data and weights) contains a label.
- **Weights**: Activate this option if the observations are weighted. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Column labels" option is activated.
Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Number of samples: Enter the number of samples to be generated.

Sample size: Enter the number of values to generate for each of the samples.

Display the report header: Deactivate this option if you want the table of sampled values to start from the first row of the Excel worksheet (situation after output to a worksheet or workbook) and not after the report header.

Example

An example showing how to generate a random normal sample is available on the Addinsoft website:

http://www.xlstat.com/demo-norm.htm

References


Variables transformation

Use this tool to quickly apply simple transformations to a set of variables.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload**: Click this button to reload the default options.
- **Delete Selections**: Click this button to delete the data selections.
- **Change Data Handling**: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab:**

**Data**: Select the data in the Excel worksheet. If headers have been selected, check that the "Column labels" option has been activated.

**Column labels**: Activate this option if the first row of the data selected (data and coding table) contains a label.

**Observation labels**: Check this option if you want to use the observation labels. If you do not check this option, labels will be created automatically (Obs1, Obs2, etc.). If a column header has been selected, check that the "Column labels" option has been activated.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.
**Sheet:** Activate this option to display the results in a new worksheet of the active workbook.

**Workbook:** Activate this option to display the results in a new workbook.

**Display the report header:** Deactivate this option if you want the results table to start from the first row of the Excel worksheet (situation after output to a worksheet or workbook) and not after the report header.

**Transformation:** Choose the transformation to apply to the data.

- **Standardize (n-1):** Choose this option to standardize the variables using the unbiased standard deviation.

- **Other:** Choose this option to use another transformation. Then click on the “Transformations” tab to choose the transformation to apply.

**Transformations** tab:

**Standardize (n):** Choose this option to standardize the variables using the biased standard deviation.

**Center:** Choose this option to center the variables.

/ **Standard deviation (n-1):** Choose this option to divide the variables by their unbiased standard deviation.

/ **Standard deviation (n):** Choose this option to divide the variables by their biased standard deviation.

**Rescale from 0 to 1:** Choose this option to rescale the data from 0 to 1.

**Rescale from 0 to 100:** Choose this option to rescale the data from 0 to 100.

**Binarize (0/1):** Choose this option to convert all values that are not 0 to 1, and leave the 0s unchanged.

**Sign (-1/0/1):** Choose this option to convert all values that are negative to -1, all positive values to 1, and leave the 0s unchanged.

**Arcsin:** Choose this option to transform the data to their arc-sine.
**Box-Cox transformation**: Activate this option to improve the normality of the sample; the Box-Cox transformation is defined by the following equation:

\[
Y_i = \begin{cases} 
\frac{X_i^\lambda - 1}{\lambda}, & (X_i > 0, \ \lambda \neq 0) \text{ or } (X_i \geq 0, \ \lambda > 0) \\
\ln(X_i), & X_i > 0, \ \lambda = 0 
\end{cases}
\]

XLSTAT accepts a fixed value of $\lambda$, or it can find the value that maximizes the likelihood of the sample, assuming the transformed sample follows a normal distribution.

**Winsorize**: Choose this transformation to remove data that are not within an interval defined by two percentiles: let $p_1$ and $p_2$ be two values comprised between 0 and 1, such that $p_1 < p_2$. If a value $x$ from the sample is lower than $q_1$, the quantile that corresponds to $p_1$ obtained from the sample, or greater than $q_2$ the quantile that corresponds to $p_2$, then the value is transformed to $q_1$ in the first case, or to $q_2$ in the second case.

**Missing data** tab:

**Do not accept missing data**: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

**Remove observations**: Activate this option to remove the observations that contain missing data.

**Ignore missing data**: Activate this option to ignore missing data.

**Estimate missing data**: Activate this option to estimate the missing data by using the mean of the variables.
Missing data

Use this tool to treat missing values prior to using other XLSTAT analyses.

Description

Most of XLSTAT functions include a missing value treatment. However, only few approaches are available. This tool allows you to complete or clean your dataset using advanced missing value treatment.

There are three types of missing value (Allison, 2001): data missing completely at random (MCAR), data missing at random (MAR) and data not missing at random (NMAR).

Data is missing completely at random (MCAR) if the event that leads to a missing data is independent of observable variables and of unobservable parameters. It should occur entirely at random. When data are MCAR, the analyses performed on the data are unbiased.

Data is missing at random (MAR) when the event that leads to a missing data is related to a particular variable, but it is not related to the value of the variable that has missing data. This is the most common case.

Data is not missing at random (NMAR) when data is missing for a particular reason. An example of this is the filtered questions in a questionnaire (the question is only intended for some respondents, the others are missing).

The methods available in this tool correspond to the MCAR and MAR cases.

Different methods are available depending on your needs and data:

- For quantitative data, XLSTAT allows you to:
  - Remove observations with missing value.
  - Use a mean imputation method.
  - Use a nearest neighbor approach.
  - Use the NIPALS algorithm.
  - Use an MCMC multiple imputation algorithm.

- For qualitative data, XLSTAT allows you to:
  - Remove the observations with missing value.
  - Use a mode imputation method.
Use a nearest neighbor approach.

**NIPALS algorithm**

The NIPALS method is a method presented by H. Wold (1973) to allow principal component analysis with missing values. The NIPALS algorithm is applied on the dataset and the obtained PCA model is used to predict the missing values.

**Multiple imputations**

XLSTAT proposes a multiple imputation algorithm based on the Markov Chain Monte Carlo (MCMC) approach also called fully conditional specification (Van Buuren, 2007).

The algorithm works as follows:

1. Initial values of the missing values are obtained sampled from a normal distribution with mean and standard error equal to the mean and standard error obtained on available data.

2. For each variable of the dataset with missing values, an imputation method based on sampling and OLS regression is applied. The used model is a regression model with the studied variable as dependent variable and all the other variables as independent variables. Disturbance using data sampled from different distributions are also used. New imputed values are obtained using this model.

These two steps are repeated until the number of imputation is reached. The average value of each imputed missing value is taken.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
: Click this button to reload the default options.

: Click this button to delete the data selections.

: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

**Data**: Select your dataset. If a column header has been selected, check that the "Variable labels" option has been activated.

**Type of data**: Select the type of data you wish to complete.

**Method**: Select the method for missing data treatment to be used.

**Number of imputation**: If you selected the multiple imputation method, enter the number of imputation to be used.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the options in a new workbook.

**Variable labels**: Activate this option if the first row (column mode) or first column (rows mode) of the selected data contain labels.

**Observation labels**: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

**Weights**: Activate this option if you want to carry out a weighted least squares regression. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.
**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the selected samples before and after imputation.

**Results**

The first table displays the descriptive statistics associated with the data. If an imputation method has been applied, a second descriptive statistics table will appear with the statistics associated with the data after imputation.

Then, the data table is displayed. Imputed missing values are shown in bold.

**Example**

An example showing how to run apply the multiple imputation method is available at:

[http://www.xlstat.com/demo-missing.htm](http://www.xlstat.com/demo-missing.htm)

**References**


Raking a survey

Use this tool to compute raking weights using supplementary qualitative variables.

Description

When working on surveys, one might need to rake the surveyed sample. That often means looking for raking weights. Sometimes, a survey sample may cover a segment of the population with different proportions than in the population itself. The basic idea is to find weights applied to the observations that will give the same proportions in the survey sample as in the population. To do that, we need to use auxiliary variables that are measured both on the survey sample and on the population.

Four methods are available in XLSTAT:

- The raking ratio method (Deming and Stefan, 1940).
- The logit method: It is close to the raking ratio but lower and upper bounds for the weights are used. These bounds have to be specified by the user.
- The linear method.
- The truncated linear method: It is the linear method with bounds for the final weights.

The weights are calculated using an algorithm called the generalized raking procedure developed by Deville, Särndall and Sautory (1993). The difference from one method to another is the function to be optimized.

Raking weights computation

XLSTAT lets you use four methods to obtain raking weights. Classical raking ratio is one of them.

Let $S$ be a sample with $n$ observations and $p$ qualitative variables called auxiliary. Let $x_{ij}$ be the value of the $i$th observation for the $j$th variable. Let $X_j$ be the marginal control totals in the population for the $j$th variable. Let $d_i$ be the initial weights before raking and $w_i$ the final weights after raking.

We are looking for weights $w_i$ close to $d_i$ satisfying the following equations:

$$\sum_{k=1}^{n} w_{ik} x_{kj} = X_j \quad \forall j = 1, ..., p$$

We choose a distance function $G(u)$ with $u = w_i / d_i$, where this function has to be convex and positive. We have an optimization problem that can be solved using Lagrange multipliers method (lambda).
The problem is the following:

\[
\min_{w_k} \sum_{k=1}^{n} d_k G\left(\frac{w_k}{d_k}\right)
\]

s.c. \[\sum_{k=1}^{n} w_k x_k = X\]

With Lagrangian equal to:

\[
L = \sum_{k=1}^{n} d_k G\left(\frac{w_k}{d_k}\right) - \lambda \left(\sum_{k=1}^{n} w_k x_k - X\right)
\]

We have:

\[
w_k = d_k F(x'_k \lambda)
\]

\[
\sum_{k=1}^{n} d_k F(x'_k \lambda) x_k = X
\]

With:

\[
F(u) = g^{-1}(u) \quad \text{with} \quad g(u) = \frac{dG(u)}{du}
\]

A numerical method like Newton’s method can be used to solve this problem.

The F() functions are:

- For the raking ratio method, we have:
  \[F(u) = \exp(u)\]

- For the logit method, we have (with a lower bound L and an upper bound U):
  \[F(u) = \frac{L(U-1) + U(1-L)\exp(Au)}{U-1 + (1-L)\exp(Au)}\]
  \[A = \frac{U-L}{(1-L)(U-1)}\]

- For the linear method, we have:
  \[F(u) = 1 + u\]

- For the truncated linear method, we have:
  \[F(u) = 1 + u \in [L; U]\]

The algorithm is iterative and continues until the resulting weights are stable using the following criterion:
$$\max_{k=1}^n \left| \frac{w_k^{(i+1)}}{d_k} - \frac{w_k^{(i)}}{d_k} \right| < \varepsilon$$

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload defaults**: Click this button to reload the default options.
- **Delete data**: Click this button to delete the data selections.
- **Change data format**: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab:**

**Data to be raked on**: Select the auxiliary qualitative variables. If headers have been selected, check that the "Sample labels" option has been activated.

**Marginal control totals**: Select the marginal control totals for all the variables to be raked on. They should be represented in column with one row for each modality of the qualitative variable. Each column should sum to the same total. Columns have to be selected in the same orders as the variables to be raked on.

**Format:**
- **Values**: Activate this option if the marginal control totals represent real value in the global population.

- **Percentages**: Activate this option if the marginal control totals are percentages of the global population. If this option is selected, you have to indicate the population size.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Sample labels**: Activate this option if the first row of the selected data (data, sub-samples, weights) contains a label.

**Initial weights**: Check this option if the observations are initially weighted. If you do not check this option, the weights will be considered as N/n. Weights must be greater than 0. If a column header has been selected, check that the "Sample labels" option is activated.

**Observation labels**: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 ...).

**Options** tab:

**Estimation method**: Select the method you want to use. For logit and truncated linear, you have to give the bounds (lower bound < 1 and upper bound > 1).

**Stop conditions**:

- **Iterations**: Enter the maximum number of iterations for the algorithm. The calculations are stopped when the maximum number of iterations has been exceeded. Default value: 50.

- **Convergence**: Enter the maximum value of the evolution in the convergence criterion from one iteration to another which, when reached, means that the algorithm is considered to have converged. Default value: 0.00001.

**Missing data** tab:
Remove observations: Activate this option to ignore the observations that contain missing data.

Estimate missing data: Activate this option to estimate the missing data by using the mode of the variables.

Outputs tab:

Descriptive statistics: Activate this option to display the descriptive statistics of the sample before and after raking.

Show data in output: Activate this option to display the initial auxiliary variable in the final weights table.

Weights ratio: Activate this option to display the weights ratio for each observation in the final weights table.

List of combines: Activate this option to display the table with the list of all combines of the categories and their associated frequency and weights ratio.

Details of iterations: Activate this option to display the table with the details for each iteration of the algorithm (Lagrange multipliers and stopping criterion).

Results

Summary statistics (before raking): This table displays for each modality of the auxiliary variables, the frequency and the percentages in the sample and in the population using marginal control totals.

Final weights: This table displays final raked weights. If the corresponding options are selected, initial data and weights ratios are also displayed.

Summary statistics (after raking): This table displays for each modality of the auxiliary variables, the frequency and the percentages in the sample with final weighting, and in the population using marginal control totals.

List of combines: This table displays all the combines of the categories of the auxiliary variables with their frequency and their weights ratio.

Details of iterations: This table displays the details for each iteration with the Lagrange multipliers and the stopping criterion.
Example

An example showing how to rake a sample is available on the Addinsoft website at:

http://www.xlstat.com/demo-raking.htm

References


Create a contingency table

Use this tool to create a contingency table from two or more qualitative variables. A chi-square test is optionally performed.

Description

A contingency table is an efficient way to summarize the relation (or correspondence) between two categorical variables V1 and V2. It has the following structure:

<table>
<thead>
<tr>
<th>V1 \ V2</th>
<th>Category 1</th>
<th>...</th>
<th>Category j</th>
<th>...</th>
<th>Category m2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category 1</td>
<td>n(1,1)</td>
<td>...</td>
<td>n(1,j)</td>
<td>...</td>
<td>n(1,m2)</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Category i</td>
<td>n(i,1)</td>
<td>...</td>
<td>n(i,j)</td>
<td>...</td>
<td>n(i,m2)</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Category m1</td>
<td>n(m1,1)</td>
<td>...</td>
<td>n(m1,j)</td>
<td>...</td>
<td>n(m1,m2)</td>
</tr>
</tbody>
</table>

where $n(i,j)$ is the frequency of observations that show both characteristic i for variable V1, and characteristic j for variable V2.

To create a contingency table from two qualitative variables V1 and V2, the first transformation consists of recoding the two qualitative variables V1 and V2 as two disjunctive tables Z1 and Z2 or indicator (or dummy) variables. For each category of a variable there is a column in the respective disjunctive table. Each time the category c of variable V1 occurs for an observation i, the value of $Z1(i,c)$ is set to one (the same rule is applied to the V2 variable). The other values of Z1 and Z2 are zero. The contingency table of the two variables is the table $Z1'Z2$ (where ' indicates matrix transpose).

The Chi-square distance has been suggested to measure the distance between two categories. The Pearson chi-square statistic, which is the sum of the Chi-square distances, is used to test the independence between rows and columns. It has asymptotically a Chi-square distribution with $(m1-1)(m2-1)$ degrees of freedom.

Inertia is a measure inspired from physics that is often used in Correspondence Analysis, a method that is used to analyse in depth contingency tables. The inertia of a set of points is the weighted mean of the squared distances to the center of gravity. In the specific case of a
contingency table, the total inertia of the set of points (one point corresponds to one category) can be written as:

$$
\phi^2 = \frac{X^2}{n} = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \left( \frac{n_{ij} - \frac{n_i n_j}{n^2}}{n} \right)^2,
$$

with $n_i = \sum_{j=1}^{m_2} n_{ij}$ and $n_j = \sum_{i=1}^{m_1} n_{ij}$

and where $n$ is the sum of the frequencies in the contingency table. We can see that the inertia is proportional to the Pearson chi-square statistic computed on the contingency table.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- ![Reload](grid.png): Click this button to reload the default options.
- ![Delete](grid.png): Click this button to delete the data selections.
- ![Change](grid.png): Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab**:

**Row variable(s)**: Select the data that correspond to the variable(s) that will be used to construct the rows of the contingency table(s).
**Column variable(s):** Select the data that correspond to the variable(s) that will be used to construct the columns of the contingency table(s).

**Range:** Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Activate this option to display the results in a new worksheet of the active workbook.

**Workbook:** Activate this option to display the results in a new workbook.

**Variable labels:** Activate this option if the first row of the data selections (row and column variables, weights) includes a header.

**Weights:** Activate this option if the observations are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

**Options tab:**

**Sort the categories alphabetically:** Activate this option so that the categories of all the variables are sorted alphabetically.

**Variable-Category labels:** Activate this option to create the labels of the contingency table using both the variable name of the name of the categories. If the option is not activated, the labels are only based on the categories.

**Chi-square test:** Activate this option to display the statistics and the interpretation of the Chi-square test of independence between rows and columns.

**Significance level (%):** Enter the significance level for the test.

**Missing data tab:**

**Do not accept missing data:** Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

**Remove observations:** Activate this option to ignore the observations that contain missing data.

**Group missing values into a new category:** Activate this option to group missing data into a new category of the corresponding variable.
**Outputs** tab:

**List of combines**: Activate this option to display the table that lists all the possible combines between the two variables that are used to create a contingency table, and the corresponding frequencies.

**Contingency table**: Activate this option to display the contingency table.

**Inertia by cell**: Activate this option to display the inertia for each cell of the contingency table.

**Chi-square by cell**: Activate this option to display the contribution to the chi-square of each cell of the contingency table.

**Significance by cell**: Activate this option to display a table indicating, for each cell, if the actual value is equal (=), lower (<) or higher (>) than the theoretical value, and to run a test (Fisher’s exact test of on a 2x2 table having the same total frequency as the complete table, and the same marginal sums for the cell of interest), in order to determine if the difference with the theoretical value is significant or not.

**Observed frequencies**: Activate this option to display the table of the observed frequencies. This table is almost identical to the contingency table, except that the marginal sums are also displayed.

**Theoretical frequencies**: Activate this option to display the table of the theoretical frequencies computed using the marginal sums of the contingency table.

**Proportions or percentages / Row**: Activate this option to display the table of proportions or percentages computed by dividing the values of the contingency table by the marginal sums of each row.

**Proportions or percentages / Column**: Activate this option to display the table of proportions or percentages computed by dividing the values of the contingency table by the marginal sums of each column.

**Proportions or percentages / Total**: Activate this option to display the table of proportions or percentages computed by dividing the values of the contingency table by the sum of all the cells of the contingency table.

**Charts** tab:

**3D view of the contingency table**: Activate this option to display the 3D bar chart corresponding to the contingency table.
Full disjunctive tables

Use this tool to create a full disjunctive table from one or more qualitative variables.

Description

A disjunctive table is a drill-down of a table defined by n observations and q qualitative variables into a table defined by n observations and p indicators where p is the sum of the numbers of categories of the q variables: each variable Q(j) is broken down into a sub-table with q(j) columns where column k contains 1's for observations corresponding to the k'th category and 0 for the other observations.

Dialog box

Click this button to start the computations.

Click this button to close the dialog box without doing any computation.

Click this button to display the help.

Click this button to reload the default options.

Click this button to delete the data selections.

Data: Select the data in the Excel worksheet. If headers have been selected, check that the “Variable labels” option has been activated.

Variable labels: Check this option if the first line of the selected data contains a label.

Observation labels: Check this option if you want to use the available line labels. If you do not check this option, line labels will be created automatically (Obs1, Obs2, etc.). If a column header has been selected, check that the “Variable labels” option has been activated.
**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Descriptive statistics**: Activate this option to display descriptive statistics for the variables selected.

**Display the report header**: Deactivate this option if you want the full disjunctive table to start from the first row of the Excel worksheet (situation after output to a worksheet or workbook) and not after the report header.

**Example**

Input table:

<table>
<thead>
<tr>
<th></th>
<th>Q1</th>
<th>Q2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obs1</td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>Obs2</td>
<td>B</td>
<td>D</td>
</tr>
<tr>
<td>Obs3</td>
<td>B</td>
<td>E</td>
</tr>
<tr>
<td>Obs4</td>
<td>A</td>
<td>D</td>
</tr>
</tbody>
</table>

Full disjunctive table:

<table>
<thead>
<tr>
<th></th>
<th>Q1-A</th>
<th>Q1-B</th>
<th>Q2-C</th>
<th>Q2-D</th>
<th>Q2-E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obs1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Obs2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Obs3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Obs4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
**Discretization**

Use this tool to discretize a numerical variable. Several discretization methods are available.

---

**Description**

Discretizing a numerical variable means transforming it into an ordinal variable. This process is used a lot in marketing where it is often referred to as segmentation.

XLSTAT makes available several discretization methods that are more or less automatic. The number of classes (or intervals, or segments) to generate is either fixed by the user (for example with the method of equal ranges), or by the method itself (for example, with the 80-20 option where two classes are created).

The Fisher’s classification algorithm can be very slow when the size of dataset exceeds 1000. This method generates a number of classes that is lower or equal to the number of classes requested by the user, as the algorithm is able to automatically merge similar classes.

---

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- ****: Click this button to reload the default options.
- ****: Click this button to delete the data selections.
- ****: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If
the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

**Observations/variables table**: Select a table comprising N objects described by P descriptors. If column headers have been selected, check that the "Variable labels" option has been activated.

**Method**: Select the discretization method:

- **Constant range**: Choose this method to create classes that have the same range. Then enter the value of the range. You can optionally specify the "minimum" that corresponds to the lower bound of the first interval. This value must be lower or equal to the minimum value of the series. If the minimum is not specified, the lower bound will be set to the minimum value of the series.

- **Intervals**: Use this method to create a given number of intervals with the same range. The range of the intervals is determined by the difference between the maximum and minimum values of the series. You can optionally specify the "minimum" that corresponds to the lower bound of the first interval. This value must be lower or equal to the minimum value of the series. If the minimum is not specified, the lower bound will be set to the minimum value of the series.

- **Equal frequencies**: Choose this method so that all the classes contain as much as possible the same number of observations. Then, enter the number of intervals (or classes) to generate.

- **Automatic (Fisher)**: Use this method to create the classes using the Fisher's algorithm. When the size of dataset exceeds 1000, the computations can be very slow. You need to enter the number of intervals (or classes) to generate. However, this method generates a number of classes that is lower or equal to the number of classes required by the user, as the algorithm is able to automatically merge similar classes.

- **Automatic (k-means)**: Choose this method to create classes (or intervals) using the k-means algorithm. Then, enter the number of intervals (or classes) to generate.

- **Intervals (user defined)**: Choose this option to select a column containing in increasing order the lower bound of the first interval, and the upper bound of all the intervals.

- **80-20**: Use this method to create two classes, the first containing the 80 first % of the series, the data being sorted in increasing order, the second containing the remaining 20%.
- **20-80**: Use this method to create two classes, the first containing the 20 first % of the series, the data being sorted in increasing order, the second containing the remaining 80%.

- **80-15-5 (ABC)**: Use this method to create two classes, the first containing the 80 first % of the series, the data being sorted in increasing order, the second containing the next 15%, and the third containing the remaining 5%. This method is sometimes referred to as “ABC classification”.

- **5-15-80**: Use this method to create two classes, the first containing the 5 first % of the series, the data being sorted in increasing order, the second containing the next 15%, and the third containing the remaining 80%.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet in the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Variable labels**: Check this option if the first line of the selected data contains a label.

**Observation labels**: Check this option if you want to use the available line labels. If you do not check this option, line labels will be created automatically (Obs1, Obs2, etc.). If a column header has been selected, check that the "Variable labels" option has been activated.

**Display the report header**: Deactivate this option if you do not want to display the report header.

**Options** tab:

**Weights**: Check this option if the observations are weighted. If you do not check this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

- **Standardize the weights**: if you check this option, the weights are standardized such that their sum equals the number of observations.

**Missing data** tab:
Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations:

- For the corresponding sample: Activate this option to ignore an observation which has a missing value only for the variables that have a missing value.

- For all samples: Activate this option to ignore an observation which has a missing value for all selected variables.

Estimate missing data: Activate this option to estimate the missing data by using the mean of the variable.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Centroids: Activate this option to display the table of centroids of the classes.

Central objects: Activate this option to display the coordinates of the nearest object to the centroid for each class.

Results by class: Activate this option to display a table giving the statistics and the objects for each of the classes.

Results by object: Activate this option to display a table giving the class each object is assigned to in the initial object order.

Charts tab:

Histograms: Activate this option to display the histograms of the samples. For a theoretical distribution, the density function is displayed.

- Bars: Choose this option to display the histograms with a bar for each interval.

- Continuous lines: Choose this option to display the histograms with a continuous line.

Cumulative histograms: Activate this option to display the cumulated histograms of the samples.

- Based on the histogram: Choose this option to display cumulative histograms based on the same interval definition as the histograms.
- **Empirical cumulative distribution**: Choose this option to display cumulative histograms which actually correspond to the empirical cumulative distribution of the sample.

**Ordinate of the histograms**: Choose the quantity to be used for the histograms: density, frequency or relative frequency.

## Results

**Summary statistics**: This table displays for the selected variables, the number of observations, the number of missing values, the number of non-missing values, the mean and the standard deviation.

A **histogram and the corresponding empirical cumulative distribution function** are displayed if the corresponding options are activated. The statistics of the intervals are then displayed.

**Class centroids**: This table shows the class centroids for the various descriptors.

**Distance between the class centroids**: This table shows the Euclidean distances between the class centroids for the various descriptors.

**Central objects**: This table shows the coordinates of the nearest object to the centroid for each class.

**Distance between the central objects**: This table shows the Euclidean distances between the class central objects for the various descriptors.

**Results by class**: The descriptive statistics for the classes (number of objects, sum of weights, within-class variance, minimum distance to the centroid, maximum distance to the centroid, mean distance to the centroid) are displayed in the first part of the table. The second part shows the objects.

**Results by object**: This table shows the assignment class for each object in the initial object order.

## References


Data management

Use this tool to manage tables of data. Four functions are included in this tool: deduping, grouping, and joining (inner and outer). These features are common in databases, but are not included in Excel.

Description

Deduping

It is sometimes necessary to dedupe a table. Some observations might be mistakenly duplicated (or repeated) when they come from different sources, or because of input errors.

Grouping

Grouping is useful when you want to aggregate data. For example, imagine a table that contains all your sales records (one column with the customer id, and one with the sales value), and which you want to transform to have one record per customer, and the corresponding sum of sales. XLSTAT allows you to aggregate the data and to obtain the summary table within seconds. The sum is only one of the available possibilities.

Joining

Joining is common task in database management. It allows to merge two tables “horizontally” on the basis of a common information named the “key”. For example, imagine you measured some chemical indicators on 150 sites. Then you want to add geographical information on the sites where the data were collected. Your geographical table contains information on 1000 sites, including the 150 sites of interest. In order to avoid the tedious work of manually merging the two tables, a join will allow you to obtain within seconds the merged table that includes both the collected data and the geographical information.

One distinguishes two main types of joins:

- Inner joins: the merged table includes only keys that are common to both input tables.
- Outer joins: the merged table includes all keys that are available in the first, the second or both input tables.
**Filtering (Keep/Remove)**

This tool allows you to select a table and create a new table that includes (Keep) or excludes (Remove) the rows, for which the value in a given column, matches a value contained in a user defined list.

**Stack / Unstack**

This tool enables you to transform a table organized with one column per group into a table with 2 columns, one with the value of the variable and one with the associated group. The opposite operation is also possible (unstack). That can be useful to transform data organized by column into a dataset that can be easily used as input in an ANOVA model.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload**: Click this button to reload the default options.
- **Delete**: Click this button to delete the data selections.
- **Change data handling**: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:
**Data:** This field is displayed if the selected method is “Dedupe” or “Group”. Select the data that correspond to the table that you want to dedupe or to aggregate.

**Observation labels:** This field is displayed only for the “Dedupe” method. Select the column (column mode) or row (row mode) where the observations labels are available. If you do not check this option, labels will be created automatically (Obs1, Obs2, etc.). If a column header has been selected, check that the "Variable labels" option has been activated.

**Table 1:** This field is displayed if the data management method is “Join”. Select the data that correspond to the first input table to use in the join procedure.

**Table 2:** This field is displayed if the data management method is “Join”. Select the data that correspond to the second input table to use in the join procedure.

**Guess types:** this option is displayed only for the “Group” method. Activate this option if you want that XLSTAT guesses the types of the variables of the selected table. If you uncheck this option, XLSTAT will prompt you to confirm or modify the type of the variables.

**Method:** select the data management method to use:

- Dedupe
- Group
- Join (Inner)
- Join (Outer)
- Filter (Keep)
- Filter (Remove)
- Stack
- Unstack

**Range:** Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Check this option to display the results in a new worksheet in the active workbook.

**Workbook:** Check this option to display the results in a new workbook.
Variable labels: Check this option if the first row of the selected data (data and observation labels) contains a label.

Operation: This option is only available if the method is “Group”. Select the operation to apply to the data when aggregating them.

Outputs tab:
This tab is only displayed if the selected method is “Dedupe” or “Group”.

Descriptive statistics: Activate this option to display descriptive statistics for the selected variables.

The following options are only displayed if the selected method is “Dedupe”:

Deduped table: Activate this option to display the deduped table.

- Frequencies: Activate this option to display in the last column of the deduped table, the frequencies of each observation in the input table (1 corresponds to non-repeated observations; values equal or greater than 2 correspond to duplicated observations).

Duplicates: Activate this option to display the duplicates that have been removed from the original table in order to obtain the deduped table.

Missing data tab:
This tab is only displayed if the selected method is “Group”.

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Ignore missing data: Activate this option to ignore missing data.
Coding

Use this tool to code or recode a table into a new table, using a coding table that contains the initial values and the corresponding new codes.

Dialog box

OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

Reload default options: Click this button to reload the default options.

Delete data selections: Click this button to delete the data selections.

Data: Select the data in the Excel worksheet. If headers have been selected, check that the "Column labels" option has been activated.

Coding table: Select a two-column table that contains in the first column the initial values, and in the second column the codes that will replace the values. If headers have been selected, check that the "Column labels" option has been activated.

Column labels: Activate this option if the first row of the data selected (data and coding table) contains a label.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.
Display the report header: Deactivate this option if you want the results table to start from the first row of the Excel worksheet (situation after output to a worksheet or workbook) and not after the report header.
Presence/absence coding

Use this tool to convert a table of lists (or attributes) into a table of presences/absences showing the frequencies of the various elements for each of the lists.

Description

This tool is used, for example, to convert a table containing \( p \) columns corresponding to \( p \) lists of objects into a table with \( p \) rows and \( q \) columns where \( q \) is the number of different objects contained in the \( p \) lists, and where for each cell of the table, there is a 1 if the object is present and a 0 if it is absent.

For example, in ecology, if we have \( p \) species measurements with, for each measurement, the different species found in columns, we will obtain a two-way table showing the presence or absence of each of the species for each of the measurements.

Dialog box

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.

- Click this button to reload the default options.
- Click this button to delete the data selections.

**Data**: Select the data in the Excel worksheet.

**Column labels**: Activate this option if the first row of the selected data contains a label.

Presence/absence coding by:

- **Rows**: Choose this option if each row corresponds to a list.
- **Columns**: Choose this option if each column corresponds to a list.
**Range:** Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Activate this option to display the results in a new worksheet of the active workbook.

**Workbook:** Activate this option to display the results in a new workbook.

**Display the report header:** Deactivate this option if you want the results table to start from the first row of the Excel worksheet (situation after output to a worksheet or workbook) and not after the report header.

**Example**

Input table:

<table>
<thead>
<tr>
<th></th>
<th>List1</th>
<th>List2</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1</td>
<td>E1</td>
<td>E3</td>
</tr>
<tr>
<td>E1</td>
<td>E1</td>
<td></td>
</tr>
<tr>
<td>E2</td>
<td>E4</td>
<td></td>
</tr>
<tr>
<td>E1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Presence/absence table:

<table>
<thead>
<tr>
<th></th>
<th>E1</th>
<th>E2</th>
<th>E3</th>
<th>E4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liste1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Liste2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Coding by ranks

Use this tool to recode a table with n observations and p quantitative variables into a table containing ranks, the latter being determined variable by variable.

Description

This tool is used to recode a table with n observations and p quantitative variables into a table containing ranks, the ranks being determined variable by variable. Coding in ranks lets you convert a table of continuous quantitative variables into discrete quantitative variables if only the order relationship is relevant and not the values themselves.

Two strategies are possible for taking tied values into account: either they are assigned to the mean rank or they are assigned to the lowest rank of the tied values.

Dialog box

- OK: Click this button to start the computations.
- Cancel: Click this button to close the dialog box without doing any computation.
- Help: Click this button to display the help.
- : Click this button to reload the default options.
- : Click this button to delete the data selections.

Data: Select the data in the Excel worksheet.

Variable labels: Check this option if the first line of the selected data contains a label.

Observation labels: Check this option if you want to use the available line labels. If you do not check this option, line labels will be created automatically (Obs1, Obs2, etc.). If a column header has been selected, check that the “Variable labels” option has been activated.
**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Take ties into account**: Activate this option to take account of the presence of tied values and adapt the rank of tied values as a consequence.

*Mean ranks*: Choose this option to replace the rank of tied values by the mean of the ranks.

*Minimum*: Choose this option to replace the rank of tied values by the minimum of their ranks.

**Display the report header**: Deactivate this option if you want the sampled table to start from the first row of the Excel worksheet (situation after output to a worksheet or workbook) and not after the report header.

**Example**

Initial table:

<table>
<thead>
<tr>
<th></th>
<th>V1</th>
<th>V2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obs1</td>
<td>1.2</td>
<td>12</td>
</tr>
<tr>
<td>Obs2</td>
<td>1.6</td>
<td>11</td>
</tr>
<tr>
<td>Obs3</td>
<td>1.2</td>
<td>10</td>
</tr>
<tr>
<td>Obs4</td>
<td>1.4</td>
<td>10.5</td>
</tr>
</tbody>
</table>

Recoded table (using mean ranks for ties):

<table>
<thead>
<tr>
<th></th>
<th>R1</th>
<th>R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obs1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Obs2</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Obs3</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Recoded table (using the lowest ranks for ties):

<table>
<thead>
<tr>
<th></th>
<th>R1</th>
<th>R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obs1</td>
<td>1.5</td>
<td>4</td>
</tr>
<tr>
<td>Obs2</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Obs3</td>
<td>1.5</td>
<td>1</td>
</tr>
<tr>
<td>Obs4</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>
Descriptive statistics and Univariate plots

Use this tool to calculate descriptive statistics and display univariate plots (Box plots, Scattergrams, etc) for a set of quantitative and/or qualitative variables.

Description

Before using advanced analysis methods like, for example, discriminant analysis or multiple regression, you must first of all reveal the data in order to identify trends, locate anomalies or simply have available essential information such as the minimum, maximum or mean of a data sample.

XLSTAT offers you a large number of descriptive statistics and charts which give you a useful and relevant preview of your data.

Although you can select several variables (or samples) at the same time, XLSTAT calculates all the descriptive statistics for each of the samples independently.

Descriptive statistics for quantitative data:

Let's consider a sample made up of N items of quantitative data \{y_1, y_2, \ldots, y_N\} whose respective weights are \{W_1, W_2, \ldots, W_N\}.

- **Number of observations**: The number N of values in the selected sample.
- **Number of missing values**: The number of missing values in the sample analyzed. In the subsequent statistical calculations, values identified as missing are ignored. We define \(n\) to be the number of non-missing values, and \{x_1, x_2, \ldots, x_n\} to be the sub-sample of non-missing values whose respective weights are \{w_1, w_2, \ldots, w_n\}.
- **Sum of weights**: The sum of the weights, \(S_w\). When all weights are 1, or when weights are "standardized", \(S_w=n\).
- **Minimum**: The minimum of the series analyzed.
- **Maximum**: The maximum of the series analyzed.
- **Frequency of minimum**: The frequency of the minimum of the series.
- **Frequency of maximum**: The frequency of the maximum of the series.
- **Range**: The range is the difference between the minimum and maximum of the series.
1st quartile*: The first quartile Q1 is defined as the value for which 25% of the values are less.

Median*: The median Q2 is the value for which 50% of the values are less.

3rd quartile*: The third quartile Q3 is defined as the value for which 75% of the values are less.

Sum*: The weighted sum of the values is defined by:

\[ S = \sum_{i=1}^{n} w_i x_i \]

Mean*: The mean of the sample is defined by \( \mu = S / Sw \).

Variance (n)*: The variance of the sample defined by:

\[ s(n)^2 = \frac{\sum_{i=1}^{n} w_i (x_i - \mu)^2}{Sw} \]

Note 1: When all the weights are 1, the variance is the sum of the square deviation to the mean divided by n, hence its name.

Note 2: The variance (n) is a biased estimate of the variance which assumes that the sample is a good representation of the total population. The variance (n-1) is, on the other hand, calculated taking into account an approximation associated with the sampling.

Variance (n-1)*: The estimated variance of the sample defined by:

\[ s(n-1)^2 = \frac{\sum_{i=1}^{n} w_i (x_i - \mu)^2}{Sw - Sw/n} \]

Note 1: When all the weights are 1, the variance is the sum of the square deviation to the mean divided by n-1, hence its name.

Note 2: The variance (n) is a biased estimate of the variance which assumes that the sample is a good representation of the total population. The variance (n-1) is, on the other hand, calculated taking into account an approximation associated with the sampling.

Standard deviation (n)*: The standard deviation of the sample defined by \( s(n) \).

Standard deviation (n-1)*: The standard deviation of the sample defined by \( s(n-1) \).

Variation coefficient: this coefficient is only calculated if the mean of the sample is non-zero. It is defined by \( CV = s(n) / \mu \). This coefficient measures the dispersion of a
sample relative to its mean. It is used to compare the dispersion of samples whose scales or means differ greatly.

- **Skewness (Pearson)***: The Pearson skewness coefficient is defined by:

\[
\gamma_1 = \frac{\mu_3}{s(n)^3} \quad \text{with} \quad \mu_3 = \frac{\sum_{i=1}^{n} w_i (x_i - \mu)^3}{S_w}
\]

This coefficient gives an indication of the shape of the distribution of the sample. If the value is negative (or positive respectively), the distribution is concentrated on the left (or right respectively) of the mean.

- **Skewness (Fisher)***: The Fisher skewness coefficient is defined by:

\[
G_1 = \frac{\sqrt{S_w (S_w - S_w/n)} \gamma_1}{S_w - 2S_w/n}
\]

Unlike the previous, this coefficient is not biased on the assumption that the data is normally distributed. This coefficient gives an indication of the shape of the distribution of the sample. If the value is negative (or positive respectively), the distribution is concentrated on the left (or right respectively) of the mean.

- **Skewness (Bowley)***: The Bowley skewness coefficient is defined by:

\[
A(B) = \frac{Q_1 - 2Q_2 + Q_3}{Q_3 - Q_1}
\]

- **Kurtosis (Pearson)***: The Pearson kurtosis coefficient is defined by:

\[
\gamma_2 = \frac{\mu_4}{s^4(n)} - 3 \quad \text{with} \quad \mu_4 = \frac{\sum_{i=1}^{n} w_i (x_i - \mu)^4}{S_w}
\]

This coefficient, sometimes called *excess kurtosis*, gives an indication of the shape of the distribution of the sample. If the value is negative (or positive respectively), the peak of the distribution of the sample is more flattened out (or respectively less) than that of a normal distribution.

- **Kurtosis (Fisher)***: The Fisher kurtosis coefficient is defined by:

\[
G_2 = \frac{S_w - S_w/n}{(S_w - 2S_w/n)(S_w - 3S_w/n)} ((S_w+S_w/n) \gamma_2 + 6)
\]

\[
= \frac{S_w - S_w/n}{(S_w - 2S_w/n)(S_w - 3S_w/n)} \left( \frac{\mu_4}{s^4(n)} - 3(S_w - sw/n) \right)
\]
Unlike the previous, this coefficient is not biased on the assumption that the data is normally distributed. This coefficient, sometimes called *excess kurtosis*, gives an indication of the shape of the distribution of the sample. If the value is negative (or positive respectively), the peak of the distribution of the sample is more flattened out (or respectively less) than that of a normal distribution.

- **Standard error of the mean**: this statistic is defined by:
\[
 s_\mu = \sqrt{\frac{s(n-1)^2}{Sw}}
\]

- **Lower bound on mean (x% or significance level \(\alpha=1-x/100\))**: this statistic corresponds to the lower bound of the confidence interval at x% of the mean. This statistic is defined by:
\[
 L_\mu = \mu - s_\mu \left| t_{(\alpha/2)} \right|
\]

- **Upper bound on mean (x% or significance level \(\alpha=1-x/100\))**: this statistic corresponds to the upper bound of the confidence interval at x% of the mean. This statistic is defined by:
\[
 U_\mu = \mu + s_\mu \left| t_{(\alpha/2)} \right|
\]

- **Standard error of the variance**: this statistic is defined by:
\[
 s_\sigma = s(n-1)^2 \sqrt{\frac{2}{Sw-1}}
\]

- **Lower bound on mean (x% or significance level \(\alpha=1-x/100\))**: this statistic corresponds to the lower bound of the confidence interval at x% of the variance. This statistic is defined by:
\[
 L_\sigma = s_\sigma \sqrt{\chi_{1-(\alpha/2)}}
\]

- **Upper bound on mean (x% or significance level \(\alpha=1-x/100\))**: this statistic corresponds to the upper bound of the confidence interval at x% of the variance. This statistic is defined by:
\[
 U_\sigma = s_\sigma \sqrt{\chi_{(\alpha/2)}}
\]

- **Standard error (Skewness (Fisher))**: The standard error of the Fisher’s skewness coefficient is defined by:
\[
 se(G_1) = \sqrt{\frac{6Sw(Sw-1)}{(Sw-2)(Sw+1)(Sw+3)}}
\]
• **Standard error (Kurtosis (Fisher))**: The standard error of the Fisher’s kurtosis coefficient is defined by:

\[
se(G_z) = \sqrt{\frac{4(Sw^2 - 1)(se(G_t))^2}{(Sw - 3)(Sw + 5)}}
\]

• **Mean absolute deviation**: as for standard deviation or variance, this coefficient measures the dispersion (or variability) of the sample. It is defined by:

\[
e(\mu) = \frac{\sum_{i=1}^{n} w_i |x_i - \mu|}{Sw}
\]

• **Median absolute deviation**: this statistic is the median of absolute deviations to the median.

• **Geometric mean**: this statistic is only calculated if all the values are strictly positive. It is defined by:

\[
\mu_G = \exp\left(\frac{1}{Sw} \sum_{i=1}^{n} w_i \ln(x_i)\right)
\]

If all the weights are equal to 1, we have:

\[
\mu_G = \sqrt[n]{\prod_{i=1}^{n} x_i}
\]

• **Geometric standard deviation**: this statistic is defined by:

\[
\sigma_G = \exp\left(\frac{1}{Sw} \sum_{i=1}^{n} w_i (\ln(x_i) - \ln(\mu_G))^2\right)
\]

• **Harmonic mean**: this statistic is defined by:

\[
\mu_H = \frac{Sw}{\sum_{i=1}^{n} w_i x_i}
\]

(*) Statistics followed by an asterisk take the weight of observations into account.

**Descriptive statistics for qualitative data:**
For a sample made up of N qualitative values, we define:

- **Number of observations**: The number N of values in the selected sample.

- **Number of missing values**: The number of missing values in the sample analyzed. In the subsequent statistical calculations, values identified as missing are ignored. We define \( n \) to be the number of non-missing values, and \( \{w_1, w_2, \ldots w_n\} \) to be the sub-sample of weights for the non-missing values.

- **Sum of weights\(^*\)**: The sum of the weights, \( S_w \). When all the weights are 1, \( S_w = n \).

- **Mode\(^*\)**: The mode of the sample analyzed. In other words, the most frequent category.

- **Frequency of mode\(^*\)**: The frequency of the category to which the mode corresponds.

- **Category**: The names of the various categories present in the sample.

- **Frequency by category\(^*\)**: The frequency of each of the categories.

- **Relative frequency by category\(^*\)**: The relative frequency of each of the categories.

- **Lower bound on frequencies (x\% or significance level a=1-x/100)\(^*\)**: this statistic corresponds to the lower bound of the confidence interval at x\% of the frequency per category.

- **Upper bound on frequencies (x\% or significance level a=1-x/100)\(^*\)**: this statistic corresponds to the upper bound of the confidence interval at x\% of the frequency per category.

- **Proportion per category\(^*\)**: The proportion of each of the categories.

- **Lower bound on proportions (x\% or significance level a=1-x/100)\(^*\)**: this statistic corresponds to the lower bound of the confidence interval at x\% of the proportion per category.

- **Upper bound on proportions (x\% or significance level a=1-x/100)\(^*\)**: this statistic corresponds to the upper bound of the confidence interval at x\% of the proportion per category.

\(^*\) Statistics followed by an asterisk take the weight of observations into account.

Several types of chart are available for quantitative and qualitative data:

**Charts for quantitative data:**

- **Box plots**: These univariate representations of quantitative data samples are sometimes called "box and whisker diagrams". It is a simple and quite complete representation since
in the version provided by XLSTAT the minimum, 1st quartile, median, mean and 3rd quartile are displayed together with both limits (the ends of the "whiskers") beyond which values are considered anomalous. The mean is displayed with a red +, and a black line corresponds to the median. Limits are calculated as follows:

- **Lower limit**: \( L_{inf} = X(i) \) such that \( \{X(i) - [Q1 - 1.5 (Q3 - Q1)]\} \) is minimum and \( X(i) \geq Q1 - 1.5 (Q3 - Q1) \).

- **Upper limit**: \( L_{sup} = X(i) \) such that \( \{X(i) - [Q3 + 1.5 (Q3 - Q1)]\} \) is minimum and \( X(i) \leq Q3 + 1.5 (Q3 - Q1) \).

Values that are outside the \( [Q1 - 3 (Q3 - Q1); Q3 + 3 (Q3 - Q1)] \) interval are displayed with the * symbol. Values that are in the \( [Q1 - 3 (Q3 - Q1); Q1 - 1.5 (Q3 - Q1)] \) or the \( [Q3 + 1.5 (Q3 - Q1); Q3 + 3 (Q3 - Q1)] \) intervals are displayed with the "o" symbol.

XLSTAT allows producing “notched” box plots. The limits of the notch allow to visualize a 95% confidence interval around the median. The limits are given by:

- **Lower limit**: \( N_{inf} = \text{Median} - \left[ \frac{1.58 (Q3 - Q1)}{\text{SquareRoot(n)}} \right] \)

- **Upper limit**: \( N_{sup} = \text{Median} + \left[ \frac{1.58 (Q3 - Q1)}{\text{SquareRoot(n)}} \right] \).

These formulae given by McGill et al. (1978) derive from the assumption that the medians are normally distributed and coming from equal size samples. If the sample sizes are indeed similar notched box plots allow to tell whether the samples have different medians or not and to compare their variability using the size of the notch.

XLSTAT allows to make the box plots width vary with the sample size. The width is proportional to the square root of the sample size.

- **Scattergrams**: These univariate representations give an idea of the distribution and possible plurality of the modes of a sample. All points are represented together with the mean and the median.

- **Strip plots**: These diagrams represent the data from the sample as strips. For a given interval, the thicker or more tightly packed the strips, the more data there is.

- **P-P Charts (normal distribution)**: P-P charts (for Probability-Probability) are used to compare the empirical distribution function of a sample with that of a normal variable for the same mean and deviation. If the sample follows a normal distribution, the data will lie along the first bisector of the plan.

- **Q-Q Charts (normal distribution)**: Q-Q charts (for Quantile-Quantile) are used to compare the quantities of the sample with that of a normal variable for the same mean and deviation. If the sample follows a normal distribution, the data will lie along the first bisector of the plan.
Charts for qualitative data:

- **Bar charts**: Check this option to represent the frequencies or relative frequencies of the various categories of qualitative variables as bars.

- **Pie charts**: Check this option to represent the frequencies or relative frequencies of the various categories of qualitative variables as pie charts.

- **Double pie charts**: These charts are used to compare the frequencies or relative frequencies of sub-samples with those of the complete sample.

- **Doughnuts**: this option is only checked if a column of sub-samples has been selected. These charts are used to compare the frequencies or relative frequencies of sub-samples with those of the complete sample.

- **Stacked bars**: this option is only checked if a column of sub-samples has been selected. These charts are used to compare the frequencies or relative frequencies of sub-samples with those of the complete sample.

**Dialog box**

The dialog box is made up of several tabs corresponding to the various options for controlling the calculations and displaying the results. A description of the various components of the dialog box is given below.

- **OK**: Click this button to start the computations.

- **Cancel**: Click this button to close the dialog box without doing any computation.

- **Help**: Click this button to display the help.

- **Reload**: Click this button to reload the default options.

- **Delete**: Click this button to delete the data selections.

- **Row/Column**: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.
**General** tab:

**Quantitative data:** Check this option to select the samples of quantitative data you want to calculate descriptive statistics for.

**Qualitative data:** Check this option to select the samples of qualitative data you want to calculate descriptive statistics for.

**Subsamples:** Check this option to select a column showing the names or indexes of the subsamples for each of the observations.

- **Variable-Category labels:** Activate this option to use variable-category labels when displaying outputs for the quantitative variables. Variable-Category labels include the variable name as a prefix and the category name as a suffix.

**Range:** Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Check this option to display the results in a new worksheet in the active workbook.

**Workbook:** Check this option to display the results in a new workbook.

**Sample labels:** Check this option if the first line of the selections (quantitative data, qualitative data, sub-samples, and weights) contains a label.

**Weights:** Check this option if the observations are weighted. If you do not check this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Sample labels" option is activated.

- **Standardize the weights:** if you check this option, the weights are standardized such that their sum equals the number of observations.

**Options** tab:

**Descriptive statistics:** Check this option to calculate and display descriptive statistics.

**Charts:** Check this option to display the charts.

**Normalize:** Check this option to standardize the data before carrying out the analysis.

**Rescale from 0 to 100:** Check this option to arrange the data on a scale of 0 to 100.

**Compare to total sample:** this option is only checked if a column of sub-samples has been selected. Check this option so that the descriptive statistics and charts are also displayed for the total sample.
Confidence interval: Enter the size of the confidence interval (in %).

Outputs tab:

Quantitative Data: Activate the options for the descriptive statistics you want to calculate. The various statistics are described in the description section.

- All: Click this button to select all.
- None: Click this button to deselect all.
- Display vertically: Check this option so that the table of descriptive statistics is displayed vertically (one line per descriptive statistic).

Qualitative Data: Activate the options for the descriptive statistics you want to calculate. The various statistics are described in the description section.

- All: Click this button to select all.
- None: Click this button to deselect all.
- Display vertically: Check this option so that the table of descriptive statistics is displayed vertically (one line per descriptive statistic).

Charts (1) tab:

This tab deals with the quantitative data.

Chart types sub-tab:

Box plots: Check this option to display box plots (or box-and-whisker plots). See the description section for more details.

Scattergrams: Check this option to display scattergrams. The mean (red +) and the median (red line) are always displayed.

Strip plots: Check this option to display strip plots. On these charts, a strip corresponds to an observation.

Stem-and-leaf plots: Check this option to display stem-and-leaf plots.

Normal P-P plots: Check this option to display P-P plots.

Normal Q-Q Charts: Check this option to display Q-Q plots.

Options sub-tab:
These options concern box plots, scattergrams and strip plots

**Horizontal**: Check this option to display box plots, scattergrams and strip plots horizontally.

**Vertical**: Check this option to display box plots, scattergrams and strip plots vertically.

**Group plots**: Check this option to group together the various box plots, scattergrams and strip plots on the same chart to compare them. You can enter “**Dimensions**”, the maximum number of box plots a group can contain (maximum is 20 with Excel 2003 and lower, and 40 with Excel 2007 and later).

- **Categories**: Check this option if you want to group categories in case you have selected subsamples.

- **Variables**: Check this option if you want to group variables, even in the case where you have selected subsamples. Check the **Grey line** option to separate variables with grey lines on the plots.

**Notched**: Check this option if you want to display notched box plots.

**Adapt width**: Check this option if you want that the width of the box plots depends on the sample size.

**Minimum/Maximum**: Check this option to systematically display the points corresponding to the minimum and maximum (box plots).

**Outliers**: Check this option to display the points corresponding to outliers (box plots) with a hollowed-out circle.

**Labels position**: Select the position where the labels have to be placed on the box plots, scattergrams and strip plots.

**Legend**: Activate this option to display the legend describing the statistics used on the box plot.

**Charts (2) tab:**

This tab deals with the qualitative data.

**Bar charts**: Check this option to represent the frequencies or relative frequencies of the various categories of qualitative variables as bars.

**Pie charts**: Check this option to represent the frequencies or relative frequencies of the various categories of qualitative variables as pie charts.

- **Doubles**: this option is only checked if a column of sub-samples has been selected. These charts are used to compare the frequencies or relative frequencies of sub-samples with those of the complete sample.
**Doughnuts:** this option is only checked if a column of sub-samples has been selected. These charts are used to compare the frequencies or relative frequencies of sub-samples with those of the complete sample.

**Stacked bars:** this option is only checked if a column of sub-samples has been selected. These charts are used to compare the frequencies or relative frequencies of sub-samples with those of the complete sample.

**Values used:** choose the type of data to be displayed:

- **Frequencies:** choose this option to make the scale of the plots correspond to the frequencies of the categories.

- **Relative frequencies:** choose this option to make the scale of the plots correspond to the relative frequencies of the categories.

**References**


Variable characterization

Use this tool to characterize elements (quantitative variables, qualitative ones or categories of qualitative variable) exploring the links they share with characterizing variables (quantitative variables, qualitative variables or categories of qualitative variables).

Description

Here are the various characterizing possibilities of the procedure:

1) Characterization of a quantitative variable:

i) With other quantitative variables:

Characterization of a quantitative variable by other quantitative variables is carried out using the correlation coefficient. For each characterizing quantitative variable, a test is performed to determine whether the latter is significantly different from 0. The Pearson correlation test is implemented in a parametric context while the Spearman correlation test is preferred in a non-parametric one. The more significantly different from 0 the correlation coefficient, the stronger the link between the 2 quantitative variables.

ii) With qualitative variables:

Characterization of a quantitative variable by qualitative variables is carried out using the correlation ratio. For each qualitative variable (with k categories), a test is performed to determine whether the latter is significantly different from 0. In a parametric framework, the Student t-test (k=2) (resp. Fisher F-test (k>2)) is implemented while the Wilcoxon (k=2) (resp. Kruskal-Wallis (k>2)) test is preferred in a non-parametric one. The more the correlation ratio significantly different from zero, the stronger the link between the quantitative variable to characterize and the characterizing qualitative variable.

iii) With categories:

Characterization of a quantitative variable by categories is carried out using a mean comparison test. For each category, it consists in determining whether the mean of the quantitative variable to characterize in the group whose members share this category is significantly different from the mean of the quantitative variable to characterize considering the whole sample.

2) Characterization of a qualitative variable (with k categories):

i) With quantitative variables:
Characterization of a qualitative variable by quantitative variables is carried out using the correlation ratio. For each quantitative variable, a test is performed to determine whether it is significantly different from 0. In a parametric framework, the Student t-test (k=2) (respectively Fisher F-test (k>2)) is implemented while the Wilcoxon (k=2) (respectively Kruskal-Wallis (k>2)) test is preferred in a non-parametric one. The more the correlation ratio significantly different from zero, the stronger the link between the quantitative variable to characterize and the characterizing qualitative variable.

ii) With other qualitative variables:

Characterization of a qualitative variable by other qualitative variables is carried out using a test for independence. For each characterizing qualitative variable, a test is performed to determine whether it is statistically independent from the qualitative variable to characterize. In a parametric framework, the Chi-square test for independence is used while the Fisher’s exact test is preferred in a non-parametric one.

3) Characterization of a category:

i) With quantitative variables:

Characterization of a category by quantitative variables is carried out using a mean comparison test. For each characterizing quantitative variable, a test is implemented to determine whether the mean of that quantitative variable in the group whose members share the category is significantly different from its mean in the whole sample.

ii) With other categories:

Characterization of a category by other categories is carried out using a proportion comparison test. For each characterizing category, a test is performed to determine if the proportion of individuals sharing both category to characterize and characterizing category is significantly different from the theoretical expected proportion. If the observed proportion is greater than the theoretical one, the category is over-represented in the group whose members share the category to characterize. Conversely, if it is smaller, the category is under-represented.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

**OK** Click this button to start the computations.
Click this button to close the dialog box without doing any computation.

Click this button to display the help.

Click this button to reload the default options.

Click this button to delete the data selections.

Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Y / Element(s) to characterize:

Quantitative variable(s): Activate this option if you want to characterize one or several quantitative variables. Then, select the response variable(s) you want to characterize. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

Qualitative variable(s): Activate this option if you want to characterize one or several qualitative variables. Then, select the response variable(s) you want to characterize. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

Categories: Activate this option if you want to characterize the categories of the qualitative variable(s) previously selected.

X / Characterizing elements:

Quantitative Variable(s): Activate this option if you want characterizing quantitative variable(s). Then, select the quantitative variables in the Excel worksheet. The data selected must be of type numeric. If the variable header has been selected, check that the "Variable labels" option has been activated.

Qualitative Variable(s): Activate this option if you want characterizing qualitative variable(s). Then, select the qualitative variables in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Variable labels" option has been activated.
**Categories:** Activate this option if you want to use the categories of the qualitative variable(s) previously selected as characterizing elements.

**Range:** Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Check this option to display the results in a new worksheet in the active workbook.

**Workbook:** Check this option to display the results in a new workbook.

**Variable labels:** Check this option if the first line of the selections (quantitative data, qualitative date, and weights) contains a label.

**Observation weights:** Activate this option if the observations are weighted. If you do not activate this option, the weights will all be taken as 1. Weights must be greater than or equal to 0. A weight of 2 is equivalent to repeating the same observation twice. If a column header has been selected, check that the "Variable labels" option has been activated.

**Options** tab:

We first introduce the common options to all of the characterizations.

**Selection:** Activate this option if you want to select the characterizing elements to display according to one of these 3 criteria:

**p-value threshold (%):** Activate this option if you want to only display characterizing elements whose p-value associated to the corresponding test is lower than the threshold. Then, enter the value of the threshold in the corresponding cell.

**Test value threshold:** Activate this option if you want to only display characterizing elements whose test value of the associated statistical test is greater (in absolute value) than the (minimal) threshold. Then, enter the value of the threshold in the corresponding cell.

Remark: For the characterization of a quantitative variable with a qualitative variable with more than two categories, the test value and p-value evolve in the same manner so that the threshold is a maximal one.

**Number:** Activate this option if you want to only display a specific number characterizing elements (those with the lowest p-value). Then, enter the value of the number in the corresponding cell.

**Significance level (%)** Enter the significance level in the corresponding cell.
The remaining options differ according to the combination: element to characterize/characterizing element chosen in the General tab. The different cases are reported below:

1) Characterisation of a quantitative variable:

i) With other quantitative variables:

Characterizing Quantitative variable(s):

Quantitative variable(s) to keep:

**Positive Correlations**: Activate this option if you want to keep the positively correlated quantitative variables.

**Negative Correlations**: Activate this option if you want to keep the negatively correlated quantitative variables.

Test:

**Parametric**: Activate this option if you want that a parametric test be performed: the Chi-square independence test.

**Non-parametric**: Activate this option if you want that a non-parametric test be performed: the Fisher’s exact test.

ii) With qualitative variables:

Test:

**Parametric**: Activate this option if you want that a parametric test be performed: Student t-test if k=2 (resp. Fisher f-test if k>2).

**Non-parametric**: Activate this option if you want that a non-parametric test be performed: Wilcoxon test if k=2 (resp. Kruskal-Wallis test if k>2).

iii) With categories:

Characterizing categories:

**Min relative weight (%)**: Activate this option if you want to only display the characterizing categories with a relative weight (calculated as: number of individuals sharing the category / number of individuals) greater than a specific threshold. Then, enter the threshold value in the corresponding cell.

Categories to keep:

**Greater than the mean**: Activate this option if you want to keep the characterizing categories such that the mean of the quantitative variable to characterize in the group whose members share the category is greater than the whole population mean for this variable.
**Lower than the mean**: Activate this option if you want to keep the characterizing categories such the mean of the quantitative variable to characterize in the group whose members share the category is lower than the whole population mean for this variable.

2) Characterization of a qualitative variable (with k categories):

i) With quantitative variables:

Test:

**Parametric**: Activate this option if you want that a parametric test be performed: Student t-test if \( k=2 \) (resp. Fisher F-test if \( k>2 \)).

**Non-parametric**: Activate this option if you want that a non-parametric test be performed: Wilcoxon test if \( k=2 \) (resp. Kruskal-Wallis test if \( k>2 \)).

ii) With other qualitative variables:

Test:

**Parametric**: Activate this option if you want that a parametric test be performed: Chi-square independence test.

**Non-parametric**: Activate this option if you want that a non-parametric test be performed: Fisher’s exact test.

3) Characterization of a category:

i) With quantitative variables:

Quantitative variable(s) to keep:

**Greater than the mean**: Activate this option if you want to keep the quantitative variable(s) such that their mean in the group whose members share the category is greater than their mean in the whole population.

**Lower than the mean**: Activate this option if you want to keep the quantitative variable(s) such that their mean in the group whose members share the category is lower than their mean in the whole population.

ii) With other categories:

Characterizing categories:

**Min relative weight (%)**: Activate this option if you want to only display the characterizing categories with a relative weight (calculated as: number of individuals sharing the category /
number of individuals) greater than a specific threshold. Then, enter the threshold value in the corresponding cell.

Categories to keep:

**Overrepresented**: Activate this option if you want to keep the characterizing categories overrepresented in the group whose members share the category to characterize.

**Under-represented**: Activate this option if you want to keep the characterizing categories under-represented in the group whose members share the category to characterize.

**Missing data** tab:

**Remove observations**: Activate this option to ignore an observation that has a missing value.

**Estimate missing data**: Activate this option to estimate the missing data by using the mean of the sample.

**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the variables selected.

**Results**

**Summary statistics**: This table displays descriptive statistics for all the variables selected. For the quantitative variables, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed. For qualitative variables, including the dependent variable, the categories with their respective frequencies and percentages are displayed.

Depending on the selected elements, a specific table of results is displayed. The table is sorted in increasing order of the p-value (from the best characterizing element to the worst) and built taking into account the specifications of the **Options** tab. The different cases are reported below:

1) Characterization of a quantitative variable:

i) With other quantitative variables:

For each quantitative variable to characterize, the columns of the table respectively correspond to: label of the characterizing quantitative variables, value of the correlation coefficient, value of the statistical test, associated p-value.
ii) With qualitative variables:

For each quantitative variable to characterize, the columns of the table respectively correspond to: label of the characterizing qualitative variables, value of the correlation ratio, value of the statistical test, associated p-value.

iii) With categories:

For each quantitative variable to characterize, the columns of the table respectively correspond to: label of the qualitative variables associated with the characterizing categories, label of the characterizing categories, relative weight of the characterizing categories, mean value of the quantitative variable to characterize in the group whose members share the characterizing category, standard deviation of the quantitative variable to characterize in the group whose members share the characterizing category, value of the statistical test, associated p-value.

2) Characterization of a qualitative variable (with k categories):

i) With quantitative variables:

For each quantitative variable to characterize, the columns of the table respectively correspond to: label of the characterizing quantitative variables, value of the correlation ratio, value of the statistical test, associated p-value.

ii) With other qualitative variables:

For each quantitative variable to characterize, the columns of the table respectively correspond to: label of the characterizing qualitative variables, value of the statistical test, associated p-value.

3) Characterization of a category:

i) With quantitative variables:

For each category to characterize, the columns of the table respectively correspond to: label of the characterizing quantitative variables, mean value of the characterizing quantitative variable in the group whose members share the category to characterize, mean value of the characterizing quantitative variable in the whole population, standard deviation of the characterizing quantitative variable in the group whose members share the category to characterize, standard deviation of the characterizing quantitative variable in the whole population, value of the statistical test, associated p-value.

ii) With other categories:

For each quantitative variable to characterize, the columns of the table respectively correspond to: label of the qualitative variables associated with the characterizing categories, label of the categories, percentage of the characterizing category in the category to characterize,
percentage of the characterizing category in the whole population, percentage of the category
to characterize in the characterizing category, value of the statistical test, associated p-value.

A bar chart representing the p-values is also displayed with each table.

Example

A tutorial on variable characterization is available on the Addinsoft website:
http://www.xlstat.com/demo-demod.htm

References

Dunod, 181-184.

Quantiles estimation

Use this tool to calculate quantiles and display univariate plots (Box plots, Scattergrams, etc) for a set of quantitative variables.

Description

Quantiles (or percentiles) can be very useful in statistics. A percentile is a quantile based on a 0 to 100 scale.

XLSTAT offers you five methods to calculate quantiles. Furthermore, two types of confidence intervals are available.

While you can select several samples at the same time, XLSTAT calculates all the descriptive statistics for each sample independently.

Definition of a quantile

Let 0 < p < 1. The p-quantile of a variable X is given by:

\[ P(X \leq x) \geq p \text{ and } P(X \geq x) \geq 1 - p \]

Quantiles are useful because they are less sensitive to outliers and skewed distributions.

Methods for quantile computation

Five different methods are available in XLSTAT. Let's consider a sample made up of N items of quantitative data \( \{x_1, x_2, \ldots, x_N\} \) whose respective weights are \( \{W_1, W_2, \ldots, W_N\} \). Let \( x(1), \ldots, x(N) \) be the ordered data.

Let \( y \) be the p-quantile, \( j \) be the integer part of \( N*p \) and \( g \) be the fractional part. We have:

\[ g = N*p - j \]

We have:

1. Weighted average at \( x(Np) \):

\[ y = (1 - g)x_{(j)} + gx_{(j+1)} \]

where \( x(0) \) is replaced by \( x(1) \).
2- Observation numbered closest to Np:

\[ y = x_{(j)} \text{ if } g < 1/2 \]
\[ y = x_{(j)} \text{ if } g = 1/2 \text{ and } j \text{ is even} \]
\[ y = x_{(j+1)} \text{ if } g = 1/2 \text{ and } j \text{ is odd} \]
\[ y = x_{(j+1)} \text{ if } g > 1/2 \]

3- Empirical distribution function:

\[ y = x_{(j)} \text{ if } g = 0 \]
\[ y = x_{(j+1)} \text{ if } g > 0 \]

4- Weighted average aimed at \( x((N+1)p) \): In that case, we take \( (N+1)p = j + g \),

\[ y = \left(1 - g\right)x_{(j)} + gx_{(j+1)} \]

where \( x(N+1) \) is replaced by \( x(N) \).

5- Empirical distribution function with averaging:

\[ y = \frac{1}{2}\left(x_{(j)} + x_{(j+1)}\right) \text{ if } g = 0 \]
\[ y = x_{(j+1)} \text{ if } g > 0 \]

When weights are associated to the selected variable, the only method available is:

\[ y = \begin{cases} 
  x_{(i)} & \text{if } w_{(i)} > pW \\
  \frac{1}{2}\left(x_{(i)} + x_{(i+1)}\right) & \text{if } \sum_{j=1}^{i}w_{(j)} = pW \\
  x_{(i+1)} & \text{if } \sum_{j=1}^{i-1}w_{(j)} < pW < \sum_{j=1}^{i+1}w_{(j)} 
\end{cases} \]

where \( w(i) \) is the weight associated to \( x(i) \) and \( W = \sum_{j=1}^{N}w_{j} \).

Confidence intervals:

You can obtain confidence intervals associated to the quantiles. Two intervals are available:

1- Confidence interval based on the normal distribution:
The 100*(1-alpha) % confidence interval for the p-quantile is:

\[
\left[ Np + z_{\alpha/2} \sqrt{Np(1-p) + 0.5} ; Np + z_{1-\alpha/2} \sqrt{Np(1-p) + 0.5} \right]
\]

This kind of interval is valid if the data has a normal distribution and if the sample size is large (>20 observations).

2- Distribution free confidence interval:

The 100*(1-alpha) % confidence interval for the p-quantile is:

\[
\left[ x_{(l)} ; x_{(u)} \right]
\]

\(l\) and \(u\) are nearly symmetric around \([Np]+1\) where \([Np]\) is the integer part of \(N*p\). \(x(l)\) and \(x(u)\) are the closest to \(x([N+1]p)\) and satisfy:

\[
Q(u-1,n,p) - Q(l-1,n,p) \geq 1 - \alpha
\]

where \(Q(k,n,p)\) is the cumulative binomial probability:

\[
Q(k,n,p) = \sum_{i=1}^{k} \binom{n}{i} p^i (1-p)^{n-i}
\]

If weights are selected, confidence intervals cannot be computed.

Charts:

- **Cumulative histogram**: XLSTAT lets you create cumulative histograms by using the empirical cumulative distribution.

- **Box plots**: These univariate representations of quantitative data samples are sometimes called "box and whisker diagrams". It is a simple representation since in the version provided by XLSTAT the 1st quartile, median and 3rd quartile are displayed together with both limits (the ends of the "whiskers") beyond which values are considered anomalous. The red line corresponds to the median. Limits are calculated as follows:

  **Lower limit**: \(L_{\text{inf}} = X(i)\) such that \(\{X(i) - [Q1 - 1.5 (Q3 - Q1)]\}\) is minimum and \(X(i) \geq Q1 - 1.5 (Q3 - Q1)\).

  **Upper limit**: \(L_{\text{sup}} = X(i)\) such that \(\{X(i) - [Q3 + 1.5 (Q3 - Q1)]\}\) is minimum and \(X(i) \leq Q3 + 1.5 (Q3 - Q1)\).
Values that are outside the \(Q1 - 3 (Q3 - Q1); Q3 + 3 (Q3 - Q1)\) interval are displayed with the * symbol; values that are in the \([Q1 - 3 (Q3 - Q1); Q1 - 1.5 (Q3 - Q1)]\) or the \([Q3 + 1.5 (Q3 - Q1); Q3 + 3 (Q3 - Q1)]\) intervals are displayed with the "o" symbol.

- **Scattergrams**: These univariate representations give an idea of the distribution and possible plurality of the modes of a sample. All points are represented together with the median.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload**: Click this button to reload the default options.
- **Delete**: Click this button to delete the data selections.
- **Data handling**: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General** **tab**:

**Data**: Check this option to select the samples you want to calculate quantiles for.

**Estimation method**: Choose the method you want to use to calculate the quantiles. A description of the methods can be found in the description section of this help. The default method is weighted average.

**Confidence interval**:
- **Normal based**: Check this option if you want to display confidence interval based on the normal distribution. See the description section for more details.

- **Distribution free**: Check this option if you want to display distribution free confidence interval. See the description section for more details.

**Range**: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Check this option to display the results in a new worksheet in the active workbook.

**Workbook**: Check this option to display the results in a new workbook.

**Sample labels**: Check this option if the first line of the selections (quantitative data, qualitative data, sub-samples, and weights) contains a label.

**Weights**: Check this option if the observations are weighted. If you do not check this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Sample labels" option is activated.

**Sub-sample**: Check this option to select a column showing the names or indexes of the sub-samples for each of the observations.

**Missing data** tab:

**Remove observations**: Activate this option to ignore an observation that has a missing value.

**Estimate missing data**: Activate this option to estimate the missing data by using the mean of the sample.

**Outputs** tab:

**Descriptive statistics**: Check this option to calculate and display descriptive statistics.

**Charts** tab:

**Empirical cumulative distribution**: Activate this option to display the cumulative histograms that actually correspond to the empirical cumulative distribution of the sample.

**Box plots**: Check this option to display box plots (or box-and-whisker plots). See the description section for more details.
**Scattergrams**: Check this option to display scattergrams. The median (red line) is always displayed.

**Show quantile on charts (%)**: Check this option and enter the percentile to compute the associated value and display it on the charts.

## Results

**Summary statistics**: This table displays for the selected samples, the number of observations, the number of missing values, the number of non-missing values, the mean and the standard deviation.

**Table of quantiles**: This table displays percentiles for common values (1, 5, 10, 25, 50, 75, 90, 95, 99) and their associated confidence interval.

## Example

An example showing how to compute percentiles is available on the Addinsoft website at [http://www.xlstat.com/demo-qua.htm](http://www.xlstat.com/demo-qua.htm)

## References


Histograms

Use this tool to create a histogram from a sample of continuous or discrete quantitative data.

Description

The histogram is one of the most frequently used display tools as it gives a very quick idea of the distribution of a sample of continuous or discrete data.

Intervals definition

One of the challenges in creating histograms is defining the intervals, as for a determined set of data, the shape of the histogram depends solely on the definition of the classes. Between the two extremes of the single class comprising all the data and giving a single bar and the histogram with one value per class, there are as many possible histograms as there are data partitions.

To obtain a visually and operationally satisfying result, defining classes may require several attempts.

The most traditional method consists of using classes defined by intervals of the same width, the lower bound of the first interval being determined by the minimum value or a value slightly less than the minimum value.

To make it easier to obtain histograms, XLSTAT lets you create histograms either by defining the number of intervals, their width or by specifying the intervals yourself. The intervals are considered as closed for the lower bound and open for the upper bound.

Cumulative histogram

XLSTAT lets you create cumulative histograms either by cumulating the values of the histogram or by using the empirical cumulative distribution. The use of the empirical cumulative distribution is recommended for a comparison with a distribution function of a theoretical distribution.

Comparison to a theoretical distribution

XLSTAT lets you compare the histogram with a theoretical distribution whose parameters have been set by you. However, if you want to check if a sample follows a given distribution, you
can use the distribution fitting tool to estimate the parameters of the distribution and if necessary check if the hypothesis is acceptable.

XLSTAT provides the following distributions:

- **Arcsine ($\alpha$):** the density function of this distribution (which is a simplified version of the Beta type I distribution) is given by:
  
  \[ f(x) = \frac{\sin(\pi x)}{\pi x} \left(\frac{x}{1-x}\right)^\alpha, \text{ with } 0 < \alpha < 1, x \in [0,1] \]
  
  We have \( \text{E}(X) = \alpha \) and \( \text{V}(X) = \alpha(1-\alpha)/2 \)

- **Bernoulli ($p$):** the density function of this distribution is given by:
  
  \[ P(X = 1) = p, \ P(X = 0) = 1 - p \ \text{ with } p \in [0,1] \]
  
  We have \( \text{E}(X) = p \) and \( \text{V}(X) = p(1-p) \)
  
  The Bernoulli, named after the Swiss mathematician Jacob Bernoulli (1654-1705), allows to describe binary phenomena where only events can occur with respective probabilities of \( p \) and \( 1-p \).

- **Beta ($\alpha, \beta$):** the density function of this distribution (also called Beta type I) is given by:
  
  \[ f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}, \ \text{ with } \alpha, \beta > 0, x \in [0,1] \text{ and } B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} \]
  
  We have \( \text{E}(X) = \alpha/(\alpha+\beta) \) and \( \text{V}(X) = \alpha\beta/[(\alpha+\beta+1)(\alpha+\beta)^2] \)

- **Beta4 ($\alpha, \beta, c, d$):** the density function of this distribution is given by:
  
  \[ f(x) = \frac{1}{B(\alpha, \beta)} \left(\frac{x-c}{d-c}\right)^{\alpha-1} \left(\frac{d-x}{d-c}\right)^{\beta-1}, \ \text{ with } \alpha, \beta > 0, x \in [c,d] \]
  
  \[ c, d \in \mathbb{R}, \text{ and } B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} \]
  
  We have \( \text{E}(X) = c+(c-d)\alpha/(\alpha+\beta) \) and \( \text{V}(X) = (c-d)^2\alpha\beta/[(\alpha+\beta+1)(\alpha+\beta)^2] \)
  
  Pour the type I beta distribution, \( X \) takes values in the \([0,1]\) range. The beta4 distribution is obtained by a variable transformation such that the distribution is on a \([c,d]\) interval where \( c \) and \( d \) can take any value.

- **Beta ($a, b$):** the density function of this distribution (also called Beta type I) is given by:
\[ f(x) = \frac{1}{B(a,b)} x^{a-1} (1-x)^{b-1}, \quad \text{with } a,b>0, \ x \in [0,1] \text{ and } B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \]

\[ \text{E}(X) = a/(a+b) \text{ and } \text{V}(X) = ab/[(a+b+1)(a+b)^2] \]

- **Binomial (n, p):** the density function of this distribution is given by:
  \[ P(X = x) = C_n^x p^x (1-p)^{n-x}, \quad \text{with } x \in \mathbb{N}, \ n \in \mathbb{N}^+, \ p \in [0,1] \]

  \[ \text{E}(X) = np \text{ and } \text{V}(X) = np(1-p) \]

  \( n \) is the number of trials, and \( p \) the probability of success. The binomial distribution is the distribution of the number of successes for \( n \) trials, given that the probability of success is \( p \).

- **Negative binomial type I (n, p):** the density function of this distribution is given by:
  \[ P(X = x) = C_{n+x-1}^x p^n (1-p)^x, \quad \text{with } x \in \mathbb{N}, \ n \in \mathbb{N}^+, \ p \in [0,1] \]

  \[ \text{E}(X) = n(1-p)/p \text{ and } \text{V}(X) = n(1-p)/p^2 \]

  \( n \) is the number of successes, and \( p \) the probability of success. The negative binomial type I distribution is the distribution of the number of unsuccessful trials necessary before obtaining \( n \) successes.

- **Negative binomial type II (k, p):** the density function of this distribution is given by:
  \[ P(X = x) = \frac{\Gamma(k+x)}{x!\Gamma(k)(1+p)^{k+x}} p^x, \quad \text{with } x \in \mathbb{N}, \ k, p>0 \]

  \[ \text{E}(X) = kp \text{ and } \text{V}(X) = kp(p+1) \]

  The negative binomial type II distribution is used to represent discrete and highly heterogeneous phenomena. As \( k \) tends to infinity, the negative binomial type II distribution tends towards a Poisson distribution with \( \lambda = kp \).

- **Chi-square (df):** the density function of this distribution is given by:
  \[ f(x) = \frac{(1/2)^{df/2}}{\Gamma(df/2)} x^{df/2-1} e^{-x/2}, \quad \text{with } x > 0, \ df \in \mathbb{N}^* \]

  \[ \text{E}(X) = df \text{ and } \text{V}(X) = 2df \]

  The Chi-square distribution corresponds to the distribution of the sum of \( df \) squared standard normal distributions. It is often used for testing hypotheses.

- **Erlang (k, \lambda):** the density function of this distribution is given by:
\[ f(x) = \lambda^k x^{k-1} \frac{e^{-\lambda x}}{(k-1)!}, \quad \text{with } x \geq 0 \text{ and } k, \lambda > 0 \text{ and } k \in \mathbb{N} \]

\[ \text{E}(X) = k/\lambda \quad \text{and} \quad \text{V}(X) = k/\lambda^2 \]

k is the shape parameter and \( \lambda \) is the rate parameter.

This distribution, developed by the Danish scientist A. K. Erlang (1878-1929) when studying the telephone traffic, is more generally used in the study of queuing problems.

Note: When \( k=1 \), this distribution is equivalent to the exponential distribution. The Gamma distribution with two parameters is a generalization of the Erlang distribution to the case where \( k \) is a real and not an integer (for the Gamma distribution the scale parameter \( \beta = 1/\lambda \) is used).

- **Exponential(\( \lambda \))**: the density function of this distribution is given by:
  \[ f(x) = \lambda \exp(-\lambda x), \quad \text{with } x > 0 \text{ and } \lambda > 0 \]
  \[ \text{E}(X) = 1/\lambda \quad \text{and} \quad \text{V}(X) = 1/\lambda^2 \]

  The exponential distribution is often used for studying lifetime in quality control.

- **Fisher (\( df_1, df_2 \))**: the density function of this distribution is given by:
  \[ f(x) = \frac{1}{xB\left(df_1/2, df_2/2\right)}\left(\frac{df_1x}{df_1x+df_2}\right)^{df_1/2}\left(1-\frac{df_1x}{df_1x+df_2}\right)^{df_2/2}, \]
  \[ \text{with } x \geq 0 \text{ and } df_1, df_2 \in \mathbb{N}^* \]
  \[ \text{E}(X) = df_2/(df_2 - 2) \text{ if } df_2 > 0, \text{ and } \text{V}(X) = 2df_2^2(df_1+df_2 - 2)/[df_1(df_1 - 2)^2(df_2 - 4)] \]

  Fisher's distribution, from the name of the biologist, geneticist and statistician Ronald Aylmer Fisher (1890-1962), corresponds to the ratio of two Chi-square distributions. It is often used for testing hypotheses.

- **Fisher-Tippett (\( \beta, \mu \))**: the density function of this distribution is given by:
  \[ f(x) = \frac{1}{\beta} \exp\left(-\frac{x-\mu}{\beta} - \exp\left(-\frac{x-\mu}{\beta}\right)\right), \quad \text{with } \beta > 0 \]
  \[ \text{E}(X) = \mu + \beta \gamma \quad \text{and} \quad \text{V}(X) = (\pi \beta^2/6) \text{ where } \gamma \text{ is the Euler-Mascheroni constant.} \]

  The Fisher-Tippett distribution, also called the Log-Weibull or extreme value distribution, is used in the study of extreme phenomena. The Gumbel distribution is a special case of the Fisher-Tippett distribution where \( \beta = 1 \) and \( \mu = 0 \).

- **Gamma (\( k, \beta, \mu \))**: the density of this distribution is given by:
\[ f(x) = (x - \mu)^{k-1} e^{-(x-\mu)/\beta} \frac{e^{-(x-\mu)/\beta}}{\beta^k \Gamma(k)}, \quad \text{with } x > \mu \text{ and } k, \beta > 0 \]

\[ E(X) = \mu + k\beta \text{ and } V(X) = k\beta^2 \]

\( k \) is the shape parameter of the distribution and \( \beta \) the scale parameter.

- **GEV \((\beta, k, \mu)\):** the density function of this distribution is given by:

\[ f(x) = \frac{1}{\beta} \left( 1 - k \frac{x - \mu}{\beta} \right)^{1/k-1} \exp\left(-\left(1 - k \frac{x - \mu}{\beta}\right)^{1/k}\right), \quad \text{with } \beta > 0 \]

We have \( E(X) = \mu + \frac{\beta}{k} \Gamma(1 + k) \) and \( V(X) = \left(\frac{\beta}{k}\right)^2 \left(\Gamma(1 + 2k) - \Gamma^2(1 + k)\right) \)

The GEV (Generalized Extreme Values) distribution is much used in hydrology for modeling flood phenomena. \( k \) lies typically between -0.6 and 0.6.

- **Gumbel:** the density function of this distribution is given by:

\[ f(x) = \exp\left(-x - \exp(-x)\right) \]

\( E(X) = \gamma \) and \( V(X) = \pi^2/6 \) where \( \gamma \) is the Euler-Mascheroni constant (0.5772156649…).

The Gumbel distribution, named after Emil Julius Gumbel (1891-1966), is a special case of the Fisher-Tippett distribution with \( \beta=1 \) and \( \mu=0 \). It is used in the study of extreme phenomena such as precipitations, flooding and earthquakes.

- **Logistic \((\mu, s)\):** the density function of this distribution is given by:

\[ f(x) = \frac{e^{-s(x-\mu)}}{s\left(1 + e^{-s(x-\mu)}\right)}, \quad \text{with } \mu \in \mathbb{R}, \text{ and } s > 0 \]

We have \( E(X) = \mu \) and \( V(X) = (\pi s)^2/3 \)

- **Lognormal \((\mu, \sigma)\):** the density function of this distribution is given by:

\[ f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-(\ln(x)-\mu)^2/2\sigma^2}, \quad \text{with } x, \sigma > 0 \]

\( E(X) = \exp(\mu + \sigma^2/2) \) and \( V(X) = [\exp(\sigma^2)-1]\exp(2\mu + \sigma^2) \)

- **Lognormal2 \((m, s)\):** the density function of this distribution is given by:
\[ f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x) - \mu)^2}{2\sigma^2}}, \quad \text{with } x, \sigma > 0 \]

\[ \mu = \ln(m) - \ln(1+\frac{s^2}{m^2})/2 \] and \[ \sigma^2 = \ln(1+\frac{s^2}{m^2}) \]

E(X) = m and V(X) = s^2

This distribution is just a reparametrization of the Lognormal distribution.

- **Normal (\(\mu, \sigma\)):** the density function of this distribution is given by:

  \[ f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad \text{with } \sigma > 0 \]

  E(X) = \(\mu\) and V(X) = \(\sigma^2\)

- **Standard normal:** the density function of this distribution is given by:

  \[ f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \]

  E(X) = 0 and V(X) = 1

  This distribution is a special case of the normal distribution with \(\mu=0\) and \(\sigma=1\).

- **Pareto (a, b):** the density function of this distribution is given by:

  \[ f(x) = \frac{ab^a}{x^{a+1}}, \quad \text{with } a, b > 0 \text{ and } x \geq b \]

  E(X) = \(ab/(a-1)\) and V(X) = \(ab^2/[((a-1)^2(a-2))]\)

  The Pareto distribution, named after the Italian economist Vilfredo Pareto (1848-1923), is also known as the Bradford distribution. This distribution was initially used to represent the distribution of wealth in society, with Pareto’s principle that 80% of the wealth was owned by 20% of the population.

- **PERT (a, m, b):** the density function of this distribution is given by:
\[ f(x) = \frac{1}{B(\alpha, \beta)} \frac{(x-a)^{\alpha-1}(b-x)^{\beta-1}}{(b-a)^{\alpha+\beta-1}}, \text{ with } \alpha, \beta > 0, x \in [a, b] \]

\[ a, b \in \mathbb{R}, \text{ and } B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} \]

\[ \alpha = \frac{4m + b - 5a}{b - a} \]

\[ \beta = \frac{5b - a - 4m}{b - a} \]

We have \( E(X) = (b-a)\alpha/(\alpha+\beta) \) and \( V(X) = (b-a)^2\alpha\beta/[(\alpha+\beta+1)(\alpha+\beta)] \)

The PERT distribution is a special case of the beta4 distribution. It is defined by its definition interval \([a, b]\) and \(m\) the most likely value (the mode). PERT is an acronym for Program Evaluation and Review Technique, a project management and planning methodology. The PERT methodology and distribution were developed during the project held by the US Navy and Lockheed between 1956 and 1960 to develop the Polaris missiles launched from submarines. The PERT distribution is useful to model the time that is likely to be spent by a team to finish a project. The simpler triangular distribution is similar to the PERT distribution in that it is also defined by an interval and a most likely value.

- **Poisson:** the density function of this distribution is given by:

\[ P(X = x) = \frac{\exp(-\lambda)\lambda^x}{x!}, \text{ with } x \in \mathbb{N} \text{ and } \lambda > 0 \]

\( E(X) = \lambda \) and \( V(X) = \lambda \)

Poisson's distribution, discovered by the mathematician and astronomer Siméon-Denis Poisson (1781-1840), pupil of Laplace, Lagrange and Legendre, is often used to study queuing phenomena.

- **Student:** the density function of this distribution is given by:

\[ f(x) = \frac{\Gamma\left(\frac{df+1}{2}\right)}{\sqrt{\pi df} \Gamma\left(\frac{df}{2}\right)} \left(1 + \frac{x^2}{df}\right)^{-\frac{(df+1)/2}{2}}, \text{ with } df > 0 \]

\( E(X) = 0 \text{ if } df > 1 \text{ and } V(X) = \frac{df}{(df-2)} \text{ if } df > 2 \)

The English chemist and statistician William Sealy Gosset (1876-1937), used the nickname Student to publish his work, in order to preserve his anonymity (the Guinness brewery forbade its employees to publish following the publication of confidential information by another researcher). The Student’s t distribution is the
distribution of the mean of \( \text{df} \) variables standard normal variables. When \( \text{df}=1 \), Student's distribution is a Cauchy distribution with the particularity of having neither expectation nor variance.

- **Trapezoidal (a, b, c, d):** the density function of this distribution is given by:

\[
\begin{align*}
    f(x) &= \frac{2(x-a)}{(d+c-b-a)(b-a)}, \quad x \in [a,b] \\
    f(x) &= \frac{2}{(d+c-b-a)}, \quad x \in [b,c] \\
    f(x) &= \frac{2(d-x)}{(d+c-b-a)(d-c)}, \quad x \in [a,b] \\
    f(x) &= 0, \quad x < a, x > d
\end{align*}
\]

with \( a < m < b \)

We have \( E(X) = \frac{(d^2+c^2-b^2-a^2+cd-ab)}{[3(d+c-b-a)]} \)

and \( V(X) = \frac{[(c+d)(c^2+d^2)-(a+b)(a^2+b^2)]}{[6(d+c-b-a)]}-E^2(X) \)

This distribution is useful to represent a phenomenon for which we know that it can take values between two extreme values (a and d), but that it is more likely to take values between two values (b and c) within that interval.

- **Triangular (a, m, b):** the density function of this distribution is given by:

\[
\begin{align*}
    f(x) &= \frac{2(x-a)}{(b-a)(m-a)}, \quad x \in [a,m] \\
    f(x) &= \frac{2(b-x)}{(b-a)(b-m)}, \quad x \in [m,b] \\
    f(x) &= 0, \quad x < a, x > b
\end{align*}
\]

with \( a < m < b \)

We have \( E(X) = \frac{(a+m+b)}{3} \) and \( V(X) = \frac{(a^2+m^2+b^2-ab-am-bm)}{18} \)

- **TriangularQ (q1, m, q2, p1, p2):** the density function of this distribution is a reparametrization of the Triangular distribution. A first step requires estimating the a and b parameters of the triangular distribution, from the q1 and q2 quantiles to which percentages p1 and p2 correspond. Once this is done, the distribution functions can be computed using the triangular distribution functions.

- **Uniform (a, b):** the density function of this distribution is given by:
\[ f(x) = \frac{1}{b-a}, \text{ with } b > a \text{ and } x \in [a, b] \]

\[ E(X) = (a+b)/2 \text{ and } V(X) = (b-a)^2/12 \]

The uniform (0,1) distribution is much used for simulations. As the cumulative distribution function of all the distributions is between 0 and 1, a sample taken in a Uniform (0,1) distribution is used to obtain random samples in all the distributions for which the inverse can be calculated.

- **Uniform discrete (a, b):** the density function of this distribution is given by:
  \[ f(x) = \frac{1}{b-a+1}, \text{ with } b > a, \ (a, b) \in N, \ x \in N, \ x \in [a, b] \]
  We have \( E(X) = (a+b)/2 \) and \( V(X) = [(b-a+1)^2 -1]/12 \)
  The uniform discrete distribution corresponds to the case where the uniform distribution is restricted to integers.

- **Weibull \((\beta)\):** the density function of this distribution is given by:
  \[ f(x) = \beta x^{\beta-1} \exp\left(-x^\beta\right), \text{ with } x > 0 \text{ and } \beta > 0 \]
  We have \( E(X) = \Gamma\left(\frac{1}{\beta} + 1\right) \) and \( V(X) = \Gamma\left(\frac{2}{\beta} + 1\right) - \Gamma^2\left(\frac{1}{\beta} + 1\right) \)
  \( \beta \) is the shape parameter for the Weibull distribution.

- **Weibull \((\beta, \gamma)\):** the density function of this distribution is given by:
  \[ f(x) = \frac{\beta}{\gamma} \left(\frac{x}{\gamma}\right)^{\beta-1} \exp\left(-\left(\frac{x}{\gamma}\right)\right), \text{ with } x > 0, \text{ and } \beta, \gamma > 0 \]
  We have \( E(X) = \gamma \Gamma\left(\frac{1}{\beta} + 1\right) \) and \( V(X) = \gamma^2 \left[ \Gamma\left(\frac{2}{\beta} + 1\right) - \Gamma^2\left(\frac{1}{\beta} + 1\right) \right] \)
  \( \beta \) is the shape parameter of the distribution and \( \gamma \) the scale parameter. When \( \beta=1 \), the Weibull distribution is an exponential distribution with parameter \( 1/\gamma \).

- **Weibull \((\beta, \gamma, \mu)\):** the density function of this distribution is given by:
  \[ f(x) = \frac{\beta}{\gamma} \left(\frac{x-\mu}{\gamma}\right)^{\beta-1} \exp\left(-\left(\frac{x-\mu}{\gamma}\right)^{\beta}\right), \text{ with } x > \mu, \text{ and } \beta, \gamma > 0 \]
We have $E(X) = \mu + \gamma \Gamma \left( \frac{1}{\beta} + 1 \right)$ and $V(X) = \gamma^2 \left[ \Gamma \left( \frac{2}{\beta} + 1 \right) - \Gamma^2 \left( \frac{1}{\beta} + 1 \right) \right]$

The Weibull distribution, named after the Swede Ernst Hjalmar Waloddi Weibull (1887-1979), is much used in quality control and survival analysis. $\beta$ is the shape parameter of the distribution and $\gamma$ the scale parameter. When $\beta=1$ and $\mu=0$, the Weibull distribution is an exponential distribution with parameter $1/\gamma$.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- ![Reload](image): Click this button to reload the default options.
- ![Delete](image): Click this button to delete the data selections.
- ![Change Data](image): Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab:**

**Data**: Select the quantitative data. If several samples have been selected, XLSTAT will carry out the calculations for each of the samples independently while allowing you to superimpose histograms if you want (see Charts tab). If headers have been selected, check that the "Sample labels" option has been activated.

**Data type**: 

- **Continuous**: Choose this option so that XLSTAT considers your data to be continuous.
**Discrete**: Choose this option so that XLSTAT considers your data to be discrete.

**Subsamples**: Activate this option then select a column (column mode) or a row (row mode) containing the sample identifiers. The use of this option gives one histogram per subsample and therefore allows to compare the distribution of data between the subsamples. If a header has been selected, check that the "Sample labels" option has been activated.

- **Variable-Category labels**: Activate this option to use variable-category labels when displaying outputs. Variable-Category labels include the variable name as a prefix and the category name as a suffix.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Sample labels**: Activate this option if the first row of the selected data (data, sub-samples, weights) contains a label.

**Weights**: Check this option if the observations are weighted. If you do not check this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Sample labels" option is activated.

**Options** tab:

**Intervals**: Choose one of the following options to define the intervals for the histogram:

- **Number**: Choose this option to enter the number of intervals to create.
- **Width**: Choose this option to define a fixed width for the intervals.
- **User defined**: Select a column containing in increasing order the lower bound of the first interval, and the upper bound of all the intervals.
- **Minimum**: Activate this option to enter the value of the lower value of the first interval. This value must be lower or equal to the minimum of the series.

**Missing data** tab:
Remove observations:

- **For the corresponding sample**: Activate this option to ignore an observation which has a missing value only for samples which have a missing value.

- **For all samples**: Activate this option to ignore an observation which has a missing value for all selected samples.

**Estimate missing data**: Activate this option to estimate the missing data by using the mean of the sample.

**Outputs** tab:

**Descriptive statistics**: Activate this option to display the descriptive statistics of the samples.

**Charts** tab:

**Histograms**: Activate this option to display the histograms of the samples. For a theoretical distribution, the density function is displayed.

- **Bars**: Choose this option to display the histograms with a bar for each interval.

- **Continuous lines**: Choose this option to display the histograms with a continuous line.

**Cumulative histograms**: Activate this option to display the cumulative histograms of the samples.

- **Based on the histogram**: Choose this option to display cumulative histograms based on the same interval definition as the histograms.

- **Empirical cumulative distribution**: Choose this option to display cumulative histograms which actually correspond to the empirical cumulative distribution of the sample.

**Ordinate of the histograms**: Choose the quantity to be used for the histograms: density, frequency or relative frequency.

**Display a distribution**: Activate this option to compare histograms of samples selected with a density function and/or to compare the histograms of samples selected with a distribution function. Then choose the distribution to be used and enter the values of the parameters if necessary.
Results

Summary statistics: This table displays for the selected samples, the number of observations, the number of missing values, the number of non-missing values, the mean and the standard deviation.

Histograms: The histograms are displayed. If desired, you can change the color of the lines, scales, titles as with any Excel chart.

Descriptive statistics for the intervals: This table displays for each interval its lower bound, upper bound, the frequency (number of values of the sample within the interval), the relative frequency (the number of values divided by the total number of values in the sample), and the density (the ratio of the frequency to the size of the interval).

Example

An example showing how to create a histogram is available on the Addinsoft website at

http://www.xlstat.com/demo-histo.htm

References


Normality tests

Use this tool to check if a sample can be considered to follow a normal distribution. The
distribution fitting tool enables the parameters of the normal distribution to be estimated but the
tests offered are not as suitable as those given here.

Description

Assuming a sample is normally distributed is common in statistics. But checking that this is
actually true is often neglected. For example, the normality of residuals obtained in linear
regression is rarely tested, even though it governs the quality of the confidence intervals
surrounding parameters and predictions.

XLSTAT offers four tests for testing the normality of a sample:

The Shapiro-Wilk test which is best suited to samples of less than 5000 observations;

The Anderson-Darling test proposed by Stephens (1974) is a modification of the Kolmogorov-
Smirnov test and is suited to several distributions including the normal distribution for cases
where the parameters of the distribution are not known and have to be estimated;

The Lilliefors test is a modification of the Kolmogorov-Smirnov test and is suited to normal
cases where the parameters of the distribution, the mean and the variance are not known and
have to be estimated;

The Jarque-Bera test which is more powerful the higher the number of values.

In order to check visually if a sample follows a normal distribution, it is possible to use P-P
plots and Q-Q plots:

P-P Plots (normal distribution): P-P plots (for Probability-Probability) are used to compare the
empirical distribution function of a sample with that of a sample distributed according to a
normal distribution of the same mean and variance. If the sample follows a normal distribution,
the points will lie along the first bisector of the plan.

Q-Q Plots (normal distribution): Q-Q plots (for Quantile-Quantile) are used to compare the
quantities of the sample with those of a sample distributed according to a normal distribution of
the same mean and variance. If the sample follows a normal distribution, the points will lie
along the first bisector of the plan.
**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload default options**: Click this button to reload the default options.
- **Delete data selections**: Click this button to delete the data selections.
- **Change data handling**: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General**

**Data**: Select the quantitative data. If several samples have been selected, XLSTAT carries out normality tests for each of the samples independently. If headers have been selected, check that the "Sample labels" option has been activated.

**Weights**: Activate this option if the observations are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Sample labels" option is activated.

**Shapiro-Wilk test**: Activate this option to perform a Shapiro-Wilk test.

**Anderson-Darling test**: Activate this option to perform an Anderson-Darling test.

**Lilliefors test**: Activate this option to carry out a Lilliefors test.

**Jarque-Bera test**: Activate this option to carry out a Jarque-Bera test.
**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Sample labels**: Activate this option if the first row of the selected data (data, sub-samples, weights) contains a label.

**Significance level (%)**: Enter the significance level for the tests.

**Subsamples**: Activate this option then select a column (column mode) or a row (row mode) containing the sample identifiers. The use of this option gives one series of tests per subsample. If a header has been selected, check that the "Sample labels" option has been activated.

**Missing data** tab:

Remove observations:

- **For the corresponding sample**: Activate this option to ignore an observation which has a missing value only for samples which have a missing value.

- **For all samples**: Activate this option to ignore an observation which has a missing value for all selected samples.

**Estimate missing data**: Activate this option to estimate the missing data by using the mean of the sample.

**Outputs** tab:

**Descriptive statistics**: Activate this option to display the descriptive statistics of the samples.

**Charts** tab:

**P-P plots**: Activate this option to display Probability-Probability plots based on the normal distribution.

**Q-Q Plots**: Activate this option to display Quantile-Quantile plots based on the normal distribution.
Results

For each test requested, the statistics relating to the test are displayed including, in particular, the p-value which is afterwards used in interpreting the test by comparing with the chosen significance threshold.

If requested, P-P and Q-Q plots are then displayed.

Example

An example showing how to test the normality of a sample is available on the Addinsoft website:

http://www.xlstat.com/demo-norm.htm

References


Resampling

Use this tool to calculate descriptive statistics using resampling methods (bootstrap, jackknife…) for a set of quantitative variables.

Description

Resampling methods have become more and more popular since computational power has increased. It is a well-known approach to nonparametric statistics. The principle is very simple: from your original sample, randomly draw a new sample and recalculate statistics. Repeating this step many times gives you the empirical distribution of the statistic, from which you obtain the standard error, and confidence intervals.

With XLSTAT, you can apply these methods on a selected number of descriptive statistics for quantitative data.

Three resampling methods are available:

- **Bootstrap**: It is the most famous approach; it has been introduced by Efron and Tibisharni (1993). It is a statistical method for estimating the sampling distribution of an estimator by sampling with replacement from the original sample. The number of samples has to be given.

- **Random without replacement**: Subsamples are drawn randomly from the original sample. The size of the subsample has to be specified.

- **Jackknife**: The sampling procedure is based on suppressing one observation to the original sample (of size N). Each subsample has N-1 observations and the process is repeated N times. It is less robust than the bootstrap.

Although you can select several variables (or samples) at the same time, XLSTAT calculates all the descriptive statistics for each of the samples independently.

**Descriptive statistics for quantitative data:**

Let’s consider a sample made up of N items of quantitative data \{y_1, y_2, \ldots, y_N\} whose respective weights are \{W_1, W_2, \ldots, W_N\).

- **Sum**: The weighted sum of the values is defined by:

\[ S = \sum_{i=1}^{n} w_i x_i \]
• **Mean**: The mean of the sample is defined by $\mu = S / Sw$.

• **Variance (n)**: The variance of the sample defined by:

$$s(n)^2 = \frac{\sum_{i=1}^{n} w_i (x_i - \mu)^2}{Sw}$$

Note 1: When all the weights are 1, the variance is the sum of the square deviation to the mean divided by n, hence its name.

Note 2: The variance (n) is a biased estimate of the variance which assumes that the sample is a good representation of the total population. The variance (n-1) is, on the other hand, calculated taking into account an approximation associated with the sampling.

• **Variance (n-1)**: The estimated variance of the sample defined by:

$$s(n-1)^2 = \frac{\sum_{i=1}^{n} w_i (x_i - \mu)^2}{Sw - Sw/n}$$

Note 1: When all the weights are 1, the variance is the sum of the square deviation to the mean divided by n-1, hence its name.

Note 2: The variance (n) is a biased estimate of the variance which assumes that the sample is a good representation of the total population. The variance (n-1) is, on the other hand, calculated taking into account an approximation associated with the sampling.

• **Standard deviation (n)**: The standard deviation of the sample defined by $s(n)$.

• **Standard deviation (n-1)**: The standard deviation of the sample defined by $s(n-1)$.

• **Median**: The median $Q2$ is the value for which 50% of the values are less.

• **1st quartile**: The first quartile $Q1$ is defined as the value for which 25% of the values are less.

• **3rd quartile**: The third quartile $Q3$ is defined as the value for which 75% of the values are less.

• **Variation coefficient**: this coefficient is only calculated if the mean of the sample is non-zero. It is defined by $CV = s(n) / \mu$. This coefficient measures the dispersion of a sample relative to its mean. It is used to compare the dispersion of samples whose scales or means differ greatly.

• **Standard error of the mean**: this statistic is defined by:
\[ s_{\mu} = \sqrt{\frac{s(n-1)^2}{Sw}} \]

- **Mean absolute deviation***: as for standard deviation or variance, this coefficient measures the dispersion (or variability) of the sample. It is defined by:

\[ e(\mu) = \frac{\sum_{i=1}^{n} w_i |x_i - \mu|}{Sw} \]

- **Median absolute deviation***: this statistic is the median of absolute deviations to the median.

- **Geometric mean***: this statistic is only calculated if all the values are strictly positive. It is defined by:

\[ \mu_G = \exp\left(\frac{1}{Sw} \sum_{i=1}^{n} w_i \ln(x_i)\right) \]

If all the weights are equal to 1, we have:

\[ \mu_G = \sqrt[n]{\prod_{i=1}^{n} x_i} \]

- **Geometric standard deviation***: this statistic is defined by:

\[ \sigma_G = \exp\left(\frac{1}{Sw} \sum_{i=1}^{n} w_i \left(\ln(x_i) - \ln(\mu_G)\right)^2\right) \]

- **Harmonic mean***: this statistic is defined by:

\[ \mu_H = \frac{Sw}{\sum_{i=1}^{n} \frac{w_i}{x_i}} \]

(\*) Statistics followed by an asterisk take the weight of observations into account.

**Statistics obtained after resampling:**

Let S be one of the preceding statistics, during the resample procedure it has been computed B times. In the case of bootstrap and random without replacement, we have:

- **Mean**: It is the mean on the B samples:
\[ \hat{\mu}^* (S) = \frac{1}{B} \sum_{i=1}^{B} \hat{S}_i \]

where \( \hat{S}_i \) is the estimated value of \( S \) for sample \( i \).

- **Standard error:**
  \[
  \hat{\sigma}^* (S) = \sqrt{\frac{1}{B-1} \sum_{i=1}^{B} (\hat{S}_i - \hat{\mu}^* (S))^2}
  \]

- **Standard bootstrap confidence interval:** It is defined by:
  \[
  \left[ S \pm u_{1-\alpha/2} \hat{\sigma}^* (S) \right]
  \]
  where \( u \) is the 1-alpha/2 percentile of the normal distribution and 1-alpha is the confidence degree. This type of interval depends on a parametric distribution.

- **Simple percentile confidence interval:** Confidence interval limits are obtained using the alpha/2 and 1-alpha/2 percentiles of the empirical distribution of \( S \).

- **Bias corrected percentile confidence interval:** Confidence interval limits are also obtained using percentiles of the empirical distribution of \( S \), but with a small difference. These limits are noted \( S[a1] \) and \( S[a2] \). Let \( p \) be the proportion of \( S_i \) lower than \( S \) (value of the statistic on the original sample). \( U_p \) is the percentile associated to the normal distribution with probability \( p \). Then we have: \( a_1 = \Phi \left( 2u_p + u_{a/2} \right) \) and \( a_2 = \Phi \left( 2u_p + u_{1-a/2} \right) \). For more details on this approach please refer to Efron and Tibshirani (1993).

*For the Jackknife:*

- **Mean:** \( \hat{\mu}^* (S) = \frac{1}{n} \sum_{i=1}^{n} \hat{S}_{(-i)} \) where \( S(-i) \) is obtained on the sample without observation \( i \).

- **Standard error:** \( \hat{\sigma}^* (S) = \sqrt{\frac{n-1}{n} \sum_{i=1}^{n} (\hat{S}_{(-i)} - \hat{\mu}^* (S))^2} \)
**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload**: Click this button to reload the default options.
- **Delete**: Click this button to delete the data selections.
- **Data Handling**: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

- **Quantitative data**: Select the samples of quantitative data you want to calculate descriptive statistics for.
- **Method**: Choose the resampling method you want to use.
  - **Bootstrap**: Check this button to apply the bootstrap method.
  - **Random without replacement**: Check this button to apply the random without replacement method.
  - **Jackknife**: Check this button to apply the Jackknife approach.

- **Sample size**: Enter the size of the subsample in the case of random without replacement.
- **Number of sample**: Enter the number of sample in the case of the bootstrap and the random without replacement.

- **Range**: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.
**Sheet**: Check this option to display the results in a new worksheet in the active workbook.

**Workbook**: Check this option to display the results in a new workbook.

**Sample labels**: Check this option if the first line of the selections (quantitative data, qualitative date, sub-samples, and weights) contains a label.

**Weights**: Check this option if the observations are weighted. If you do not check this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Sample labels" option is activated.

**Missing data** tab:

**Remove observations**: Activate this option to ignore an observation that has a missing value.

**Estimate missing data**: Activate this option to estimate the missing data by using the mean of the sample.

**Outputs** tab:

**Quantitative Data**: Activate the options for the descriptive statistics you want to calculate. The various statistics are described in the description section.

- **All**: Click this button to select all.
- **None**: Click this button to deselect all.
- **Display vertically**: Check this option so that the table of descriptive statistics is displayed vertically (one line per descriptive statistic).

**Confidence interval**: Enter the size of the confidence interval (in %).

**Standard bootstrap confidence interval**: Activate this option to display the standard bootstrap confidence interval.

**Simple percentile confidence interval**: Activate this option to display the simple percentile confidence interval.

**Bias corrected percentile confidence interval**: Activate this option to display the bias corrected percentile confidence interval.

**Resampled statistics**: Activate this option to display the resampled statistics.

**Resampled data**: Activate this option to display the resampled data.
Charts tab:

Histograms: Activate this option to display the histograms of the samples. For a theoretical
distribution, the density function is displayed.

- Bars: Choose this option to display the histograms with a bar for each interval.
- Continuous lines: Choose this option to display the histograms with a continuous line.

Cumulative histograms: Activate this option to display the cumulative histograms of the
samples.

- Based on the histogram: Choose this option to display cumulative histograms based
  on the same interval definition as the histograms.
- Empirical cumulative distribution: Choose this option to display cumulative
  histograms that actually correspond to the empirical cumulative distribution of the
  sample.

Ordinate of the histograms: Choose the quantity to be used for the histograms: density,
frequency or relative frequency.

Results

Summary statistics: This table displays for the selected samples, the number of
observations, the number of missing values, the number of non-missing values, the mean and
the standard deviation.

Resampling: This table displays for the selected statistics, the mean, the standard error and
the confidence interval obtained with resampling.

Resampled statistics: This table displays the resampled statistics for each of the B samples.

Resampled data: This table displays the B samples obtained by resampling the initial data.

Histograms: The histograms are displayed. If desired, you can change the color of the lines,
scales, titles as with any Excel chart.

Descriptive statistics for the intervals: This table displays for each interval its lower bound,
upper bound, the frequency (number of values of the sample within the interval), the relative
frequency (the number of values divided by the total number of values in the sample), and the
density (the ratio of the frequency to the size of the interval).
**Example**

An example showing how to create apply bootstrap is available on the Addinsoft website at

http://www.xlstat.com/demo-resample.htm

**References**


Similarity/dissimilarity matrices (Correlations, ...)

Use this tool to calculate a proximity index between the rows or the columns of a data table. The most classic example of the use of this tool is in calculating a correlation or covariance matrix between quantitative variables.

Description

This tool offers a large number of proximity measurements between a series of objects whether they are in rows (usually the observations) or in columns (usually the variables).

The correlation coefficient is a measurement of the similarity of the variables: the more the variables are similar, the higher the correlation coefficient.

Similarities and dissimilarities

The proximity between two objects is measured by measuring at what point they are similar (similarity) or dissimilar (dissimilarity).

The indexes offered depend on the nature of the data:

- **Quantitative data:**
  The similarity coefficients proposed by the calculations from the quantitative data are as follows: Cosine, Covariance (n-1), Covariance (n), Inertia, Gower coefficient, Kendall correlation coefficient, Pearson correlation coefficient, Spearman correlation coefficient.

  The dissimilarity coefficients proposed by the calculations from the quantitative data are as follows: Bhattacharya's distance, Bray and Curtis' distance, Canberra's distance, Chebychev's distance, Chi² distance, Chi² metric, Chord distance, Squared chord distance, Euclidian distance, Geodesic distance, Kendall's dissimilarity, Mahalanobis distance, Manhattan distance, Ochiai's index, Pearson's dissimilarity, Spearman's dissimilarity.

- **Binary data:**
  The similarity and dissimilarity (pay simple transformation) coefficients proposed by the calculations from the binary data are as follows: Dice coefficient (also known as the Sorensen coefficient), Jaccard coefficient, Kulczinski coefficient, Pearson Phi, Ochiai coefficient, Rogers & Tanimoto coefficient, Sokal & Michener's coefficient (simple matching coefficient), Sokal & Sneath's coefficient (1), Sokal & Sneath's coefficient (2).
• Qualitative data:

The similarity coefficients proposed by the calculations from the qualitative data are as follows: Cooccurrence, Percent agreement.

The dissimilarity coefficients proposed by the calculations from the qualitative data are as follows: Percent disagreement

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- Click this button to start the computations.
- Click this button to close the dialog box without doing any computation.
- Click this button to display the help.
- Click this button to reload the default options.
- Click this button to delete the data selections.

General tab:

Data: Select a table comprising N objects described by P descriptors. If column headers have been selected, check that the "Column labels" option has been activated.

Data type: Choose the type of the selected data.

Note (1): in the case where the selected data type is « Binary », if the input data are not binary, XLSTAT asks you if they should be automatically transformed into binary data (all values that are not equal to 0 are replaced by 1s).

Note (2): in the case where the selected data type is « Qualitative », whatever the true type of the data, they are considered as qualitative.
**Row weights:** Activate this option if the rows are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Column labels" option is activated.

**Proximity type:** similarities / dissimilarities: Choose the proximity type to be used. The data type and proximity type determine the list of possible indexes for calculating the proximity matrix.

Note: to calculate a classical correlation coefficient (also called Pearson's correlation coefficient) you must select data types "quantitative", "similarities" and "Pearson's correlation coefficient".

**Range:** Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Activate this option to display the results in a new worksheet of the active workbook.

**Workbook:** Activate this option to display the results in a new workbook.

**Column labels:** Activate this option if the first row of the data selections (Observations/variables table, row labels, row weights, column weights) contains a label.

**Row labels:** Activate this option if observations labels are available. Then select the corresponding data. If the "Column labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Compute proximities for:

**Columns:** Activate this option to measure proximities between columns.

**Rows:** Activate this option to measure proximities between rows.

**Missing data** tab:

**Do not accept missing data:** Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

**Remove observations:** Activate this option to remove the observations with missing data.
**Estimate missing data**: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

- **Nearest neighbour**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.

**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the variables selected.

**Proximity matrix**: Activate this option to display the proximity matrix.

**Flag similar objects**: Activate this option to identify similar objects in the proximity matrix.

**List similar objects**: Activate this option to display the list of similar objects.

**Dissimilarity threshold**: Enter the threshold value of the index from which you consider objects to be similar. If the index chosen is a similarity, the values will be considered as being similar if they are greater than this value. If the index chosen is a dissimilarity, the values will be considered as being similar if they are less than this value.

**Cronbach’s Alpha**: Activate this option to calculate Cronbach’s alpha coefficient.

**Bartlett’s sphericity test**: Activate this option to calculate Bartlett’s sphericity test (only for Pearson correlation or covariance).

**Significance level (%)**: Enter the significance level for the sphericity test.

**Results**

**Summary statistics**: This table shows the descriptive statistics for the samples.

**Proximity matrix**: This table displays the proximities between the object for the chosen index. If the "Identify similar objects" option has been activated and the dissimilarity threshold has been exceeded, the values for the similar objects are displayed in bold.
List of similar objects: If the "List similar objects" option has been checked and at least one pair of objects has a similarity beyond the threshold, the list of similar objects is displayed.

Example

An example showing how to compute a dissimilarity matrix is available on the Addinsoft website:

http://www.xlstat.com/demo-mds.htm

References


Biserial correlation

Use this tool to compute the biserial correlation between on one hand, one or more quantitative variables, and on the other hand, one or more qualitative binary variables.

Description

This tool allows computing the biserial correlation between one hand one or more quantitative variables and the other one or more binary variables. The biserial correlation introduced by Pearson (1909), between a quantitative variable and a binary variable is given by:

$$ r = \frac{\hat{\mu}_2 - \hat{\mu}_1}{\hat{\sigma}_{n}} \sqrt{p_1p_2} $$

Where $\hat{\mu}_1$ and $\hat{\mu}_2$ are the estimated means for the two possible values of the binary variable, $\hat{\sigma}_{n}$ is the biased standard deviation estimated on all the data, and $p_1$ and $p_2$ are the proportions corresponding to the two values of the binary variable ($p_1+p_2=1$). As for the Pearson correlation, the biserial correlation coefficient varies between -1 and 1. 0 corresponds to no association (the means of the quantitative variable for the two categories of the qualitative variable are identical).

XLSTAT allows testing if the $r$ value that has been obtained is different from 0 or not.

For the two-tailed test, the null $H_0$ and alternative $H_a$ hypotheses are as follows:

- $H_0 : r = 0$
- $H_a : r \neq 0$

In the left one-tailed test, the following hypotheses are used:

- $H_0 : r = 0$
- $H_a : r < 0$

In the right one-tailed test, the following hypotheses are used:

- $H_0 : r = 0$
- $H_a : r > 0$
Two methods to compute the p-value are proposed by XLSTAT. The user can choose between a p-value computed using on a large sample approximation, and a p-value computed using Monte Carlo resamplings. The second method is recommended.

To compute the p-value using the large sample approximation, we use the following result:

If \( n \) is the full sample size, the statistic defined by

\[
t = r \sqrt{\frac{n - 2}{1 - r^2}}
\]

follows a Student distribution with \( n-2 \) degrees of freedom under the null hypothesis.

Note: the XLSTAT_Biserial spreadsheet function can be used to compute the biserial correlation between a quantitative variable and a binary variable.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload default options**: Click this button to reload the default options.
- **Delete data selections**: Click this button to delete the data selections.
- **Change data handling**: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab:**
**Qualitative variables**: Activate this option to select one or more quantitative variables. If a column header has been selected, check that the "Variable labels" option is activated.

**Qualitative variables**: Activate this option to select one or more binary qualitative variables. If a column header has been selected, check that the "Variable labels" option is activated.

**Weights**: Check this option if the observations are weighted. If you do not check this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Variable labels**: Activate this option if the first row of the data selections (variables, weights, observations labels) includes a header.

**Options** tab:

**Alternative hypothesis**: Choose the alternative hypothesis to be used for the test (see the description section for more information).

**Significance level (%)**: Enter the significance level for the test (default value: 5%).

**Asymptotic p-value**: Activate this option if you want XLSTAT to calculate the p-value based on the asymptotic approximation (see the description section for more information).

**Monte Carlo method**: Activate this option if you want XLSTAT to calculate the p-value based on Monte Carlo permutations, and select the number of random permutations to perform or the maximum time to spend.

**Missing data** tab:

**Do not accept missing data**: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

**Remove the observations**: Activate this option to remove observations with missing data.
• **For the corresponding variable**: Activate this option to ignore an observation which has a missing value only for the variables that have a missing value.

• **For all variables**: Activate this option to ignore an observation which has a missing value for all selected variables.

**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the selected variables.

**Results**

**Descriptive statistics**: The table of descriptive statistics shows the simple statistics for all the variables selected. The number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the quantitative variables. For qualitative variables, the categories with their respective frequencies and percentages are displayed.

The biserial correlation is then given for each pair (quantitative variable, qualitative variable). The p-values are then displayed if they have been requested. The details for the test are given only when the correlation is calculated one quantitative variable and one qualitative variable.

**Example**

An example showing how to compute the biserial correlation is available on the Addinsoft website. To download this data, go to:

http://www.xlstat.com/demo-biserial.htm

**References**


**Pearson K. (1909)**. On a New Method of Determining Correlation between a measured Character A and a Character B, of which only the Percentage of cases wherein B exceeds (or falls short of) a given Intensity is recorded for each grade of A. *Biometrika*, 7, 96-105.
*Journal of General Psychology, 8*, 463-465.
**Multicolinearity statistics**

Use this tool to identify multicolinearities between your variables.

**Description**

Variables are said to be multicolinear if there is a linear relationship between them. This is an extension of the simple case of colinearity between two variables. For example, for three variables $X_1$, $X_2$ and $X_3$, we say that they are multicolinear if we can write:

$$X_1 = aX_2 + bX_3$$

where $a$ and $b$ are real numbers.

Principle Component Analysis (PCA) can detect the presence of multicolinearities within the data (a number of non-null factors less than the number of variables indicates the presence of a multicolinearity), but it cannot identify the variables which are responsible.

To detect the multicolinearities and identify the variables involved, linear regressions must be carried out on each of the variables as a function of the others. We then calculate

The R² of each of the models. If the R² is 1, then there is a linear relationship between the dependent variable of the model (the Y) and the explanatory variables (the Xs).

The **tolerance** for each of the models. The tolerance is $(1-R^2)$. It is used in several methods (linear regression, logistic regression, discriminant factorial analysis) as a criterion for filtering variables. If a variable has a tolerance less than a fixed threshold (the tolerance is calculated by taking into account variables already used in the model), it is not allowed to enter the model as its contribution is negligible and it risks causing numerical problems.

The VIF (Variance Inflation Factor) which is equal to the inverse of the tolerance.

Detect multicolinearities within a group of variables can be useful especially in the following cases:

To identify structures within the data and take operational decisions (for example, stop the measurement of a variable on a production line as it is strongly linked to others which are already being measured),

To avoid numerical problems during certain calculations. Certain methods use matrix inversions. The inverse of a $(p \times p)$ matrix can be calculated if it is of rank $p$ (or regular). If it is of lower rank, in other words, if there are linear relationships between its columns, then it is singular and cannot be inverted.
**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- ![Reload](image): Click this button to reload the default options.
- ![Delete](image): Click this button to delete the data selections.
- ![Change Data Handling](image): Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab:**

- **Observations/variables table**: Select a table with N observations and P variables. If column headers have been selected, check that the "Variable labels" option has been activated.
- **Variable labels**: Activate this option if the first row of the selection includes a header.
- **Weights**: Check this option if the observations are weighted. If you do not check this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

- **Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.
- **Sheet**: Activate this option to display the results in a new worksheet of the active workbook.
- **Workbook**: Activate this option to display the results in a new workbook.
**Missing data** tab:

**Do not accept missing data**: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

**Remove the observations**: Activate this option to remove observations with missing data.

**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the selected variables.

**Correlations**: Activate this option to display the correlations matrix.

**R²**: Activate this option to display the R-squares.

**Tolerance**: Activate this option to display the tolerances.

**VIF**: Activate this option to display the VIFs.

**Charts** tab:

**Bar charts**: Activate this option to display the bar charts of the following statistics:

- **R²**
- Tolerance
- VIF

**Results**

The results comprise the descriptive statistics of the variables selected, the correlation matrix of the variables and the multicolinearity statistics (R², Tolerance and VIF). Bar charts are used to locate the variables which are more multi-correlated than the others.

When the tolerance is 0, the VIF has infinite value and is not displayed.
References

Contingency tables (descriptive statistics)

Use this tool to compute a variety of descriptive statistics on a contingency table. A chi-square test is optionally performed. Additional tests on contingency tables are available in the “Tests on contingency tables” section.

Description

A contingency table is an efficient way to summarize the relation (or correspondence) between two categorical variables V1 and V2. It has the following structure:

<table>
<thead>
<tr>
<th>V1 \ V2</th>
<th>Category 1</th>
<th>...</th>
<th>Category j</th>
<th>...</th>
<th>Category m2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category 1</td>
<td>n(1,1)</td>
<td></td>
<td>n(1,j)</td>
<td></td>
<td>n(1,m2)</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
<td>...</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>Category i</td>
<td>n(i,1)</td>
<td></td>
<td>n(i,j)</td>
<td></td>
<td>n(i,m2)</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
<td>...</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>Category m1</td>
<td>n(m1,1)</td>
<td></td>
<td>n(m1,j)</td>
<td></td>
<td>n(m1,m2)</td>
</tr>
</tbody>
</table>

where n(i,j) is the frequency of observations that show both characteristic i for variable V1, and characteristic j for variable V2.

The Chi-square distance has been suggested to measure the distance between two categories. The Pearson chi-square statistic, which is the sum of the Chi-square distances, is used to test the independence between rows and columns. It has asymptotically a Chi-square distribution with (m1-1)(m2-1) degrees of freedom.

Inertia is a measure inspired from physics that is often used in Correspondence Analysis, a method that is used to analyse in depth contingency tables. The inertia of a set of points is the weighted mean of the squared distances to the center of gravity. In the specific case of a contingency table, the total inertia of the set of points (one point corresponds to one category) can be written as:

\[
\phi^2 = \frac{\chi^2}{n} = \sum_{i=1}^{m1} \sum_{j=1}^{m2} \left( \frac{n_{ij} - \frac{n_i n_j}{n^2}}{n_i n_j} \right)^2, \quad \text{with} \quad n_i = \sum_{j=1}^{m2} n_{ij} \quad \text{and} \quad n_j = \sum_{i=1}^{m1} n_{ij}
\]
and where $n$ is the sum of the frequencies in the contingency table. We can see that the inertia is proportional to the Pearson chi-square statistic computed on the contingency table.

**Bootstrap confidence intervals**

XLSTAT allows you to obtain bootstrap confidence interval around the theoretical frequency of each pair of categories in a contingency table. It offers an alternative to the classical Chi-square by cell.

The method is as follow:

1- Build a dataset with two qualitative variables using the value of the contingency table.

2- Randomly draw with replacement $N$ observations from the dataset for both variables independently.

3- Build a contingency table with the new dataset.

4- Repeat 2 and 3 as many times as specified by the user.

5- Compute mean, standard error, confidence interval and percentile confidence intervals for each pair of categories.

Pairs with observed value out of the confidence interval show a significant difference between the two categories (Amiri *et al*. 2011).

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload defaults**: Click this button to reload the default options.
- **Delete selections**: Click this button to delete the data selections.
Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

**Contingency table**: Select the data that correspond to the contingency table. If row and column labels are included, make sure that the "Labels included" option is checked.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Labels included**: Activate this option if the row and column labels are selected.

**Options** tab:

**Chi-square test**: Activate this option to display the statistics and the interpretation of the Chi-square test of independence between rows and columns.

**Significance level (%)**: Enter the significance level for the test.

**Bootstrap confidence interval**: Activate this option to display the bootstrap confidence interval around the theoretical value for each pair of categories of the contingency table.

**Number of samples**: Enter the number of samples to be used to compute bootstrap confidence intervals.

**Missing data** tab:

**Do not accept missing data**: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

**Replace missing data by 0**: Activate this option if you consider that missing data are equivalent to 0.
Replace missing data by their expected value: Activate this option if you want to replace the missing data by the expected value. The expectation is given by:

\[ E(n_{ij}) = \frac{n_i n_j}{n} \]

where \( n_i \) is the row sum, \( n_j \) is the column sum, and \( n \) is the grand total of the table before replacement of the missing data.

Outputs tab:

List of combines: Activate this option to display the table that lists all the possible combines between the two variables that are used to create a contingency table, and the corresponding frequencies.

Contingency table: Activate this option to display the contingency table.

Inertia by cell: Activate this option to display the inertia for each cell of the contingency table.

Chi-square by cell: Activate this option to display the contribution to the chi-square of each cell of the contingency table.

Significance by cell: Activate this option to display a table indicating, for each cell, if the actual value is equal (=), lower (<) or higher (>) than the theoretical value, and to run a test (Fisher’s exact test of on a 2x2 table having the same total frequency as the complete table, and the same marginal sums for the cell of interest), in order to determine if the difference with the theoretical value is significant or not.

Observed frequencies: Activate this option to display the table of the observed frequencies. This table is almost identical to the contingency table, except that the marginal sums are also displayed.

Theoretical frequencies: Activate this option to display the table of the theoretical frequencies computed using the marginal sums of the contingency table.

Proportions or percentages / Row: Activate this option to display the table of proportions or percentages computed by dividing the values of the contingency table by the marginal sums of each row.

Proportions or percentages / Column: Activate this option to display the table of proportions or percentages computed by dividing the values of the contingency table by the marginal sums of each column.

Proportions or percentages / Total: Activate this option to display the table of proportions or percentages computed by dividing the values of the contingency table by the sum of all the cells of the contingency table.
Raw data: Activate this option to display the raw data table, meaning the observations/variables table, having \( N \) rows and 2 columns.

Charts tab:

3D view of the contingency table: Activate this option to display the 3D bar chart corresponding to the contingency table.

References

**XLSTAT-Pivot**

Use this module to turn an individuals/variables table into a dynamic pivot table optimized to let you understand and analyze the issue phenomenon corresponding to one of the variables describing the individuals.

**Description**

XLSTAT-Pivot is a unique solution that allows you to quickly create intelligent pivot tables.

XLSTAT-Pivot is based on classification trees using the CHAID algorithm in order to find the most relevant explanatory variables of a response variable.

A *pivot table* (or contingency table, or two-way table) is a synthetic representation of occurrences observed on an N-size population for crosses of all the different categories of two variables.

A *dynamic pivot table* allows to take more than two variables into account and to organize the table structure into a hierarchy. The table is said to be dynamic in the sense that software functionalities allow to navigate among the hierarchy and to create a focused view on particular classes of particular variables.

XLSTAT-Pivot allows you to create dynamic pivot tables whose structure is optimized with respect to a target variable. Numeric continuous or discrete explanatory variables are automatically sliced into classes that contribute to optimize the quality of the table.

The target variable can be a qualitative variable, or a quantitative variable.

XLSTAT-Pivot uses classification trees to discretize the quantitative variables and to identify the contributions of the variables (see the chapter on classification trees). The CHAID method is used because it suits well the dynamic pivot table representation.

When you run XLSTAT-Pivot you will see successively two dialog boxes:

- The first dialog box lets you select the data and a few options.

- The second dialog box allows you to select the dimensions that you want to use in the pivot table (up to four variables may be selected). To help you select the variables the explanatory power of each variable are displayed. A specific score is used for that purpose (see bellow for a detailed description).

XLSTAT-Pivot offers some other options (default options should give the best results). For example, you can choose to use an external method for discretizing the quantitative
explanatory variables. A sensitivity index is also available in order to better fit your needs in term of complexity of the tree generated.

**Explanatory variables score index**

In order to evaluate the contribution of the variables on the response variable, an index has been used. It will differ depending on the type of response variable.

In the case of a **quantitative response variable**, the score index for each variable (quantitative or qualitative), as defined by Breiman et al. (1984), is:

\[
Score(var_i) = \sum_{j \in T} i^2 I(var_i \in Node_j)
\]

With

\[
i^2 = \frac{v_i v_j}{v_i + v_j} \left( \bar{y}_i - \bar{y}_j \right)^2
\]

with \( i \) and \( j \) being the two nodes separating the studied node, \( T \) being the tree and \( I(var_i \in Node_j) \) the indicator function.

The weights \( v \) are computed with:

\[
v_i = \frac{n_i}{N} \left( 1 - \frac{n_i}{N} \right), \quad n_i \text{ being the number of observation associated to the leaf and } N \text{ being the number of observations associated to the studied node.}
\]

In the case of a **qualitative response variable**, the score index for each variable (quantitative or qualitative), as defined by Breiman et al. (1984), is:

\[
Score(var_i) = \sum_{j \in T} i^2 I(var_i \in Node_j)
\]

With

\[
i^2 = \frac{v_i v_j}{v_i + v_j} \sum_{k=1}^{n-\text{nod}} \left( p_{ik} - p_{jk} \right)^2
\]

with \( i \) and \( j \) being the two nodes separating the studied node, \( T \) being the tree and

\[
I(var_i \in Node_j) = \begin{cases} 1 & \text{if node } j \text{ is associated to variable } i \\ 0 & \text{if not} \end{cases}
\]
The weights \( w \) are computed with: \( w_i = \frac{n_i}{N} \left( 1 - \frac{n_i}{N} \right) \), \( n_i \) being the number of observation associated to the leaf and \( N \) being the number of observations associated to the studied node. The probabilities are the probabilities of having modality \( k \) of the response variable in each leaf.

**Sensitivity index associated to the tree**

Building a classification tree requires to set a number of parameters (maximum depth, leaf size, the thresholds of grouping and separation ...). To simplify the use of XLSTAT-Pivot, a sensitivity index was developed. It takes values between 0 and 1.

When this index is close to 0, then the building of the tree is not sensitive to small differences. The number of intervals in the discretization of the quantitative variables will be lower and the size of the tree will be small. It is therefore the strongest contributions that will be revealed in the pivot table.

When this index is close to 1, then the building of the tree is very sensitive to small differences. The number of intervals in the discretization of the quantitative variables will be larger and the size of the tree will be large. All contributions will be revealed in the pivot table (but sometime too many).

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload**: Click this button to reload the default options.
- **Delete**: Click this button to delete the data selections.
Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

**Y / Response variable**: Select the response variable you want to model. If a column header has been selected, check that the "Variable labels" option has been activated.

Choose the type of response variable you have selected:

- **Quantitative**: If you select this option, you must select a quantitative variable.
- **Qualitative**: If you select this option, you must select a qualitative variable. You must then select a target category which will be used for the outputs of the pivot table. A new box with the list of the categories of the response variable will appear on the right.

**X / Explanatory variables**

**Quantitative**: Activate this option if you want to include one or more quantitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The data selected may be of the numerical type. If a variable header has been selected, check that the "Variable labels" option has been activated.

**Qualitative**: Activate this option if you want to include one or more qualitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If a variable header has been selected, check that the "Variable labels" option has been activated.

**Weights**: Check this option if the observations are weighted. If you do not check this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Sample labels" option is activated.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.
**Variable labels**: Activate this option if the first row of the data selections (response and explanatory variables, weights, observations labels) includes a header.

**Observation labels**: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

**Options** tab:

**Sensitivity**: Enter the value of the sensitivity parameter. When it is close to 1, the classification tree is large. When it is close to 0, the classification tree is small. For a detailed description, please refer to the description part of this chapter. The default value is 0.5.

**Discretization – quantitative variables**: this option is enabled only if quantitative explanatory variables have been selected.

- **Automatic**: Activate this option to use the automatic discretization within the tree algorithm (this is the default option).
- **Equal width**: Activate this option to discretize the quantitative variable using equal width intervals.
- **Equal frequency**: Activate this option to discretize the quantitative variable using equal frequency intervals.
- **User defined**: Activate this option to discretize the quantitative variable using user defined interval. Select a table with one row for each bound of the intervals and one column for each variable.

**Missing data** tab:

**Remove observations**: Activate this option to remove the observations with missing data.

**Estimate missing data**: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbour**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.
**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the variables selected.

**Discretization**: Activate this option to display the discretized explanatory variables.

**Contributions**: Activate this option to display the contributions table and the corresponding bar chart.

**Pivot table**: Activate this option to display the dynamic pivot table.

**Results**

**Descriptive statistics**: The table of descriptive statistics shows the simple statistics for all the variables selected. The number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the quantitative variables. For qualitative variables, including the response variable, the categories with their respective frequencies and percentages are displayed.

The next table presents the **discretized explanatory variables**.

The next table presents the **variables contributions** (raw, % relative and cumulated contribution). This table allows you to quickly see which variables have the greater impact on the target variable. A bar chart of the contributions is also displayed. This histogram is an Excel chart that you can modify to suit your needs.

The most important result provided by XLSTAT-Pivot is the **dynamic pivot table**. Each cell corresponds to a unique combination of the values of the explanatory variables. It is described by the following 4 values, that can be displayed or not according to the user preferences:

- **Target average**: Percentage of the cases where the target category of the response variable is present in the case of a qualitative variable; average of the target variable calculated on the sub-population corresponding to the combination in the case of continuous variable;

- **Target size**: Count of the occurrences of the target category for the response variable in the case of qualitative variable;

- **Population size %**: Percentage of the overall population corresponding to the combination;

- **Population size**: Population size corresponding to the combination.
Example

An example based on data collected for a population census in the United States is permanently available on the Addinsoft web site. To download this data, go to:

http://www.xlstat.com/demo-pivot.htm

References

Scatter plots

Use this tool to create 2- or 3-dimensional plots (the 3rd dimension being represented by the size of the point), or indeed 4-dimensional plots (a qualitative variable can be selected). This tool is also used to create matrices of plots to enable a study of a series of 2-dimensional plots to be made at the same time.

Note: XLSTAT-3DPlot can create plots with much more impact thanks to its large number of options with the possibility of representing data on a third axis.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload default options**: Click this button to reload the default options.
- **Delete data selections**: Click this button to delete the data selections.
- **Change data handling**: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab**:

**X**: In this field select the data to be used as coordinates along the X-axis.

**Y**: In this field select the data to be used as coordinates along the Y-axis.

**Z**: Check this option to select the values which will determine the size of the points on the charts.
• **Use bubbles**: Check this option to use charts with MS Excel bubbles.

**Groups**: Check this option to select the values which correspond to the identifier of the group to which each observation belongs. On the chart, the color of the point depends on the group.

**Range**: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Check this option to display the results in a new worksheet in the active workbook.

**Workbook**: Check this option to display the results in a new workbook.

**Variable labels**: Check this option if the first line of the selected data (X, Y, Z, Groups, Weights and observation labels) contains a label.

**Observation labels**: Check this option if you want to use the available line labels. If you do not check this option, labels will be created automatically (Obs1, Obs2, etc.). If a column header has been selected, check that the "Variable labels" option has been activated.

**Weights**: Check this option if the observations are weighted. If you do not check this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

**Options** tab:

**Matrix of plots**: Check this option to display all possible combinations of variables in pairs in the form of a two-entry table with Y-variables in rows and Y-variables in columns.

• **Histograms**: Activate this option so that if the X and Y variables are identical, XLSTAT displays a histogram instead of a X/X plot.

• **Q-Q plots**: Activate this option so that if the X and Y variables are identical, XLSTAT displays a Q-Q plot instead of a X/X plot.

**Frequencies**: Check this option to display the frequencies for each point on the charts.

• **Only if >1**: Check this option if you only want frequencies strictly greater than zero to be displayed.

**Confidence ellipses**: Activate this option to display confidence ellipses. The confidence ellipses correspond to a 95% confidence interval for a bivariate normal distribution with the same means and the same covariance matrix as the variables represented in abscissa and ordinates.

**Legend**: Check this option if you want the chart legend to be displayed.
Example

A tutorial on using Scatter plots is available on the XLSTAT website on the following page:

http://www.xlstat.com/demo-scatter.htm

References


Parallel coordinates plots

Use this tool to visualize multidimensional data (described by P quantitative and Q qualitative variables) on a single two dimensional chart.

Description

This visualization method is useful for data analysis when you need to discover or validate groups. For example, this method could be used after Agglomerative Hierarchical Clustering.

If you consider N observations described by P quantitative and Q qualitative variables, the chart consists of P+Q vertical axes each representing a variable, and N lines corresponding to each observation. A line crosses an axis at the value corresponding to the value that the observation takes for the variable corresponding to the axis.

If the number of observations is too high, the visualization might be not very efficient or even impossible due to the Excel restrictions (maximum of 255 data series). In that case, it is recommended to use the random sampling option.

Dialog box

OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

Reload default options.

Delete data selections.

Change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.
**General** tab:

**Quantitative Data:** Check this option to select the samples of quantitative data you want to calculate descriptive statistics for.

**Qualitative Data:** Check this option to select the samples of qualitative data you want to calculate descriptive statistics for.

**Weights:** Check this option if the observations are weighted. If you do not check this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

**Groups:** Check this option to select the values which correspond to the identifier of the group to which each observation belongs. On the chart, the color of the point depends on the group.

**Range:** Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Check this option to display the results in a new worksheet in the active workbook.

**Workbook:** Check this option to display the results in a new workbook.

**Variable labels:** Check this option if the first line of the selected data (quantitative data, qualitative data, weights, groups and observation labels) contains a label.

**Observation labels:** Check this option if you want to use the available line labels. If you do not check this option, line labels will be created automatically (Obs1, Obs2, etc.). If a column header has been selected, check that the "Variable labels" option has been activated.

**Rescale:** Check this option so that all variables are represented on the same scale of 0% to 100% (for numeric variables 0 corresponds to the minimum and 100 to the maximum; for all nominal variables, the categories are regularly spaced and classified in alphabetic order.)

**Options** tab:

**Display as many lines as possible:** Check this option to display as many lines as possible (the maximum is 250 due to the limitations of Excel).

**Display the descriptive statistics lines:** Check this option to display lines for the following statistics only:

- Minimum and maximum
- Median
- **First quantile (%)**: enter the value of the first quantile (2.5% by default).
- **Second quantile (%)**: enter the value of the second quantile (97.5% by default).
- Mode (for qualitative variables)

**Example**

A tutorial on generating a parallel coordinates plot is available on the Addinsoft website at the following address:

http://www.xlstat.com/demo-pcor.htm

**References**


Ternary diagrams

Use this tool to create ternary diagrams to represent within a triangle a set of points that have their coordinates in a three-dimensional space, with the constraint that the sum of the coordinates is constant.

Description

This visualization method is particularly useful in domains where one works with three elements with varying proportions, for example in chemistry or petrology.

This tool lets you quickly create a ternary diagram representing points and the projection lines connecting each point to each axis.

There are two approaches for ternary graphs:

- Either segments corresponding to the orthogonal projection of the points on the axes give the information on the relative proportions of the three elements.

- Or the projection parallel to the axis A onto the axis B corresponds to the coordinate of the point along the axis B, where B is after A when turning counterclockwise.

XLSTAT currently allows only the second approach.

Dialog box

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload Options**: Click this button to reload the default options.
- **Delete Data Selections**: Click this button to delete the data selections.
- **Data Handling**: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.
**General tab:**

* X: Check this option to select the data corresponding to the first element.
* Y: Check this option to select the data corresponding to the second element.
* Z: Activate this option to select the quantitative data that correspond to the third element.

**Range:** Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Check this option to display the results in a new worksheet in the active workbook.

**Workbook:** Check this option to display the results in a new workbook.

**Variable labels:** Check this option if the first line of the selected data contains a label.

**Options tab:**

**Constant:** Check this option to display as many lines as possible (the maximum is 250 due to the limitations of Excel).

**Charts tab:**

* X1/X2/X3 | Min / Max: You can modify the min and the Max for each variable. However you must take into account that the (Max-Min) for each dimension is the same.

* Number of segments: Enter the number of segments into which you want to divide each axis of the ternary chart.

* Projection lines: Activate this option to display dotted red lines between the points and their coordinate on each axis.

* Lines between axes: Activate this option to display the lines between the axes.

* Link to input data: Activate this option to link the chart to the input data. If you check this option, a change in the input data is immediately reflected on the ternary diagramm.
Example

A tutorial on generating ternary plots is available on the Addinsoft website at the following address:

http://www.xlstat.com/demo-ternary.htm
2D plots for contingency tables

Use this tool to create a 2-dimensional plot based on a contingency table.

Description

This visualization tool allows to quickly generate a 2D plot showing the relative importance of the various combinations that you can obtain when creating a two-way contingency table (also called cross-tabs).

This tool can work directly on raw data (weighted or not) or on a contingency table.

Dialog box

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- ![Reload](image): Click this button to reload the default options.
- ![Delete](image): Click this button to delete the data selections.
- ![Direction](image): Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab:**

- **Contingency table**: If the data format is "contingency table", select the data that correspond to the contingency table. If row and column labels are included, make sure that the "Labels included" option is checked.
Qualitative variable(1): If the data format is "qualitative variables", select the data that correspond to the qualitative variable that will be used to construct the rows of the contingency table, and that will be used for the ordinates axis of the plot.

Qualitative variable(2): If the data format is "qualitative variables", select the data that correspond to the qualitative variable that will be used to construct the columns of the contingency table, and that will be used for the abscissa axis of the plot.

Z: If the data format is "qualitative variables", check this option to select the values which will weigh the observations and modify the size of the points on the plot.

Data format: Select the data format.
- Contingency table: Activate this option if your data correspond to a contingency table.
- Qualitative variables: Activate this option if your data are available as two qualitative variables to be used to create a contingency table.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Labels included: Activate this option if the row and column labels of the contingency table are selected.

Variable labels: Activate this option if the first row of the data selections includes a header.

Display title: Check this option, to display a title on the plot.

Options tab:

Use bubbles: Check this option to use MS Excel bubbles.

Shape: Select the shape you want to use.
- Circle
- Square
**Rescale**: Choose the interval of sizes to use when displaying the points. The minimum must be between 2 and 71, and the maximum between 3 and 72.

**Example**

A tutorial on generating a 2D plot for a contingency table is available at:

Error bars

Use this tool to easily create Excel charts with error bars that can be different for each point.

Description

This tool has is there to get around a failure of Excel: if it is possible to add error bars on different types of graphs, this operation is tedious if the bounds are not the same for all points. With this tool you can create a chart with error bars at once.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- OK: Click this button to start the computations.
- Cancel: Click this button to close the dialog box without doing any computation.
- Help: Click this button to display the help.
- : Click this button to reload the default options.
- : Click this button to delete the data selections.
- : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

X: Select in this field the data to be used as coordinates for the x-axis. If you select several columns (column mode) or several rows (row mode), you must then select the same number of columns (or rows) for Y, the lower bounds, and the upper bounds. However, if you select a single column (or row), you can then select one or more columns (or rows) for Y, the lower bounds and upper bounds.
Y: Select in this field that data to be used as coordinates on the y-axis. See above the constraints that apply to the number of columns.

**Lower bound:** Activate this option if you want to add lower bounds on the graph. Then select in this field the data to be used as lower bounds. The number of columns (column mode) or rows (row mode) to be selected must be equal to that of Y.

**Upper bound:** Activate this option if you want to add lower bounds on the graph. Then select in this field the data to be used as upper bounds. The number of columns (column mode) or rows (row mode) to be selected must be equal to that of Y.

**Range:** Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Check this option to display the results in a new worksheet in the active workbook.

**Workbook:** Check this option to display the results in a new workbook.

**Variable labels:** Check this option if the first line of the selected data (X, Y, Z, Groups, Weights and observation labels) contains a label.

**Charts** tab:

**Chart type:** select the type of chart you want to display:

- Bar chart.
- Curve.
- Scatter plot.

**Example**

A tutorial on how to a chart with error bars is available on the Addinsoft website

[http://www.xlstat.com/demo-err.htm](http://www.xlstat.com/demo-err.htm)
Plot a function

Use this tool to create a chart and plot a function on it, or to add a function to an existing chart.

Description

This tool allows you to plot a function of the type \( y = f(x) \) on an existing or new chart. The syntax of the function must respect the conventions imposed by Excel for functions used in spreadsheets. In addition, the abscissa must be identified by \( X1 \).

Examples:

<table>
<thead>
<tr>
<th>Function</th>
<th>XLSTAT Syntax</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y=x^2 )</td>
<td>( X1^2 )</td>
</tr>
<tr>
<td>( Y=\ln(x) )</td>
<td>( \text{LN}(X1) )</td>
</tr>
<tr>
<td>( Y=e^x )</td>
<td>( \text{EXP}(X1) )</td>
</tr>
<tr>
<td>( Y=</td>
<td>x</td>
</tr>
<tr>
<td>( Y=x ) if ( x&lt;0 ), ( Y=2x ) if ( x\geq0 )</td>
<td>( \text{IF}(X1&lt;0,X1,2*\text{X1}) )</td>
</tr>
</tbody>
</table>

In addition, you can as well use XLSTAT worksheet functions. For example, to plot the normal cumulative distribution function, enter \( \text{XLSTAT\_CDFNormal}(X1) \).

Dialog box

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without making any changes.
- **Help**: Click this button to display the help.
- **Reload Default Options**: Click this button to reload the default options.
Click this button to delete the data selections.

**General** tab:

**Function Y =**: Enter the function that you want to plot, while respecting the syntax defined in the Description section.

**Minimum**: Enter the minimum value for which the function must be evaluated and plotted.

**Maximum**: Enter the maximum value for which the function must be evaluated and plotted.

**Number of points**: Enter the number of points at which the function must be evaluated between the minimum and maximum values. This option allows you to adjust the quality of the graph. For a function with many inflection points, too few points might give a graph of poor quality. Too many points may also degrade the quality of the display.

**Range**: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Check this option to display the results in a new worksheet in the active workbook.

**Workbook**: Check this option to display the results in a new workbook.

**Active chart**: Activate this option to add the function to the chart that is currently being selected.

**Example**

An example showing how to create a chart with a function is available on the Addinsoft website at:

AxesZoomer

Use this tool to change the minimum and maximum values on the X- and Y-axes of a plot.

Dialog box

Important: before running this tool, you must select a scatter plot or curve.

Apply: Click this button to apply changes to the plot.
Done: Click on this button to close the dialog box.
Help: Click this button to display help.

Min X: Enter the minimum value of the X-axis.
Max X: Enter the maximum value of the X-axis.
Min Y: Enter the maximum value of the X-axis.
Max Y: Enter the maximum value of the Y-axis.
EasyLabels

Use this tool to add labels, formatted if required, to a series of values on a chart.

Dialog box

Important: before running this tool, you must select a scatter plot or curve, or a series of points on a plot.

OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without making any changes.

Help: Click this button to display the help.

: Click this button to reload the default options.

: Click this button to delete the data selections.

: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that the labels are in a column. If the arrow points to the right, XLSTAT considers that the labels are in a row.

Labels: Select the labels to be added to the series of values selected on the plot.

Header in the first cell: Check this option if the first cell of the labels selected is a header and not a label.

Use the text properties: Check this option if you want the text format used in the cells containing the labels to also be applied to the text of labels in the chart:

- **Font**: Check this option to use the same character font.
- **Size**: Check this option to use the same size of font.
- **Style**: Check this option to use the same font style (normal, bold, italic).
- **Color**: Check this option to use the same font color.
**Use the cell properties:** Check this option if you want the format applied to the cells containing the labels to also be applied to the labels in the chart:

- **Border:** Check this option to use the same border.
- **Pattern:** Check this option to use the same pattern.

**Use the point properties:** Check this option if you want the label color to be the same as the color of the points:

- **Inside color:** Check this option to use the color inside the points.
- **Border color:** Check this option to use the border color of the points.
Reposition labels

Use this tool to change the position of observation labels on a chart.

Dialog box

- **Cancel**: Click this button to close the dialog box without making any changes.
- **Help**: Click this button to display the help.
- **Reload options**: Click this button to reload the default options.
- **Delete data selections**: Click this button to delete the data selections.

**Corners**: Check this option to place labels in the direction of the corner of the quadrant in which the point is located.

Distance to point:

- **Automatic**: Check this option for XLSTAT to automatically determine the most appropriate distance to the point.
- **User defined**: Check this option to enter the value (in pixels) of the distance between the label and the point.

**Above**: Check this option to place labels above the point.

**Right**: Check this option to place labels to the right of the point.

**Below**: Check this option to place labels below the point.

**Left**: Check this option to place labels to the left of the point.

**Apply only to the selected series**: Check this option to only change the position of labels for the series selected.
EasyPoints

Use this tool to modify the size, the color or the shape of the points that are displayed in an Excel chart.

Dialog box

Important: before running this tool, you must select a scatter plot or curve, or a series of points on a plot.

OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without making any changes.

Help: Click this button to display the help.

Reload: Click this button to reload the default options.

Delete: Click this button to delete the data selections.

Up/Down: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that the labels are in a column. If the arrow points to the right, XLSTAT considers that the labels are in a row.

Size: Activate this option and select the cells that give the size to be applied to the points. The size of the points is determined by the values in the cells.

Header in the first cell: Check this option if the first cell of the labels selected is a header and not a label.

Rescale: Choose the interval of sizes to use when displaying the points. The minimum must be between 2 and 71, and the maximum between 3 and 72.

Shapes and/or color: Activate this option to change the shape of the points and/or the color to be applied to the points. Select the cells and which color (if the Use the cell properties that tell which shape should be used for each point: 1 corresponds to a square, 2 to a diamond, 3 to a triangle, 4 to an x, 5 to a star (*), 6 to a point (.), 7 to a dash (-), 8 to a plus (+) and 9 to a circle (o). The color of the border of the points depends on the color of the bottom border of the cells and the inside color of the points depends on the background color of the cells (Note: the default color of the cells is “none”, so you need to set it to white to obtain white points).
Change shapes: Check this option if you want the shapes to be changed depending on the values selected in the “Shapes and or color” field.

Use the cell properties: Check this option if you want the format applied to the cells to also be applied to the points in the chart:

- **Border**: Check this option to use the cell borders as the foreground color.
- **Background**: Check this option to use the cell color as the background color.

Example

An example describing how to use the EasyPoints tool is available on the Addinsoft website at:

http://www.xlstat.com/demo-easyp.htm
Orthonormal plots

Use this tool to adjust the minimum and maximum of the X- and Y- axes so that the plot becomes orthonormal. This tool is particularly useful if you have enlarged an orthonormal plot produced by XLSTAT (for example after a PCA) and you want to ensure the plot is still orthonormal.

Note: an orthonormal plot is where a unit on the X-axis appears the same size as a unit on the Y-axis. Orthonormal plots avoid interpretation errors due to the effects of dilation or overwriting.

Dialog box

- **Apply**: Click this button to apply the transformation to the plot.
- **Cancel**: Click this button to close the dialog box without making any changes.
- **Done**: Click on this button to close the dialog box.
- **Help**: Click this button to display the help.
Plot transformations

Use this tool to apply one or more transformations to the points in a plot.

Dialog box

Important: before running this tool, you must select a scatter plot or curve.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- ![Reload](reload.png): Click this button to reload the default options.
- ![Delete](delete.png): Click this button to delete the data selections.

Symmetry:

- **Horizontal axis**: Check this option to apply a symmetry around the X-axis.
- **Vertical axis**: Check this option to apply a symmetry around the Y-axis.

  Note: if you select both the previous options, the symmetry applied will be a central symmetry.

Translation:

- **Horizontal**: Check this option to enter the number of units for a horizontal translation.
- **Vertical**: Check this option to enter the number of units for a vertical translation.

Rotation:

- **Angle (°)**: enter the angle in degrees for the rotation to be applied.
- **Right**: if this option is activated, a clockwise rotation is applied.
- **Left**: if this option is activated, an anti-clockwise rotation is applied.

Rescaling:
- **Factor**: enter the scaling factor to be applied to the data.

**Range**: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Check this option to display the results in a new worksheet in the active workbook.

**Workbook**: Check this option to display the results in a new workbook.

**Display the new coordinates**: Check this option to display the coordinates once all the transformations have been applied.

**Update Min and Max on the new plot**: Check this option for XLSTAT to automatically adjust the minimum and maximum of the X- and Y- axes, once the transformations have been carried out, so that all points are visible.

**Orthonormal plot**: Check this option for XLSTAT to automatically adjust the minimum and maximum of the X- and Y- axes, once the transformations have been carried out, so that the plot becomes orthonormal.
Merge plots

Use this tool to merge multiple plots into one.

Dialog box

Important: before using this tool, you must select at least two plots of the same type (e.g. two scatter plots).

OK: Click this button to merge the plots.

Cancel: Click on this button to close the dialog box.

Help: Click this button to display help.

Reload options: Click this button to reload the default options.

Delete data selections: Click this button to delete the data selections.

Display title: Check this option, to display a title on the merged plot.

- Title of the first chart: Check this option to use the title of the first chart.

- New title: Check this option to enter a title for the merged plot.

Orthonormal plot: Check this option for XLSTAT to verify that the plot resulting from the merged plots is orthonormal.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet in the active workbook.

Workbook: Activate this option to display the results in a new workbook.

New chart sheet: Check this option to display the plot resulting from the merge in a new chart sheet.
**Display the report header**: clear this option to stop the previous report header for the chart from being displayed.
Factor analysis

Factor analysis highlights, where possible, the existence of underlying factors common to the quantitative variables measured in a set of observations.

Description

The factor analysis method dates from the start of the 20th century (Spearman, 1904) and has undergone a number of developments, several calculation methods having been put forward. This method was initially used by psychometricians, but its field of application has little by little spread into many other areas, for example, geology, medicine and finance.

Today, there are two main types of factor analysis:

Exploratory factor analysis (or EFA)

Confirmatory factor analysis (or CFA)

It is EFA which will be described below and which is used by XLSTAT. It is a method which reveals the possible existence of underlying factors which give an overview of the information contained in a very large number of measured variables. The structure linking factors to variables is initially unknown and only the number of factors may be assumed.

CFA in its traditional guise uses a method identical to EFA but the structure linking underlying factors to measured variables is assumed to be known. A more recent version of CFA is linked to models of structural equations.

Going from p variables to k factors

Spearman's historical example, even if the subject of numerous criticisms and improvements, may still be used to understand the principle and use of the method. By analyzing correlations between scores obtained by children in different subjects, Spearman wanted to form a hypothesis that the scores depended ultimately on one factor, intelligence, with a residual part due to an individual, cultural or other effect.

Thus the score obtained by an individual (i) in subject (j) could be written as $x(i,j) = \mu + b(j)F + e(i,j)$, where $\mu$ is the average score in the sample studied and $F$ the individual's level of intelligence (the underlying factor) and $e(i,j)$ the residual.

Generalizing this structure to p subjects (the input variables) and to k underlying factors, we obtain the following model:

$$x = \mu + Af + u$$
where \( \mathbf{x} \) is a vector of dimension \((p \times 1)\), \( \mathbf{\mu} \) in the mean vector, \( \mathbf{\Lambda} \) is the matrix \((p \times k)\) of the factor loadings and \( \mathbf{f} \) and \( \mathbf{u} \) are the random vectors of dimensions \((k \times 1)\) and \((p \times 1)\) respectively are assumed to be independent. The elements of \( \mathbf{f} \) are called common factors, and those of \( \mathbf{u} \) specific factors.

If we set the norm of \( \mathbf{f} \) to 1, then the covariance matrix for the input variables from expression (1) is written as:

\[
\Sigma = \mathbf{\Lambda}\mathbf{\Lambda}' + \Psi
\]

Thus the variance of each of the variables can be divided into two parts: The communality (as it arises from the common factors),

\[
h_i^2 = \sum_{j=1}^{k} \lambda_{ij}^2,
\]

and \( \Psi_{ii} \) the specific variance or unique variance (as it is specific to the variable in question).

It can be shown that the method used to calculate matrix \( \mathbf{\Lambda} \), an essential challenge in factorial analysis, is independent of the scale. It is therefore equivalent to working from the covariance matrix or correlation matrix.

The challenge of factorial analysis is to find matrices \( \mathbf{\Lambda} \) and \( \Psi \), such that equation (2) can be at least approximately verified.

Note: factor analysis is sometimes included with Principle Component Analysis (PCA) as PCA is a special case of factor analysis (where \( k \), the number of factors, equals \( p \), the number of variables). Nevertheless, these two methods are not generally used in the same context. Indeed, PCA is first and foremost used to reduce the number of dimensions while maximizing the unchanged variability in order to obtain independent (non-correlated) factors or for visualizing data in a 2- or 3-dimensional space. Whereas, factor analysis is used to identify a latent structure and for possibly reducing afterwards the number of variables measured if they are redundant with respect to the latent factors.

**Extracting Factors**

Three methods of extracting latent factors are offered by XLSTAT:

**Principle components**: this method is also used in Principle Component Analysis (PCA). It is only offered here in order to make a comparison between the results of the three methods bearing in mind that the results from the module dedicated to PCA are more complete.

**Principal factors**: this method is probably the most used. It is an iterative method which enables the communalities to be gradually converged. The calculations are stopped when the maximum change in the communalities is below a given threshold or when a maximum
The number of iterations is reached. The initial communalities can be calculated according to various methods.

**Maximum likelihood:** this method was first put forward by Lawley (1940). The proposal to use the Newton-Raphson algorithm (iterative method) dates from Jennrich (1969). It was afterwards improved and generalized by Jöreskog (1977). This method assumes that the input variables follow a normal distribution. The initial communalities are calculated according to the method proposed by Jöreskog (1977). As part of this method, an adjustment test is calculated. The statistic used for the test follows a Chi² distribution to \((p-k)^2 / 2 - (p+k) / 2\) degrees of freedom where \(p\) is the number of variables and \(k\) the number of factors.

**Number of factors**

Determining the number of factors to select is one of the challenges of factor analysis. The "automatic" method offered by XLSTAT is uniquely based on the spectral decomposition of the correlation matrix and the detection of a threshold from which the contribution made by information (in the sense of variability) is not significant.

The likelihood maximum method offers an adjustment test to help determine the correct number of principle factors for the principle factor method. For the principal factors method, defining the number of factors is more difficult?

The Kaiser-Guttman rule suggests that only those factors with associated eigenvalues which are strictly greater than 1 should be kept. The number of factors to be kept corresponds to the first turning point found on the curve. Crossed validation methods have been suggested to achieve this aim.

**Anomalies (Heywood cases)**

Communalities are by definition the squares of correlations. They must therefore be between 0 and 1. However, it may happen that the iterative algorithms (principal factors method or likelihood maximum method) will produce solutions with communalities equal to 1 (Heywood cases), or greater than 1 (ultra Heywood cases). There may be many reasons for these anomalies (too many factors, not enough factors, etc.). When this happens, XLSTAT sets the communalities to 1 and adapts the elements of \(\Lambda\) in consequence.

**Rotations**

Once the results have been obtained, they may be transformed in order to make them more easy to interpret, for example by trying to arrange that the coordinates of the variables on the factors are either high (in absolute value), or near to zero. There are two main families of rotations:
Orthogonal rotations can be used when the factors are not correlated (hence orthogonal). The methods offered by XLSTAT are Varimax, Quartimax, Equamax, Parsimax and Orthomax. Varimax rotation is the most used. It ensures that for each factor there are few high factor loadings and few that are low. Interpretation is thus made easier as, in principle, the initial variables will mostly be associated with one of the factors.

Oblique transformations can be used when the factors are correlated (hence oblique). The methods offered by XLSTAT are Quartimin and Oblimin.

The Promax method, also offered by XLSTAT, is a mixed procedure since it consists initially of a Varimax rotation followed by an oblique rotation so that the high factor loadings and low factor loadings are the same but with the low values even lower.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload default options**: Click this button to reload the default options.
- **Delete data selections**: Click this button to delete the data selections.
- **Change data handling**: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab**

The main data entry field is used to select one of three types of table:

- **Observations/variables table / Correlation matrix / Covariance matrix**: Choose the option appropriate to the format of your data, and then select the data. If your data correspond to a
table comprising N observations described by P quantitative variables select the Observations/variables option. If column headers have been selected, check that the "Variable labels" option has been activated. If you select a correlation or covariance matrix, and if you include the variable names in the first row of the selection, you must also select them in the first column.

**Correlation**: Choose the type of matrix to be used by factor analysis.

**Extraction method**: Choose the factor extraction method to be used, The three possible methods are (see the description section for more details):

- Principal components
- Principal factors
- Maximum likelihood

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Variable labels**: Activate this option if the first row of the data selections (input table, weights, observations labels) includes a header. Where the selection is a correlation or covariance matrix, if this option is activated, the first column must also include the variable labels.

**Observation labels**: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

**Weights**: Activate this option if the observations are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

**Options** tab:

**Number of factors**:

- **Automatic**: Activate this option to make XLSTAT determine the number of factors automatically.
• **User defined** Activate this option to tell XLSTAT the number of factors to use in the calculations.

**Initial communalities**: Choose this calculation method for the initial communalities (this option is only visible for the principle factors methods):

- **Squared multiple correlations**: The initial communalities are based on a variable's level of dependence with regard to the other variables.
- **Random**: The initial communalities are drawn from the interval \([0 ; 1]\).
- **1**: The initial communalities are set to 1.
- **Maximum**: The initial communalities are set to the maximum value of the squares of the multiple correlations.

**Stop conditions**:

- **Iterations**: Enter the maximum number of iterations for the algorithm. The calculations are stopped when the maximum number of iterations has been exceeded. Default value: 50.
- **Convergence**: Enter the maximum value of the evolution in the communalities from one iteration to another which, when reached, means that the algorithm is considered to have converged. Default value: 0.0001.

**Rotation**: Activate this option if you want to apply a rotation to the factor coordinate matrix.

- **Number of factors**: Enter the number of factors the rotation is to be applied to.
- **Method**: Choose the rotation method to be used. For certain methods a parameter must be entered (Gamma for Orthomax, Tau for Oblimin, and the power for Promax).
- **Kaiser normalization**: Activate this option to apply Kaiser normalization during the rotation calculation.

**Missing data tab**:

- **Do not accept missing data**: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.
- **Remove the observations**: Activate this option to remove observations with missing data.
Pairwise deletion: Activate this option to remove observations with missing data only when
the variables involved in the calculations have missing data. For example, when calculating the
correlation between two variables, an observation will only be ignored if the data
_corresponding to one of the two variables is missing.

Estimate missing data: Activate this option to estimate the missing data before the
calculation starts.

- **Mean or mode**: Activate this option to estimate the missing data by using the mean
  (quantitative variables) or the mode (qualitative variables) for the corresponding
  variables.

- **Nearest neighbour**: Activate this option to estimate the missing data for an observation
  by searching for the nearest neighbour to the observation.

Outputs tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the selected
variables.

**Correlations**: Activate this option to display the correlation or covariance matrix depending on
the type of options chosen in the "General" tab. If the **Test significance** option is activated,
the significant correlations at the selected **significance threshold** are displayed in bold.

**Kaiser-Meyer-Olkin**: Activate this option to compute the Kaiser-Meyer-Olkin Measure of
Sampling Adequacy.

**Cronbach's Alpha**: Activate this option to compute the Cronbach's alpha coefficient.

**Eigenvalues**: Activate this option to display the table and chart (scree plot) of eigenvalues.

**Factor pattern**: Activate this option to display factor loadings (coordinates of variables in the
factor space).

**Factor/Variable correlations**: Activate this option to display correlations between factors and
variables.

**Factor pattern coefficients**: Activate this option if you want the coefficients of the factor
pattern to be displayed. Multiplying the (standardized) coordinates of the observations in the
initial space by these coefficients gives the coordinates of the observations in the factor space.

**Factor structure**: Activate this option to display correlations between factors and variables
after rotation.
**Variables charts:** Activate this option to display charts representing the variables in the new space.

- **Vectors:** Activate this option to display the initial variables in the form of vectors.

**Correlations charts:** Activate this option to display charts showing the correlations between the factors and initial variables.

- **Vectors:** Activate this option to display the initial variables in the form of vectors.

**Observations charts:** Activate this option to display charts representing the observations in the new space.

- **Labels:** Activate this option to have observation labels displayed on the charts. The number of labels displayed can be changed using the filtering option.

**Colored labels:** Activate this option to show labels in the same color as the points.

**Filter:** Activate this option to modulate the number of observations displayed:

- **Random:** The observations to display are randomly selected. The “Number of observations” N to display must then be specified.

- **N first rows:** The N first observations are displayed on the chart. The “Number of observations” N to display must then be specified.

- **N last rows:** The N last observations are displayed on the chart. The “Number of observations” N to display must then be specified.

- **Group variable:** If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to display.

**Results**

**Descriptive statistics:** The table of descriptive statistics shows the simple statistics for all the variables selected. This includes the number of observations, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased).

**Correlation/Covariance matrix:** This table shows the data to be used afterwards in the calculations. The type of correlation depends on the option chosen in the "General" tab in the dialog box. For correlations, significant correlations are displayed in bold.
Measure of Sample Adequacy of Kaiser-Meyer-Olkin: this table gives the value of the KMO measure for each individual variable and the overall KMO measure. The KMO measure ranges between 0 and 1. A low value corresponds to the case where it is not possible to extract synthetic factors (or latent variables). In other words, observations do not bring out the model that one could imagine (the sample is "inadequate"). Kaiser (1974) recommends not to accept a factor model if the KMO is less than 0.5. If the KMO is between 0.5 and 0.7 then the quality of the sample is mediocre, it is good for a KMO between 0.7 and 0.8, very good between 0.8 and 0.9 and excellent beyond.

Cronbach's Alpha: If this option has been activated, the value of Cronbach's Alpha is displayed.

Maximum change in communality at each iteration: This table is used to observe the maximum change in communality for the last 10 iterations. For the maximum likelihood method, the evolution of a criterion which is proportional to the opposite of the likelihood maximum is also displayed.

Goodness of fit test: The goodness of fit test is only displayed when the likelihood maximum method has been chosen.

Reproduced correlation matrix: This matrix is the product of the factor loadings matrix with its transpose.

Residual correlation matrix: This matrix is calculated as the difference between the variables correlation matrix and the reproduced correlation matrix.

Eigenvalues: This table shows the eigenvalues associated with the various factors together with the corresponding percentages and cumulative percentages.

Eigenvectors: This table shows the eigenvectors.

Factor pattern: This table shows the factor loadings (coordinates of variables in the vector space, also called factor pattern). The corresponding chart is displayed.

Factor/Variable correlations: This table displays the correlations between factors and variables.

Factor pattern coefficients: This table displays the coefficients of the factor pattern to be displayed. Multiplying the (standardized) coordinates of the observations in the initial space by these coefficients gives the coordinates of the observations in the factor space.

Where a rotation has been requested, the results of the rotation are displayed with the rotation matrix first applied to the factor loadings. This is followed by the modified variability percentages associated with each of the axes involved in the rotation. The coordinates of the variables and observations after rotation are displayed in the following tables.

Factor structure: This table shows the correlations between factors and variables after rotation.
Example

A tutorial on how to use Factor analysis is available on the Addinsoft website:

http://www.xlstat.com/demo-fa.htm

References


Principal Component Analysis (PCA)

Use Principle Component Analysis to analyze a quantitative observations/variables table or a correlation or covariance matrix. This method is used to:

Study and visualize correlations between variables.

Obtain non-correlated factors which are linear combinations of the initial variables.

Visualize observations in a 2- or 3-dimensional space.

Description

Principle Component Analysis (PCA) is one of the most frequently used multivariate data analysis methods. Given a table of quantitative data (continuous or discrete) in which n observations (observations, products, etc.) are described by p variables (the descriptors, attributes, measurements, etc.), if p is quite high, it is impossible to grasp the structure of the data and the nearness of the observations by merely using univariate statistical analysis methods or even a correlation matrix.

Uses of PCA

There are several uses for PCA, including:

The study and visualization of the correlations between variables to hopefully be able to limit the number of variables to be measured afterwards;

Obtaining non-correlated factors which are linear combinations of the initial variables so as to use these factors in modeling methods such as linear regression, logistic regression or discriminant analysis.

Visualizing observations in a 2- or 3-dimensional space in order to identify uniform or atypical groups of observations.

Principle of PCA

PCA can be considered as a projection method which projects observations from a p-dimensional space with p variables to a k-dimensional space (where k < p) so as to conserve the maximum amount of information (information is measured here through the total variance of the scatter plots) from the initial dimensions. If the information associated with the first 2 or 3 axes represents a sufficient percentage of the total variability of the scatter plot, the
observations will be able to be represented on a 2-3-dimensional chart, thus making interpretation much easier.

Correlations or covariance

PCA is used to calculate matrices to project the variables in a new space using a new matrix which shows the degree of similarity between the variables. It is common to use the Pearson correlation coefficient or the covariance as the index of similarity, Pearson correlation and covariance have the advantage of giving positive semi-defined matrices whose properties are used in PCA. However other indexes may be used. XLSTAT provides Spearman and Kendall or polychoric correlations for ordinal data (tetrachoric correlations are a special case of polychoric correlations which use binary data).

Traditionally, a correlation coefficient rather than the covariance is used as using a correlation coefficient removes the effect of scale: thus a variable which varies between 0 and 1 does not weigh more in the projection than a variable varying between 0 and 1000. However in certain areas, when the variables are supposed to be on an identical scale or we want the variance of the variables to influence factor building, covariance is used.

Where only a similarity matrix is available rather than a table of observations/variables, or where you want to use another similarity index, you can carry out a PCA starting from the similarity matrix. The results obtained will only concern the variables as no information on the observations was available.

Note: where PCA is carried out on a correlation matrix, it is called normalized PCA.

Interpreting the results

Representing the variables in a space of k factors enables the correlations between the variables and between the variables and factors to be visually interpreted with certain precautions.

Indeed if the observations or variables are being represented in the factor space, two points a long distance apart in a k-dimensional space may appear near in a 2-dimensional space depending on the direction used for the projection (see diagram below).
We can consider that the projection of a point on an axis, a plan or a 3-dimensional space is reliable if the sum of the squared cosines on the representation axis is near to 1. The squared cosines are displayed in the results given by XLSTAT in order to avoid any incorrect interpretation.

If the factors are afterwards to be used with other methods, it is useful to study the relative contribution (expressed as a percentage or a proportion) of the different variables in building each of the factor axes so as to make the results obtained afterwards easy to interpret. The contributions are displayed in the results given by XLSTAT.

**Number of factors**

Two methods are commonly used for determining the number of factors to be used for interpreting the results:

The *scree test* (Cattell, 1966) is based on the decreasing curve of eigenvalues. The number of factors to be kept corresponds to the first turning point found on the curve.

We can also use the cumulative variability percentage represented by the factor axes and decide to use only a certain percentage.

**Graphic representations**

One of the advantages of PCA is that it simultaneously provides the best view of the variables and data and biplots combining both (see below). However these representations are only reliable if the sum of the variability percentages associated with the axes of the representation space are sufficiently high. If this percentage is high (for example 80%), the representation can be considered as reliable. If the percentage is reliable, it is recommended to produce representations on several axis pairs in order to validate the interpretation made on the first two factor axes.
Biplots

After carrying out a PCA, it is possible to simultaneously represent both observations and variables in the factor space. The first work on this subject dates from Gabriel (1971). Gower (1996) and Legendre (1998) synthesized the previous work and extended this graphical representation technique to other methods. The term biplot is reserved for simultaneous representations which respect the fact that the projection of observations on variable vectors must be representative of the input data for the same variables. In other words, the projected points on the variable vectors must respect the order and the relative distances of the observations for that same variable, in the input data.

The simultaneous representation of observations and variables cannot be produced directly by taking the coordinates of the variables and observations in the factor space. A transformation is required in order to make the interpretation precise. Three methods using the graphic representation are available depending on the type of interpretation desired:

**Correlation biplot:** This type of biplot interprets the angles between the variables as these are directly linked to the correlations between the variables. The position of two observations projected onto a variable vector can be used to determine their relative level for this variable. The distance between the two observations is an approximation of the Mahalanobis distance in the k-dimensional factor space. Lastly, the projection of a variable vector in the representation space is an approximation of the standard deviation of the variable (the length of the vector in the k-dimensional factor space is equal to the standard deviation of the variable).

**Distance biplot:** A distance biplot is used to interpret the distances between the observations as these are an approximation of their Euclidean distance in the p-dimensional variable space. The position of two observations projected onto a variable vector can be used to determine their relative level for this variable. Lastly, the length of a variable vector in the representation space is representative of the variable's level of contribution to building this space (the length of the vector is the square root of the sum of the contributions).

**Symmetric biplot:** This biplot was proposed by Jobson (1992) and is half-way between the two previous biplots. If neither the angles nor the distances can be interpreted, this representation may be chosen as it is a compromise between the two.

XLSTAT lets you adjust the lengths of the variable vectors so as to improve the readability of the charts. However, if you use this option with a correlation biplot, the projection of a variable vector will no longer be an approximation of the standard deviation of the variable.
**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload options**: Click this button to reload the default options.
- **Delete data selections**: Click this button to delete the data selections.
- **Data handling options**: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab**:

The main data entry field is used to select one of three types of table:

**Observations/variables table / Correlation matrix / Covariance matrix**: Choose the option appropriate to the format of your data, and then select the data. If your data correspond to a table comprising N observations described by P quantitative variables, select the Observations/variables option. If column headers have been selected, check that the "Variable labels" option has been activated. If you select a correlation or covariance matrix, and if you include the variable names in the first row of the selection, you must also select them in the first column.

**PCA type**: Choose the type of matrix to be used for PCA. The difference between the Pearson (n) and the Pearson (n-1) options, only influences the way the variables are standardized, and the difference can only be noticed on the coordinates of the observations.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.
Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (observations/variables table, weights, observation labels) includes a header. Where the selection is a correlation or covariance matrix, if this option is activated, the first column must also include the variable labels.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

Options tab:

Filter factors: You can activate one of the following two options in order to reduce the number of factors for which results are displayed.

- Minimum %: Activate this option then enter the minimum percentage of the total variability that the chosen factors must represent.

- Maximum Number: Activate this option to set the number of factors to take into account.

Rotation: Activate this option if you want to apply a rotation to the factor coordinate matrix.

- Number of factors: Enter the number of factors the rotation is to be applied to.

- Method: Choose the rotation method to be used. For certain methods a parameter must be entered (Gamma for Orthomax, Tau for Oblimin, and the power for Promax).

- Kaiser normalization: Activate this option to apply Kaiser normalization during the rotation calculation.

Supplementary data tab:

Supplementary observations: Activate this option if you want to calculate the coordinates and represent additional observations. These observations are not taken into account for the factor axis calculations (passive observations as opposed to active observations). Several methods for selecting supplementary observations are provided:
- **Random**: The observations are randomly selected. The “Number of observations” N to display must then be specified.

- **N last rows**: The last N observations are selected for validation. The “Number of observations” N to display must then be specified.

- **N first rows**: The first N observations are selected for validation. The “Number of observations” N to display must then be specified.

- **Group variable**: If you choose this option, you must then select an indicator variable set to 0 for active observations and 1 for passive observations.

**Supplementary variables**: Activate this option if you want to calculate coordinates afterwards for variables which were not used in calculating the factor axes (passive variables as opposed to active variables).

- **Quantitative**: Activate this option if you have supplementary quantitative variables. If column headers were selected for the main table, ensure that a label is also present for the variables in this selection.

- **Qualitative**: Activate this option if you have supplementary qualitative variables. If column headers were selected for the main table, ensure that a label is also present for the variables in this selection.
  - **Color observations**: Activate this option so that the observations are displayed in different colors depending on the value of the first qualitative variable.
  - **Display the centroids**: Activate this option to display the centroids that correspond to the categories of the supplementary qualitative variables.
  - **Confidence ellipses**: Activate this option to display confidence ellipses around the centroids. The confidence ellipses correspond to a x% confidence interval (where x is determined using the significance level entered in the Options tab) for a bivariate normal distribution with the same means and the same covariance matrix as the factor scores for each group.

**Data options** tab:

**Missing data**: 

- **Do not accept missing data**: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

- **Remove the observations**: Activate this option to remove observations with missing data.

- **Pairwise deletion**: Activate this option to remove observations with missing data only when the variables involved in the calculations have missing data. For example, when calculating the
correlation between two variables, an observation will only be ignored if the data corresponding to one of the two variables is missing.

**Estimate missing data:** Activate this option to estimate the missing data before the calculation starts.

- **Mean or mode:** Activate this option to estimate the missing data by using the mean (quantitative variables) or the mode (qualitative variables) for the corresponding variables.
- **Nearest neighbour:** Activate this option to estimate the missing data for an observation by searching for the nearest neighbour to the observation.

**Groups:**

**By group analysis:** Activate this option and select the data that describe to which group each observation belongs, if you want that XLSTAT performs the analysis on each group separately.

**Filter:** Activate this option and select the data that describe to which group each observation belongs, if you want that XLSTAT performs the analysis for some groups that you will be able to select in a separate dialog box during the computations. If the “By group analysis” option is also activated, XLSTAT will perform the analysis for each group separately, only for the selected subset of groups.

**Outputs** tab:

**Descriptive statistics:** Activate this option to display descriptive statistics for the variables selected.

**Correlations:** Activate this option to display the correlation or covariance matrix depending on the type of options chosen in the "General" tab.

- **Test significance:** Where a correlation was chosen in the "General" tab in the dialog box, activate this option to test the significance of the correlations.
- **Bartlett's sphericity test:** Activate this option to perform the Bartlett sphericity test.
- **Significance level (%):** Enter the significance level for the above tests.
- **Kaiser-Meyer-Olkin:** Activate this option to compute the Kaiser-Meyer-Olkin Measure of Sampling Adequacy.

**Eigenvalues:** Activate this option to display the table and chart (scree plot) of eigenvalues.
**Factor loadings:** Activate this option to display the coordinates of the variables in the factor space.

**Variables/Factors correlations:** Activate this option to display correlations between factors and variables.

**Factor scores:** Activate to display the coordinates of the observations (factor scores) in the new space created by PCA.

**Contributions:** Activate this option to display the contribution tables for the variables and observations.

**Squared cosines:** Activate this option to display the tables of squared cosines for the variables and observations.

**Charts** tab:

**Variables** sub-tab:

**Correlations charts:** Activate this option to display charts showing the correlations between the components and initial variables.

- **Vectors:** Activate this option to display the initial variables in the form of vectors.
- **Colored labels:** Activate this option to show labels in the same color as the points.

**Filter:** Activate this option to modulate the number of variables displayed:

- **Random:** The observations to display are randomly selected. The “Number of variables” N to display must then be specified.
- **N first variables:** The first N variables are displayed on the chart. The “Number of variables” N to display must then be specified.
- **N last variables:** The last N variables are displayed on the chart. The “Number of variables” N to display must then be specified.
- **Group variable:** If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the variables to display.
- **Sum(Cos2)>:** Only the variables for which the sum of squared cosines (communalities) are bigger than a value to enter are displayed on the plots.

**Observations** sub-tab:

**Observations charts:** Activate this option to display charts representing the observations in the new space.
• **Labels**: Activate this option to have observation labels displayed on the charts. The number of labels displayed can be changed using the filtering option.
  
  – **Colored labels**: Activate this option to show labels in the same color as the points.

**Filter**: Activate this option to modulate the number of observations displayed:

• **Random**: The observations to display are randomly selected. The “Number of observations” N to display must then be specified.

• **N first rows**: The first N observations are displayed on the chart. The “Number of observations” N to display must then be specified.

• **N last rows**: The last N observations are displayed on the chart. The “Number of observations” N to display must then be specified.

• **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to display.

• **Sum(Cos2)>**: Only the observations for which the sum of squared cosines (communalities) are bigger than a value to enter are displayed on the plots.

**Biplots** sub-tab:

**Biplots**: Activate this option to display charts representing the observations and variables simultaneously in the new space.

• **Vectors**: Activate this option to display the initial variables in the form of vectors.

• **Labels**: Activate this option to have observation labels displayed on the biplots. The number of labels displayed can be changed using the filtering option.

**Type of biplot**: Choose the type of biplot you want to display. See the description section for more details.

• **Correlation biplot**: Activate this option to display correlation biplots.

• **Distance biplot**: Activate this option to display distance biplots.

• **Symmetric biplot**: Activate this option to display symmetric biplots.

• **Coefficient**: Choose the coefficient whose square root is to be multiplied by the coordinates of the variables. This coefficient lets you to adjust the position of the variable points in the biplot in order to make it more readable. If set to other than 1, the length of the variable vectors can no longer be interpreted as standard deviation (correlation biplot) or contribution (distance biplot).
Results

**Descriptive statistics**: The table of descriptive statistics shows the simple statistics for all the variables selected. This includes the number of observations, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased).

**Correlation/Covariance matrix**: This table shows the data to be used afterwards in the calculations. The type of correlation depends on the option chosen in the "General" tab in the dialog box. For correlations, significant correlations are displayed in bold.

**Bartlett’s sphericity test**: The results of the Bartlett sphericity test are displayed. They are used to confirm or reject the hypothesis according to which the variables do not have significant correlation.

**Measure of Sample Adequacy of Kaiser-Meyer-Olkin**: this table gives the value of the KMO measure for each individual variable and the overall KMO measure. The KMO measure ranges between 0 and 1. A low value corresponds to the case where it is not possible to extract synthetic factors (or latent variables). In other words, observations do not bring out the model that one could imagine (the sample is "inadequate"). Kaiser (1974) recommends not to accept a factor model if the KMO is less than 0.5. If the KMO is between 0.5 and 0.7 then the quality of the sample is mediocre, it is good for a KMO between 0.7 and 0.8, very good between 0.8 and 0.9 and excellent beyond.

**Eigenvalues**: The eigenvalues and corresponding chart (scree plot) are displayed. The number of eigenvalues is equal to the number of non-null eigenvalues.

If the corresponding output options have been activated, XLSTAT afterwards displays the factor loadings in the new space, then the correlations between the initial variables and the components in the new space. The correlations are equal to the factor loadings in a normalized PCA (on the correlation matrix).

If supplementary variables have been selected, the corresponding coordinates and correlations are displayed at the end of the table.

**Contributions**: Contributions are an interpretation aid. The variables which had the highest influence in building the axes are those whose contributions are highest.

**Squared cosines**: As in other factor methods, squared cosine analysis is used to avoid interpretation errors due to projection effects. If the squared cosines associated with the axes used on a chart are low, the position of the observation or the variable in question should not be interpreted.
The **factor scores** in the new space are then displayed. If supplementary data have been selected, these are displayed at the end of the table.

**Contributions**: This table shows the contributions of the observations in building the principal components.

**Squared cosines**: This table displays the squared cosines between the observation vectors and the factor axes.

Where a rotation has been requested, the results of the rotation are displayed with the **rotation matrix** first applied to the factor loadings. This is followed by the modified variability percentages associated with each of the axes involved in the rotation. The coordinates, contributions and cosines of the variables and observations after rotation are displayed in the following tables.

**Example**

A tutorial on how to use Principal Component Analysis is available on the Addinsoft website:

http://www.xlstat.com/demo-pca.htm

A tutorial on how to use Principal Component Analysis and apply filters based on communalities (squared cosines) is available on the Addinsoft website:

http://www.xlstat.com/demo-pcafilter.htm

**References**


**Discriminant Analysis (DA)**

Use discriminant analysis to explain and predict the membership of observations to several classes using quantitative or qualitative explanatory variables.

**Description**

Discriminant Analysis (DA) is an old method (Fisher, 1936) which in its classic form has changed little in the past twenty years. This method, which is both explanatory and predictive, can be used to:

- Check on a two- or three-dimensional chart if the groups to which observations belong are distinct,
- Show the properties of the groups using explanatory variables,
- Predict which group an observation will belong to

DA may be used in numerous applications, for example in ecology and the prediction of financial risks (credit scoring).

**Linear or quadratic model**

Two models of DA are used depending on a basic assumption: if the covariance matrices are assumed to be identical, linear discriminant analysis is used. If, on the contrary, it is assumed that the covariance matrices differ in at least two groups, then the quadratic model is used. The Box test is used to test this hypothesis (the Bartlett approximation enables a $\chi^2$ distribution to be used for the test). We start with linear analysis then, depending on the results from the Box test, carry out quadratic analysis if required.

**Multicollinearity issues**

With linear and still more with quadratic models, we can face problems of variables with a null variance or multicollinearity between variables. XLSTAT has been programmed so as to avoid these problems. The variables responsible for these problems are automatically ignored either for all calculations or, in the case of a quadratic model, for the groups in which the problems arise. Multicollinearity statistics are optionally displayed so that you can identify the variables which are causing problems.
**Stepwise methods**

As for linear and logistic regression, efficient stepwise methods have been proposed. They can, however, only be used when quantitative variables are selected as the input and output tests on the variables assume them to be normally distributed. The stepwise method gives a powerful model which avoids variables which contribute only little to the model.

**Classification table, ROC curve and cross-validation**

Among the numerous results provided, XLSTAT can display the classification table (also called confusion matrix) used to calculate the percentage of well-classified observations. When only two classes (or categories or modalities) are present in the dependent variable, the ROC curve may also be displayed.

The ROC curve (*Receiver Operating Characteristics*) displays the performance of a model and enables a comparison to be made with other models. The terms used come from signal detection theory.

The proportion of well-classified positive events is called the *sensitivity*. The *specificity* is the proportion of well-classified negative events. If you vary the threshold probability from which an event is to be considered positive, the sensitivity and specificity will also vary. The curve of points (1-specificity, sensitivity) is the ROC curve.

Let's consider a binary dependent variable which indicates, for example, if a customer has responded favorably to a mail shot. In the diagram below, the blue curve corresponds to an ideal case where the n% of people responding favorably corresponds to the n% highest probabilities. The green curve corresponds to a well-discriminating model. The red curve (first bisector) corresponds to what is obtained with a random Bernoulli model with a response probability equal to that observed in the sample studied. A model close to the red curve is therefore inefficient since it is no better than random generation. A model below this curve would be disastrous since it would be less even than random.
The area under the curve (or AUC) is a synthetic index calculated for ROC curves. The AUC corresponds to the probability such that a positive event has a higher probability given to it by the model than a negative event. For an ideal model, AUC=1 and for a random model, AUC = 0.5. A model is usually considered good when the AUC value is greater than 0.7. A well-discriminating model must have an AUC of between 0.87 and 0.9. A model with an AUC greater than 0.9 is excellent.

The results of the model as regards forecasting may be too optimistic: we are effectively trying to check if an observation is well-classified while the observation itself is being used in calculating the model. For this reason, cross-validation was developed: to determine the probability that an observation will belong to the various groups, it is removed from the learning sample, then the model and the forecast are calculated. This operation is repeated for all the observations in the learning sample. The results thus obtained will be more representative of the quality of the model. XLSTAT gives the option of calculating the various statistics associated with each of the observations in cross-validation mode together with the classification table and the ROC curve if there are only two classes.

Lastly, you are advised to validate the model on a validation sample wherever possible. XLSTAT has several options for generating a validation sample automatically.

**Discriminant analysis and logistic regression**

Where there are only two classes to predict for the dependent variable, discriminant analysis is very much like logistic regression. Discriminant analysis is useful for studying the covariance structures in detail and for providing a graphic representation. Logistic regression has the advantage of having several possible model templates, and enabling the use of stepwise
selection methods including for qualitative explanatory variables. The user will be able to compare the performances of both methods by using the ROC curves.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- ![Reload](image)
  - Click this button to reload the default options.
- ![Delete](image)
  - Click this button to delete the data selections.
- ![Change](image)
  - Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab**:

**Y / Dependent variables**:

**Qualitative**: Select the qualitative variable(s) you want to model. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

**X / Explanatory variables**:

**Quantitative**: Activate this option if you want to include one or more quantitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The data selected may be of the numerical type. If a variable header has been selected, check that the "Variable labels" option has been activated.
Qualitative: Activate this option if you want to include one or more qualitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If a variable header has been selected, check that the "Variable labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 ...).

Observation weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will all be considered as 1. XLSTAT uses these weights for calculating degrees of freedom. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option has been activated.

Options tab:

Tolerance: Enter the value of the tolerance threshold below which a variable will automatically be ignored.

Equality of covariance matrices: Activate this option if you want to assume that the covariance matrices associated with the various classes of the dependent variable are equal.

Prior probabilities: Activate this option if you want to take prior possibilities into account. The probabilities associated with each of the classes are equal to the frequency of the classes. Note: this option has no effect if the prior possibilities are equal for the various groups.

Filter factors: You can activate one of the two following options in order to reduce the number of factors used in the model:
• **Minimum %**: Activate this option and enter the minimum percentage of total variability that the selected factors should represent.

• **Maximum number**: Activate this option to set the maximum number of factors to take into account.

**Significance level (%)**: Enter the significance level for the various tests calculated.

**Model selection**: Activate this option if you want to use one of the four selection methods provided:

• **Stepwise (Forward)**: The selection process starts by adding the variable with the largest contribution to the model. If a second variable is such that its entry probability is greater than the **entry threshold value**, then it is added to the model. After the third variable is added, the impact of removing each variable present in the model after it has been added is evaluated. If the probability of the calculated statistic is greater than the **removal threshold value**, the variable is removed from the model.

• **Stepwise (Backward)**: This method is similar to the previous one but starts from a complete model.

• **Forward**: The procedure is the same as for stepwise selection except that variables are only added and never removed.

• **Backward**: The procedure starts by simultaneously adding all variables. The variables are then removed from the model following the procedure used for stepwise selection.

**Classes weight correction**: If the number of observations for the various classes for the dependent variables are not uniform, there is a risk of penalizing classes with a low number of observations in establishing the model. To get over this problem, XLSTAT has two options:

• **Automatic**: Correction is automatic. Artificial weights are assigned to the observations in order to obtain classes with an identical sum of weights.

• **Corrective weights**: You can select the weights to be assigned to each observation.

**Validation tab**:

**Validation**: Activate this option if you want to use a sub-sample of the data to validate the model.

**Validation set**: Choose one of the following options to define how to obtain the observations used for the validation:
- **Random**: The observations are randomly selected. The “Number of observations” N must then be specified.

- **N last rows**: The N last observations are selected for the validation. The “Number of observations” N must then be specified.

- **N first rows**: The N first observations are selected for the validation. The “Number of observations” N must then be specified.

- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to use for the validation.

**Prediction** tab:

**Prediction**: Activate this option if you want to select data to use them in prediction mode. If activate this option, you need to make sure that the prediction dataset is structured as the estimation dataset: same variables with the same order in the selections. On the other hand, variable labels must not be selected: The first row of the selections listed below must correspond to data.

**Quantitative**: Activate this option to select the quantitative explanatory variables. The first row must not include variable labels.

**Qualitative**: Activate this option to select the qualitative explanatory variables. The first row must not include variable labels.

**Observations labels**: Activate this option if observations labels are available. Then select the corresponding data. If this option is not activated, the observations labels are automatically generated by XLSTAT (PredObs1, PredObs2 …).

**Missing data** tab:

**Remove observations**: Activate this option to remove the observations with missing data.

**Estimate missing data**: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

- **Nearest neighbour**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.

**Outputs** tab:
**Descriptive statistics**: Activate this option to display descriptive statistics for the variables selected.

**Correlations**: Activate this option to display the correlation matrix.

**Multicollinearity statistics**: Activate this option to display the table of multicollinearity statistics.

**Covariance matrices**: Activate this option to display the inter-class, intra-class, intra-class total, and total covariance matrices.

**SSCP matrices**: Activate this option to display the inter-class, intra-class total, and total SSCP (Sums of Squares and Cross Products) matrices.

**Distance matrices**: Activate this option to display the matrices of distances between groups.

**Canonical correlations and functions**: Activate this option for canonical correlations and functions.

**Classification functions**: Activate this option to display classification functions.

**Eigenvalues**: Activate this option to display the table and chart (scree plot) of eigenvalues.

**Eigenvectors**: Activate this option to display the eigenvector table.

**Variables/Factors correlations**: Activate this option to display correlations between factors and variables.

**Factor scores**: Activate this option to display the coordinates of the observations in the factor space. The prior and posterior classes for each observation, the probabilities of assignment for each class and the distances of the observations from their centroid are also displayed in this table.

**Confusion matrix**: Activate this option to display the table showing the numbers of well- and badly-classified observations for each of the classes.

**Cross-validation**: Activate this option to display cross-validation results (probabilities for observations and confusion matrix).

---

**Charts** tab:

**Correlation charts**: Activate this option to display the charts involving correlations between the factors and input variables.

- **Vectors**: Activate this option to display the input variables with vectors.

**Observations charts**: Activate this option to display the charts that allow visualizing the observations in the new space.
• **Labels**: Activate this option to display the observations labels on the charts. The number of labels can be modulated using the filtering option.

• **Display the centroids**: Activate this option to display the centroids that correspond to the categories of the dependent variable.

• **Confidence ellipses**: Activate this option to display confidence ellipses. The confidence ellipses correspond to a x% confidence interval (where x is determined using the significance level entered in the Options tab) for a bivariate normal distribution with the same means and the same covariance matrix as the factor scores for each category of the dependent variable.
  
  o **Use covariance hypothesis**: Activate this option to base the computation of the ellipses on the hypothesis that covariance matrices are equal or not.

**Centroids and confidence circles**: Activate this option to display a chart with the centroids and the confidence circles around the means.

**Colored labels**: Activate this option to display the labels with the same color as the corresponding points. If this option is not activated the labels are displayed in black.

**Filter**: Activate this option to modulate the number of observations displayed:

  • **Random**: The observations to display are randomly selected. The “Number of observations” N to display must then be specified.

  • **N first rows**: The N first observations are displayed on the chart. The “Number of observations” N to display must then be specified.

  • **N last rows**: The N last observations are displayed on the chart. The “Number of observations” N to display must then be specified.

  • **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to display.

**Results**

**Descriptive statistics**: The table of descriptive statistics shows the simple statistics for all the variables selected. The number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the quantitative variables. For qualitative variables, including the dependent variable, the categories with their respective frequencies and percentages are displayed.

**Correlation matrix**: This table displays the correlations between the explanatory variables.
**Means by class:** This table provides the means of the various explanatory variables for the various classes of the dependent variable.

**Sum of weights, prior probabilities and logarithms of determinants for each class:** These statistics are used, among other places, in the posterior calculations of probabilities for the observations.

**Multicolinearity:** This table identifies the variables responsible for the multicolinearity between variables. As soon as a variable is identified as being responsible for a multicolinearity (its tolerance is less than the limit tolerance set in the "options" tab in the dialog box), it is not included in the multicolinearity statistics calculation for the following variables. Thus in the extreme case where two variables are identical, only one of the two variables will be eliminated from the calculations. The statistics displayed are the tolerance (equal to 1-R²), its inverse and the VIF (Variance inflation factor).

**SSCP matrices:** The SSCP (Sums of Squares and Cross Products) matrices are proportional to the covariance matrices. They are used in the calculations and check the following relationship: SSCP total = SSCP inter + SSCP intra total.

**Covariance matrices:** The inter-class covariance matrix (equal to the unbiased covariance matrix for the means of the various classes), the intra-class covariance matrix for each of the classes (unbiased), the total intra-class covariance matrix, which is a weighted sum of the preceding ones, and the total covariance matrix calculated for all observations (unbiased) are displayed successively.

**Box test:** The Box test is used to test the assumption of equality for intra-class covariance matrices. Two approximations are available, one based on the Chi² distribution, and the other on the Fisher distribution. The results of both tests are displayed.

**Kullback’s test:** The Kullback test is used to test the assumption of equality for intra-class covariance matrices. The statistic calculated is approximately distributed according to a Chi² distribution.

**Mahalanobis distances:** The Mahalanobis distance is used to measure the distance between classes taking account of the covariance structure. If we assume the intra-class variance matrices are equal, the distance matrix is calculated by using the total intra-class covariance matrix which is symmetric. If we assume the intra-class variance matrices are not equal, the Mahalanobis distance between classes i and j is calculated by using the intra-class covariance matrix for class i which is symmetric. The distance matrix is therefore asymmetric.

**Fisher’s distances:** If the covariance matrices are assumed to be equal, the Fisher distances between the classes are displayed. They are calculated from the Mahalanobis distance and are used for a significance test. The matrix of p-values is displayed so as to identify which distances are significant.
**Generalized squared distances:** If the covariance matrices are not assumed to be equal, the table of generalized squared distances between the classes is displayed. The generalized distance is also calculated from the Mahalanobis distances and uses the logarithms of the determinants of the covariance matrices together with the logarithms of the prior probabilities if required by the user.

**Wilks’ Lambda test (Rao’s approximation):** The test is used to test the assumption of equality of the mean vectors for the various classes. When there are two classes, the test is equivalent to the Fisher test mentioned previously. If the number of classes is less than or equal to three, the test is exact. The Rao approximation is required from four classes to obtain a statistic approximately distributed according to a Fisher distribution.

**Unidimensional test of equality of the means of the classes:** These tests are used to test the assumption of equality of the means between classes variable by variable. Wilk's univariate lambda is always between 0 and 1. A value of 1 means the class means are equal. A low value is interpreted as meaning there are low intra-class variations and therefore high inter-class variations, hence a significant difference in class means.

**Pillai's trace:** The test is used to test the assumption of equality of the mean vectors for the various classes. It is less used than Wilk's Lambda test and also uses the Fisher distribution for calculating p-values.

**Hotelling-Lawley trace:** The test is used to test the assumption of equality of the mean vectors for the various classes. It is less used than Wilk's Lambda test and also uses the Fisher distribution for calculating p-values.

**Roy's greatest root:** The test is used to test the assumption of equality of the mean vectors for the various classes. It is less used than Wilk's Lambda test and also uses the Fisher distribution for calculating p-values.

**Eigenvalues:** This table shows the eigenvalues associated with the various factors together with the corresponding discrimination percentages and cumulative percentages. In discriminant analysis, the number of non-null eigenvalues is equal to at most (k-1) where k is the number of classes. The scree plot is used to display how the discriminant power is distributed between the discriminant factors. The sum of the eigenvalues is equal to the Hotelling trace.

**Bartlett's test on significany of eigenvalues:** This table displays for each eigenvalue, the Bartlett statistic and the corresponding p-value which is computed using the asymptotic Chi-square approximation. The Bartlett's test allows to test the null hypothesis H0 that all the p eigenvalues are equal to zero. If it is rejected for the greatest eigenvalue then the test is performed again until H0 cannot be rejected. This test is known as conservative, meaning that it tends to confirm H0 in some cases where it should not. You can however use this test to check how many factorial axes you should consider (see Jobson, 1992).
**Eigenvectors:** This table shows the eigenvectors afterwards used in the canonical correlations, canonical functions and observation coordinates (*scores*) calculations.

**Variables/Factors correlations:** The calculation of correlations between the scores in the initial variable space and in the discriminant factor space is used to display the relationship between the initial variables and the factors in a correlation circle. The correlation circle is an aid in interpreting the representation of the observations in factor space.

**Canonical correlations:** The canonical correlations associated with each factor are the square roots of \( L(i) / (1- L(i)) \) where \( L(i) \) is the eigenvalue associated with factor \( i \). Canonical correlations are also a measurement of the discriminant power of the factors. Their sum is equal to the Pilai’s trace.

**Canonical discriminant function coefficients:** These coefficients can be used to calculate the coordinates of an observation in discriminant factor space from its coordinates in the initial variable space.

**Standardized canonical discriminant function coefficients:** These coefficients are the same as the previous, but are standardized. Thus comparing them gives a measure of the relative contribution of the initial variables to the discrimination for a given factor.

**Functions at the centroids:** This table gives the evaluation of the discriminant functions for the mean points for each of the classes.

**Classification functions:** The classification functions can be used to determine which class an observation is to be assigned to using values taken for the various explanatory variables. If the covariance matrices are assumed equal, these functions are linear. If the covariance matrices are assumed unequal, these functions are quadratic. An observation is assigned to the class with the highest classification function.

**Prior and posterior classification, membership probabilities, scores and squared distances:** This table shows for each observation its membership class defined by the dependent variable, the membership class as deduced by the membership probabilities, the probabilities of membership of each of the classes, the coordinates in discriminant factor space and the squared distances of the observations from the centroids of each of the classes.

**Confusion matrix for the estimation sample:** The confusion matrix is deduced from prior and posterior classifications together with the overall percentage of well-classified observations. Where the dependent variable only comprises two classes, the ROC curve is displayed (see the description section for more details).

**Cross-validation:** Where cross-validation has been requested, the table containing the information for the observations and the confusion matrix are displayed (see the description section for more details).
Example

A tutorial on how to use Discriminant Analysis is available on the Addinsoft website:

http://www.xlstat.com/demo-da.htm

References


Correspondence Analysis (CA)

Use this tool to visualize the links between the categories of two qualitative variables. The variables can be available as an observations/variables table, as a contingency table, or as a more general type of two-way table.

Description

Correspondence Analysis is a powerful method that allows studying the association between two qualitative variables. The research of J.-P. Benzécri that started in the early sixties led to the emergence of the method. His disciples worked on several developments of the basic method. For example, M.J. Greenacre’s book (1984) contributed to the popularity of the method throughout the world. The work of C. Lauro from the University of Naples led to a non-symmetrical variant of the method.

Measuring the association between two qualitative variables is a complex subject that first requires transforming the data: it is not possible to compute a correlation coefficient using the data directly, as one could do with quantitative variables.

The first transformation consists of recoding the two qualitative variables V1 and V2 as two disjunctive tables $Z_1$ and $Z_2$ or indicator (or dummy) variables. For each category of a variable there is a column in the respective disjunctive table. Each time the category $c$ of variable V1 occurs for an observation $i$, the value of $Z_1(i,c)$ is set to one (the same rule is applied to the V2 variable). The other values of $Z_1$ and $Z_2$ are zero. The generalization of this idea to more than two variables is called Multiple Correspondence Analysis. When there are only two variables, it is sufficient to study the contingency table of the two variables, that is the table $Z_1^TZ_2$.

The Chi-square distance has been suggested to measure the distance between two categories. To represent the distance between two categories it is not necessary to use from the $X_1$ and $X_2$ disjunctive tables. It is enough to start from the contingency table that algebraically corresponds to the $X_1^TX_2$ product.

The contingency table has the following structure:

<table>
<thead>
<tr>
<th>V1 \ V2</th>
<th>Category 1</th>
<th>...</th>
<th>Category j</th>
<th>...</th>
<th>Category m2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category 1</td>
<td>n(1,1)</td>
<td>...</td>
<td>n(1,j)</td>
<td>...</td>
<td>n(1,m2)</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Category i</td>
<td>n(i,1)</td>
<td>...</td>
<td>n(i,j)</td>
<td>...</td>
<td>n(i,m2)</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Where \( n(i,j) \) is the frequency of observations that show both characteristic \( i \) for variable \( V1 \), and characteristic \( j \) for variable \( V2 \).

Inertia is a measure inspired from physics that is often used in Correspondence Analysis. The inertia of a set of points is the weighted mean of the squared distances to the center of gravity. In the specific case of CA, the total inertia of the set of points (one point corresponds to one category) can be written as:

\[
\varphi^2 = \frac{\chi^2}{n} = \sum_{i=1}^{m1} \sum_{j=1}^{m2} \left( \frac{n_{ij} - n_i n_j}{n_i^2} \right)^2,
\]

with \( n_i = \sum_{j=1}^{m2} n_{ij} \) and \( n_j = \sum_{i=1}^{m1} n_{ij} \)

and where \( n \) is the sum of the frequencies in the contingency table. We can see that the inertia is proportional to the Pearson chi-square statistic computed on the contingency table.

The aim of CA is to represent as much of the inertia on the first principal axis as possible, a maximum of the residual inertia on the second principal axis and so on until all the total inertia is represented in the space of the principal axes. One can show that the number of dimensions of the space is equal to \( \min(m1, m2) - 1 \).

**Non-Symmetrical Correspondence Analysis (NSCA)** developed by Lauro and D'Ambra (1984) analyzes the association between the rows and columns of a contingency table while introducing the notion of dependency between the rows and the columns, which leads to an asymmetry in their treatment. The example the authors used in their first article on this subject corresponds to the analysis of a contingency table that contains the prescriptions of 6 drugs for 7 different diseases for 69 patients. There is here an obvious dependency of the drugs on the disease. In order to take into account the dependency, Goodman and Kruskal’s tau (1954) was suggested. The tau coefficient that corresponds to the case where the rows depend on the columns can be written as:

\[
\tau_{h/RC} = \frac{\sum_{j=1}^{m2} n_j / n \sum_{i=1}^{m1} \left( n_{ij} / n_{ij} - n_i / n \right)^2}{1 - \sum_{i=1}^{m1} \left( n_i / n \right)^2}
\]

As with the total inertia, it is possible to compute a representation space for the categories, such that the proportion of the Goodman and Kruskal’s tau represented on the chart is maximized on the first axes.
Greenacre (1984) defined a framework (the generalized singular value decomposition) that allows computing both CA and the related method of NSCA in a similar way.

An alternative approach using the **Hellinger distance** was proposed by Rao (1995). The Hellinger distance only depends on the profiles of the concerned pair and does not depend on the sample sizes on which the profiles are estimated. Therefore, the Hellinger distance approach might be a good alternative to the classical CA when average column profiles are not relevant (e.g. when columns represent populations of individuals classified according to row categories) or if some categories have low frequencies. Computations follows the unified approach described by Cuadras and Cuadras (2008). The inertia is generalized by the following formula:

\[
\varphi^2(\alpha, 1/2) = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \left( \left( \frac{n_{ij}/n}{n_i/n} \right)^\alpha - 1 \right)^2 \left( \left( \frac{n_{ij}/n}{n_j/n} \right)^\beta - 1 \right) , \text{ with } n_i = \sum_{j=1}^{m_2} n_{ij} \text{ and } n_j = \sum_{i=1}^{m_1} n_{ij}
\]

In the case of the Correspondence Analysis using the Hellinger distance, \( \alpha = 1/2 \) and \( \beta = 1/2 \).

**Note:** In the case of classical Correspondence Analysis, \( \alpha = 1 \) and \( \beta = 1/2 \) and we find the earlier formula of \( \varphi^2 \).

The **analysis of a subset of categories** is a new method that has been recently developed by Greenacre (2006). It allows parts of tables to be analyzed while maintaining the margins of the whole table and thus the same weights and chi-square distances of the whole table, simplifying the analysis of large tables by breaking down the interpretation into parts.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**
  - Click this button to start the computations.

- **Cancel**
  - Click this button to close the dialog box without doing any computation.

- **Help**
  - Click this button to display the help.
Click this button to reload the default options.

Click this button to delete the data selections.

Click these buttons to change the way XLSTAT loads the data:

- Case where the data are in a contingency table or a more general two-way table: if the arrow points down XLSTAT allows you to select data by columns or by range. If the arrow points to the right, XLSTAT allows you to select data by rows or by range.

- Case where the data are in an observations/variables table: if the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab:**

The first selection field lets you alternatively select two types of tables:

**Two-way table:** Select this option if your data correspond to a two-way table where the cells contain the frequencies corresponding to the various categories of two qualitative variables (in this case it is more precisely a contingency table), or other type of values.

**Observations/variables table:** Select this option if your data correspond to N observations described by 2 qualitative variables. This type of table typically corresponds to a survey with 2 questions. During the computations, XLSTAT will automatically transform this table into a contingency table.

**Range:** Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Activate this option to display the results in a new worksheet of the active workbook.

**Workbook:** Activate this option to display the results in a new workbook.

**Labels included:** This option is visible if the selected table is a contingency table or a more general two-way table. Activate this option if the labels of the columns and rows are included in the selection.

**Variable labels:** This option is visible only if you selected the observations/variables table format. Activate this option if the first row contains the variable labels (case of an observations/variables table) or the category labels (case of a disjunctive table).
Weights: This option is visible only if you selected the observations/variables table format. Activate this option if you want to weight the observations. If you do not activate this option, the weights are considered to be equal to 1. The weights must be greater or equal to 0. If the “Variable labels” option is activated make sure that the header of the selection has also been selected.

Options tab:

Advanced analysis: This option is active only in the case where the input is a contingency table or a more general two-way table. The possible options are:

- **Supplementary data**: If you select this option you may then enter the number of supplementary rows and/or columns. Supplementary rows and columns are passive data that are not taken into account for the computation of the representation space. Their coordinates are computed a posteriori. Notice that supplementary data should be the last rows and/or columns of the data table.

- **Subset analysis**: If you select this option you can then enter the number of rows and/or columns to exclude from the subset analysis. See the description section for more information on this topic. Notice that the excluded data should be the last rows and/or columns of the data table.

Non-symmetrical analysis: This option allows computing a non-symmetrical correspondence analysis, as proposed by Lauro et al. (1984).

- **Rows depend on columns**: Select this option if you consider that the row variable depends on the column variable and if you want to analyze the association between both while taking into account this dependency.

- **Columns depend on rows**: Select this option if you consider that the column variable depends on the row variable and if you want to analyze the association between both while taking into account this dependency.

Distance: This options allows computing a correspondence analysis based on the Hellinger distance as proposed by Rao (1995).

- **Chi-Square**: Select this option to compute classical correspondence analysis (CA).

- **Hellinger**: Select this option to compute correspondence analysis based on Hellinger distance (HD). This option is not available if the non-symmetrical option has been selected.

To summarize, three approaches of the correspondence analysis are proposed:

- Classical correspondence analysis (CA): Do not select the “non-symmetrical analysis” option and select Chi-Square distance.
- Non-symmetrical correspondence analysis (NSCA): Select the “non-symmetrical analysis” option and select Chi-Square distance.

- Correspondence analysis using the Hellinger distance (HD): Do not select the “non-symmetrical analysis” option and select Hellinger distance.

**Test of independence**: Activate this option if you want XLSTAT to compute a test of independence based on the chi-square statistic.

- **Significance level (%)**: Enter the value of the significance level for the test (default value: 5%).

**Filter factors**: You can activate one of the two following options in order to reduce the number of factors displayed:

- **Minimum %**: Activate this option and then enter the minimum percentage that should be reached to determine the number of factors to display.

- **Maximum number**: Activate this option to set the maximum number of factors to take into account when displaying the results.

**Missing data** tab:

**Options for contingency tables and other two-way tables**:

- **Do not accept missing data**: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

- **Replace missing data by 0**: Activate this option if you consider that missing data are equivalent to 0.

- **Replace missing data by their expected value**: Activate this option if you want to replace the missing data by the expected value. The expectation is given by:

  \[ E(n_{ij}) = \frac{n_i n_j}{n} \]

  where \( n_i \) is the row sum, \( n_j \) is the column sum, and \( n \) is the grand total of the table before replacement of the missing data.

**Options for the observations/variables tables**:

- **Do not accept missing data**: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

- **Remove observations**: Activate this option to ignore the observations that contain missing data.
**Group missing values into a new category:** Activate this option to group missing data into a new category of the corresponding variable.

**Outputs** tab:

**Options specific to the observations/variables tables:**

**Descriptive statistics:** Activate this option to display the descriptive statistics for the two selected variables.

**Disjunctive table:** Activate this option to display the full disjunctive table that corresponds to the qualitative variables.

**Sort the categories alphabetically:** Activate this option so that the categories of all the variables are sorted alphabetically.

**Common options:**

**Contingency table:** Activate this option to display the contingency table.

- **3D view of the contingency table / two-way table:** Activate this option to display the 3D bar chart corresponding to the contingency table or to the two-way table.

**Inertia by cell:** Activate this option to display the inertia for each cell of the contingency table.

**Row and column profiles:** Activate this option to display the row and column profiles.

**Eigenvalues:** Activate this option to display the table and the scree plot of the eigenvalues.

**Chi-square (or Hellinger) distances:** Activate this option to display the chi-square (or Hellinger) distances between the row points and between the column points.

**Principal coordinates:** Activate this option to display the principal coordinates of the row points and the column points.

**Standard coordinates:** Activate this option to display the standard coordinates of the row points and the column points.

**Contributions:** Activate this option to display the contributions of the row points and the column points to the principal axes.

**Squared cosines:** Activate this option to display the squared cosines of the row points and the column points to the principal axes.

**Table for 3D visualization:** Activate this option to display the table for 3D visualization.
Charts tab:

Maps sub-tab:

Symmetric plots: Activate this option to display the plots for which the row points and the column points play a symmetrical role. These maps are based on the principal coordinates of the row points and the column points.

- **Rows and columns**: Activate this option to display a chart on which the row points and the column points are displayed.
- **Rows**: Activate this option to display a chart on which only the row points are displayed.
- **Columns**: Activate this option to display a chart on which only the column points are displayed.

Asymmetric plots: Activate this option to display the plots for which the row points and the column points play an asymmetrical role. These plots use on the one hand the principal coordinates and on the other hand the standard coordinates.

- **Rows**: Activate this option to display a chart where the row points are displayed using their principal coordinates, and the column points are displayed using their standard coordinates.
- **Columns**: Activate this option to display a chart where the row points are displayed using their standard coordinates, and the column points are displayed using their principal coordinates.
- **Vectors**: Activate this option to display the vectors for the standard coordinates on the asymmetric charts.
- **Length factor**: Activate this option to modulate the length of the vectors.

Contribution biplots: Activate this option to display the contribution biplots for which the row points and the column points play an asymmetrical role. These plots use on the one hand the principal coordinates and on the other hand the contribution coordinates that take into account the weights.

- **Rows**: Activate this option to display a chart where the row points are displayed using their principal coordinates, and the column points are displayed using their contribution coordinates.
- **Columns**: Activate this option to display a chart where the row points are displayed using their contribution coordinates, and the column points are displayed using their principal coordinates.

Options specific to the observations/variables tables:
Row options sub-tab:

Filter rows: Activate this option to modulate the number of rows displayed:

- **Random**: The rows to display are randomly selected. The “Number of rows” N to display must then be specified.
- **N first rows**: The first N rows are displayed on the chart. The “Number of rows” N to display must then be specified.
- **N last rows**: The last N rows are displayed on the chart. The “Number of rows” N to display must then be specified.
- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the rows to display.
- **Sum(Cos2)>**: Only the rows for which the sum of squared cosines on the given dimensions are larger than a value to enter are displayed on the plots.

Resize row points with: Activate this option to resize the row points:

- **Cos2**: The sizes of the row points are proportional to the sum of squared cosines on the given dimensions.
- **Contribution**: The sizes of the row points are proportional to the sum of contributions on the given dimensions.

Confidence ellipses: Activate this option to display confidence ellipses to identify the row categories that contribute to the dependency between the row and column categories of the contingency table.

Row labels: Activate this option to display the labels of the row categories on the charts.

- **Colored labels**: Activate this option to display the labels with the same color as the corresponding points. If this option is not activated the labels are displayed in black.

Column options sub-tab:

Filter columns: Activate this option to modulate the number of columns displayed:

- **Random**: The columns to display are randomly selected. The “Number of columns” N to display must then be specified.
- **N first columns**: The first N columns are displayed on the chart. The “Number of columns” N to display must then be specified.
- **N last columns**: The last N columns are displayed on the chart. The “Number of columns” N to display must then be specified.
- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the columns to display.
• **Sum(Cos^2)>**: Only the columns for which the sum of squared cosines on the given dimensions are larger than a value to enter are displayed on the plots.

**Resize column points with**: Activate this option to resize the column points:

• **Cos^2**: The sizes of the column points are proportional to the sum of squared cosines on the given dimensions.

• **Contribution**: The sizes of the column points are proportional to the sum of contributions on the given dimensions.

**Confidence ellipses**: Activate this option to display confidence ellipses to identify the column categories that contribute to the dependency between the row and column categories of the contingency table.

**Column labels**: Activate this option to display the labels of the column categories on the charts.

• **Colored labels**: Activate this option to display the labels with the same color as the corresponding points. If this option is not activated the labels are displayed in black.

### Results

**Descriptive statistics**: This table is displayed only if the input data correspond to an observations/variables table.

**Disjunctive table**: This table is displayed only if the input data correspond to an observations/variables table. This table is an intermediate table that allows to obtain the contingency table that corresponds to the two selected variables.

**Contingency table**: The contingency table is displayed at this stage. The **3D bar chart** that follows corresponds to the table.

**Inertia by cell**: This table displays the inertia that corresponds to each cell of the contingency table.

**Test of independence between rows and columns**: This test allows us to determine if we can reject the null hypothesis that rows and columns of the table are independent. A detailed interpretation of this test is displayed below the table that summarizes the test statistic.

**Eigenvalues and percentages of inertia**: The eigenvalues and the corresponding scree plot are displayed. Only the non-trivial eigenvalues are displayed. If a filtering has been requested in the dialog box, it is not applied to this table, but only to the results that follow. Note: the sum of the eigenvalues is equal to the total inertia. To each eigenvalue corresponds a principal axis which accounts for a certain percentage of inertia. This allows us to measure the cumulative percentage of inertia for a given set of dimensions.
A series of results is displayed afterwards, first for the row points, then for the column points:

Weights, distances and squared distances to the origin, inertias and relative inertias: This table gives basic statistics for the points.

Profiles: This table displays the profiles.

Chi-square (or Hellinger) distances: This table displays the chi-square (or Hellinger) distances between the profile points.

Principal coordinates: This table displays the principal coordinates which are used later to represent projections of the profile points in symmetric and asymmetric plots.

Standard coordinates: This table displays the standard coordinates which are used later to represent projections of unit profile points in asymmetric plots.

Contributions: The contributions are helpful for interpreting the plots. The categories that have influenced the most the calculation of the axes are those that have the higher contributions. An approach consists of restricting the interpretation to the categories whose contribution to a given axis is higher than the corresponding relative weight that is displayed in the first column.

Squared cosines: As with other data analysis methods, the analysis of the squared cosines allows us to avoid misinterpretations of the plots that are due to projection effects. If, for a given category, the cosines are low on the axes of interest, then any interpretation of the position of the category is hazardous.

The plots (or maps) are the ultimate goal of Correspondence Analysis, because they allow us to considerably accelerate our understanding of the association patterns in the data table.

Symmetric plots: These plots are exclusively based on the principal coordinates. Depending on the choices made in the dialog box, a symmetric plot mixing row points and column points, a plot with only the row points, and a plot with only the column points, are displayed. The percentage of inertia that corresponds to each axis and the percentage of inertia cumulated over the two axes are displayed on the map. If the “confidence ellipses” option was selected, confidence ellipses are draw around the points. The confidence ellipses allow the identification of the categories that contribute to the association structure between the variables. The ellipses reflect the information contained in dimensions non-represented on the map.

Asymmetric plots: These plots use the principal coordinates the standard coordinates for the rows and columns or vice versa. The percentage of inertia that corresponds to each axis and the percentage of inertia cumulated over the two axes are displayed on the map. In an “asymmetric row plot”, on can study the way the row points are positioned relatively to the column vectors. The latter indicate directions: if two row points are displayed in the same
direction as a column vector, the row point that is the furthest in the column vector direction is the one that is more associated with the columns.

**Contribution biplots:** These plots use the principal coordinates for the rows and the contribution coordinates for the columns or vice versa. The percentage of inertia that corresponds to each axis and the percentage of inertia cumulated over the two axes are displayed on the map. In an “contribution row plot”, one can study the way the row points are positioned relatively to the column vectors while the length of the column vectors take into account their contribution to the building of the biplot.

**Example**

A tutorial on how to use Correspondence Analysis is available on the Addinsoft website:

[http://www.xlstat.com/demo-ca.htm](http://www.xlstat.com/demo-ca.htm)

**References**


Multiple Correspondence Analysis (MCA)

Use this tool to visualize the links between the categories of two or more qualitative variables.

Description

Multiple Correspondence Analysis (MCA) is a method that allows studying the association between two or more qualitative variables. MCA is to qualitative variables what Principal Component Analysis is to quantitative variables. One can obtain maps where it is possible to visually observe the distances between the categories of the qualitative variables and between the observations.

Multiple Correspondence Analysis (MCA) can also be understood as a generalization of Correspondence Analysis (CA) to the case where there are more than two variables. While it is possible to summarize a table with n observations and p (p>2) qualitative variables in a table whose structure is close to a contingency table, it is much more common in MCA to start from an observations/variables table (for example, from a survey where p questions were submitted to n individuals). XLSTAT also allows the user to start from a full disjunctive table (indicator matrix).

The generation of the disjunctive table is, in any case, a preliminary step of the MCA computations. The p qualitative variables are broken down into p disjunctive tables Z1, Z2, ..., Zp, composed of as many columns as there are categories in each of the variables. Each time a category c of the jth variable corresponds to an observation i, one sets the value of Zj(i,c) to one. The other values of Zj are zero. The p disjunctive tables are concatenated into a full disjunctive table.

A series of transformations allows the computing of the coordinates of the categories of the qualitative variables, as well as the coordinates of the observations in a representation space that is optimal for a criterion based on inertia. In the case of MCA one can show that the total inertia is equal to the average number of categories minus one. As a matter of fact, the inertia does not only depend on the degree of association between the categories but is seriously inflated. Greenacre (1993) suggested an adjusted version of inertia, inspired from Joint Correspondence Analysis (JCA). This adjustment allows us to have higher and more meaningful percentages for the maps.

The analysis of a subset of categories is a method that has very recently been developed by Greenacre and Pardo (2006). It allows us to concentrate the analysis on some categories only, while still taking into account all the available information in the input table. XLSTAT allows you to select the categories that belong to the subset.
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

Reload: Click this button to reload the default options.

Delete: Click this button to delete the data selections.

Change data handling: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

The first selection field lets you alternatively select two types of tables:

Observations/variables table: Select this option if your data correspond to a table with N observations described by P qualitative variables. If the headers of the columns have also been selected, make sure the “Variable labels” option is activated.

Disjunctive table: Select this option if your data correspond to a disjunctive table. If the headers of the columns have also been selected, make sure the “Variable labels” option is activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.
Variable labels: Activate this option if the first row contains the variable labels (case of an observations/variables table) or the category labels (case of a disjunctive table).

Weights: Activate this option if you want to weight the observations. If you do not activate this option, the weights are considered to be equal to 1. The weights must be greater or equal to 0. If the “Variable labels” option is activated make sure that the header of the selection has also been selected.

Options tab:

Advanced analysis:

- Supplementary data: If you select this option, the “Supplementary data” tab is activated, and you can then modify the corresponding options.

- Subset analysis: If you select this option, XLSTAT will ask you to select during the computations the categories that belong to the subset to analyze.

Sort the categories alphabetically: Activate this option so that the categories of all the variables are sorted alphabetically.

Variable-Category labels: Activate this option to use variable-category labels when displaying outputs. Variable-Category labels include the variable name as a prefix and the category name as a suffix.

Filter factors: You can activate one of the three following options in order to reduce the number of factors displayed:

- Minimum %: Activate this option and then enter the minimum percentage that should be reached to determine the number of factors to display.

- Maximum number: Activate this option to set the maximum number of factors to take into account when displaying the results.

- 1/p: Activate this option to only take into account the factors which eigenvalue is greater than 1/p, where p is the number of variables. This is the default option.

Supplementary data tab:

Supplementary observations: Activate this option if you want to compute the coordinates and to display supplementary observations. These observations are not taken into account for the first phase of the computations. They are passive observations. Several methods are available to identify the supplementary observations:
• **Random**: The observations are randomly selected. The “Number of observations” N must then be specified.

• **N last rows**: The N last observations are selected for the validation. The “Number of observations” N must then be specified.

• **N first rows**: The N first observations are selected for the validation. The “Number of observations” N must then be specified.

• **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to use as supplementary observations.

**Supplementary variables**: Activate this option if you want to compute a posteriori the coordinates of variables that are not taken into account for the computing of the principal axes (passive variables, as opposed to active variables).

• **Quantitative**: Activate this option if you want to include quantitative supplementary variables. If the headers of the columns of the main table have been selected, you also need to select headers here.

• **Qualitative**: Activate this option if want to include qualitative supplementary variables. If the headers of the columns of the main table have been selected, you also need to select headers here.

**Missing data** tab:

• **Do not accept missing data**: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

• **Remove observations**: Activate this option to ignore the observations that contain missing data.

• **Group missing values into a new category**: Activate this option to group missing data into a new category of the corresponding variable.

• **Replace missing data**: Activate this option to replace missing data. When a missing data corresponds to a quantitative supplementary variable, they are replaced by the mean of the variable. When a missing data corresponds to a qualitative variable of the initial table (active variables) or to a qualitative supplementary variable (passive variable), a new “Missing” category is create for the variable.

**Outputs** tab:
Descriptive statistics: Activate this option to display the descriptive statistics for the selected variables.

Disjunctive table: Activate this option to display the full disjunctive table that corresponds to the selected qualitative variables.

Burt table: Activate this option to display the Burt table.

Display results for:

- **Observations**: Activate this option to display the results that concern the observations.
- **Variables**: Activate this option to display the results that concern the variables.

Eigenvalues: Activate this option to display the table and the scree plot of the eigenvalues.

Principal coordinates: Activate this option to display the principal coordinates.

Standard coordinates: Activate this option to display the standard coordinates.

Contributions: Activate this option to display the contributions.

Squared cosines: Activate this option to display the squared cosines.

Test values: Activate this option to display the test values for the variables.

- **Significance level (%)**: Enter the significance level used to determine if the test values are significant or not.

Charts tab:

3D view of the Burt table: Activate this option to display a 3D visualization of the Burt table.

Symmetric plots: Activate this option to display the symmetric observations and variables plots.

- **Observations and variables**: Activate this option to display a plot that shows both the observations and variables.
- **Observations**: Activate this option to display a plot that shows only the observations.
- **Variables**: Activate this option to display a plot that shows only the variables.
Asymmetric plots: Activate this option to display plots for which observations and variables play an asymmetrical role. These plots are based on the principal coordinates for the observations and the standard coordinates for the variables.

- **Observations**: Activate this option to display an asymmetric plot where the observations are displayed using their principal coordinates, and where the variables are displayed using their standard coordinates.

- **Variables**: Activate this option to display an asymmetric plot where the variables are displayed using their principal coordinates, and where the observations are displayed using their standard coordinates.

Labels: Activate this option to display the labels of the categories on the charts.

- **Colored labels**: Activate this option to display the labels with the same color as the corresponding points. If this option is not activated the labels are displayed in black.

Vectors: Activate this option to display the vectors for the standard coordinates on the asymmetric charts.

- **Length factor**: Activate this option to modulate the length of the vectors.

Filter: Activate this option to modulate the number of observations displayed:

- **Random**: The observations to display are randomly selected. The “Number of observations” N to display must then be specified.

- **N first rows**: The N first observations are displayed on the chart. The “Number of observations” N to display must then be specified.

- **N last rows**: The N last observations are displayed on the chart. The “Number of observations” N to display must then be specified.

- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to display.

**Dialog box (subset categories)**

This dialog is displayed if you selected the **Advanced analysis / Subset analysis** option in the MCA dialog box.
The list of categories that corresponds to the complete set of active qualitative variables is displayed so that you can select the subset of categories on which the analysis will be focused.

**All**: Click this button to select all the categories.

**None**: Click this button to deselect all the categories.

### Results

**Descriptive statistics**: This table is displayed only if the input data correspond to an observations/variables table.

**Disjunctive table**: This table is displayed only if the input data correspond to an observations/variables table. This table is an intermediary table that allows us to obtain the contingency table that corresponds to the two selected variables.

**Burt table**: The Burt table is displayed only if the corresponding option is activated in the dialog box. The 3D bar chart that follows is the graphical visualization of this table.

**Eigenvalues and percentages of inertia**: The eigenvalues, the percentages of inertia, the percentages of adjusted inertia and the corresponding scree plot are displayed. Only the non-trivial eigenvalues are displayed. If a filtering has been requested in the dialog box, it is not applied to this table, but only to the results that follow.

A series of results is displayed afterwards, first for the variables, then for the observations:

**Principal coordinates**: This table displays the principal coordinates which are used later to represent projections of profile points in symmetric and asymmetric plots.

**Standard coordinates**: This table displays the standard coordinates which are used later to represent projections of unit profile points in asymmetric plots.

**Contributions**: The contributions are helpful for interpreting the plots. The categories that have influenced the most the calculation of the axes are those that have the higher contributions. A shortcut consists of restricting the analysis to the categories which contribution on a given axis is higher than the corresponding relative weight that is displayed in the first column.
**Squared cosines**: As with other data analysis methods, the analysis of the squared cosines allows us to avoid misinterpretations of the plots that are due to projection effects. If, for a given category, the cosines are low on the axes of interest, then any interpretation of the position of the category is hazardous.

The plots (or maps) are the ultimate goal of Multiple Correspondence Analysis, because they considerably facilitate our interpretation of the data.

**Symmetric plots**: These plots are exclusively based on the principal coordinates. Depending on the choices made in the dialog box, a symmetric plot mixing observations and variables, a plot showing only the categories of the variables, and a plot showing only the observations, are displayed. The percentage of adjusted inertia that corresponds to each axis and the percentage of adjusted inertia cumulated over the two axes are displayed on the map.

**Asymmetric plots**: These plots use the principal coordinates for the categories of the variables and the standard coordinates for the observations and vice versa. The percentage of adjusted inertia that corresponds to each axis and the percentage of adjusted inertia cumulated over the two axes are displayed on the map. On an "asymmetric observations plot", one can study the way the observations are positioned relatively to the category vectors. The later indicate directions: if two observations are displayed in the same direction as a category vector, the observation that is the furthest in the category vector direction is more likely to have selected that category of response.

**Example**

A tutorial on how to use Multiple Correspondence Analysis is available on the Addinsoft website:

http://www.xlstat.com/demo-mca.htm

**References**


Multidimensional Scaling (MDS)

Use multidimensional scaling to represent in a two- or three-dimensional space the observations for which only a proximity matrix (similarity or dissimilarity) is available.

Description

Multidimensional Scaling (MDS) is used to go from a proximity matrix (similarity or dissimilarity) between a series of $N$ objects to the coordinates of these same objects in a $p$-dimensional space. $p$ is generally fixed at 2 or 3 so that the objects may be visualized easily. For example, with MDS, it is possible to reconstitute the position of towns on a map very precisely from the distances in kilometers (the dissimilarity in this case being the Euclidean distance) between the towns, modulo a rotation and a symmetry.

This example is only intended to demonstrate the performance of the method and to give a general understanding of how it is used. Practically, MDS is often used in psychometry (perception analysis) and marketing (distances between products obtained from consumer classifications) but there are applications in a large number of domains.

If the starting matrix is a similarity matrix (a similarity is greater the nearer the objects are), it will automatically be converted into a dissimilarity matrix for the calculations. The conversion is carried out by subtracting the matrix data from the value of the diagonal.

There are two types of MDS depending on the nature of the dissimilarity observed:

- **Metric MDS**: The dissimilarities are considered as continuous and giving exact information to be reproduced as closely as possible. There are a number of sub-models:
  - Absolute MDS: the distances obtained in the representation space must correspond as closely as possible to the distances observed in the starting dissimilarity matrix.
  - Ratio MDS: the distances obtained in the representation space must correspond as closely as possible to the distances observed in the initial matrix using a near proportionality factor (the factor being identical for all pairs of distances).
  - Interval MDS: the distances obtained in the representation space must correspond as closely as possible to the distances observed in the initial matrix using a near linear relationship (the linear relationship being identical for all pairs of distances).
  - Polynomial MDS: the distances obtained in the representation space must correspond as closely as possible to the distances observed in the initial matrix
using a near 2nd-degree polynomial relationship (the polynomial relationship being identical for all pairs of distances).

Note: the absolute model is used to compare distances in the representation space with those in the initial space. The other models have the advantage of speeding up the calculations.

- **Non metric MDS**: with this type of MDS, only the order of the dissimilarities counts. In other words, the MDS algorithm does not have to try to reproduce the dissimilarities but only their order. Two models are available:
  - Ordinal (1): the order of the distances in the representation space must correspond to the order of the corresponding dissimilarities. If there are two dissimilarities of the same rank, then there are no restrictions on the corresponding distances. In other words, dissimilarities of the same rank need not necessarily give equal distances in the representation space.
  - Ordinal (2): the order of the distances in the representation space must correspond to the order of the corresponding dissimilarities. If dissimilarities exist in the same rank, the corresponding distances must be equal.

The MDS algorithms aim to reduce the difference between the disparity matrix from the models and the distance matrix obtained in the representation configuration. For the absolute model, the disparity is equal to the dissimilarity of the starting matrix. The difference is measured through the Stress, several variations of which have been proposed:

- **Raw Stress**:
  \[
  \sigma_r = \sum_{i<j} w_{ij} (D_{ij} - d_{ij})^2
  \]
  where \( D_{ij} \) is the disparity between individuals i and j, and \( d_{ij} \) is the Euclidean distance on the representation for the same individuals. \( w_{ij} \) is the weight of the ij proximity (value is 1 by default).

- **Normalized Stress**:
  \[
  \sigma_n = \frac{\sum_{i<j} w_{ij} (D_{ij} - d_{ij})^2}{\sum_{i<j} w_{ij} D_{ij}^2}
  \]

- **Kruskal’s stress 1**:
  \[
  \sigma_1 = \sqrt{\frac{\sum_{i<j} w_{ij} (D_{ij} - d_{ij})^2}{\sum_{i<j} w_{ij} d_{ij}^2}}
  \]

- **Kruskal’s stress 2**:
\[ \sigma^2 = \frac{\sum_{i<j} w_{ij} (D_{ij} - d_{ij})^2}{\sum_{i<j} w_{ij} (d_{ij} - \bar{d})^2} \]

where \( \bar{d} \) is the average of the distances on the representation.

Note: for a given number of dimensions, the weaker the stress, the better the quality of the representation. Furthermore, the higher the number of dimensions, the weaker the stress.

To find out whether the result obtained is satisfactory and to determine which is the correct number of dimensions needed to give a faithful representation of the data, the evolution in the stress with the number of dimensions and the point from which the stress stabilizes may be observed. The Shepard diagram is used to observe any ruptures in the ordination of the distances. The more the chart looks linear, the better the representation. For the absolute model, for an ideal representation, the points must be aligned along the first bisector.

There are several MDS algorithms including, in particular, ALSCAL (Takane et al. 1977) and SMACOF (Scaling by MAjorizing a CONvex Function) which minimizes the "Normalized Stress" (de Leeuw, 1977). XLSTAT uses the SMACOF algorithm.

**Dialog box**

The dialog box is made up of several tabs corresponding to the various options for controlling the calculations and displaying the results. A description of the various components of the dialog box is given below.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload**: Click this button to reload the default options.
- **Delete**: Click this button to delete the data selections.
Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT allows you to select data by columns or by range. If the arrow points to the right, XLSTAT allows you to select data by rows or by range.

**General** tab:

The main data entry field is used to select one of two types of table:

**Data**: Select a similarity or dissimilarity matrix. If only the lower or upper triangle is available, the table is accepted. If differences are detected between the lower and upper parts of the selected matrix, XLSTAT warns you and offers to change the data (by calculating the average of the two parts) to continue with the calculations.

**Dissimilarities / Similarities**: Choose the option that corresponds to the type of your data.

**Model**: Select the model to be used. See description for more details.

**Dimensions**: Enter the minimum and maximum number of dimensions for the object representation space. The algorithm will be repeated for all dimensions between the two boundaries.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet in the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Labels included**: Activate this option if you have included row and column labels in the selection.

**Weights**: Activate this option if the data are weighted. You then select a weighting matrix (without selecting labels for rows and columns). If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0.

**Options** tab:

**Stress**: Choose the type of stress to be used for returning the results, given that the SMACOF algorithm minimizes the raw stress. See the description section for more details.

Initial configuration:
• **Random**: Activate this option to make XLSTAT generate the starting configuration randomly. Then enter the number of times the algorithm is to be repeated from a new randomly-generated configuration. The default value for the number of repetitions is 100. Note: the configuration displayed in the results is the repetition for which the best result was found.

• **User defined**: Activate this option to select an initial configuration which the algorithm will use as a basis for carrying out optimization.

Stop conditions:

• **Iterations**: Enter the maximum number of iterations for the SMACOF algorithm. Stress Optimization is stopped when the maximum number if iterations has been exceeded. Default value: 100.

• **Convergence**: Enter the minimum value of evolution in stress from one iteration to another which, when reached, means that the algorithms is considered to have converged. Default value: 0.00001.

**Missing data** tab:

**Do not accept missing data**: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

**Ignore missing data**: If you activate this option, XLSTAT does not include proximities corresponding to missing data when minimizing stress.

**Outputs** tab:

**Distances**: Activate this option to display the matrix of Euclidean distances corresponding to the optimum configuration.

**Disparities**: Activate this option to display the disparity matrix corresponding to the optimum configuration.

**Residual distances**: Activate this option to display the matrix of residual distances corresponding to the difference between the distance matrix and the disparity matrix.

**Charts** tab:

**Evolution of stress**: Activate this option to display the stress evolution chart according to the number of dimensions in the configuration.

**Configuration**: Activate this option to display the configuration representation chart. This chart is only displayed for the configuration in a two-dimensional space if this has been calculated.
- **Labels**: Activate this option if you want object labels to be displayed.
- **Colored labels**: Activate this option to show labels in the same color as the points.
- **Shepard diagram**: Activate this option to display the Shepard diagram.

**Results**

**Stress after minimization**: This table shows the final stress obtained, the number of iterations required and the level of convergence reached for the dimensions considered. Where multiple dimensions were considered, a chart is displayed showing the stress evolution as a function of the number of dimensions.

The results which follow are displayed for each of the dimensions considered.

**Configuration**: This table shows the coordinates of objects in the representation space. If this is a two-dimensional space, a graphic representation of the configuration is provided. If you have XLSTAT-3DPlot, you can also display a three-dimensional configuration.

**Distances measured in the representation space**: This table shows the distances between objects in the representation space.

**Disparities computed using the model**: This table shows the disparities calculated according to the model chosen (absolute, interval, etc.).

**Residual distances**: These distances are the difference between the dissimilarities of the starting matrix and the distances measured in the representation space.

**Comparative table**: This table is used to compare dissimilarities, disparities and distances and the ranks of these three measurements for all paired combinations of objects.

**Shepard diagram**: This chart compares the disparities and the distances to the dissimilarities. For a metric model, the representation is better the more the points are aligned with the first bisector of the plan. For a non-metric model, the model is better the more regularly the line of dissimilarities/disparities increases. Furthermore, the performance of the model can be evaluated by observing if the (dissimilarity/distance) points are near to the (dissimilarity/disparity) points.

**Example**

A tutorial on how to use Multidimensional Scaling is available on the Addinsoft website:

http://www.xlstat.com/demo-mds.htm
References


k-means clustering

Use k-means clustering to make up homogeneous groups of objects (classes) on the basis of their description by a set of quantitative variables.

Description

k-means clustering was introduced by McQueen in 1967. Other similar algorithms had been developed by Forgey (1965) (moving centers) and Friedman (1967).

k-means clustering has the following advantages in particular:

An object may be assigned to a class during one iteration then change class in the following iteration, which is not possible with Agglomerative Hierarchical Clustering for which assignment is irreversible.

By multiplying the starting points and the repetitions, several solutions may be explored.

The disadvantage of this method is that it does not give a consistent number of classes or enable the proximity between classes or objects to be determined.

The k-means and AHC methods are therefore complementary.

Note: if you want to take qualitative variables into account in the clustering, you must first perform a Multiple Correspondence Analysis (MCA) and consider the resulting coordinates of the observations on the factorial axes as new variables.

Principle of the k-means method

k-means clustering is an iterative method which, wherever it starts from, converges on a solution. The solution obtained is not necessarily the same for all starting points. For this reason, the calculations are generally repeated several times in order to choose the optimal solution for the selected criterion.

For the first iteration, a starting point is chosen which consists in associating the center of the k classes with k objects (either taken at random or not). Afterwards the distance between the objects and the k centers is calculated and the objects are assigned to the centers they are nearest to. Then the centers are redefined from the objects assigned to the various classes. The objects are then reassigned depending on their distances from the new centers. And so on until convergence is reached.

Classification criteria
Several classification criteria may be used to reach a solution. XLSTAT offers four criteria to be minimized.

**Trace(W):** The $W$ trace, *pooled SSCP matrix*, is the most traditional criterion. Minimizing the $W$ trace for a given number of classes amounts to minimizing the total within-class variance, in other words minimizing the heterogeneity of the groups. This criterion is sensitive to effects of scale. In order to avoid giving more weight to certain variables and not to others, the data must be normalized beforehand. Moreover, this criterion tends to produce classes of the same size.

**Determinant(W):** The determinant of $W$, *pooled within covariance matrix*, is a criterion considerably less sensitive to effects of scale than the $W$ trace criterion. Furthermore, group sizes may be less homogeneous than with the trace criterion.

**Wilks lambda:** The results given by minimizing this criterion are identical to that given by the determinant of $W$. Wilk's lambda criterion corresponds to the division of determinant(W) by determinant(T) where T is the total inertia matrix. Dividing by the determinant of T always gives a criterion between 0 and 1.

**Trace(W) / Median:** If this criterion is chosen, the class centroid is not the mean point of the class but the median point which corresponds to an object of the class. The use of this criterion gives rise to longer calculations.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

**OK**: Click this button to start the computations.

**Cancel**: Click this button to close the dialog box without doing any computation.

**Help**: Click this button to display the help.

**Reload**: Click this button to reload the default options.

**Delete**: Click this button to delete the data selections.

**General** tab:
**Observations/variables table**: Select a table comprising N objects described by P descriptors. If column headers have been selected, check that the "Variable labels" option has been activated.

**Column weights**: Activate this option if the columns are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Column labels" option is activated.

**Row weights**: Activate this option if the rows are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Column labels" option is activated.

**Classification criterion**: Choose the classification criterion (see the description section for more details).

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Column labels**: Activate this option if the first row of the data selections (Observations/variables table, row labels, row weights, column weights) contains a label.

**Row labels**: Activate this option if observations labels are available. Then select the corresponding data. If the "Column labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 ...).

**Number of classes**: Enter the number of classes to be created by the algorithm.

**Options** tab:

**Cluster rows**: Activate this option if you want to create classes of objects in rows described by descriptors in columns.

**Cluster columns**: Activate this option if you want to create classes of objects in columns described by descriptors in rows.
**Center**: Activate this option if you want to center the data before starting the calculations.

**Reduce**: Activate this option if you want to reduce the data before starting the calculations.

You can then select whether you want to apply the transformation on the rows or the columns.

**Stop conditions:**

- **Iterations**: Enter the maximum number of iterations for the k-means algorithm. The calculations are stopped when the maximum number of iterations has been exceeded. Default value: 500.

- **Convergence**: Enter the minimum value of evolution for the chosen criterion from one iteration to another which, when reached, means that the algorithm is considered to have converged. Default value: 0.00001.

**Initial partition**: Use these options to choose the way the first partition is chosen, in other words, the way objects are assigned to classes in the first iteration of the clustering algorithm.

- **N classes by data order**: Objects are assigned to classes depending on their order.

- **Random**: Objects are assigned to classes randomly.

- **User defined**: Objects are assigned to classes according to an indicator variable defined by the user. The user must in this case select a column indicator variable containing as many rows as objects (with an optional header), and the classes must be defined by the values 1 to k where k is the number of classes. If the "Column labels" option is activated you need to include a header in the selection.

- **Defined by centers**: The user has to select the k centers corresponding to the k classes. The number of rows must be equal to the number of classes and the number of columns equal to the number of columns in the data table. If the "Column labels" option is activated you need to include a header in the selection.

**Missing data** tab:

- **Do not accept missing data**: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

- **Remove observations**: Activate this option to remove the observations with missing data.

- **Estimate missing data**: Activate this option to estimate missing data before starting the computations.
- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

- **Nearest neighbour**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.

**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the variables selected.

**Results in the original space**: Activate this option to display the results in the original space. If the center/reduce options are activated and this option is not activated, the results are provided in the standardized space.

**Optimization summary**: Activate this option to display the optimization summary.

**Centroids**: Activate this option to display the table of centroids of the classes.

**Central objects**: Activate this option to display the coordinates of the nearest object to the centroid for each class.

**Results by class**: Activate this option to display a table giving the statistics and the objects for each of the classes.

**Results by object**: Activate this option to display a table giving the class each object is assigned to in the initial object order.

**Charts** tab:

**Evolution of the criterion**: Activate this option for the evolution chart of the chosen criterion.

**Profile plot**: Activate this option to display a plot that allows to compare the means of the different classes that have been create.

**Results**

**Summary statistics**: This table displays for the descriptors of the objects, the number of observations, the number of missing values, the number of non-missing values, the mean and the standard deviation.
**Optimization summary**: This table shows the evolution of the within-class variance. If several repetitions have been requested, the results for each repetition are displayed.

**Statistics for each iteration**: Activate this option to see the evolution of miscellaneous statistics calculated as the iterations for the repetition proceed, given the optimum result for the chosen criterion. If the corresponding option is activated in the Charts tab, a chart showing the evolution of the chosen criterion as the iterations proceed is displayed.

Note: if the values are standardized (option in the Options tab), the results for the optimization summary and the statistics for each iteration are calculated in the standardized space. On the other hand, the following results are displayed in the original space if the "Results in the original space" option is activated.

**Variance decomposition for the optimal classification**: This table shows the within-class variance, the inter-class variance and the total variance.

**Class centroids**: This table shows the class centroids for the various descriptors.

**Distance between the class centroids**: This table shows the Euclidean distances between the class centroids for the various descriptors.

**Central objects**: This table shows the coordinates of the nearest object to the centroid for each class.

**Distance between the central objects**: This table shows the Euclidean distances between the class central objects for the various descriptors.

**Results by class**: The descriptive statistics for the classes (number of objects, sum of weights, within-class variance, minimum distance to the centroid, maximum distance to the centroid, mean distance to the centroid) are displayed in the first part of the table. The second part shows the objects.

**Results by object**: This table shows the assignment class for each object in the initial object order.

**Profile plot**: This chart allows to compare the means of the different classes that have been create.

**Example**

A tutorial on k-means clustering is available on the Addinsoft website:

http://www.xlstat.com/demo-cluster2.htm
References


Agglomerative Hierarchical Clustering (AHC)

Use Agglomerative Hierarchical Clustering to make up homogeneous groups of objects (classes) on the basis of their description by a set of variables, or from a matrix describing the similarity or dissimilarity between the objects.

Description

Agglomerative Hierarchical Clustering (AHC) is a classification method which has the following advantages:

You work from the dissimilarities between the objects to be grouped together. A type of dissimilarity can be chosen which is suited to the subject studied and the nature of the data.

One of the results is the dendrogram which shows the progressive grouping of the data. It is then possible to gain an idea of a suitable number of classes into which the data can be grouped.

The disadvantage of this method is that it is slow. Furthermore, the dendrogram can become unreadable if too much data is used.

Principle of AHC

Agglomerative Hierarchical Clustering (AHC) is an iterative classification method whose principle is simple.

The process starts by calculating the dissimilarity between the N objects. Then two objects which when clustered together minimize a given agglomeration criterion, are clustered together thus creating a class comprising these two objects. Then the dissimilarity between this class and the N-2 other objects is calculated using the agglomeration criterion. The two objects or classes of objects whose clustering together minimizes the agglomeration criterion are then clustered together. This process continues until all the objects have been clustered.

These successive clustering operations produce a binary clustering tree (dendrogram), whose root is the class that contains all the observations. This dendrogram represents a hierarchy of partitions.

It is then possible to choose a partition by truncating the tree at a given level, the level depending upon either user-defined constraints (the user knows how many classes are to be obtained) or more objective criteria.
**Similarities and dissimilarities**

The proximity between two objects is measured by measuring at what point they are similar (similarity) or dissimilar (dissimilarity). If the user chooses a similarity, XLSTAT converts it into a dissimilarity as the AHC algorithm uses dissimilarities. The conversion for each object pair consists in taking the maximum similarity for all pairs and subtracting from this the similarity of the pair in question.

The **similarity** coefficients proposed are as follows: Cooccurrence, Cosine, Covariance (n-1), Covariance (n), Dice coefficient (also known as Sorensen coefficient), General similarity, Gower coefficient, Inertia, Jaccard coefficient, Kendall correlation coefficient, Kulczinski coefficient, Ochiai coefficient, Pearson’s correlation coefficient, Pearson Phi, Percent agreement, Rogers & Tanimoto coefficient, Sokal & Michener coefficient (or simple matching coefficient), Sokal & Sneath coefficient (1), Sokal & Sneath coefficient (2), Spearman correlation coefficient.

The **dissimilarity** coefficients proposed: Bhattacharya's distance, Bray and Curtis' distance, Canberra's distance, Chebychev's distance, Chi² distance, Chi² metric, Chord distance, Squared chord distance, Dice coefficient, Euclidian distance, Geodesic distance, Jaccard coefficient, Kendall dissimilarity, Kulczinski coefficient, Mahalanobis distance, Manhattan distance, Ochiai coefficient, Pearson's dissimilarity, Pearson's Phi, General dissimilarity, Rogers & Tanimoto coefficient, Sokal & Michener's coefficient, Sokal & Sneath's coefficient (1), Sokal & Sneath coefficient (2), Spearman dissimilarity.

Note: some of the abovementioned coefficients should be used with binary data only. If the data are not binary, XLSTAT asks you if it should automatically transform the data into binary data.

**Agglomeration methods**

To calculate the dissimilarity between two groups of objects A and B, different strategies are possible. XLSTAT offers the following methods:

**Simple linkage**: The dissimilarity between A and B is the dissimilarity between the object of A and the object of B that are the most similar. Agglomeration using simple linkage tends to contract the data space and to flatten the levels of each step in the dendrogram. As the dissimilarity between two elements of A and of B is sufficient to link A and B, this criterion can lead to very long clusters (chaining effect) while they are not homogeneous.

**Complete linkage**: The dissimilarity between A and B is the largest dissimilarity between an object of A and an object of B. Agglomeration using complete linkage tends to dilate the data space and to produce compact clusters.

**Unweighted pair-group average linkage**: The dissimilarity between A and B is the average of the dissimilarities between the objects of A and the objects of B. Agglomeration using
Unweighted pair-group average linkage is a good compromise between the two preceding criteria, and provides a fair representation of the data space properties.

**Weighted pair-group average linkage:** The average dissimilarity between the objects of A and of B is calculated as the sum of the weighted dissimilarities, so that equal weights are assigned to both groups. As with unweighted pair-group average linkage, this criterion provides a fairly good representation of the data space properties.

**Flexible linkage:** This criterion uses a $\beta$ parameter that varies between [-1, +1]; this can generate a family of agglomeration criteria. For $\beta = 0$ the criterion is weighted pair-group average linkage. When $\beta$ is near to 1, chain-like clusters result, but as $\beta$ decreases and becomes negative, more and more dilatation is obtained.

**Ward’s method:** This method aggregates two groups so that within-group inertia increases as little as possible to keep the clusters homogeneous. This criterion, proposed by Ward (1963), can only be used in cases with quadratic distances, i.e. cases of Euclidian distance and Chi-square distance.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload**: Click this button to reload the default options.
- **Delete**: Click this button to delete the data selections.

**General** tab:

**Observations/variables table/ Proximity matrix:** Choose the option which corresponds to the format of your data, then select the data. For the **Observations/variables table** option, select a table comprising N objects described by P quantitative descriptors. For a **Proximity matrix**, select a squared matrix giving the proximities between the objects. If column headers
have been selected, check that the "Variable labels" option has been activated. For a proximity matrix, if column labels have been selected, row labels must also be selected.

**Proximity type: similarities / dissimilarities**: Choose the proximity type to be used. The data type and proximity type determine the list of possible indexes for calculating the proximity matrix.

**Agglomeration method**: Choose the agglomeration method (see the description section for more details).

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet in the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Column labels**: Activate this option if the first row of the data selections (Observations/variables table, row labels, row weights, column weights) contains a label. Where the selection is a proximity matrix, if this option is activated, the first column must also include the object labels.

**Row labels**: Activate this option if observations labels are available. Then select the corresponding data. If the "Column labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

**Column weights**: Activate this option if the columns are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Column labels" option is activated.

**Row weights**: Activate this option if the rows are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Column labels" option is activated.

**Options** tab:

**Cluster rows**: Activate this option if you want to create classes of objects in rows described by data in columns.

**Cluster columns**: Activate this option if you want to create classes of objects in columns described by data in rows.
Center: Activate this option is you want to center the data before starting the calculations.

Reduce: Activate this option is you want to reduce the data before starting the calculations.

You can then select whether you want to apply the transformation on the rows or the columns.

Truncation: Activate this option is you want XLSTAT to automatically define the truncation level, and therefore the number of classes to retain, or if you want to define the number of classes to create, or the level at which the dendrogram is to be truncated.

Within-class variances: Activate this option to select the within-class variances. This option is only active if object weights have been selected (row weights if you are clustering rows, column weights if you are clustering columns). This option can be used if you previously clustered the objects using another method (k-means for example) and want to use a method such as unweighted pair group averages to cluster the groups previously obtained. If a column header has been selected, check that the “Column labels” option is activated.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- Mean or mode: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

- Nearest neighbour: Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Proximity matrix: Activate this option to display the proximity matrix.

Node statistics: Activate this option to display the statistics for dendrogram nodes.
**Centroids**: Activate this option to display the table of centroids of the classes.

**Central objects**: Activate this option to display the coordinates of the nearest object to the centroid for each class.

**Results by class**: Activate this option to display a table giving the statistics and the objects for each of the classes.

**Results by object**: Activate this option to display a table giving the class each object is assigned to in the initial object order.

**Charts** tab:

**Levels bar chart**: Activate this option to display the diagram of levels showing the impact of successive clusterings.

**Dendrogram**: Activate this option to display the dendrogram.

- **Horizontal**: Choose this option to display a horizontal dendrogram.
- **Vertical**: Choose this option to display a vertical dendrogram.
- **Full**: Activate this option to display the full dendrogram (all objects are represented).
- **Truncated**: Activate this option to display the truncated dendrogram (the dendrogram starts at the level of the truncation).
- **Labels**: Activate this option to display object labels (full dendrogram) or classes (truncated dendrogram) on the dendrogram.
- **Colors**: Activate this option to use colors to represent the different groups on the full dendrogram.

**Profile plot**: Activate this option to display a plot that allows to compare the means of the different classes that have been create.

**Results**

**Summary statistics**: This table displays for the descriptors of the objects, the number of observations, the number of missing values, the number of non-missing values, the mean and the standard deviation.

**Node statistics**: This table shows the data for the successive nodes in the dendrogram. The first node has an index which is the number of objects increased by 1. Hence it is easy to see
at any time if an object or group of objects is clustered with another object or group of objects at
the level of a new node in the dendrogram.

**Levels bar chart:** This table displays the statistics for dendrogram nodes.

**Dendrograms:** The full dendrogram displays the progressive clustering of objects. If truncation
has been requested, a broken line marks the level the truncation has been carried out. The
truncated dendrogram shows the classes after truncation.

**Class centroids:** This table shows the class centroids for the various descriptors.

**Distance between the class centroids:** This table shows the Euclidean distances between
the class centroids for the various descriptors.

**Central objects:** This table shows the coordinates of the nearest object to the centroid for
each class.

**Distance between the central objects:** This table shows the Euclidean distances between
the class central objects for the various descriptors.

**Results by class:** The descriptive statistics for the classes (number of objects, sum of
weights, within-class variance, minimum distance to the centroid, maximum distance to the
centroid, mean distance to the centroid) are displayed in the first part of the table. The second
part shows the objects.

**Results by object:** This table shows the assignment class for each object in the initial object
order.

**Profile plot:** This chart allows to compare the means of the different classes that have been
create.

---

**Example**

A tutorial on agglomerative hierarchical clustering is available on the Addinsoft website:

http://www.xlstat.com/demo-cluster.htm

---

**References**

Singapore.


Gaussian Mixture Models

Use Gaussian mixture models to cluster multidimensional data according to their distribution.

Description

Gaussian mixture models allow data to be modeled by a set of Gaussian distributions. Usually, these models are used in a clustering framework and each Gaussian is supposed to correspond to one group.

Mixture model

Denote $x = \{x_1, \ldots, x_n\}$ a vector of size $n$, where $x_i \in \mathbb{R}^d$. Assume that each $x_i$ is distributed according to a probability distribution function $f$:

$$f(x_i; \theta) = \sum_{k=1}^{K} \pi_k h(x_i; \nu_k),$$

where $\pi_k$ is the mixture proportion of the group $k$ ($\forall k \in \{1, \ldots, K\}$, $0 < \pi_k < 1$ and $\sum_{k=1}^{K} \pi_k = 1$) and $\theta$ represents the model parameters. The function $h(\cdot; \nu_k)$ is a probability distribution of dimension $d$ with parameter $\nu_k$. For instance, for Gaussian mixture models, $h$ is a Gaussian with mean $\mu_k$ and variance $\Sigma_k$, hence $\nu_k = \left( \mu_k, \Sigma_k \right)$.

Note that, for a mixture distribution, there is a label vector $z = \{z_1, \ldots, z_n\}$ with $z_i = \{z_{i1}, \ldots, z_{ik}\}$ defined such that:

$$
\begin{cases}
  z_{ik} = 1 \text{ if } x_i \text{ is assigned to the k-th component} \\
  z_{ik} = 0 \text{ otherwise}
\end{cases}
$$

This vector is often unknown and in a clustering context or density estimation context, the estimation of each $z_i$ is of main interest.

Inference of the model parameters

Due to the latent variables $z$, the estimation of mixture models parameters cannot be done by directly maximizing the log-likelihood. This optimization requires an iterative algorithm such as
the EM (Dempster et al. (1977)) or the SEM, its stochastic version proposed by McLachlan and Peel (2000).

Once the parameters have been estimated, the vector of labels is directly obtained by assigning each \( x_i \) to the component providing the highest posterior probability \( \hat{\tau}_{ik} \) given by:

\[
\hat{\tau}_{ik} = \frac{\hat{\pi}_k h(x_i; \hat{\theta}_k)}{\sum_{j=1}^{K} \hat{\pi}_j h(x_i; \hat{\theta}_j)}.
\]

For a clustering purpose, Celeux and Govaert (1992) proposed the CEM (Classification EM) algorithm which is a k-means-like algorithm and can be viewed as a classifying version of the EM. Contrary to the EM and the SEM, the CEM algorithm maximizes the quantity

\[
\sum_{j=1}^{n} \sum_{k=1}^{K} z_{ik} \log \left[ \pi_k h(x_i; \nu_k) \right]
\]

and not the log-likelihood.

**Model selection (Choice of the number of components)**

The number of components of a mixture model is often unknown. Several criteria such as the BIC (Bayesian Information Criterion, Schwarz (1978)) or the AIC (Akaike Information Criterion, Akaike (1974)) can be used. These criteria are based on a penalization of the observed log-likelihood \( L(x; \theta) \). In 2000, Biernacki et al. proposed the ICL (Integrated Completed Likelihood) which aims at penalizing the complete log-likelihood \( L(x, z; \theta) \). This criterion can be written as a BIC criterion penalized by an entropy term:

\[
-\sum_{j=1}^{n} \sum_{k=1}^{K} \hat{z}_{ik} \log \tau_{ik}.
\]

For assessing the number of components of a mixture, we can try to find the model which provides well-separated clusters. Proposed by Celeux and Soromengo (1996), the NEC is an entropy-based criterion which measures the overlap of the mixture components:

\[
NEC_k = \frac{E_k}{L_k - L_1}
\]

where \( E_k \) is the entropy of the mixture model with \( k \) components and \( L_k \) its complete log-likelihood (calculated on the ML estimates). This criterion can also be used as a diagnostic tool. For a given number of components \( K' \), if \( NEC_{K'} \leq 1 \), we can say that there is a clustering structure in the data.

**Parsimonious Gaussian mixture models**
In the Gaussian mixture models context, the number of parameters can be large and the quantity of data available can be insufficient to achieve a reliable estimate. A classical approach is to reduce the number of parameters by applying constraints on the variance-covariance matrix $\Sigma_k$. Bandfield and Raftery (1993) and Celeux and Govaert (1995) proposed to express the matrix $\Sigma_k$ in term of its eigenvalue decomposition:

$$\Sigma_k = \lambda_k D_k A_k D_k^\top,$$

where $\lambda_k = \left| \Sigma_k \right|^\frac{1}{2}$ is the volume of the $k$-th component, $D_k$ the matrix of eigenvectors and $A_k$ is a diagonal matrix composed by the eigenvalues of $\Sigma_k$ organized in decreasing order such that $|A_k| = 1$. These two matrices $D_k$ and $A_k$ allow to control the orientation and the shape of the component.

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of parameters</th>
<th>Model name</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_k D_k A_k D_k^\top$</td>
<td>$a + Kb$</td>
<td>VVV</td>
</tr>
<tr>
<td>$\lambda DAD^\top$</td>
<td>$a + b$</td>
<td>EEE</td>
</tr>
<tr>
<td>$\lambda D_k AD_k^\top$</td>
<td>$a + Kb - (K - 1)d$</td>
<td>EEV</td>
</tr>
<tr>
<td>$\lambda D_k A_k D_k^\top$</td>
<td>$a + Kb - (K - 1)$</td>
<td>EVV</td>
</tr>
<tr>
<td>$\lambda_k D_k AD_k^\top$</td>
<td>$a + Kb - (K - 1)(d - 1)$</td>
<td>VEV</td>
</tr>
<tr>
<td>$\lambda B$</td>
<td>$a + d$</td>
<td>EEI</td>
</tr>
<tr>
<td>$\lambda B_k$</td>
<td>$a + Kd - K + 1$</td>
<td>EVI</td>
</tr>
<tr>
<td>$\lambda_k B_k$</td>
<td>$a + Kd$</td>
<td>VVI</td>
</tr>
<tr>
<td>$\lambda_k B$</td>
<td>$a + d + K - 1$</td>
<td>VEI</td>
</tr>
<tr>
<td>$\lambda I$</td>
<td>$a + 1$</td>
<td>EII</td>
</tr>
<tr>
<td>$\lambda_k I$</td>
<td>$a + d$</td>
<td>VII</td>
</tr>
<tr>
<td>$\lambda_k DAD$</td>
<td>$a + b + K - 1$</td>
<td>VEE</td>
</tr>
<tr>
<td>$\lambda DA_k D$</td>
<td>$a + Kb + (K - 1)(d - 1)$</td>
<td>EVE</td>
</tr>
<tr>
<td>$\lambda_k DA_k D$</td>
<td>$a + Kb + (K - 1)d$</td>
<td>VVE</td>
</tr>
</tbody>
</table>
Thus, in the multidimensional case, we have 28 different models. In the one-dimensional case, only two models are available (equal variance or not).

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- ![Reload](image)
  - Click this button to reload the default options.
- ![Delete](image)
  - Click this button to delete the data selections.
- ![Data Change](image)
  - Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab:**

**Observations/variables table**: Select a table with N objects described by P descriptors. If column headers have been selected, check that the "Column labels" option has been activated.

**Row weights**: Activate this option if the rows are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Column labels" option is activated.
**Partial labeling**: Activate this option if you want to specify that some rows are constraint to be included in a specific group. If you do not activate this option, all the rows’ groups will be considered as unknown. Group identifier must be integers greater than or equal to 1. If a column header has been selected, check that the "Column labels" option is activated.

**Data dimension**: you can either do a one-dimensional (column by column) or multidimensional analysis.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Column labels**: Activate this option if the first row of the data selections (Observations/variables row weights) contains a label.

**Row labels**: Activate this option if observations labels are available. Then select the corresponding data. If the “Column labels” option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2…).

**Options(1) tab:**

**Inference algorithm**: Select the algorithm used for inference.

- **EM**: Usual EM algorithm proposed by Dempster et al. (1977) is used. It is the default algorithm.

- **SEM**: stochastic version of the EM algorithm. A stochastic step is added to the classical EM which aims at assigning the observations to the clusters. This algorithm can lead to empty classes and disturb parameters estimation.

- **CEM**: classifying version of the EM algorithm. A classification step is added to the classical EM which aims at assigning the observations to the clusters. This algorithm can lead to empty classes and disturb parameters estimation.

**Selection criteria**: Select the criterion to estimate the number of clusters.

- **BIC**: Bayesian Information Criterion. It is the default criterion.
- **AIC**: Akaike Information Criterion. This criterion tends to overestimate the number of components.

- **ICL**: Integrated Complete Likelihood. This criterion searches for the model which provides well-separated clusters. Usually, the selected number of clusters is larger than the number obtained with BIC.

- **NEC**: Normalized Entropy Criterion. The NEC is not defined for a model with one component. This criterion is devoted to choose the number of components rather than the model parameterization.

**Initialization**: Select the method to initialize the inference algorithm.

- **Random**: Objects are assigned to classes randomly. The algorithm is run as many times as specified by the number of repetitions until convergence of the algorithm. The best estimate from all repetitions is retained.

- **Short runs**: Objects are assigned to classes randomly. The algorithm is run as many times as specified by the number of repetitions with a maximum number of 5 iterations. The best estimate from all repetitions is retained to initialize the algorithm.

- **K-means**: Objects are assigned to classes according to the k-means algorithm.

**Number of repetitions**: Specify the number of repetitions when the initialization method is Random or Short EM.

**Stop conditions**:

- **Iterations**: Enter the maximum number of iterations for inference algorithm. The calculations are stopped when the maximum number if iterations has been exceeded. Default value: 500.

- **Convergence**: Enter the minimum value of evolution for the chosen criterion from one iteration to another which, when reached, means that the algorithms is considered to have converged. Default value: 0.00001

**Options(2) tab**:

- **Mixture models**: Select the model(s) you want to use to fit the data. The best model will be retained according the selection criteria.

- **Number of classes**: Select the minimum and maximum number of classes. The minimum number must be greater than or equal to 1 and the maximum number lower than the number of data points. Default values are 2 and 5.

- **Equal proportions**: Activate this option to constrain mixture proportion to be equal.
**Missing data** tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

**Outputs** tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Evolution of the criterion: Activate this option to display the evolution table of the chosen criterion.

Posterior probabilities: Activate this option to display the table of posterior probabilities for each cluster.

MAP classification: Activate this option to display the table of classification obtained by the MAP rule.

**Charts** tab:

Evolution of the criterion: Activate this option to display the evolution chart of the chosen criterion.

MAP classification: Activate this option to display the classification chart obtained by the MAP rule.

Fitted model: Activate this option to display the selected model.

Cumulative density function: Activate this option to display both the empirical and the estimated cdf. This chart is a diagnostic tool. If the two cdf are similar, the mixture model fits well. This chart is only available in the one-dimensional case.

Q-Q plot: Activate this option to display the Q-Q plot of the empirical distribution against the estimated mixture distribution. This chart is a diagnostic tool. If the points in the Q-Q plot approximately lie on the line $y=x$, we can consider the two distributions as similar.
Results

Summary statistics: This table displays for the descriptors of the objects, the number of observations, the number of missing values, the number of non-missing values, the mean and the standard deviation.

Evolution of the criterion: This table displays the values of the criterion for each selected model. A chart is also displayed.

Estimated parameters: Three tables are displayed: the mixture proportions, the means and the variance for each cluster.

Characteristics of the selected model: This table shows some characteristics of the selected model (BIC, AIC, ICL, Log-likelihood, NEC, Entropy, DDL).

Posterior probabilities: The posterior probabilities of belonging to each cluster are displayed in this table.

MAP classification: This table displays the assignment of each observation according to the MAP rule. A chart also displays this classification.

Adjusted model: The fitted model is represented on this chart.

Cumulative density function: This chart allows to compare the empirical cdf to the estimated one.

Q-Q plot: This chart allows to display the quantiles of the empirical distribution against those of the estimated mixture distribution.

Example

A tutorial on Gaussian mixture models is available on the Addinsoft website:

http://www.xlstat.com/demo-gmm.htm

References


Univariate clustering

Use univariate clustering to optimally cluster objects in k homogeneous classes, based on their description using a single quantitative variable.

Description

Univariate clustering clusters N one-dimensional observations (described by a single quantitative variable) into k homogeneous classes.

Homogeneity is measured here using the sum of the within-class variances. To maximize the homogeneity of the classes, we therefore try to minimize the sum of the within-class variances.

The algorithm used here is very fast and uses the method put forward by W.D. Fisher (1958). This method can be seen as a process of turning a quantitative variable into a discrete ordinal variable. There are many applications, e.g. in mapping applications for creating color scales or in marketing for creating homogeneous segments.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- OK: Click this button to start the computations.
- Cancel: Click this button to close the dialog box without doing any computation.
- Help: Click this button to display the help.
- : Click this button to reload the default options.
- : Click this button to delete the data selections.

General tab:
Observations/variables table: Select a table comprising N objects described by P descriptors. If column headers have been selected, check that the "Variable labels" option has been activated.

Row weights: Activate this option if the rows are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Column labels" option is activated.

Number of classes: Enter the number of classes to be created by the algorithm.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet in the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column labels: Activate this option if the first row of the data selections (Observations/variables table, row labels, row weights, column weights) contains a label.

Row labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Column labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- Mean or mode: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

- Nearest neighbour: Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.

Outputs tab:
Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Centroids: Activate this option to display the table of centroids of the classes.

Central objects: Activate this option to display the coordinates of the nearest object to the centroid for each class.

Results by class: Activate this option to display a table giving the statistics and the objects for each of the classes.

Results by object: Activate this option to display a table giving the class each object is assigned to in the initial object order.

Results

Summary statistics: This table displays for the descriptor of the objects, the number of observations, the number of missing values, the number of non-missing values, the mean and the standard deviation.

Class centroids: This table shows the class centroids for the various descriptors.

Distance between the class centroids: This table shows the Euclidean distances between the class centroids for the various descriptors.

Central objects: This table shows the coordinates of the nearest object to the centroid for each class.

Distance between the central objects: This table shows the Euclidean distances between the class central objects for the various descriptors.

Results by class: The descriptive statistics for the classes (number of objects, sum of weights, within-class variance, minimum distance to the centroid, maximum distance to the centroid, mean distance to the centroid) are displayed in the first part of the table. The second part shows the objects.

Results by object: This table shows the assignment class for each object in the initial object order.
References

Association rules

Use this tool to discover association rules for a set of items or objects.

Description

In 1994, Rakesh Agrawal and Ramakrishnan Sikrant have proposed the Apriori algorithm to identify associations between items in the form of rules. This algorithm is used when the volume of data to be analyzed is important. As the number of items can be several tens of thousands, combinatorics are such that all the rules can not be studied. It is therefore necessary to limit the search for rules to the most important ones. The quality measurements are probabilistic values which limit the combinatorial explosion during the two phases of the algorithm, and allow the sorting of the results.

Definition

Items: Depending on the application field, they can be products, objects, patients, events.

Transaction: Identified with a unique identifier, it is a set of items with a minimum of one item. Items can belong to several transactions.

Itemset: A group of items. Itemsets can be found in one or more transactions.

Support: The probability to find item or itemset X in a transaction. Estimated by the number of times an item or itemset is found across all the available transactions. This value lies between 0 and 1.

Rule: A rule defines a relationship between two itemsets X and Y that have no items in common. X→Y means that if we have X in a transaction, then we can have Y in the same transaction.

Support of a rule: The probability to find items or itemsets X and Y in a transaction. Estimated by the number of times both items or itemsets are found across all the available transactions. This value lies between 0 and 1.

Confidence of a rule: The probability to find item or itemset Y in a transaction, knowing item or itemset X is in the transaction. Estimated by the observed corresponding frequency (number of times X and Y are found across all transactions divided by the number of times X is found). This value lies between 0 and 1.

Lift of a rule: The lift of a rule, which is symmetric (Lift(X→Y)=Lift(Y→X)), is the support of the itemset grouping X and Y, divided by the support of X and the support of Y. This value can be any positive real. A lift greater than 1 implies a positive effect of X on Y (or Y on X) and
therefore the significance of the rule. A value of 1 means there is no effect and it is as if the
items or itemsets are independent. A lift lower than 1, means there is a negative effect of X on
Y or reciprocally. As if they were excluding each other.

Let \( I = \{i_1, \ldots, i_m\} \) be a set of items. Let \( T = \{t_1, \ldots, t_n\} \) be a set of transactions, such that \( t_i \) is a
subset of \( I \).

An association rule \( R \) is written in the following way:

\[
R : X \rightarrow Y, \quad X \in T, \quad Y \in T, \quad X \cap Y = \emptyset
\]

The support for a subset of \( I \) is given by: \( \text{support}(X) = \Pr(X) \)

The confidence of a rule \( (R: X \rightarrow Y) \) is given by: \( \text{confidence}(R) = \Pr(Y \mid X) \)

The lift of a rule \( (R: X \rightarrow Y) \) is given by:

\[
\text{lift}(R) = \frac{\sup port(X \cup Y)}{\sup port(X) \sup port(Y)}
\]

**Apriori algorithm**

This algorithm involves two steps:

1. Generation of subsets of \( I \) with a support greater than a minimum support.

2. Generation of association rules from the subsets of \( I \) whose confidence is greater than a
fixed minimum confidence.

**Hierarchies and multilevel approach**

XLSTAT proposes to take into account a hierarchy for grouping the items and study the
existing rules at different levels. The proposed method can generate association rules for
which the causes or consequences belong either to the same level of the hierarchy or to two
different levels.

To simplify the reading of the results, Han and Fu (1999) propose two indexes alpha and beta
between 0 and 1, to eliminate redundant and unnecessary rules.

A rule is said to be redundant if it is derived from a rule that is covering it hierarchically: the rule
\( R \) such that \( A_1, \ldots, A_n \rightarrow B_1, \ldots, B_m \) is redundant if there is a rule \( R' \) such that \( A'_1, \ldots, A'_n \rightarrow
B'_1, \ldots, B'_m \) with each \( A'_i \) and \( B'_i \) \( (i = 1 \ldots n) \) that are either the same or the parent of an
element of \( A \).

\( R \) is said to be redundant if its confidence \( \text{Conf}(R) \) lies in the interval \( [\exp(\text{Conf}(R)) - \alpha, \exp(\text{Conf}(R)) + \alpha] \)
with \( \exp(\text{Conf}(R)) = (\text{support}(B_1) / \text{support}(B'_1)) \ast \ldots \ast (\text{support}(B_m) / \text{support}(B'_m)) \ast \phi(R') \).
A rule is said to be useless if it does not provide more information than a rule with the same consequence and with fewer items as antecedents: Let R be the rule (A, B -> C), and R' the rule (A -> C). R is considered useless if its confidence Conf(R) is in the interval [Conf (R') - beta, Conf (R') + beta].

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- OK: Click this button to start the computations.
- Cancel: Click this button to close the dialog box without doing any computation.
- Help: Click this button to display the help.
- : Click this button to reload the default options.
- : Click this button to delete the data selections.
- : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

**Items:** Select a table of items and specify the data format. If column headers were selected, please check the option "Labels included" is activated. The available **data formats** are:

- **Transactional:** Choose this format if your data is in two columns, one indicating the transaction (to be selected in the Transactions field), the other item. Typically with this format, there is a column with the transaction IDs, with for each transaction, as many rows as there are items in the transaction, and a column indicating the items. The transactions can be in the first column and selected in this field.

- **List:** Choose this format if your data include one line per transaction while columns contain the names of the items corresponding to the transaction. The number of items...
per transaction may vary from one line to another. The number of columns in the
selection corresponds to the maximum number of items per transaction.

- **Transactions/Variables**: Choose this format if your data correspond to one line per
transaction and to one column per variable. This format is such that all transactions
have the same number of items, which is the number of variables, and that items from a
given variable cannot be present in the same transaction.

- **Contingency table**: Choose this format if your data include one row per transaction
and a column per item, with null values if the item is not present and a number greater
than 1 if it is present.

You also have the option to select the data in a flat file by clicking the [...] button.

**Transactions**: Select a column with the transaction IDs for each item. This selection is
required if the selected format is "Transactional" and if the array of items has only one column.
If the items table has two columns, the first column is considered as corresponding to the
transaction.

**Target items / Target variables**: Activate this option to define one or more items that you
want to appear on the right side (the consequence) rules. If the data format is Transactional or
List, you can select a list of items that must be in the right part (consequence) of the rules to be
generated. If the data format is Transactions/Variables you need to select the variable that will
be considered as target. All the rules will have in the right part (consequence) a category of the
selected variable. If the data format is Contingency table you can select one or more columns
that will be used to identify the target items.

**Range**: Activate this option if you want to display the results starting from a cell in an existing
worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Labels included**: Activate this option if the first row of the selected data contains a header.

**Minimum support**: Enter the value of the minimum support (between 0 and 1) so that only
rules containing subsets with a support greater than that value are generated.

**Minimum confidence**: Enter the value of the minimum confidence (between 0 and 1) so that
only rules with a confidence greater than that value are generated.
**Minimum number of antecedents:** Enter the minimum number of antecedents for the rules to generate.

**Options** tab:

**Sort:** Select the criterion used to sort the results (confidence, support, lift or nothing).

**Multi-level** tab:

**Use hierarchical information:** Activate this option if you want to select and use hierarchical information.

**Hierarchy:** Select a hierarchical table describing the hierarchy of items and the groups that include them. An item can only belong to a group of higher order. You have the option to select the data from a flat file.

**Support for each level:** select a table of values to assign a different support for each hierarchical level.

**Cross-level analysis:** Choose this option if you want to generate the rules regardless of their level.

**Alpha (redundant rules):** Select a value between 0 and 1 to remove redundant rules. Leave 0 if you do not want to use this option.

**Beta (useless rules):** Select a value between 0 and 1 to remove unnecessary rules. Let 0 if you do not want to use this option.

**Outputs** tab:

**Influence matrix:** Activate this option to display the influence matrix calculated from the confidence of the association rules.

**Matrix of items:** Activate this option to display a table showing the relative importance of combinations of items.

**Charts** tab:

**Influence chart:** Select this option to display a 2D graph showing the relative importance of various combinations obtained by the rules of association.

**Items charts:** This chart represents the relative importance combinations of items.
Results

Association rules: This table displays the association rules obtained by the Apriori algorithm as well as different values for each rule.

Matrix of influence: This table is the crosstab between the antecedents and consequences of rules, with as value, the criterion chosen for sorting the rules (confidence, support, or lift) in the Options tab.

Influence chart: représentation 2D montrant de l'importance relative des règles d'associations.

Items matrix: This chart allows you to view the relative importance of combinations of items. This symmetric table shows the average confidence for each combination of items (row / column and column / row). It is therefore an indicator of the strength of link between the items. It is then used to make a Multidimensional scaling (MDS) to obtain the Items chart which is a graphical representation of the table.

Example

A tutorial on association rules is available on the Addinsoft website:

http://www.xlstat.com/demo-assocrules.htm

References


Distribution fitting

Use this tool to fit a distribution to a sample of continuous or discrete quantitative data.

Description

Fitting a distribution to a data sample consists, once the type of distribution has been chosen, in estimating the parameters of the distribution so that the sample is the most likely possible (as regards the maximum likelihood) or that at least certain statistics of the sample (mean, variance for example) correspond as closely as possible to those of the distribution.

Distributions

XLSTAT provides the following distributions:

- **Arcsine ($\alpha$):** the density function of this distribution (which is a simplified version of the Beta type I distribution) is given by:
  \[ f(x) = \frac{\sin(\pi \alpha)}{\pi x} \left( \frac{x}{1-x} \right)^{\alpha}, \quad \text{with } 0<\alpha<1, \ x \in [0,1] \]
  We have \( E(X) = \frac{\alpha}{\alpha+1} \) and \( V(X) = \frac{\alpha(1-\alpha)}{2(\alpha+1)^2} \)

- **Bernoulli ($p$):** the density function of this distribution is given by:
  \[ P(X = 1) = p, \ P(X = 0) = 1-p \quad \text{with } p \in [0,1] \]
  We have \( E(X)= p \) and \( V(X) = p(1-p) \)
  The Bernoulli, named after the Swiss mathematician Jacob Bernoulli (1654-1705), allows to describe binary phenomena where only events can occur with respective probabilities of $p$ and $1-p$.

- **Beta ($\alpha, \beta$):** the density function of this distribution (also called Beta type I) is given by:
  \[ f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}, \quad \text{with } \alpha, \beta>0, \ x \in [0,1] \text{ and } B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \]
  We have \( E(X) = \alpha/(\alpha+\beta) \) and \( V(X) = \alpha\beta/[(\alpha+\beta+1)(\alpha+\beta)^2] \)

- **Beta4 ($\alpha, \beta, c, d$):** the density function of this distribution is given by:
We have $E(X) = c + (c-d)\alpha/\beta$ and $V(X) = (c-d)^2\alpha\beta/[(\alpha+\beta+1)(\alpha+\beta)^2]$

Pour the type I beta distribution, $X$ takes values in the $[0,1]$ range. The beta distribution is obtained by a variable transformation such that the distribution is on a $[c, d]$ interval where $c$ and $d$ can take any value.

- **Beta ($a, b$):** the density function of this distribution (also called Beta type I) is given by:

\[
f(x) = \frac{1}{B(a, b)} x^{a-1} (1-x)^{b-1}, \quad \text{with } a, b > 0, \quad x \in [0, 1]
\]

\[
E(X) = a/(a+b) \quad \text{and} \quad V(X) = ab/[(a+b+1)(a+b)^2]
\]

- **Binomial ($n, p$):** the density function of this distribution is given by:

\[
P(X = x) = C_n^x p^x (1-p)^n-x, \quad \text{with } x \in N, \quad n \in N^+, \quad p \in [0,1]
\]

\[
E(X) = np \quad \text{and} \quad V(X) = np(1-p)
\]

$n$ is the number of trials, and $p$ the probability of success. The binomial distribution is the distribution of the number of successes for $n$ trials, given that the probability of success is $p$.

- **Negative binomial type I ($n, p$):** the density function of this distribution is given by:

\[
P(X = x) = C_{n+x-1}^x p^n (1-p)^x, \quad \text{with } x \in N, \quad n \in N^+, \quad p \in [0,1]
\]

\[
E(X) = n(1-p)/p \quad \text{and} \quad V(X) = n(1-p)/p^2
\]

$n$ is the number of successes, and $p$ the probability of success. The negative binomial type I distribution is the distribution of the number $x$ of unsuccessful trials necessary before obtaining $n$ successes.

- **Negative binomial type II ($k, p$):** the density function of this distribution is given by:

\[
P(X = x) = \frac{\Gamma(k+x)}{x!\Gamma(k)(1+p)^{k+x}} p^x, \quad \text{with } x \in N, \quad k, p > 0
\]

\[
E(X) = kp \quad \text{and} \quad V(X) = kp(p+1)
\]
The negative binomial type II distribution is used to represent discrete and highly heterogeneous phenomena. As \( k \) tends to infinity, the negative binomial type II distribution tends towards a Poisson distribution with \( \lambda = kp \).

- **Chi-square (df):** the density function of this distribution is given by:
  \[
  f(x) = \left(\frac{1}{2}\right)^{df/2} \frac{x^{df/2-1} e^{-x/2}}{\Gamma(df/2)} , \quad \text{with } x > 0, \ df \in \mathbb{N}^*
  \]
  \[
  \text{E}(X) = df \text{ and } \text{V}(X) = 2df
  \]
  The Chi-square distribution corresponds to the distribution of the sum of \( df \) squared standard normal distributions. It is often used for testing hypotheses.

- **Erlang \((k, \lambda)\):** the density function of this distribution is given by:
  \[
  f(x) = \lambda^k x^{k-1} e^{-\lambda x} \frac{1}{(k-1)!} , \quad \text{with } x \geq 0 \text{ and } k, \lambda > 0 \text{ and } k \in \mathbb{N}
  \]
  \[
  \text{E}(X) = k/\lambda \text{ and } \text{V}(X) = k/\lambda^2
  \]
  \( k \) is the shape parameter and \( \lambda \) is the rate parameter.

  This distribution, developed by the Danish scientist A. K. Erlang (1878-1929) when studying the telephone traffic, is more generally used in the study of queuing problems.

  Note: When \( k=1 \), this distribution is equivalent to the exponential distribution. The Gamma distribution with two parameters is a generalization of the Erlang distribution to the case where \( k \) is a real and not an integer (for the Gamma distribution the scale parameter \( \beta = 1/\lambda \) is used).

- **Exponential(\( \lambda \))**:
  \[
  f(x) = \lambda \exp\left(-\lambda x\right) , \quad \text{with } x > 0 \text{ and } \lambda > 0
  \]
  \[
  \text{E}(X) = 1/\lambda \text{ and } \text{V}(X) = 1/\lambda^2
  \]
  The exponential distribution is often used for studying lifetime in quality control.

- **Fisher \((df_1, df_2)\)**:
  \[
  f(x) = \frac{1}{xB\left(df_1/2, df_2/2\right)} \left(\frac{df_1 x}{df_1 x + df_2}\right)^{df_1/2} \left(1 - \frac{df_1 x}{df_1 x + df_2}\right)^{df_2/2} ,
  \]
  \[
  \text{with } x \geq 0 \text{ and } df_1, df_2 \in \mathbb{N}^*
  \]
  \[
  \text{E}(X) = df_2/(df_2-2) \text{ if } df_2>0 , \text{ and } \text{V}(X) = 2df_2^3(df_1+df_2-2)/\left[df_1(df_2-2)^2(df_2-4)\right]
  \]
Fisher’s distribution, from the name of the biologist, geneticist and statistician Ronald Aylmer Fisher (1890-1962), corresponds to the ratio of two Chi-square distributions. It is often used for testing hypotheses.

- **Fisher-Tippett ($\beta, \mu$):** the density function of this distribution is given by:
  
  $$f(x) = \frac{1}{\beta} \exp \left( -\frac{x-\mu}{\beta} - \exp \left( -\frac{x-\mu}{\beta} \right) \right), \quad \text{with } \beta > 0$$

  $$E(X) = \mu + \beta \gamma \quad \text{and} \quad V(X) = (\pi \beta)^2/6 \quad \text{where } \gamma \text{ is the Euler-Mascheroni constant.}$$

  The Fisher-Tippett distribution, also called the Log-Weibull or extreme value distribution, is used in the study of extreme phenomena. The Gumbel distribution is a special case of the Fisher-Tippett distribution where $\beta=1$ and $\mu=0$.

- **Gamma ($k, \beta, \mu$):** the density of this distribution is given by:
  
  $$f(x) = (x-\mu)^{k-1} e^{-(x-\mu)/\beta} / \beta^k \Gamma(k), \quad \text{with } x > \mu \text{ and } k, \beta > 0$$

  $$E(X) = \mu + k \beta \quad \text{and} \quad V(X) = k \beta^2$$

  $k$ is the shape parameter of the distribution and $\beta$ the scale parameter.

- **GEV ($\beta, k, \mu$):** the density function of this distribution is given by:
  
  $$f(x) = \frac{1}{\beta} \left( 1 - k \frac{x-\mu}{\beta} \right)^{1/k-1} \exp \left( - \left( 1 - k \frac{x-\mu}{\beta} \right)^{1/k} \right), \quad \text{with } \beta > 0$$

  We have $E(X) = \mu + \frac{\beta}{k} \Gamma(1+k)$ and $V(X) = \left( \frac{\beta}{k} \right)^2 \left( \Gamma(1+2k) - \Gamma^2(1+k) \right)$

  The GEV (Generalized Extreme Values) distribution is much used in hydrology for modeling flood phenomena. $k$ lies typically between -0.6 and 0.6.

- **Gumbel:** the density function of this distribution is given by:
  
  $$f(x) = \exp \left( -x - \exp \left( -x \right) \right)$$

  $$E(X) = \gamma \quad \text{and} \quad V(X) = \pi^2/6 \quad \text{where } \gamma \text{ is the Euler-Mascheroni constant (0.5772156649...).}$$

  The Gumbel distribution, named after Emil Julius Gumbel (1891-1966), is a special case of the Fisher-Tippett distribution with $\beta=1$ and $\mu=0$. It is used in the study of extreme phenomena such as precipitations, flooding and earthquakes.

- **Logistic ($\mu, s$):** the density function of this distribution is given by:
\[ f(x) = \frac{e^{-\frac{(x-\mu)}{s}}}{s \left(1 + e^{-\frac{(x-\mu)}{s}}\right)}, \quad \text{with } \mu \in R, \text{ and } s > 0 \]

We have \( E(X) = \mu \) and \( V(X) = (\pi s)^2/3 \)

- **Lognormal \((\mu, \sigma)\):** the density function of this distribution is given by:

\[ f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}, \quad \text{with } x, \sigma > 0 \]

\( E(X) = \exp(\mu + \sigma^2/2) \) and \( V(X) = [\exp(\sigma^2)-1]\exp(2\mu + \sigma^2) \)

- **Lognormal2 \((m, s)\):** the density function of this distribution is given by:

\[ f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}, \quad \text{with } x, \sigma > 0 \]

\( \mu = \ln(m) - \ln(1+s^2/m^2)/2 \) and \( \sigma^2 = \ln(1+s^2/m^2) \)

\( E(X) = m \) and \( V(X) = s^2 \)

This distribution is just a reparametrization of the Lognormal distribution.

- **Normal \((\mu, \sigma)\):** the density function of this distribution is given by:

\[ f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad \text{with } \sigma > 0 \]

\( E(X) = \mu \) and \( V(X) = \sigma^2 \)

- **Standard normal:** the density function of this distribution is given by:

\[ f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \]

\( E(X) = 0 \) and \( V(X) = 1 \)

This distribution is a special case of the normal distribution with \( \mu=0 \) and \( \sigma=1 \).

- **Pareto \((a, b)\):** the density function of this distribution is given by:

\[ f(x) = \frac{ab^a}{x^{a+1}}, \quad \text{with } a, b > 0 \text{ and } x \geq b \]

\( E(X) = ab/(a-1) \) and \( V(X) = ab^2/[(a-1)^2(a-2)] \)
The Pareto distribution, named after the Italian economist Vilfredo Pareto (1848-1923), is also known as the Bradford distribution. This distribution was initially used to represent the distribution of wealth in society, with Pareto's principle that 80% of the wealth was owned by 20% of the population.

- **PERT** (a, m, b): the density function of this distribution is given by:

\[
f(x) = \frac{1}{B(\alpha, \beta)} \frac{(x-a)^{\alpha-1}(b-x)^{\beta-1}}{(b-a)\alpha\beta}, \quad \text{with } \alpha, \beta > 0, x \in [a, b]
\]

\[
a, b \in \mathbb{R}, \text{ and } B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}
\]

\[
\alpha = \frac{4m + b - 5a}{b - a}
\]

\[
\beta = \frac{5b - a - 4m}{b - a}
\]

We have \(E(X) = (b-a)\alpha/(\alpha+\beta)\) and \(V(X) = (b-a)^2\alpha\beta/[(\alpha+\beta+1)(\alpha+\beta)^2]\)

The PERT distribution is a special case of the beta4 distribution. It is defined by its definition interval \([a, b]\) and \(m\) the most likely value (the mode). PERT is an acronym for *Program Evaluation and Review Technique*, a project management and planning methodology. The PERT methodology and distribution were developed during the project held by the US Navy and Lockheed between 1956 and 1960 to develop the Polaris missiles launched from submarines. The PERT distribution is useful to model the time that is likely to be spent by a team to finish a project. The simpler triangular distribution is similar to the PERT distribution in that it is also defined by an interval and a most likely value.

- **Poisson** (\(\lambda\)): the density function of this distribution is given by:

\[
P(X = x) = \frac{\exp(-\lambda)\lambda^x}{x!}, \quad \text{with } x \in \mathbb{N} \text{ and } \lambda > 0
\]

\(E(X) = \lambda\) and \(V(X) = \lambda\)

Poisson's distribution, discovered by the mathematician and astronomer Siméon-Denis Poisson (1781-1840), pupil of Laplace, Lagrange and Legendre, is often used to study queuing phenomena.

- **Student** (\(df\)): the density function of this distribution is given by:

\[
f(x) = \frac{\Gamma\left(\left(df + 1/2\right)\right)}{\sqrt{\pi df} \Gamma\left(df / 2\right)} \left(1 + x^2 / df\right)^{-\left(df + 1\right)/2}, \quad \text{with } df > 0
\]
E(X) = 0 if \( df > 1 \) and \( V(X) = \frac{df}{(df - 2)} \) if \( df > 2 \)

The English chemist and statistician William Sealy Gosset (1876-1937), used the nickname Student to publish his work, in order to preserve his anonymity (the Guinness brewery forbade its employees to publish following the publication of confidential information by another researcher). The Student’s t distribution is the distribution of the mean of \( df \) variables standard normal variables. When \( df = 1 \), Student’s distribution is a Cauchy distribution with the particularity of having neither expectation nor variance.

- **Trapezoidal (a, b, c, d):** the density function of this distribution is given by:
  \[
  \begin{align*}
  f(x) &= \frac{2(x-a)}{(d+c-b-a)(b-a)}, \quad x \in [a,b] \\
  f(x) &= \frac{2}{d+c-b-a}, \quad x \in [b,c] \\
  f(x) &= \frac{2(d-x)}{(d+c-b-a)(d-c)}, \quad x \in [a,b] \\
  f(x) &= 0, \quad x < a, x > d \\
  \end{align*}
  \]
  with \( a < m < b \)

  We have \( E(X) = \frac{d^2+c^2-b^2-a^2+cd-ab}{3(d+c-b-a)} \)

  and \( V(X) = \frac{(c+d)(c^2+d^2)-(a+b)(a^2+b^2)}{6(d+c-b-a)} - E^2(X) \)

  This distribution is useful to represent a phenomenon for which we know that it can take values between two extreme values (a and d), but that it is more likely to take values between two values (b and c) within that interval.

- **Triangular (a, m, b):** the density function of this distribution is given by:
  \[
  \begin{align*}
  f(x) &= \frac{2(x-a)}{(b-a)(m-a)}, \quad x \in [a,m] \\
  f(x) &= \frac{2(b-x)}{(b-a)(b-m)}, \quad x \in [m,b] \\
  f(x) &= 0, \quad x < a, x > b \\
  \end{align*}
  \]
  with \( a < m < b \)

  We have \( E(X) = \frac{a+m+b}{3} \) and \( V(X) = \frac{a^2+m^2+b^2-ab-am-bm}{18} \)

- **TriangularQ (q1, m, q2, p1, p2):** the density function of this distribution is a reparametrization of the Triangular distribution. A first step requires estimating the \( a \) and \( b \) parameters of the triangular distribution, from the \( q1 \) and \( q2 \) quantiles to which
percentages p1 and p2 correspond. Once this is done, the distribution functions can be computed using the triangular distribution functions.

- **Uniform (a, b):** the density function of this distribution is given by:

  \[ f(x) = \frac{1}{b-a}, \text{ with } b > a \text{ and } x \in [a, b] \]

  \[ E(X) = \frac{(a+b)}{2} \text{ and } V(X) = \frac{(b-a)^2}{12} \]

  The uniform (0,1) distribution is much used for simulations. As the cumulative distribution function of all the distributions is between 0 and 1, a sample taken in a Uniform (0,1) distribution is used to obtain random samples in all the distributions for which the inverse can be calculated.

- **Uniform discrete (a, b):** the density function of this distribution is given by:

  \[ f(x) = \frac{1}{b-a+1}, \text{ with } b > a, (a,b) \in N, x \in N, x \in [a, b] \]

  We have \( E(X) = \frac{(a+b)}{2} \) and \( V(X) = \frac{[(b-a+1)^2 -1]}{12} \)

  The uniform discrete distribution corresponds to the case where the uniform distribution is restricted to integers.

- **Weibull (\( \beta \)):** the density function of this distribution is given by:

  \[ f(x) = \beta x^{\beta-1} \exp(-x^\beta), \text{ with } x > 0 \text{ and } \beta > 0 \]

  We have \( E(X) = \Gamma \left( \frac{1}{\beta} + 1 \right) \) and \( V(X) = \Gamma \left( \frac{2}{\beta} + 1 \right) - \Gamma^2 \left( \frac{1}{\beta} + 1 \right) \)

  \( \beta \) is the shape parameter for the Weibull distribution.

- **Weibull (\( \beta, \gamma \)):** the density function of this distribution is given by:

  \[ f(x) = \frac{\beta}{\gamma} \left( \frac{x}{\gamma} \right)^{\beta-1} e^{-\left( \frac{x}{\gamma} \right)^\beta}, \text{ with } x > 0, \text{ and } \beta, \gamma > 0 \]

  We have \( E(X) = \gamma \Gamma \left( \frac{1}{\beta} + 1 \right) \) and \( V(X) = \gamma^2 \left[ \Gamma \left( \frac{2}{\beta} + 1 \right) - \Gamma^2 \left( \frac{1}{\beta} + 1 \right) \right] \)

  \( \beta \) is the shape parameter of the distribution and \( \gamma \) the scale parameter. When \( \beta=1 \), the Weibull distribution is an exponential distribution with parameter \( 1/\gamma \).

- **Weibull (\( \beta, \gamma, \mu \)):** the density function of this distribution is given by:
\[ f(x) = \frac{\beta}{\gamma} \left( \frac{x - \mu}{\gamma} \right)^{\beta-1} e^{-\left( \frac{x - \mu}{\gamma} \right)^{\beta}}, \text{ with } x > \mu, \text{ and } \beta, \gamma > 0 \]

We have \( E(X) = \mu + \gamma \Gamma\left(\frac{1}{\beta} + 1\right) \) and \( V(X) = \gamma^2 \left[ \Gamma\left(\frac{2}{\beta} + 1\right) - \Gamma^2\left(\frac{1}{\beta} + 1\right) \right] \)

The Weibull distribution, named after the Swede Ernst Hjalmar Waloddi Weibull (1887-1979), is much used in quality control and survival analysis. \( \beta \) is the shape parameter of the distribution and \( \gamma \) the scale parameter. When \( \beta=1 \) and \( \mu=0 \), the Weibull distribution is an exponential distribution with parameter \( 1/\gamma \).

**Fitting method**

XLSTAT offers two fitting methods:

Moments: this simple method uses the definition of the moments of the distribution as a function of the parameters to determine the latter. For most distributions, the use of the mean and the variance is sufficient. However, for certain distributions, the mean suffices (for example Poisson's distribution), or, if not, the asymmetry coefficient is also required (Weibull's distribution for example).

Likelihood: the parameters of the distribution are estimated by maximizing the likelihood of the sample. This method, more complex, has the advantage of rigor for all distributions and enables approximate standard deviations to be obtained for parameter estimators. The maximum likelihood method is offered for the negative binomial type II distribution, Fisher-Tippett distribution, GEV distribution and Weibull distribution.

For certain distributions, the moments method gives exactly the same result as the maximum likelihood method. This is particularly true for the normal distribution.

**Goodness of fit test**

Once the parameters for the chosen distribution have been estimated, the hypothesis must be tested in order to check if the phenomenon observed through the sample follows the distribution in question. XLSTAT offers two goodness of fit tests.

The **Chi-square goodness of fit test** is a parametric test using the distance (as regards Chi-square) between the histogram of the theoretical distribution (determined by the estimated parameters) and the histogram of the empirical distribution of the sample. The histograms are calculated using \( k \) intervals chosen by the user. It is shown that the statistic calculated asymptotically follows a Chi-square distribution with \( (n-k) \) degrees of freedom where \( n \) is the
number of observations in the sample. This test is better for discrete distributions and it is recommended to check that the expected frequency in each class is not less than 5.

It may happen that the Chi-square test leads to a bad fit of the distribution to the data with one class contributing much more to the Chi-square than the others. In this case, the union of the class in question with a neighbouring class is used to check if the conclusion is due only to the class in question or it is actually the fit which is incorrect.

The **Kolmogorov-Smirnov goodness of fit test** is an exact non-parametric test based on the maximum distance between a theoretical distribution function (entirely determined by the known values of its parameters) and the empirical distribution function of the sample. This test can only be used for continuous distributions.

When a parameter estimation precedes the goodness of fit test, the Kolmogorov-Smirnov test is not correct as the parameters are estimated by trying to bring the theoretical distribution as close as possible to the data. If it confirms the goodness of fit hypothesis, the Kolmogorov-Smirnov test risks being optimistic.

For the case where the distribution used is the normal distribution, Lilliefors and Stephens (see normality tests) have put forward a modified Kolmogorov-Smirnov test which allows parameters to be estimated on the sample tested.

### Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- ****: Click this button to reload the default options.
- ****: Click this button to delete the data selections.
- ****: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If
the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab:**

**Data:** Select the data for which the goodness of fit test is to be calculated. You can select several columns (columns mode) or rows (rows mode) if you want to carry out tests on several samples at the same time.

**Distribution:** Choose the probability distribution to be used for the fit and/or goodness of fit tests. See the description section for more information on the distributions offered. The **automatic** option allows to let XLSTAT identify the best fitting distribution (determined using a Kolmogorov-Smirnov test).

**Parameters:** You can choose to **enter** the parameters for the distribution, or **estimate** them. If you choose to enter the parameters, you must enter their values.

**Range:** Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Activate this option to display the results in a new worksheet of the active workbook.

**Workbook:** Activate this option to display the results in a new workbook.

**Sample labels:** Activate this option if the sample labels are on the first row (columns mode) or in the first column (rows mode) of the selected data.

**Weights:** Activate this option if the observations are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the “Variable labels” option is activated.

- **Standardize weights:** If you activate this option, the weights are standardized such that their sum equals the number of observations.

**Options tab:**

**Tests:** Choose the type of goodness of fit test (see the description section for more details on the tests).

- **Kolmogorov-Smirnov:** Activate this option to perform a Kolmogorov-Smirnov test.
- **Chi-square:** Activate this option to perform the Chi-square test.
• **Significance level (%)**: Enter the significance level for the above tests.

**Estimation method**: Choose the method of estimating the parameters of the chosen distribution (see the description section for more details on estimation methods).

• **Moments**: Activate this option to use the moments method.

• **Maximum likelihood**: Activate this option to use the maximum likelihood method. You can then change the *convergence* limit value which when reached means the algorithm is considered to have converged. Default value: 0.00001.

**Intervals**: For a Chi-square distribution, or if you want to compare the density of the distribution chosen with the sample histogram, you must choose one of the following options:

• **Number**: Choose this option to enter the number of intervals to create.

• **Width**: Choose this option to define a fixed width for the intervals.

• **User defined**: Select a column containing in increasing order the lower bound of the first interval, and the upper bound of all the intervals.

• **Minimum**: Activate this option to enter the value of the lower value of the first interval. This value must be lower or equal to the minimum of the series.

**Missing data** tab:

**Remove observations**: Activate this option to remove the observations with missing data.

**Estimate missing data**: Activate this option to estimate missing data before starting the computations.

• **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

• **Nearest neighbour**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.

**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the samples selected.
Charts tab:

**Histograms**: Activate this option to display the histograms of the samples. For a theoretical distribution, the density function is displayed.

- **Bars**: Choose this option to display the histograms with a bar for each interval.
- **Continuous lines**: Choose this option to display the histograms with a continuous line.

**Cumulative histograms**: Activate this option to display the cumulated histograms of the samples. For a theoretical distribution, the distribution function is displayed.

Results

**Summary statistics**: This table displays for the selected samples, the number of observations, the number of missing values, the number of non-missing values, the mean and the standard deviation.

**Estimated parameters**: This table displays the parameters for the distribution.

**Statistics estimated on the input data and computed using the estimated parameters of the distribution**: This table is used to compare the mean, variance, skewness and kurtosis coefficients calculated from the sample with those calculated from the values of the distribution parameters.

**Kolmogorov-Smirnov test**: The results of the Kolmogorov-Smirnov test are displayed if the corresponding option has been activated.

**Chi-square test**: The results of the Chi-square test are displayed if the corresponding option has been activated.

**Comparison between the observed and theoretical frequencies**: This table is displayed if a Chi-square test was requested.

**Descriptive statistics for the intervals**: This table is displayed if histograms have been requested. It shows for each interval the frequencies, the relative frequencies, together with the densities for the samples and distribution chosen.

Example

A tutorial on distribution fitting is available on the Addinsoft website:

http://www.xlstat.com/demo-dfit.htm
References


Linear regression

Use this tool to create a simple or multiple linear regression model for explanation or prediction.

Description

Linear regression is without doubt the most frequently used statistical method. A distinction is usually made between simple regression (with only one explanatory variable) and multiple regression (several explanatory variables) although the overall concept and calculation methods are identical.

The principle of linear regression is to model a quantitative dependent variable $Y$ though a linear combination of $p$ quantitative explanatory variables, $X_1, X_2, \ldots, X_p$. The determinist model (not taking randomness into account) is written for observation $i$ as follows:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i$$

where $y_i$ is the value observed for the dependent variable for observation $i$, $x_{ij}$ is the value taken by variable $j$ for observation $i$, and $\varepsilon_i$ is the error of the model.

The statistical framework and the hypotheses which accompany it are not required for fitting this model. Furthermore, minimization using the least squares method (the sum of squared errors $\varepsilon_i^2$ is minimized) provides an exact analytical solution. However, to be able to test the hypothesis and measure the explanatory power of the various explanatory variables in the model, a statistical framework is necessary.

The linear regression hypotheses are as follows: the errors $\varepsilon_i$ follow the same normal distribution $N(0, \sigma)$ and are independent.

The way the model with this hypothesis added is written means that, within the framework of the linear regression model, the $y_i$s are the expression of random variables with mean $\mu_i$ and variance $\sigma^2$, where

$$\mu_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

To use the various tests proposed in the results of linear regression, it is recommended to check retrospectively that the underlying hypotheses have been correctly verified. The normality of the residues can be checked by analyzing certain charts or by using a normality test. The independence of the residues can be checked by analyzing certain charts or by using the Durbin-Watson test.
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload**: Click this button to reload the default options.
- **Delete**: Click this button to delete the data selections.
- **Data Handling**: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab:**

**Y / Dependent variables:**

**Quantitative**: Select the response variable(s) you want to model. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

**X / Explanatory variables:**

**Quantitative**: Select the quantitative explanatory variables in the Excel worksheet. The data selected must be of type numeric. If the variable header has been selected, check that the "Variable labels" option has been activated.

**Qualitative**: Activate this option to perform an ANCOVA analysis. Then select the qualitative explanatory variables (the factors) in the Excel worksheet. The selected data may be of any type.
type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Variable labels" option has been activated.

- **Range:** Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

- **Sheet:** Activate this option to display the results in a new worksheet of the active workbook.

- **Workbook:** Activate this option to display the results in a new workbook.

- **Variable labels:** Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

- **Observation labels:** Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 ...).

- **Observation weights:** Activate this option if the observations are weighted. If you do not activate this option, the weights will all be taken as 1. Weights must be greater than or equal to 0. A weight of 2 is equivalent to repeating the same observation twice. If a column header has been selected, check that the "Variable labels" option has been activated.

- **Regression weights:** Activate this option if you want to carry out a weighted least squares regression. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

- **Options** tab:

  - **Fixed constant:** Activate this option to fix the constant of the regression model to a value you then enter (0 by default).

  - **Tolerance:** Activate this option to prevent the OLS regression calculation algorithm taking into account variables which might be either constant or too correlated with other variables already used in the model (0.0001 by default).

  - **Interactions / Level:** Activate this option to include interactions in the model then enter the maximum interaction level (value between 1 and 4).
**Confidence interval (%)**: Enter the percentage range of the confidence interval to use for the various tests and for calculating the confidence intervals around the parameters and predictions. Default value: 95.

**Model selection**: Activate this option if you want to use one of the four selection methods provided:

- **Best model**: This method lets you choose the best model from amongst all the models which can handle a number of variables varying from "Min variables" to "Max Variables". Furthermore, the user can choose several "criteria" to determine the best model.
  - **Criterion**: Choose the criterion from the following list: Adjusted R², Mean Square of Errors (MSE), Mallows Cp, Akaike's AIC, Schwarz's SBC, Amemiya's PC.
  - **Min variables**: Enter the minimum number of variables to be used in the model.
  - **Max variables**: Enter the maximum number of variables to be used in the model.

Note: this method can cause long calculation times as the total number of models explored is the sum of the Cn,ks for k varying from "Min variables" to "Max variables", where Cn,k is equal to n!/(k!(n-k)!). It is there recommended that the value of "Max variables" be increased gradually.

- **Stepwise**: The selection process starts by adding the variable with the largest contribution to the model (the criterion used is Student's t statistic). If a second variable is such that the probability associated with its t is less than the "Probability for entry", it is added to the model. The same for a third variable. After the third variable is added, the impact of removing each variable present in the model after it has been added is evaluated (still using the t statistic). If the probability is greater than the "Probability of removal", the variable is removed. The procedure continues until no more variables can be added or removed.

- **Forward**: The procedure is the same as for stepwise selection except that variables are only added and never removed.

- **Backward**: The procedure starts by simultaneously adding all variables. The variables are then removed from the model following the procedure used for stepwise selection.

**Validation tab**:

**Validation**: Activate this option if you want to use a sub-sample of the data to validate the model.
**Validation set:** Choose one of the following options to define how to obtain the observations used for the validation:

- **Random:** The observations are randomly selected. The “Number of observations” N must then be specified.

- **N last rows:** The N last observations are selected for the validation. The “Number of observations” N must then be specified.

- **N first rows:** The N first observations are selected for the validation. The “Number of observations” N must then be specified.

- **Group variable:** If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to use for the validation.

**Prediction** tab:

**Prediction:** Activate this option if you want to select data to use them in prediction mode. If activate this option, you need to make sure that the prediction dataset is structured as the estimation dataset: same variables with the same order in the selections. On the other hand, variable labels must not be selected: the first row of the selections listed below must correspond to data.

**Quantitative:** Activate this option to select the quantitative explanatory variables. The first row must not include variable labels.

**Qualitative:** Activate this option to select the qualitative explanatory variables. The first row must not include variable labels.

**Observations labels:** activate this option if observations labels are available. Then select the corresponding data. If this option is not activated, the observations labels are automatically generated by XLSTAT (PredObs1, PredObs2 …).

**Missing data** tab:

**Remove observations:** Activate this option to remove the observations with missing data.

- **Check for each Y separately:** Choose this option to remove the observations with missing data in the select Y (dependent) variables, only if the Y of interest has a missing data.

- **Across all Ys:** Choose this option to remove the observations with missing data in the Y (dependent) variables, even if the Y of interest has no missing data.

**Estimate missing data:** Activate this option to estimate missing data before starting the computations.
• **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

• **Nearest neighbour**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.

**Outputs** tab:

**General** sub-tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the variables selected.

**Correlations**: Activate this option to display the correlation matrix for quantitative variables (dependent or explanatory).

**Multicolinearity statistics**: Activate this option to display the multicollinearity statistics for all explanatory variables.

**Analysis of variance**: Activate this option to display the analysis of variance table.

**Type I/III SS**: Activate this option to display the Type I and Type III sum of squares tables.

**Press**: Activate this option to calculate and display Press' coefficient.

**Standardized coefficients**: Activate this option if you want the standardized coefficients (beta coefficients) for the model to be displayed.

**Predictions and residuals**: Activate this option to display the predictions and residuals for all the observations.

  • **X**: Activate this option to display the explanatory variables in the predictions and residuals table.

  • **Adjusted predictions**: Activate this option to calculate and display adjusted predictions in the table of predictions and residuals.

  • **Cook's D**: Activate this option to calculate and display Cook's distances in the table of predictions and residuals.

**Contrasts** sub-tab:

**Compute contrasts**: Activate this option to compute contrasts, then select the contrasts table, where there must be one column per contrast and one row for each coefficient of the model.
Charts tab:

Regression charts: Activate this option to display regression chart:

- **Standardized coefficients**: Activate this option to display the standardized parameters for the model with their confidence interval on a chart.

- **Predictions and residuals**: Activate this option to display the following charts.

  1. Line of regression: This chart is only displayed if there is only one explanatory variable and this variable is quantitative.

  2. Explanatory variable versus standardized residuals: This chart is only displayed if there is only one explanatory variable and this variable is quantitative.

  3. Dependent variable versus standardized residuals.

  4. Predictions for the dependent variable versus the dependent variable.

  5. Bar chart of standardized residuals.

    - **Confidence intervals**: Activate this option to have confidence intervals displayed on charts (1) and (4).

Results

**Summary statistics**: The tables of descriptive statistics show the simple statistics for all the variables selected. The number of observations, missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the dependent variables (in blue) and the quantitative explanatory variables. For qualitative explanatory variables the names of the various categories are displayed together with their respective frequencies.

**Correlation matrix**: This table is displayed to give you a view of the correlations between the various variables selected.

**Summary of the variables selection**: Where a selection method has been chosen, XLSTAT displays the selection summary. For a stepwise selection, the statistics corresponding to the different steps are displayed. Where the best model for a number of variables varying from p to q has been selected, the best model for each number or variables is displayed with the corresponding statistics and the best model for the criterion chosen is displayed in bold.

**Goodness of fit statistics**: The statistics relating to the fitting of the regression model are shown in this table:

- **Observations**: The number of observations used in the calculations. In the formulas shown below, \( n \) is the number of observations.
• **Sum of weights**: The sum of the weights of the observations used in the calculations. In the formulas shown below, $W$ is the sum of the weights.

• **DF**: The number of degrees of freedom for the chosen model (corresponding to the error part).

• **$R^2$**: The determination coefficient for the model. This coefficient, whose value is between 0 and 1, is only displayed if the constant of the model has not been fixed by the user. Its value is defined by:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} w_i (y_i - \bar{y})^2}, \quad \text{where } \bar{y} = \frac{1}{n} \sum_{i=1}^{n} w_i y_i.$$  

The $R^2$ is interpreted as the proportion of the variability of the dependent variable explained by the model. The nearer $R^2$ is to 1, the better is the model. The problem with the $R^2$ is that it does not take into account the number of variables used to fit the model.

• **Adjusted $R^2$**: The adjusted determination coefficient for the model. The adjusted $R^2$ can be negative if the $R^2$ is near to zero. This coefficient is only calculated if the constant of the model has not been fixed by the user. Its value is defined by:

$$\hat{R}^2 = 1 - \left(1 - R^2\right) \frac{W - 1}{W - p - 1}$$

The adjusted $R^2$ is a correction to the $R^2$ which takes into account the number of variables used in the model.

• **MSE**: The mean squared error (MSE) is defined by:

$$MSE = \frac{1}{W - p} \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2$$

• **RMSE**: The root mean square of the errors (RMSE) is the square root of the MSE.

• **MAPE**: The Mean Absolute Percentage Error is calculated as follows:

$$MAPE = \frac{100}{W} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

• **DW**: The Durbin-Watson statistic is defined by:

$$DW = \frac{\sum_{i=2}^{n} \left( (y_i - \hat{y}_i) - (y_{i-1} - \hat{y}_{i-1}) \right)^2}{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}$$
This coefficient is the order 1 autocorrelation coefficient and is used to check that the residuals of the model are not autocorrelated, given that the independence of the residuals is one of the basic hypotheses of linear regression. The user can refer to a table of Durbin-Watson statistics to check if the independence hypothesis for the residuals is acceptable.

- **Cp**: Mallows Cp coefficient is defined by:
  \[
  Cp = \frac{SSE}{\hat{\sigma}} + 2 p^* - W
  \]
  where SSE is the sum of the squares of the errors for the model with p explanatory variables and \(\hat{\sigma}\) is the estimator of the variance of the residuals for the model comprising all the explanatory variables. The nearer the Cp coefficient is to p*, the less the model is biased.

- **AIC**: Akaike’s Information Criterion is defined by:
  \[
  AIC = W \ln \left( \frac{SSE}{W} \right) + 2 p^*
  \]
  This criterion, proposed by Akaike (1973) is derived from the information theory and uses Kullback and Leibler’s measurement (1951). It is a model selection criterion which penalizes models for which adding new explanatory variables does not supply sufficient information to the model, the information being measured through the MSE. The aim is to minimize the AIC criterion.

- **SBC**: Schwarz’s Bayesian Criterion is defined by:
  \[
  SBC = W \ln \left( \frac{SSE}{W} \right) + \ln(W) p^*
  \]
  This criterion, proposed by Schwarz (1978) is similar to the AIC and, like this, the aim is to minimize it.

- **PC**: Amemiya’s Prediction Criterion is defined by:
  \[
  PC = \frac{(1 - R^2)(W + p^*)}{W - p^*}
  \]
  This criterion, proposed by Amemiya (1980) is used, like the adjusted R² to take account of the parsimony of the model.

- **Press**: Press’ statistic is only displayed if the corresponding option has been activated in the dialog box. It is defined by:
  \[
  Press = \sum_{i=1}^{n} w_i \left( y_i - \hat{y}_{(i)} \right)^2
  \]
where \( \hat{y}_{i(-i)} \) is the prediction for observation \( i \) when the latter is not used for estimating parameters. We then get:

\[
\text{Press RMSE} = \sqrt{\frac{\text{Press}}{W - p}}
\]

Press's RMSE can then be compared to the RMSE. A large difference between the two shows that the model is sensitive to the presence or absence of certain observations in the model.

- **Q²**: This statistic also known as the cross-validated R². It is only displayed if the Press option has been activated in the dialog box. It is defined by:

\[
Q^2 = 1 - \frac{\text{Press}}{\sum_{i=1}^{n} (y_i - \bar{y})^2}
\]

This gives the proportion of the total variance that is explained by the explanatory variables when the predictions are computed when the corresponding observation is not in the model. A large difference between the Q² and the R² shows that the model is sensitive to the presence or absence of certain observations in the model.

The **analysis of variance table** is used to evaluate the explanatory power of the explanatory variables. Where the constant of the model is not set to a given value, the explanatory power is evaluated by comparing the fit (as regards least squares) of the final model with the fit of the rudimentary model including only a constant equal to the mean of the dependent variable. Where the constant of the model is set, the comparison is made with respect to the model for which the dependent variable is equal to the constant which has been set.

If the Type I/III SS (SS: Sum of Squares) is activated, the corresponding tables are displayed.

The table of **Type I SS** values is used to visualize the influence that progressively adding explanatory variables has on the fitting of the model, as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. The sums of squares in the Type I table always add up to the model SS. Note: the order in which the variables are selected in the model influences the values obtained.

The table of **Type III SS** values is used to visualize the influence that removing an explanatory variable has on the fitting of the model, all other variables being retained, expect those were the effect is present (interactions), as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. Note: unlike Type I SS, the order in which the variables are
selected in the model has no influence on the values obtained. Type II and Type III are identical if there are no interactions or if the design is balanced.

The parameters of the model table displays the estimate of the parameters, the corresponding standard error, the Student’s t, the corresponding probability, as well as the confidence interval.

The equation of the model is then displayed to make it easier to read or re-use the model.

The table of standardized coefficients (also called beta coefficients) are used to compare the relative weights of the variables. The higher the absolute value of a coefficient, the more important the weight of the corresponding variable. When the confidence interval around standardized coefficients has value 0 (this can be easily seen on the chart of normalized coefficients), the weight of a variable in the model is not significant.

The predictions and residuals table shows, for each observation, its weight, the value of the qualitative explanatory variable, if there is only one, the observed value of the dependent variable, the model's prediction, the residuals, the studentized residuals, the confidence intervals together with the fitted prediction and Cook’s D if the corresponding options have been activated in the dialog box. Two types of confidence interval are displayed: a confidence interval around the mean (corresponding to the case where the prediction would be made for an infinite number of observations with a set of given values for the explanatory variables) and an interval around the isolated prediction (corresponding to the case of an isolated prediction for the values given for the explanatory variables). The second interval is always greater than the first, the random values being larger. If the validation data have been selected, they are displayed at the end of the table.

The charts which follow show the results mentioned above. If there is only one explanatory variable in the model, the first chart displayed shows the observed values, the regression line and both types of confidence interval around the predictions. The second chart shows the normalized residuals as a function of the explanatory variable. In principle, the residuals should be distributed randomly around the X-axis. If there is a trend or a shape, this shows a problem with the model.

The three charts displayed next show respectively the evolution of the standardized residuals as a function of the dependent variable, the distance between the predictions and the observations (for an ideal model, the points would all be on the bisector), and the standardized residuals on a bar chart. The last chart quickly shows if an abnormal number of values are outside the interval ]-2, 2[ given that the latter, assuming that the sample is normally distributed, should contain about 95% of the data.

If you have selected the data to be used for calculating predictions on new observations, the corresponding table is displayed next.
Example

A tutorial on simple linear regression is available on the Addinsoft website:
http://www.xlstat.com/demo-reg.htm

A tutorial on multiple linear regression is available on the Addinsoft website:
http://www.xlstat.com/demo-reg2.htm

References


ANOVA

Use this model to carry out ANOVA (ANalysis Of VAriance) of one or more balanced or unbalanced factors. The advanced options enable you to choose the constraints on the model and to take account of interactions between the factors. Multiple comparison tests can be calculated.

Description

Analysis of Variance (ANOVA) uses the same conceptual framework as linear regression. The main difference comes from the nature of the explanatory variables: instead of quantitative, here they are qualitative. In ANOVA, explanatory variables are often called factors.

If $p$ is the number of factors, the ANOVA model is written as follows:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_{k(i),j} + \epsilon_i \quad (1)$$

where $y_i$ is the value observed for the dependent variable for observation $i$, $k(i,j)$ is the index of the category of factor $j$ for observation $i$, and $\epsilon_i$ is the error of the model.

The hypotheses used in ANOVA are identical to those used in linear regression: the errors $\epsilon_i$ follow the same normal distribution $N(0,\sigma)$ and are independent.

The way the model with this hypothesis added is written means that, within the framework of the linear regression model, the $y_i$s are the expression of random variables with mean $\mu_i$ and variance $\sigma^2$, where

$$\mu_i = \beta_0 + \sum_{j=1}^{p} \beta_{k(i),j}$$

To use the various tests proposed in the results of linear regression, it is recommended to check retrospectively that the underlying hypotheses have been correctly verified. The normality of the residues can be checked by analyzing certain charts or by using a normality test. The independence of the residues can be checked by analyzing certain charts or by using the Durbin Watson test.

Interactions

By interaction is meant an artificial factor (not measured) which reflects the interaction between at least two measured factors. For example, if we carry out treatment on a plant, and tests are carried out under two different light intensities, we will be able to include in the model an
interaction factor treatment*light which will be used to identify a possible interaction between
the two factors. If there is an interaction between the two factors, we will observe a significantly
larger effect on the plants when the light is strong and the treatment is of type 2 while the effect
is average for weak light, treatment 2 and strong light, treatment 1 combinations.

To make a parallel with linear regression, the interactions are equivalent to the products
between the continuous explanatory values although here obtaining interactions requires
nothing more than simple multiplication between two variables. However, the notation used to
represent the interaction between factor A and factor B is A*B.

The interactions to be used in the model can be easily defined in XLSTAT.

Balanced and unbalanced ANOVA

We talk of balanced ANOVA when the numbers of categories are equal for all factors. When
the numbers of all categories for one of the factors are not equal, then the ANOVA is said to be
unbalanced. XLSTAT can handle both cases.

Nested effects

When constraints prevent us from crossing every level of one factor with every level of the
other factor, nested factors can be used. We say we have a nested effect when fewer than all
levels of one factor occur within each level of the other factor. An example of this might be if
we want to study the effects of different machines and different operators on some output
characteristic, but we can't have the operators change the machines they run. In this case,
each operator is not crossed with each machine but rather only runs one machine. XLSTAT
has an automatic device to find nested factors and one nested factor can be included in the
model.

Random effects

Random factors can be included in an ANOVA. When some factors are supposed to be
random, XLSTAT displays the expected mean squares table.

Constraints

During the calculations, each factor is broken down into a sub-matrix containing as many
columns as there are categories in the factor. Typically, this is a full disjunctive table.
Nevertheless, the breakdown poses a problem: if there are g categories, the rank of this sub-
matrix is not g but g-1. This leads to the requirement to delete one of the columns of the sub-
matrix and possibly to transform the other columns. Several strategies are available depending
on the interpretation we want to make afterwards:
1) $a_1 = 0$: the parameter for the first category is null. This choice allows us force the effect of the first category as a standard. In this case, the constant of the model is equal to the mean of the dependent variable for group 1.

2) $a_n = 0$: the parameter for the last category is null. This choice allows us force the effect of the last category as a standard. In this case, the constant of the model is equal to the mean of the dependent variable for group $g$.

3) $\sum (a_i) = 0$: the sum of the parameters is null. This choice forces the constant of the model to be equal to the mean of the dependent variable when the ANOVA is balanced.

4) $\sum (n_i a_i) = 0$: the sum of the parameters is null. This choice forces the constant of the model to be equal to the mean of the dependent variable even when the ANOVA is unbalanced.

Note: even if the choice of constraint influences the values of the parameters, it has no effect on the predicted values and on the different fitting statistics.

**Multiple Comparisons Tests**

One of the main applications of ANOVA is multiple comparisons testing whose aim is to check if the parameters for the various categories of a factor differ significantly or not. For example, in the case where four treatments are applied to plants, we want to know not only if the treatments have a significant effect, but also if the treatments have different effects.

Numerous tests have been proposed for comparing the means of categories. The majority of these tests assume that the sample is normally distributed. XLSTAT provides the main tests including:

**Tukey's HSD test:** this test is the most used (HSD: *Honestly Significant Difference*).

Fisher's LSD test: this is Student's test that tests the hypothesis that all the means for the various categories are equal (LSD: *Least Significant Difference*).

**Bonferroni's $t^*$ test:** this test is derived from Student's test and is less reliable as it takes into account the fact that several comparisons are carried out simultaneously. Consequently, the significance level of the test is modified according to the following formula:

$$\alpha' = \frac{\alpha}{g(g-1)/2}$$

where $g$ is the number of categories of the factor whose categories are being compared.

**Dunn-Sidak's test:** this test is derived from Bonferroni's test. It is more reliable in some situations.
\[ \alpha' = 1 - \left(1 - \alpha\right)^{2/(g(x-1))} \]

The following tests are more complex as they are based on iterative procedures where the results depend on the number of combinations remaining to be tested for each category.

**Newman-Keuls’s test (SNK):** this test is derived from Student's test (SNK: Student Newman-Keuls), and is very often used although not very reliable.

**Duncan’s test:** this test is little used.

**REGWQ test:** this test is among the most reliable in a majority of situations (REGW: RyanEinot-Gabriel-Welsch).

**Benjamini-Hochberg:** Use this option to control the False Discovery Rate (FDR). This p-value penalization procedure is poorly conservative.

The **Games-Howell (GH) test** can be used in one-way ANOVAs when the variances lack of homogeneity. While it can be used with unequal sample sizes, it is recommended to use it when the smallest sample has 5 elements or more, otherwise it is too liberal. The **Tamhane’s T2 test** is more conservative, but not as powerful as the GH test.

All the above tests enable comparisons to be made between all pairs of categories and belong to the MCA test family (**Multiple Comparisons of All**, or **All-Pairwise Comparisons**).

Other tests make comparisons between all categories and a control category. These tests are called MCB tests (**Multiple Comparisons with the Best**, **Comparisons with a control**). XLSTAT offers the Dunnett test which is the most used. There are three Dunnett tests:

**Two-tailed test:** the null hypothesis assumes equality between the category tested and the control category. The alternative hypothesis assumes the means of the two categories differ.

**Left one-tailed test:** the null hypothesis assumes equality between the category tested and the control category. The alternative hypothesis assumes that the mean of the control category is less than the mean of the category tested.

**Right one-tailed test:** the null hypothesis assumes equality between the category tested and the control category. The alternative hypothesis assumes that the mean of the control category is greater than the mean of the category tested.

**Robust tests of mean comparison for a one-way ANOVA**

In an analysis of variance, it may happen that the variances can not be assumed to be equal. In this case, the F test of the ANOVA is not robust enough to be used. XLSTAT offers two tests based on the F distribution but more robust than the classical F test.

These tests are:
• Welch Test or Welch ANOVA (Welch, 1951). The Welch test adjusts the denominator of the F ratio so it has the same expectation as the numerator when the null hypothesis is true, despite the heterogeneity of within-group variance. The p-value can be interpreted in the same manner as in the analysis of variance table.

• The Brown-Forsythe test or Brown-Forsythe F-ratio (1974). This test uses a different denominator for the formula of F in the ANOVA. Instead of dividing by the mean square of the error, the mean square is adjusted using the observed variances of each group. The p-value can be interpreted in the same manner as in the analysis of variance table.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

Click this button to start the computations.

Click this button to close the dialog box without doing any computation.

Click this button to display the help.

Click this button to reload the default options.

Click this button to delete the data selections.

Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Y / Dependent variables:

Quantitative: Select the response variable(s) you want to model. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.
**X / Explanatory variables:**

**Quantitative:** Activate this option to perform an ANCOVA analysis. Then select the quantitative explanatory variables in the Excel worksheet. The data selected must be of type numeric. If the variable header has been selected, check that the "Variable labels" option has been activated.

**Qualitative:** Select the qualitative explanatory variables (the factors) in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Variable labels" option has been activated.

**Range:** Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Activate this option to display the results in a new worksheet of the active workbook.

**Workbook:** Activate this option to display the results in a new workbook.

**Variable labels:** Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

**Observation labels:** Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 ...).

**Observation weights:** Activate this option if the observations are weighted. If you do not activate this option, the weights will all be taken as 1. Weights must be greater than or equal to 0. A weight of 2 is equivalent to repeating the same observation twice. If a column header has been selected, check that the "Variable labels" option has been activated.

**Regression weights:** Activate this option if you want to carry out a weighted least squares regression. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

**Options** tab:

**Fixed constant:** Activate this option to fix the constant of the regression model to a value you then enter (0 by default).
**Tolerance**: Activate this option to prevent the OLS regression calculation algorithm taking into account variables which might be either constant or too correlated with other variables already used in the model (0.0001 by default).

**Interactions / Level**: Activate this option to include interactions in the model then enter the maximum interaction level (value between 1 and 4).

**Nested effects**: Activate this option to include one nested effect in the model.

**Confidence interval (%)**: Enter the percentage range of the confidence interval to use for the various tests and for calculating the confidence intervals around the parameters and predictions. Default value: 95.

**Constraints**: Details on the various options are available in the description section.

- **a1 = 0**: Choose this option so that the parameter of the first category of each factor is set to 0.
- **an = 0**: Choose this option so that the parameter of the last category of each factor is set to 0.
- **Sum (ai) = 0**: for each factor, the sum of the parameters associated with the various categories is set to 0.
- **Sum (ni.ai) = 0**: for each factor, the sum of the parameters associated with the various categories weighted by their frequencies is set to 0.

**Random factors**: Activate this option if you want to include random factors in your model. The expected mean squares table will be automatically displayed.

**Model selection**: Activate this option if you want to use one of the four selection methods provided:

- **Best model**: This method lets you choose the best model from amongst all the models which can handle a number of variables varying from "Min variables" to "Max Variables". Furthermore, the user can choose several "criteria" to determine the best model.
  - **Criterion**: Choose the criterion from the following list: Adjusted R², Mean Square of Errors (MSE), Mallows Cp, Akaike's AIC, Schwarz's SBC, Amemiya's PC.
  - **Min variables**: Enter the minimum number of variables to be used in the model.
Max variables: Enter the maximum number of variables to be used in the model.

Note: this method can cause long calculation times as the total number of models explored is the sum of the Cn,k's for k varying from "Min variables" to "Max variables", where Cn,k is equal to n!/(n-k)!k!. It is recommended that the value of "Max variables" be increased gradually.

- **Stepwise**: The selection process starts by adding the variable with the largest contribution to the model (the criterion used is Student's t statistic). If a second variable is such that the probability associated with its t is less than the "Probability for entry", it is added to the model. The same for a third variable. After the third variable is added, the impact of removing each variable present in the model after it has been added is evaluated (still using the t statistic). If the probability is greater than the "Probability of removal", the variable is removed. The procedure continues until no more variables can be added or removed.

- **Forward**: The procedure is the same as for stepwise selection except that variables are only added and never removed.

- **Backward**: The procedure starts by simultaneously adding all variables. The variables are then removed from the model following the procedure used for stepwise selection.

**Validation** tab:

**Validation**: Activate this option if you want to use a sub-sample of the data to validate the model.

**Validation set**: Choose one of the following options to define how to obtain the observations used for the validation:

- **Random**: The observations are randomly selected. The “Number of observations” N must then be specified.

- **N last rows**: The N last observations are selected for the validation. The “Number of observations” N must then be specified.

- **N first rows**: The N first observations are selected for the validation. The “Number of observations” N must then be specified.

- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to use for the validation.

**Prediction** tab:
**Prediction**: Activate this option if you want to select data to use them in prediction mode. If activate this option, you need to make sure that the prediction dataset is structured as the estimation dataset: same variables with the same order in the selections. On the other hand, variable labels must not be selected: the first row of the selections listed below must correspond to data.

**Quantitative**: Activate this option to select the quantitative explanatory variables. The first row must not include variable labels.

**Qualitative**: Activate this option to select the qualitative explanatory variables. The first row must not include variable labels.

**Observations labels**: activate this option if observations labels are available. Then select the corresponding data. If this option is not activated, the observations labels are automatically generated by XLSTAT (PredObs1, PredObs2 …).

**Missing data** tab:

**Remove observations**: Activate this option to remove the observations with missing data.

- **Check for each Y separately**: Choose this option to remove the observations with missing data in the select Y (dependent) variables, only if the Y of interest has a missing data.

- **Across all Ys**: Choose this option to remove the observations with missing data in the Y (dependent) variables, even if the Y of interest has no missing data.

**Estimate missing data**: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

- **Nearest neighbour**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.

**Outputs** tab:

**General** sub-tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the variables selected.

**Correlations**: Activate this option to display the correlation matrix for quantitative variables (dependent or explanatory).
**Multicolinearity statistics**: Activate this option to display the multicollinearity statistics for all explanatory variables.

**Analysis of variance**: Activate this option to display the analysis of variance table.

**Type I/II/III SS**: Activate this option to display the Type I, Type II, and Type III sum of squares tables.

**Press**: Activate this option to calculate and display Press’ coefficient.

**Standardized coefficients**: Activate this option if you want the standardized coefficients (beta coefficients) for the model to be displayed.

**Predictions and residuals**: Activate this option to display the predictions and residuals for all the observations.

- **Adjusted predictions**: Activate this option to calculate and display adjusted predictions in the table of predictions and residuals.
- **Cook's D**: Activate this option to calculate and display Cook’s distances in the table of predictions and residuals.

**Welch and Brown-Forsythe tests**: Activate this option to display the Welch and Brown-Forsythe tests (see the description section of this chapter) in the case of a one-way ANOVA.

**Multiple comparisons** sub-tab:

Information on the multiple comparison tests is available in the description section.

**Apply to all factors**: Activate this option to compute the selected tests for all factors.

**Use least squares means**: Activate this option to compare the means using their least squares estimators (obtained from the parameters of the model). If this option is not activated, the means are computed using their estimation based on the data.

**Sort up**: Activate this option to sort the compared categories in increasing order, the sort criterion being their respective means. If this option is not activated, the sort is decreasing.

**Standard errors**: Activate this option to display the standard errors with the means.

- **Confidence intervals**: Activate this option to additionally display the confidence intervals around the means.

**Pairwise comparisons**: Activate this option then choose the comparison methods.

**Comparisons with a control**: Activate this option then choose the type of Dunnett test you want to carry out.
Choose the MSE: Activate this option to select the mean squared error to be taken as reference for multiple comparisons. When using random factors, using the mean squared error of the model (classical case) is not appropriate. In that case, the user should choose a mean square error associated with another term in the model (usually an interaction term). If this option is enabled, a new dialog allowing you to select the variable to use will appear.

Protected: Activate this option to prevent the multiple comparisons results from being displayed for the factor that are not significant.

Top/Bottom boxes: Activate this option to display the Top/Bottom boxes (frequency of the top and bottom values). You can choose between Top/Bottom 2 or 3 values. The top/bottom 2 box is the frequency of the highest/lowest two values. The top/bottom 3 box is the frequency of the highest/lowest three values.

Contrasts sub-tab:

Compute contrasts: Activate this option to compute contrasts, then select the contrasts table, where there must be one column per contrast and one row for each coefficient of the model.

Charts tab:

Regression charts: Activate this option to display regression chart:

- **Standardized coefficients**: Activate this option to display the standardized parameters for the model with their confidence interval on a chart.

- **Predictions and residuals**: Activate this option to display the following charts.
  1. Line of regression: This chart is only displayed if there is only one explanatory variable and this variable is quantitative.
  2. Explanatory variable versus standardized residuals: This chart is only displayed if there is only one explanatory variable and this variable is quantitative.
  3. Dependent variable versus standardized residuals.
  4. Predictions for the dependent variable versus the dependent variable.
  5. Bar chart of standardized residuals.

- Confidence intervals: Activate this option to have confidence intervals displayed on charts (1) and (4).

Means charts: Activate this option to display the charts used to display the means of the various categories of the various factors.
• **Confidence intervals**: Activate this option to display the confidence intervals around the means on the same chart.

**Results**

**Summary statistics**: The tables of descriptive statistics show the simple statistics for all the variables selected. The number of observations, missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the dependent variables (in blue) and the quantitative explanatory variables. For qualitative explanatory variables the names of the various categories are displayed together with their respective frequencies.

**Correlation matrix**: This table is displayed to give you a view of the correlations between the various variables selected.

**Summary of the variables selection**: Where a selection method has been chosen, XLSTAT displays the selection summary. For a stepwise selection, the statistics corresponding to the different steps are displayed. Where the best model for a number of variables varying from p to q has been selected, the best model for each number or variables is displayed with the corresponding statistics and the best model for the criterion chosen is displayed in bold.

**Goodness of fit statistics**: The statistics relating to the fitting of the regression model are shown in this table:

- **Observations**: The number of observations used in the calculations. In the formulas shown below, \( n \) is the number of observations.

- **Sum of weights**: The sum of the weights of the observations used in the calculations. In the formulas shown below, \( W \) is the sum of the weights.

- **DF**: The number of degrees of freedom for the chosen model (corresponding to the error part).

- **R²**: The determination coefficient for the model. This coefficient, whose value is between 0 and 1, is only displayed if the constant of the model has not been fixed by the user. Its value is defined by:

\[
R^2 = 1 - \frac{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} w_i (y_i - \bar{y})^2}, \quad \text{where} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} w_i y_i,
\]

The \( R^2 \) is interpreted as the proportion of the variability of the dependent variable explained by the model. The nearer \( R^2 \) is to 1, the better is the model. The problem with the \( R^2 \) is that it does not take into account the number of variables used to fit the model.
- **Adjusted R²**: The adjusted determination coefficient for the model. The adjusted R² can be negative if the R² is near to zero. This coefficient is only calculated if the constant of the model has not been fixed by the user. Its value is defined by:

\[ \hat{R}^2 = 1 - \left(1 - R^2\right) \frac{W - 1}{W - p - 1} \]

The adjusted R² is a correction to the R² which takes into account the number of variables used in the model.

- **MSE**: The mean squared error (MSE) is defined by:

\[ MSE = \frac{1}{W - p*} \sum_{i=1}^{W} w_i \left(y_i - \hat{y}_i\right)^2 \]

- **RMSE**: The root mean square of the errors (RMSE) is the square root of the MSE.

- **MAPE**: The *Mean Absolute Percentage Error* is calculated as follows:

\[ MAPE = \frac{100}{W} \sum_{i=1}^{W} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \]

- **DW**: The Durbin-Watson statistic is defined by:

\[ DW = \frac{\sum_{i=2}^{W} \left[ \left(y_i - \hat{y}_i\right) - \left(y_{i-1} - \hat{y}_{i-1}\right) \right]^2}{\sum_{i=1}^{W} w_i \left(y_i - \hat{y}_i\right)^2} \]

This coefficient is the order 1 autocorrelation coefficient and is used to check that the residuals of the model are not autocorrelated, given that the independence of the residuals is one of the basic hypotheses of linear regression. The user can refer to a table of Durbin-Watson statistics to check if the independence hypothesis for the residuals is acceptable.

- **Cp**: Mallows Cp coefficient is defined by:

\[ Cp = \frac{SSE}{\hat{\sigma}^2} + 2 p* - W \]

where SSE is the sum of the squares of the errors for the model with p explanatory variables and \( \hat{\sigma}^2 \) is the estimator of the variance of the residuals for the model comprising all the explanatory variables. The nearer the Cp coefficient is to p*, the less the model is biased.

- **AIC**: Akaike’s Information Criterion is defined by:
\[ AIC = W \ln \left( \frac{SSE}{W} \right) + 2p^* \]

This criterion, proposed by Akaike (1973) is derived from the information theory and uses Kullback and Leibler's measurement (1951). It is a model selection criterion which penalizes models for which adding new explanatory variables does not supply sufficient information to the model, the information being measured through the MSE. The aim is to minimize the AIC criterion.

- **SBC**: Schwarz's Bayesian Criterion is defined by:

\[ SBC = W \ln \left( \frac{SSE}{W} \right) + \ln(W) p^* \]

This criterion, proposed by Schwarz (1978) is similar to the AIC and, like this, the aim is to minimize it.

- **PC**: Amemiya's Prediction Criterion is defined by:

\[ PC = \frac{(1 - R^2)(W + p^*)}{W - p^*} \]

This criterion, proposed by Amemiya (1980) is used, like the adjusted R², to take account of the parsimony of the model.

- **Press**: Press' statistic is only displayed if the corresponding option has been activated in the dialog box. It is defined by:

\[ \text{Press} = \sum_{i=1}^{n} w_i (y_i - \hat{y}_{i(-i)})^2 \]

where \( \hat{y}_{i(-i)} \) is the prediction for observation i when the latter is not used for estimating parameters. We then get:

\[ \text{Press RMSE} = \sqrt{\frac{\text{Press}}{W - p^*}} \]

Press's RMSE can then be compared to the RMSE. A large difference between the two shows that the model is sensitive to the presence or absence of certain observations in the model.

- **Q²**: This statistic also known as the cross-validated R². It is only displayed if the Press option has been activated in the dialog box. It is defined by:
This gives the proportion of the total variance that is explained by the explanatory variables when the predictions are computed when the corresponding observation is not in the model. A large difference between the $Q^2$ and the $R^2$ shows that the model is sensitive to the presence or absence of certain observations in the model.

The analysis of variance table is used to evaluate the explanatory power of the explanatory variables. Where the constant of the model is not set to a given value, the explanatory power is evaluated by comparing the fit (as regards least squares) of the final model with the fit of the rudimentary model including only a constant equal to the mean of the dependent variable. Where the constant of the model is set, the comparison is made with respect to the model for which the dependent variable is equal to the constant which has been set.

If the Type I/II/III SS (SS: Sum of Squares) is activated, the corresponding tables are displayed.

The table of Type I SS values is used to visualize the influence that progressively adding explanatory variables has on the fitting of the model, as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. The sums of squares in the Type I table always add up to the model SS. Note: the order in which the variables are selected in the model influences the values obtained.

The table of Type II SS values is used to visualize the influence that removing an explanatory variable has on the fitting of the model, all other variables being retained, as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. Note: unlike Type I SS, the order in which the variables are selected in the model has no influence on the values obtained.

The table of Type III SS values is used to visualize the influence that removing an explanatory variable has on the fitting of the model, all other variables being retained, except those were the effect is present (interactions), as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. Note: unlike Type I SS, the order in which the variables are selected in the model has no influence on the values obtained. Type II and Type III are identical if there are no interactions or if the design is balanced.
The **parameters of the model** table displays the estimate of the parameters, the corresponding standard error, the Student’s t, the corresponding probability, as well as the confidence interval.

The **equation of the model** is then displayed to make it easier to read or re-use the model.

The table of **standardized coefficients** (also called beta coefficients) are used to compare the relative weights of the variables. The higher the absolute value of a coefficient, the more important the weight of the corresponding variable. When the confidence interval around standardized coefficients has value 0 (this can be easily seen on the chart of standardized coefficients), the weight of a variable in the model is not significant.

The **predictions and residuals** table shows, for each observation, its weight, the value of the qualitative explanatory variable, if there is only one, the observed value of the dependent variable, the model’s prediction, the residuals, the studentized residuals, the confidence intervals together with the fitted prediction and Cook’s D if the corresponding options have been activated in the dialog box. Two types of confidence interval are displayed: a confidence interval around the mean (corresponding to the case where the prediction would be made for an infinite number of observations with a set of given values for the explanatory variables) and an interval around the isolated prediction (corresponding to the case of an isolated prediction for the values given for the explanatory variables). The second interval is always greater than the first, the random values being larger. If the validation data have been selected, they are displayed at the end of the table.

The **charts** which follow show the results mentioned above. If there is only one explanatory variable in the model, the first chart displayed shows the observed values, the regression line and both types of confidence interval around the predictions. The second chart shows the standardized residuals as a function of the explanatory variable. In principle, the residuals should be distributed randomly around the X-axis. If there is a trend or a shape, this shows a problem with the model.

The **three charts** displayed next show respectively the evolution of the standardized residuals as a function of the dependent variable, the distance between the predictions and the observations (for an ideal model, the points would all be on the bisector), and the standardized residuals on a bar chart. The last chart quickly shows if an abnormal number of values are outside the interval \([-2, 2]\) given that the latter, assuming that the sample is normally distributed, should contain about 95% of the data.

If you have selected the data to be used for calculating **predictions on new observations**, the corresponding table is displayed next.

If multiple comparison tests have been requested, the corresponding results are then displayed.
When a one-way ANOVA is applied and the corresponding option is enabled, the results of the Welch and Brown-Forsythe tests are displayed. The associated statistics, the degrees of freedom and the p-values are displayed.

If several dependent variables have been selected and if the multiple comparisons option has been activated, a table showing the means for each category of each factor and across all Ys is displayed. The cells of the table are colored using a spectrum scale from blue to red. If there are more than 10 categories, only the 5 lowest and 5 highest means are colored. A chart allows visualizing the same results.

**Example**

A tutorial on one-way ANOVA and multiple comparisons tests is available on the Addinsoft website:
http://www.xlstat.com/demo-ano.htm

A tutorial on two-way ANOVA is available on the Addinsoft website:
http://www.xlstat.com/demo-ano2.htm

**References**


ANCOVA

Use this module to model a quantitative dependent variable by using quantitative and qualitative dependent variables as part of a linear model.

Description

ANCOVA (ANalysis of COVAriance) can be seen as a mix of ANOVA and linear regression as the dependent variable is of the same type, the model is linear and the hypotheses are identical. In reality it is more correct to consider ANOVA and linear regression as special cases of ANCOVA.

If p is the number of quantitative variables, and q the number of factors (the qualitative variables including the interactions between qualitative variables), the ANCOVA model is written as follows:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \sum_{j=1}^{q} \beta_{k(i,j),j} + \epsilon_i$$

(1)

where $y_i$ is the value observed for the dependent variable for observation i, $x_{ij}$ is the value taken by quantitative variable j for observation i, $k(i,j)$ is the index of the category of factor j for observation i and $\epsilon_i$ is the error of the model.

The hypotheses used in ANOVA are identical to those used in linear regression and ANOVA: the errors $\epsilon_i$ follow the same normal distribution $N(0,\sigma)$ and are independent.

Interactions between quantitative variables and factors

One of the features of ANCOVA is to enable interactions between quantitative variables and factors to be taken into account. The main application is to test if the level of a factor (a qualitative variable) has an influence on the coefficient (often called slope in this context) of a quantitative variable. Comparison tests are used to test if the slopes corresponding to the various levels of a factor differ significantly or not. A model with one quantitative variable and a factor with interaction is written:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_{k(i,1),1} + \beta_{k(i,1),2} x_{i1} + \epsilon_i$$

(2)

This can be simplified by setting

$$\gamma_{k(i,1),1} = \beta_1 + \beta_{k(i,1),2}$$

(3)

hence we get
\[ y_i = \beta_0 + \beta_{k(i,j)} + \gamma_{k(i,j)} x_{ij} + \varepsilon_i \]  

(4)

The comparison of the \( \gamma \) parameters are used to test if the factor has an affect on the slope.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload**: Click this button to reload the default options.
- **Delete**: Click this button to delete the data selections.
- **Arrow buttons**: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

- **Y / Dependent variables**: 
  - **Quantitative**: Select the response variable(s) you want to model. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.
**Quantitative**: Select the quantitative explanatory variables in the Excel worksheet. The data selected must be of type numeric. If the variable header has been selected, check that the "Variable labels" option has been activated.

**Qualitative**: Select the qualitative explanatory variables (the factors) in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Variable labels" option has been activated.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Variable labels**: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

**Observation labels**: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

**Observation weights**: Activate this option if the observations are weighted. If you do not activate this option, the weights will all be taken as 1. Weights must be greater than or equal to 0. A weight of 2 is equivalent to repeating the same observation twice. If a column header has been selected, check that the "Variable labels" option has been activated.

**Regression weights**: Activate this option if you want to carry out a weighted least squares regression. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

**Options** tab:

**Fixed constant**: Activate this option to fix the constant of the regression model to a value you then enter (0 by default).

**Tolerance**:.
Interactions / Level: Activate this option to include interactions in the model then enter the maximum interaction level (value between 1 and 4).

Confidence interval (%): Enter the percentage range of the confidence interval to use for the various tests and for calculating the confidence intervals around the parameters and predictions. Default value: 95.

Constraints: Details on the various options are available in the description section.

\( a_1 = 0 \): Choose this option so that the parameter of the first category of each factor is set to 0.

\( a_n = 0 \): Choose this option so that the parameter of the last category of each factor is set to 0.

\( \sum (a_i) = 0 \): for each factor, the sum of the parameters associated with the various categories is set to 0.

\( \sum (n_ia_i) = 0 \): for each factor, the sum of the parameters associated with the various categories weighted by their frequencies is set to 0.

Model selection: Activate this option if you want to use one of the four selection methods provided:

- **Best model:** This method lets you choose the best model from amongst all the models which can handle a number of variables varying from "Min variables" to "Max Variables". Furthermore, the user can choose several "criteria" to determine the best model.
  - **Criterion:** Choose the criterion from the following list: Adjusted R², Mean Square of Errors (MSE), Mallows Cp, Akaike's AIC, Schwarz's SBC, Amemiya's PC.
  - **Min variables:** Enter the minimum number of variables to be used in the model.
  - **Max variables:** Enter the maximum number of variables to be used in the model.

Note: this method can cause long calculation times as the total number of models explored is the sum of the \( C_n,k \) for \( k \) varying from "Min variables" to "Max variables", where \( C_n,k \) is equal to \( n! / [(n-k)!k!] \). It is there recommended that the value of "Max variables" be increased gradually.

- **Stepwise:** The selection process starts by adding the variable with the largest contribution to the model (the criterion used is Student's t statistic). If a second variable is such that the probability associated with its t is less than the "Probability for entry", it is added to the model. The same for a third variable. After the third variable is added, the impact of removing each variable present in the model after it has been added is evaluated (still using the t statistic). If the probability is greater than the "Probability of
removal", the variable is removed. The procedure continues until no more variables can be added or removed.

- **Forward**: The procedure is the same as for stepwise selection except that variables are only added and never removed.

- **Backward**: The procedure starts by simultaneously adding all variables. The variables are then removed from the model following the procedure used for stepwise selection.

**Validation tab:**

**Validation**: Activate this option if you want to use a sub-sample of the data to validate the model.

**Validation set**: Choose one of the following options to define how to obtain the observations used for the validation:

- **Random**: The observations are randomly selected. The "Number of observations" N must then be specified.

- **N last rows**: The N last observations are selected for the validation. The "Number of observations" N must then be specified.

- **N first rows**: The N first observations are selected for the validation. The "Number of observations" N must then be specified.

- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to use for the validation.

**Prediction tab:**

**Prediction**: Activate this option if you want to select data to use them in prediction mode. If activate this option, you need to make sure that the prediction dataset is structured as the estimation dataset: same variables with the same order in the selections. On the other hand, variable labels must not be selected: the first row of the selections listed below must correspond to data.

**Quantitative**: Activate this option to select the quantitative explanatory variables. The first row must not include variable labels.

**Qualitative**: Activate this option to select the qualitative explanatory variables. The first row must not include variable labels.

**Observations labels**: activate this option if observations labels are available. Then select the corresponding data. If this option is not activated, the observations labels are automatically generated by XLSTAT (PredObs1, PredObs2 ...).
Missing data tab:

Remove observations: Activate this option to remove the observations with missing data.

- Check for each Y separately: Choose this option to remove the observations with missing data in the select Y (dependent) variables, only if the Y of interest has a missing data.
- Across all Ys: Choose this option to remove the observations with missing data in the Y (dependent) variables, even if the Y of interest has no missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- Mean or mode: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- Nearest neighbour: Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.

Outputs tab:

General sub-tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Correlations: Activate this option to display the correlation matrix for quantitative variables (dependent or explanatory).

Multicolinearity statistics: Activate this option to display the multicolinearity statistics for all explanatory variables.

Analysis of variance: Activate this option to display the analysis of variance table.

Type I/II/III SS: Activate this option to display the Type I, Type II, and Type III sum of squares tables. Type II table is only displayed if it is different from Type III.

Press: Activate this option to calculate and display Press' coefficient.

Standardized coefficients: Activate this option if you want the standardized coefficients (beta coefficients) for the model to be displayed.

Predictions and residuals: Activate this option to display the predictions and residuals for all the observations.
• **Adjusted predictions**: Activate this option to calculate and display adjusted predictions in the table of predictions and residuals.

• **Cook's D**: Activate this option to calculate and display Cook's distances in the table of predictions and residuals.

**Welch and Brown-Forsythe tests**: Activate this option to display the Welch and Brown-Forsythe tests (see the description section of this chapter) in the case of a one-way ANOVA.

**Multiple comparisons** sub-tab:

Information on the multiple comparison tests is available in the description section.

**Apply to all factors**: Activate this option to compute the selected tests for all factors.

**Use least squares means**: Activate this option to compare the means using their least squares estimators (obtained from the parameters of the model). If this option is not activated, the means are computed using their estimation based on the data.

**Sort up**: Activate this option to sort the compared categories in increasing order, the sort criterion being their respective means. If this option is not activated, the sort is decreasing.

**Standard errors**: Activate this option to display the standard errors with the means.

• **Confidence intervals**: Activate this option to additionally display the confidence intervals around the means.

**Pairwise comparisons**: Activate this option then choose the comparison methods.

**Comparisons with a control**: Activate this option then choose the type of Dunnett test you want to carry out.

**Choose the MSE**: Activate this option to select the mean squared error to be taken as reference for multiple comparisons. When using random factors, using the mean squared error of the model (classical case) is not appropriate. In that case, the user should choose a mean square error associated with another term in the model (usually an interaction term). If this option is enabled, a new dialog allowing you to select the variable to use will appear.

**Comparison of slopes**: Activate this option to compare the interaction slopes between the quantitative and qualitative variables (see the description section on this subject).

**Contrasts** sub-tab:

**Compute contrasts**: Activate this option to compute contrasts, then select the contrasts table, where there must be one column per contrast and one row for each coefficient of the model.
Charts tab:

Regression charts: Activate this option to display regression chart:

- **Standardized coefficients**: Activate this option to display the standardized parameters for the model with their confidence interval on a chart.

- **Predictions and residuals**: Activate this option to display the following charts.

  1. Line of regression: This chart is only displayed if there is only one explanatory variable and this variable is quantitative.

  2. Explanatory variable versus standardized residuals: This chart is only displayed if there is only one explanatory variable and this variable is quantitative.

  3. Dependent variable versus standardized residuals.

  4. Predictions for the dependent variable versus the dependent variable.

  5. Bar chart of standardized residuals.

  - **Confidence intervals**: Activate this option to have confidence intervals displayed on charts (1) and (4).

Means charts: Activate this option to display the charts used to display the means of the various categories of the various factors.

Results

Summary statistics: The tables of descriptive statistics show the simple statistics for all the variables selected. The number of observations, missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the dependent variables (in blue) and the quantitative explanatory variables. For qualitative explanatory variables the names of the various categories are displayed together with their respective frequencies.

Correlation matrix: This table is displayed to give you a view of the correlations between the various variables selected.

Summary of the variables selection: Where a selection method has been chosen, XLSTAT displays the selection summary. For a stepwise selection, the statistics corresponding to the different steps are displayed. Where the best model for a number of variables varying from p to q has been selected, the best model for each number or variables is displayed with the corresponding statistics and the best model for the criterion chosen is displayed in bold.
Goodness of fit statistics: The statistics relating to the fitting of the regression model are shown in this table:

- **Observations**: The number of observations used in the calculations. In the formulas shown below, \( n \) is the number of observations.

- **Sum of weights**: The sum of the weights of the observations used in the calculations. In the formulas shown below, \( W \) is the sum of the weights.

- **DF**: The number of degrees of freedom for the chosen model (corresponding to the error part).

- **\( R^2 \)**: The determination coefficient for the model. This coefficient, whose value is between 0 and 1, is only displayed if the constant of the model has not been fixed by the user. Its value is defined by:

\[
R^2 = 1 - \frac{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} w_i (y_i - \bar{y})^2}, \quad \text{where} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} w_i y_i,
\]

The \( R^2 \) is interpreted as the proportion of the variability of the dependent variable explained by the model. The nearer \( R^2 \) is to 1, the better is the model. The problem with the \( R^2 \) is that it does not take into account the number of variables used to fit the model.

- **Adjusted \( R^2 \)**: The adjusted determination coefficient for the model. The adjusted \( R^2 \) can be negative if the \( R^2 \) is near to zero. This coefficient is only calculated if the constant of the model has not been fixed by the user. Its value is defined by:

\[
\hat{R}^2 = 1 - (1 - R^2)\left(\frac{W - 1}{W - p - 1}\right)
\]

The adjusted \( R^2 \) is a correction to the \( R^2 \) which takes into account the number of variables used in the model.

- **MSE**: The mean squared error (MSE) is defined by:

\[
MSE = \frac{1}{W - p} \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2
\]

- **RMSE**: The root mean square of the errors (RMSE) is the square root of the MSE.

- **MAPE**: The **Mean Absolute Percentage Error** is calculated as follows:

\[
MAPE = \frac{100}{W} \sum_{i=1}^{n} \frac{w_i |y_i - \hat{y}_i|}{y_i}
\]

- **DW**: The Durbin-Watson statistic is defined by:
This coefficient is the order 1 autocorrelation coefficient and is used to check that the residuals of the model are not autocorrelated, given that the independence of the residuals is one of the basic hypotheses of linear regression. The user can refer to a table of Durbin-Watson statistics to check if the independence hypothesis for the residuals is acceptable.

- **Cp**: Mallows Cp coefficient is defined by:

\[
C_p = SSE \left( \frac{\hat{\sigma}^2 \hat{\sigma}}{\sigma^2} \right) + 2 p^* W
\]

where SSE is the sum of the squares of the errors for the model with p explanatory variables and \( \hat{\sigma} \) is the estimator of the variance of the residuals for the model comprising all the explanatory variables. The nearer the Cp coefficient is to p*, the less the model is biased.

- **AIC**: Akaike’s Information Criterion is defined by:

\[
AIC = W \ln \left( \frac{SSE}{W} \right) + 2 p^*
\]

This criterion, proposed by Akaike (1973) is derived from the information theory and uses Kullback and Leibler’s measurement (1951). It is a model selection criterion which penalizes models for which adding new explanatory variables does not supply sufficient information to the model, the information being measured through the MSE. The aim is to minimize the AIC criterion.

- **SBC**: Schwarz’s Bayesian Criterion is defined by:

\[
SBC = W \ln \left( \frac{SSE}{W} \right) + \ln (W) p^*
\]

This criterion, proposed by Schwarz (1978) is similar to the AIC and, like this, the aim is to minimize it.

- **PC**: Amemiya’s Prediction Criterion is defined by:

\[
PC = \frac{(1 - R^2)(W + p^*)}{W - p^*}
\]

This criterion, proposed by Amemiya (1980) is used, like the adjusted R² to take account of the parsimony of the model.
• **Press**: Press' statistic is only displayed if the corresponding option has been activated in the dialog box. It is defined by:

\[
Press = \sum_{i=1}^{n} w_i (y_i - \hat{y}_{i(-t)})^2
\]

where \(\hat{y}_{i(-t)}\) is the prediction for observation \(i\) when the latter is not used for estimating parameters. We then get:

\[
\text{Press RMSE} = \sqrt{\frac{\text{Press}}{W - p}}
\]

Press's RMSE can then be compared to the RMSE. A large difference between the two shows that the model is sensitive to the presence or absence of certain observations in the model.

• **Q²**: This statistic also known as the cross-validated \(R^2\). It is only displayed if the Press option has been activated in the dialog box. It is defined by:

\[
Q^2 = 1 - \frac{\text{Press}}{\sum_{i=1}^{n} (y_i - \bar{y})^2}
\]

This gives the proportion of the total variance that is explained by the explanatory variables when the predictions are computed when the corresponding observation is not in the model. A large difference between the \(Q^2\) and the \(R^2\) shows that the model is sensitive to the presence or absence of certain observations in the model.

The **analysis of variance table** is used to evaluate the explanatory power of the explanatory variables. Where the constant of the model is not set to a given value, the explanatory power is evaluated by comparing the fit (as regards least squares) of the final model with the fit of the rudimentary model including only a constant equal to the mean of the dependent variable. Where the constant of the model is set, the comparison is made with respect to the model for which the dependent variable is equal to the constant which has been set.

If the Type I/II/III SS (SS: Sum of Squares) is activated, the corresponding tables are displayed.

The table of **Type I SS** values is used to visualize the influence that progressively adding explanatory variables has on the fitting of the model, as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. The sums of squares in the Type I table always add up to the model SS. Note: the order in which the variables are selected in the model influences the values obtained.
The table of **Type II SS** values is used to visualize the influence that removing an explanatory variable has on the fitting of the model, all other variables being retained, as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. Note: unlike Type I SS, the order in which the variables are selected in the model has no influence on the values obtained.

The table of **Type III SS** values is used to visualize the influence that removing an explanatory variable has on the fitting of the model, all other variables being retained, except those were the effect is present (interactions), as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. Note: unlike Type I SS, the order in which the variables are selected in the model has no influence on the values obtained. Type II and Type III are identical if there are no interactions or if the design is balanced.

The **parameters of the model** table displays the estimate of the parameters, the corresponding standard error, the Student's t, the corresponding probability, as well as the confidence interval.

The **equation of the model** is then displayed to make it easier to read or re-use the model.

The table of **standardized coefficients** (also called beta coefficients) are used to compare the relative weights of the variables. The higher the absolute value of a coefficient, the more important the weight of the corresponding variable. When the confidence interval around standardized coefficients has value 0 (this can be easily seen on the chart of standardized coefficients), the weight of a variable in the model is not significant.

The **predictions and residuals** table shows, for each observation, its weight, the value of the qualitative explanatory variable, if there is only one, the observed value of the dependent variable, the model's prediction, the residuals, the studentized residuals, the confidence intervals together with the fitted prediction and Cook's D if the corresponding options have been activated in the dialog box. Two types of confidence interval are displayed: a confidence interval around the mean (corresponding to the case where the prediction would be made for an infinite number of observations with a set of given values for the explanatory variables) and an interval around the isolated prediction (corresponding to the case of an isolated prediction for the values given for the explanatory variables). The second interval is always greater than the first, the random values being larger. If the validation data have been selected, they are displayed at the end of the table.

The **charts** which follow show the results mentioned above. If there is only one explanatory variable in the model, the first chart displayed shows the observed values, the regression line and both types of confidence interval around the predictions. The second chart shows the standardized residuals as a function of the explanatory variable. In principle, the residuals should be distributed randomly around the X-axis. If there is a trend or a shape, this shows a problem with the model.
The **three charts** displayed next show respectively the evolution of the normalized residuals as a function of the dependent variable, the distance between the predictions and the observations (for an ideal model, the points would all be on the bisector), and the normalized residuals on a bar chart. The last chart quickly shows if an abnormal number of values are outside the interval \([-2, 2]\) given that the latter, assuming that the sample is normally distributed, should contain about 95% of the data.

If you have selected the data to be used for calculating **predictions on new observations**, the corresponding table is displayed next.

If multiple comparison tests have been requested, the corresponding results are then displayed.

**Example**

A tutorial on ANCOVA is available on the Addinsoft website:

http://www.xlstat.com/demo-anco.htm

**References**


Repeated Measures ANOVA

Use this tool to carry out Repeated Measures ANOVA (ANalysis Of VAriance). The advanced options enable you to choose the constraints on the model and to take account of interactions between the factors. Multiple comparison tests can be calculated.

XLSTAT proposes two ways for handling repeated measures ANOVA. The classical way using least squares estimation (LS) that is based on the same model as the classical ANOVA and the alternative way that is based on the maximum likelihood estimation (REML and ML). This chapter is devoted to the first method. For details on the second method, please read the chapter on mixed models.

Description

Repeated measures Analysis of Variance (ANOVA) uses the same conceptual framework as classical ANOVA. The main difference comes from the nature of the explanatory variables. The exploratory variable is measured at different time or repetition. In ANOVA, explanatory variables are often called factors.

If p is the number of factors, the ANOVA model is written as follows:

\[ y_{it} = \beta_0 + \sum_{j=1}^{p} \beta_{k(i,j),j} + \epsilon_i \]  

where \( y_{it} \) is the value observed for the dependent variable for observation I for measure t, \( k(i,j) \) is the index of the category of factor j for observation i, and \( \epsilon_i \) is the error of the model.

The hypotheses used in ANOVA are identical to those used in linear regression: the errors \( \epsilon_i \) follow the same normal distribution \( N(0,\sigma) \) and are independent.

However, other hypotheses are necessary in the case of repeated measures ANOVA. As measures are taken from the same subjects at different times, the repetitions are correlated. In repeated measures ANOVA we assume that the covariance matrix between the ys is spherical (for example, compound symmetry is a spherical shape). We can drop this hypothesis when using the maximum likelihood based approach.

The principle of repeated measures ANOVA is simple. For each measure, a classical ANOVA model is estimated, then the sphericity of the covariance matrix between measures is tested using Mauchly’s test, Greenhouse-Geisser epsilon or Huyn-Heldt epsilon. If the sphericity hypothesis is not rejected, between- and within-subject effects can be tested.

Interactions
By interaction is meant an artificial factor (not measured) which reflects the interaction between at least two measured factors. For example, if we carry out treatment on a plant, and tests are carried out under two different light intensities, we will be able to include in the model an interaction factor treatment*light which will be used to identify a possible interaction between the two factors. If there is an interaction between the two factors, we will observe a significantly larger effect on the plants when the light is strong and the treatment is of type 2 while the effect is average for weak light, treatment 2 and strong light, treatment 1 combinations.

To make a parallel with linear regression, the interactions are equivalent to the products between the continuous explanatory values although here obtaining interactions requires nothing more than simple multiplication between two variables. However, the notation used to represent the interaction between factor A and factor B is A*B.

The interactions to be used in the model can be easily defined in XLSTAT.

**Nested effects**

When constraints prevent us from crossing every level of one factor with every level of the other factor, nested factors can be used. We say we have a nested effect when fewer than all levels of one factor occur within each level of the other factor. An example of this might be if we want to study the effects of different machines and different operators on some output characteristic, but we can't have the operators change the machines they run. In this case, each operator is not crossed with each machine but rather only runs one machine. XLSTAT has an automatic device to find nested factors and one nested factor can be included in the model.

**Constraints**

During the calculations, each factor is broken down into a sub-matrix containing as many columns as there are categories in the factor. Typically, this is a full disjunctive table. Nevertheless, the breakdown poses a problem: if there are g categories, the rank of this sub-matrix is not g but g-1. This leads to the requirement to delete one of the columns of the sub-matrix and possibly to transform the other columns. Several strategies are available depending on the interpretation we want to make afterwards:

1) \( a_1=0 \): the parameter for the first category is null. This choice allows us force the effect of the first category as a standard. In this case, the constant of the model is equal to the mean of the dependent variable for group 1.

2) \( a_n=0 \): the parameter for the last category is null. This choice allows us force the effect of the last category as a standard. In this case, the constant of the model is equal to the mean of the dependent variable for group g.

Note: even if the choice of constraint influences the values of the parameters, it has no effect on the predicted values and on the different fitting statistics.
Multiple Comparisons Tests

One of the main applications of ANOVA is multiple comparisons testing whose aim is to check if the parameters for the various categories of a factor differ significantly or not. For example, in the case where four treatments are applied to plants, we want to know not only if the treatments have a significant effect, but also if the treatments have different effects.

Numerous tests have been proposed for comparing the means of categories. The majority of these tests assume that the sample is normally distributed. XLSTAT provides the main tests including:

Tukey's HSD test: this test is the most used (HSD: Honestly Significant Difference).

Fisher's LSD test: this is Student's test that tests the hypothesis that all the means for the various categories are equal (LSD: Least Significant Difference).

Bonferroni's t* test: this test is derived from Student's test and is less reliable as it takes into account the fact that several comparisons are carried out simultaneously. Consequently, the significance level of the test is modified according to the following formula:

\[ \alpha' = \alpha \frac{g(g - 1)}{2} \]

where \( g \) is the number of categories of the factor whose categories are being compared.

Dunn-Sidak's test: this test is derived from Bonferroni's test. It is more reliable in some situations.

\[ \alpha' = 1 - \left(1 - \alpha \right)^{\frac{1}{g(g-1)}} \]

The following tests are more complex as they are based on iterative procedures where the results depend on the number of combinations remaining to be tested for each category.

Newman-Keuls's test (SNK): this test is derived from Student's test (SNK: Student Newman-Keuls), and is very often used although not very reliable.

Duncan's test: this test is little used.

REGWQ test: this test is among the most reliable in a majority of situations (REGW: Ryan-Einot-Gabriel-Welsch).

Benjamini-Hochberg: Use this option to control the False Discovery Rate (FDR). This p-value penalization procedure is poorly conservative.

The Games-Howell (GH) test can be used in one-way ANOVAs when the variances lack of homogeneity. While it can be used with unequal sample sizes, it is recommended to use it
when the smallest sample has 5 elements or more, otherwise it is too liberal. The Tamhane’s \( T_2 \) test is more conservative, but not as powerful as the GH test.

All the above tests enable comparisons to be made between all pairs of categories and belong to the MCA test family (Multiple Comparisons of All, or All-Pairwise Comparisons).

Other tests make comparisons between all categories and a control category. These tests are called MCB tests (Multiple Comparisons with the Best, Comparisons with a control). XLSTAT offers the Dunnett test which is the most used. There are three Dunnett tests:

- **Two-tailed test**: the null hypothesis assumes equality between the category tested and the control category. The alternative hypothesis assumes the means of the two categories differ.

- **Left one-tailed test**: the null hypothesis assumes equality between the category tested and the control category. The alternative hypothesis assumes that the mean of the control category is greater than the mean of the category tested.

- **Right one-tailed test**: the null hypothesis assumes equality between the category tested and the control category. The alternative hypothesis assumes that the mean of the control category is less than the mean of the category tested.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.

- Click this button to reload the default options.
- Click this button to delete the data selections.
- Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If
the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

Y / Dependent variables:

**Quantitative**: Select the response variable(s) you want to model. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

**One column for all repetitions**: activate this option if your dependent variable is organized in only one column. In that case, you have to select as explanatory variables, one variable with the name of the repetition and another variable with the name of the subject. For more details on that format, please see the chapter on mixed models.

**One column per repetition**: activate this option if your dependent variable has T columns for T repetitions. In that case, when you will have to select the factors, a factor called repetition and a factor called subject will appear.

X / Explanatory variables:

**Qualitative**: Select the qualitative explanatory variables (the factors) in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Variable labels" option has been activated.

**Quantitative**: Select one or more quantitative variables on the Excel sheet. If the variable labels have been selected, please check the option "Variable labels" is activated. When no qualitative variables are selected, then it is a repeated measures linear regression. If qualitative and quantitative variables are selected then it is a repeated measures ANCOVA. If no explanatory variables are selected, it is a one-way repeated measures ANOVA.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Variable labels**: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.
**Observation weights**: Activate this option if the observations are weighted. If you do not activate this option, the weights will all be taken as 1. Weights must be greater than or equal to 0. A weight of 2 is equivalent to repeating the same observation twice. If a column header has been selected, check that the "Variable labels" option has been activated.

**Regression weights**: Activate this option if you want to carry out a weighted least squares regression. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

**Options** tab:

**Fixed constant**: Activate this option to fix the constant of the regression model to a value you then enter (0 by default).

**Tolerance**: Activate this option to prevent the OLS regression calculation algorithm taking into account variables which might be either constant or too correlated with other variables already used in the model (0.0001 by default).

**Interactions / Level**: Activate this option to include interactions in the model then enter the maximum interaction level (value between 1 and 4).

**Nested effects**: Activate this option to include one nested effect in the model.

**Confidence interval (%)**: Enter the percentage range of the confidence interval to use for the various tests and for calculating the confidence intervals around the parameters and predictions. Default value: 95.

**Constraints**: Details on the various options are available in the description section.

**a1 = 0**: Choose this option so that the parameter of the first category of each factor is set to 0.

**an = 0**: Choose this option so that the parameter of the last category of each factor is set to 0.

**Missing data** tab:

**Remove observations**: Activate this option to remove the observations with missing data.
**Estimate missing data:** Activate this option to estimate missing data before starting the computations.

- **Mean or mode:** Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

- **Nearest neighbour:** Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.

**Outputs** tab:

**Descriptive statistics:** Activate this option to display descriptive statistics for the variables selected.

**Analysis of variance:** Activate this option to display the analysis of variance table for each repetition t.

**Type I/III SS:** Activate this option to display the Type I and Type III sum of squares tables for each ANOVA associated to repetition t.

**Standardized coefficients:** Activate this option if you want the standardized coefficients (beta coefficients) for the model to be displayed.

**Predictions and residuals:** Activate this option to display the predictions and residuals for all the observations.

**Multiple comparisons:**

Information on the multiple comparison tests is available in the description section.

**Apply to all factors:** Activate this option to compute the selected tests for all factors.

**Use least squares means:** Activate this option to compare the means using their least squares estimators (obtained from the parameters of the model). If this option is not activated, the means are computed using their estimation based on the data.

**Sort up:** Activate this option to sort the compared categories in increasing order, the sort criterion being their respective means. If this option is not activated, the sort is decreasing.

**Pairwise comparisons:** Activate this option then choose the comparison methods.

**Comparisons with a control:** Activate this option then choose the type of Dunnett test you want to carry out.
Charts tab:

Regression charts: Activate this option to display regression chart:

- **Predictions and residuals**: Activate this option to display the following charts.
  
  (1) Line of regression: This chart is only displayed if there is only one explanatory variable and this variable is quantitative.
  
  (2) Explanatory variable versus standardized residuals: This chart is only displayed if there is only one explanatory variable and this variable is quantitative.
  
  (3) Dependent variable versus standardized residuals.
  
  (4) Predictions for the dependent variable versus the dependent variable.
  
  (5) Bar chart of standardized residuals.

Means charts: Activate this option to display the charts used to display the means of the various categories of the various factors.

Factors and interactions dialog box

Once the first dialog box disappears, a second one appears, to allow you to clarify the belonging of each factor. It is necessary to select the fixed factors (fixed effects), a repeated factor and a subject factor. If you selected the *one column per repetition* layout, then a factor called repetition and a factor called subject are displayed and must respectively be selected as repeated and subject factors.

Results

Summary statistics: The tables of descriptive statistics show the simple statistics for all the variables selected. The number of observations, missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the dependent variables (in blue) and the quantitative explanatory variables. For qualitative explanatory variables the names of the various categories are displayed together with their respective frequencies.

Then for each repetition, we have:

Goodness of fit statistics: The statistics relating to the fitting of the regression model are shown in this table:
• **Observations**: The number of observations used in the calculations. In the formulas shown below, \( n \) is the number of observations.

• **Sum of weights**: The sum of the weights of the observations used in the calculations. In the formulas shown below, \( W \) is the sum of the weights.

• **DF**: The number of degrees of freedom for the chosen model (corresponding to the error part).

• **\( R^2 \)**: The determination coefficient for the model. This coefficient, whose value is between 0 and 1, is only displayed if the constant of the model has not been fixed by the user. Its value is defined by:

\[
R^2 = 1 - \frac{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} w_i (y_i - \bar{y})^2}, \quad \text{where} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} w_i y_i,
\]

The \( R^2 \) is interpreted as the proportion of the variability of the dependent variable explained by the model. The nearer \( R^2 \) is to 1, the better is the model. The problem with the \( R^2 \) is that it does not take into account the number of variables used to fit the model.

• **Adjusted \( R^2 \)**: The adjusted determination coefficient for the model. The adjusted \( R^2 \) can be negative if the \( R^2 \) is near to zero. This coefficient is only calculated if the constant of the model has not been fixed by the user. Its value is defined by:

\[
\hat{R}^2 = 1 - (1 - R^2) \frac{W - 1}{W - p - 1}
\]

The adjusted \( R^2 \) is a correction to the \( R^2 \) which takes into account the number of variables used in the model.

• **MSE**: The mean squared error (MSE) is defined by:

\[
MSE = \frac{1}{W - p} \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2
\]

• **RMSE**: The root mean square of the errors (RMSE) is the square root of the MSE.

• **MAPE**: The *Mean Absolute Percentage Error* is calculated as follows:

\[
MAPE = \frac{100}{W} \sum_{i=1}^{n} w_i \left| \frac{y_i - \hat{y}_i}{y_i} \right|
\]

• **DW**: The Durbin-Watson statistic is defined by:
This coefficient is the order 1 autocorrelation coefficient and is used to check that the residuals of the model are not autocorrelated, given that the independence of the residuals is one of the basic hypotheses of linear regression. The user can refer to a table of Durbin-Watson statistics to check if the independence hypothesis for the residuals is acceptable.

- **Cp**: Mallows Cp coefficient is defined by:

  \[
  Cp = \frac{SSE}{\hat{\sigma}} + 2p^* - W
  \]

  where \( SSE \) is the sum of the squares of the errors for the model with \( p \) explanatory variables and \( \hat{\sigma} \) is the estimator of the variance of the residuals for the model comprising all the explanatory variables. The nearer the Cp coefficient is to \( p^* \), the less the model is biased.

- **AIC**: Akaike's Information Criterion is defined by:

  \[
  AIC = W \ln \left( \frac{SSE}{W} \right) + 2p^*
  \]

  This criterion, proposed by Akaike (1973) is derived from the information theory and uses Kullback and Leibler's measurement (1951). It is a model selection criterion which penalizes models for which adding new explanatory variables does not supply sufficient information to the model, the information being measured through the MSE. The aim is to minimize the AIC criterion.

- **SBC**: Schwarz's Bayesian Criterion is defined by:

  \[
  SBC = W \ln \left( \frac{SSE}{W} \right) + \ln(W)p^*
  \]

  This criterion, proposed by Schwarz (1978) is similar to the AIC and, like this, the aim is to minimize it.

- **PC**: Amemiya's Prediction Criterion is defined by:

  \[
  PC = \frac{(1 - R^2)(W + p^*)}{W - p^*}
  \]

  This criterion, proposed by Amemiya (1980) is used, like the adjusted \( R^2 \) to take account of the parsimony of the model.
The analysis of variance table is used to evaluate the explanatory power of the explanatory variables. Where the constant of the model is not set to a given value, the explanatory power is evaluated by comparing the fit (as regards least squares) of the final model with the fit of the rudimentary model including only a constant equal to the mean of the dependent variable. Where the constant of the model is set, the comparison is made with respect to the model for which the dependent variable is equal to the constant which has been set.

If the Type I/III SS (SS: Sum of Squares) option is activated, the corresponding tables are displayed.

The table of **Type I SS** values is used to visualize the influence that progressively adding explanatory variables has on the fitting of the model, as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. The sums of squares in the Type I table always add up to the model SS. Note: the order in which the variables are selected in the model influences the values obtained.

The table of **Type III SS** values is used to visualize the influence that removing an explanatory variable has on the fitting of the model, all other variables being retained, as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. Note: unlike Type I SS, the order in which the variables are selected in the model has no influence on the values obtained. While Type II SS depends on the number of observations per cell (cell means combination of categories of the factors), Type III does not and is therefore preferred.

The parameters of the model table displays the estimate of the parameters, the corresponding standard error, the Student’s t, the corresponding probability, as well as the confidence interval.

The table of **standardized coefficients** (also called beta coefficients) are used to compare the relative weights of the variables. The higher the absolute value of a coefficient, the more important the weight of the corresponding variable. When the confidence interval around standardized coefficients has value 0 (this can be easily seen on the chart of standardized coefficients), the weight of a variable in the model is not significant.

The predictions and residuals table shows, for each observation, its weight, the value of the qualitative explanatory variable, if there is only one, the observed value of the dependent variable, the model’s prediction, the residuals, the confidence. Two types of confidence interval are displayed: a confidence interval around the mean (corresponding to the case where the prediction would be made for an infinite number of observations with a set of given values for the explanatory variables) and an interval around the isolated prediction (corresponding to the case of an isolated prediction for the values given for the explanatory variables). The second interval is always greater than the first, the random values being larger. If the validation data have been selected, they are displayed at the end of the table.
The **charts** which follow show the results mentioned above. If there is only one explanatory variable in the model, the first chart displayed shows the observed values, the regression line and both types of confidence interval around the predictions. The second chart shows the standardized residuals as a function of the explanatory variable. In principle, the residuals should be distributed randomly around the X-axis. If there is a trend or a shape, this shows a problem with the model.

The **three charts** displayed next show respectively the evolution of the standardized residuals as a function of the dependent variable, the distance between the predictions and the observations (for an ideal model, the points would all be on the bisector), and the standardized residuals on a bar chart. The last chart quickly shows if an abnormal number of values are outside the interval \([-2, 2]\) given that the latter, assuming that the sample is normally distributed, should contain about 95% of the data.

If multiple comparison tests have been requested, the corresponding results are then displayed.

Finally, tables associated to the repeated measures analysis are displayed:

**Mauchly’s sphericity test** can be used to test the sphericity of the covariance matrix between repetitions. It has a small power and should not be trusted with small samples. In this table, Greenhouse-Geisser and Huynh-Feldt epsilons can also be found. The closer they are to one, the more spherical the covariance matrix is.

**The test of within-subject effects** is then displayed. It shows which factor has a significant effect across repetition.

**The test of between-subject effects** is then displayed. It shows which factor has an effect which is significantly different from one subject to another and not from one repetition to another.

**Example**

A tutorial on repeated measures ANOVA is available on the Addinsoft website: [http://www.xlstat.com/demo-anorep2.htm](http://www.xlstat.com/demo-anorep2.htm)
References


Mixed Models

Use this tool to build ANOVA models with repeated factors, random components or repeated measures.

Description

Mixed models are complex models based on the same principle as general linear models. They make it possible to take into account, on the one hand, the concept of repeated measurement and, on the other hand, that of random factor. The explanatory variables could be as well quantitative as qualitative. Within the framework of the mixed models, the explanatory variables are often called factors. XLSTAT uses mixed models to carry out repeated measures ANOVA.

A mixed model is written as follows:

\[ y = X\beta + Z\gamma + \varepsilon \]  (1)

where \( y \) is the dependent variable, \( X \) gathers all fixed effects (these factors are the classical OLS regression variables or the ANOVA factors), \( \beta \) is a vector of parameters associated with the fixed factors, \( Z \) is a matrix gathering all the random effects (factors that cannot be set as fixed), \( \gamma \) is a vector of parameters associated with the random effects and \( \varepsilon \) is an error vector. The main difference between general linear model and mixed model is that

\[ \gamma \sim N(0, G(\theta_G)) \] \quad and \quad \[ \varepsilon \sim N(0, R(\theta_R)) \].

We have:

\[
E\begin{bmatrix} \gamma \\ \varepsilon \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{and} \quad Var\begin{bmatrix} \gamma \\ \varepsilon \end{bmatrix} = \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix}
\]

The variance of \( y \) is written as follows:

\[ Var(y) = V(\theta) = Z'GZ + R \], where \( \theta \) is a vector of parameters associated with the unknown parameters of \( G \) and \( R \rightarrow y \sim N\left(X\beta, V(\theta)\right) \).

According to the model to be estimated, the matrices \( R \) and \( G \) will have different forms:

- For a classical linear model, we have: \( Z = 0 \) and \( R = \sigma^2 I_n \).

- For a repeated measures ANOVA, we have: \( Z = 0 \) and \( \text{cov}(\varepsilon) = R(\theta) \), where \( R \) is a square-block matrix with a user-defined design. Each block gathers the covariance between different measures on the same subject (which are correlated). Explanatory variables are all qualitative.
- For a random component model, we have $\text{cov}(\gamma) = G$, where $G$ is a matrix with a user-defined design.

The following table shows the designs implemented in XLSTAT for the R and G matrices (dimension $p \times p$):

<table>
<thead>
<tr>
<th>Covariance structure</th>
<th>Number of parameters</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variance components</td>
<td>Number of random factors (if no random factor = 1)</td>
<td>$\sigma_{ij} = \sigma_k^2 I(i = j) \ k$ is the random effect associated with the $i^{th}$ line</td>
</tr>
<tr>
<td>Autoregressive(1)</td>
<td>2</td>
<td>$\sigma_{ij} = \sigma^2 \rho^{j-i}$</td>
</tr>
<tr>
<td>Compound symmetry</td>
<td>2</td>
<td>$\sigma_{ij} = \sigma_1 + \sigma^2 I(i = j)$</td>
</tr>
<tr>
<td>Unstructured</td>
<td>$p(p+1)/2$</td>
<td>$\sigma_{ij} = \sigma_{ij}$</td>
</tr>
<tr>
<td>Toeplitz</td>
<td>$p$</td>
<td>$\sigma_{ij} = \sigma_{</td>
</tr>
<tr>
<td>Toeplitz(q)</td>
<td>$\text{Min}(p,q)$</td>
<td>$\sigma_{ij} = \sigma_{</td>
</tr>
</tbody>
</table>

Parameters estimation is performed by using the maximum likelihood approach. There exist two methods: the classical maximum likelihood (ML) and the restricted maximum likelihood (REML). The latter is the default in XLSTAT. The likelihood function is given by:

$$l_{\text{REML}}(G, R) = -\frac{1}{2} \log |V| - \frac{1}{2} \log |X'V^{-1}X| - \frac{1}{2} r'V^{-1}r - \frac{n - p}{2} \log(2\pi)$$

(2)

where $r = y - X \hat{\beta}$. The parameters are obtained by using the first and second derivatives of $l_{\text{REML}}(G, R)$. For the details of these matrices, one can see Wolfinger, Tobias and Sall (1994). The use of an analytical method to obtain the $\theta$ parameters is not possible. XLSTAT does not profile the variance during the computation and initial values of the covariance matrix are the variances obtained with the general linear model. We thus use the iterative Newton-Raphson algorithm in order to obtain an estimate of $\theta$. Once $\theta$ is obtained, the coefficients $\beta$ and $\gamma$ are calculated by solving the following equation system:

$$
\begin{bmatrix}
X'\hat{R}^{-1}X & X'\hat{R}^{-1}Z \\
Z'\hat{R}^{-1}X & Z'\hat{R}^{-1}Z + \hat{G}^{-1}
\end{bmatrix}
\begin{bmatrix}
\hat{\beta} \\
\hat{\gamma}
\end{bmatrix}
= 
\begin{bmatrix}
X'\hat{R}^{-1}y \\
Z'\hat{R}^{-1}y
\end{bmatrix}
$$

(3)
We obtain:

\[
\hat{\beta} = \left( X'\hat{V}^{-1}X \right)^{-1} X'\hat{V}^{-1}y \\
\hat{\gamma} = \hat{GZ}'\hat{V}^{-1} \left( y - X\hat{\beta} \right)
\]  

(4)

where \( \hat{\cdot} \) is the generalized inverse of the matrix. The interpretation of the model is made in the same way than in the linear case.

**Data format**

Within the framework of mixed models, the data will have a specific format:

- If there are no repeated measurements, then there will be one column per variable associated to each fixed effect and one column per variable associated to each random effect.

- If there are repeated measurements, all the repetitions will have to be one after the other. We cannot have one column for each repetition. We thus define a factor identifying each repetition and another factor identifying the subject to be treated in each repetition. Thus, for a data set with 3 repetitions and 2 subjects and an explanatory variable \( X \) measured at times \( T_1 \) and \( T_2 \) for the two subjects, we have the following table:

<table>
<thead>
<tr>
<th>fact.rep</th>
<th>fact.suj</th>
<th>( X )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>( x_1^{T_1} )</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>( x_1^{T_2} )</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>( x_1^{T_2} )</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>( x_2^{T_2} )</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>( x_1^{T_2} )</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>( x_2^{T_2} )</td>
</tr>
</tbody>
</table>

XLSTAT makes it possible to select a repeated factor and a subject factor. These factors must be qualitative and are necessary for repeated measures ANOVA, and available for mixed models.

**Interactions**

By interaction is meant an artificial factor (not measured) that reflects the interaction between at least two measured factors. For example, if we carry out treatment on a plant, and tests are carried out under two different light intensities, we will be able to include in the model an interaction factor treatment*light which will be used to identify a possible interaction between
the two factors. If there is an interaction between the two factors, we will observe, for example, a significantly higher effect on the plants when the light is strong and the treatment is of type 2 while the effect is average for low light and treatment 2 or strong light and treatment 1 combinations.

To make a parallel with linear regression, the interactions are equivalent to the products between the continuous explanatory values. For qualitative variables it is a little more complex, and constraints must be defined to avoid multicolinearities in the model (see below). However, the notation used to represent the interaction between factor A and factor B is $A*B$.

The interactions to be used in the model can be easily defined in XLSTAT.

**Constraints**

During the calculations, each factor is broken down into a sub-matrix containing as many columns as there are categories in the factor. Typically, this is a full disjunctive table. Nevertheless, the breakdown poses a problem: if there are $g$ categories, the rank of this sub-matrix is not $g$ but $g-1$. This leads to the requirement to delete one of the columns of the sub-matrix and possibly to transform the other columns. Several strategies are available depending on the interpretation we want to make afterwards:

1) $a_1=0$: the parameter for the first category is null. This choice allows us force the effect of the first category as a standard. In this case, the constant of the model is equal to the mean of the dependent variable for group 1.

2) $a_n=0$: the parameter for the last category is null. This choice allows us force the effect of the last category as a standard. In this case, the constant of the model is equal to the mean of the dependent variable for group $g$.

Note: even if the choice of constraint influences the values of the parameters, it has no effect on the predicted values and on the different fitting statistics.

**Inference and tests**

XLSTAT allows computing the type I, II and III tests of the fixed effects. The principle of these tests is the same one as in the case of the linear model. Nevertheless, their calculation differs slightly. All these tests are based on the following $F$ statistics:

$$F = \frac{\hat{\beta}'L'\left( L(X\hat{\beta}^{-1}X)^{-}L' \right)^{-}L\hat{\beta}}{r},$$

where $L$ is a specific matrix associated with each fixed effect and it differs depending on the type of test. We have $r = rank\left( L(X\hat{\beta}^{-1}X)^{-}L' \right)$. A p-value is obtained using the Fisher distribution with Num. DF and Den. DF degrees of freedom.
We have $Num.DF = rank(L)$ and Den. DF depends on the estimated model. XLSTAT uses:

- The contain method if a random effect is selected, we have: $Den.DF = N - rank(XZ)$.

- The residual method if no random effect is selected, we have: $Den.DF = n - rank(X)$.

**Multiple Comparisons Tests (only for repeated measures ANOVA)**

As in classical ANOVA, in repeated measures ANOVA multiple comparisons can be performed. It is aimed at checking whether the various categories of a factor differ significantly or not. For example, in the case where four treatments are applied to plants, we want to know not only if the treatments have a significant effect, but also if the treatments have different effects.

Numerous tests have been proposed for comparing the means of categories. The majority of these tests assume that the sample is normally distributed. XLSTAT provides the main tests. In the case of repeated measures ANOVA, standard deviations are obtained using the maximum likelihood estimates. For more details on the tests, please see the description section of the ANOVA help.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- Click this button to start the computations.
- Click this button to close the dialog box without doing any computation.
- Click this button to display the help.
- Click this button to reload the default options.
- Click this button to delete the data selections.
Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

**Y / Dependent variables:**

**Quantitative:** Select the response variable(s) you want to model. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

**X / Explanatory variables:**

**Quantitative:** Activate this option to perform an ANCOVA analysis. Then select the quantitative explanatory variables in the Excel worksheet. The data selected must be of type numeric. If the variable header has been selected, check that the "Variable labels" option has been activated.

**Qualitative:** Select the qualitative explanatory variables (the factors) in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Variable labels" option has been activated.

**Range:** Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Activate this option to display the results in a new worksheet of the active workbook.

**Workbook:** Activate this option to display the results in a new workbook.

**Variable labels:** Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

**Observation weights:** Activate this option if the observations are weighted. If you do not activate this option, the weights will all be taken as 1. Weights must be greater than 0. A weight of 2 is equivalent to repeating the same observation twice. If a column header has been selected, check that the "Variable labels" option has been activated.

**Regression weights:** Activate this option if you want to include weights in the model’s equation. If you do not activate this option, the weights will be considered as 1. Weights must
be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

**Options** tab:

**Fixed constant**: Activate this option to fix the constant of the regression model to a value you then enter (0 by default).

**Tolerance**: Activate this option to prevent the OLS regression calculation algorithm taking into account variables which might be either constant or too correlated with other variables already used in the model (0.0001 by default).

**Interactions / Level**: Activate this option to include interactions in the model then enter the maximum interaction level (value between 1 and 4).

**Confidence interval (%)**: Enter the percentage range of the confidence interval to use for the various tests and for calculating the confidence intervals around the parameters and predictions. Default value: 95.

**Constraints**: Details on the various options are available in the description section.

**a1 = 0**: Choose this option so that the parameter of the first category of each factor is set to 0.

**an = 0**: Choose this option so that the parameter of the last category of each factor is set to 0.

**Repeated measures**: Activate this option if you want to include a repeated factor in your model.

**Covariance structure**: Choose the covariance structure you want to use for the R matrix. XLSTAT offers: Autoregressive(1), Compound Symmetry, Toeplitz, Toeplitz(q), Unstructured and Variance Components. Details on the various options are available in the description section.

**Random effect** (only with mixed models): Activate this option if you want to include a random effect in your model.

**Covariance structure**: Choose the covariance structure you want to use for the R matrix. XLSTAT offers: Autoregressive(1), Compound Symmetry, Toeplitz, Toeplitz(q), Unstructured and Variance Components. Details on the various options are available in the description section.

**Estimation method**: choose between REML and ML to estimate your model. Details are available in the description section.
**Missing data** tab:

**Remove observations**: Activate this option to remove the observations with missing data.

**Estimate missing data**: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

- **Nearest neighbour**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.

**Outputs** tab:

**General**:

**Descriptive statistics**: Activate this option to display descriptive statistics for the variables selected.

**Correlations**: Activate this option to display the correlation matrix for quantitative variables (dependent or explanatory).

**Goodness of fit statistics**: Activate this option to display the table of goodness of fit statistics for the model.

**Covariance parameters**: Activate this option to display the table of the covariance parameters.

**Null model likelihood ratio test**: Activate this option to display the results of the null model likelihood ratio test.

**Fixed effects coefficients**: Activate this option to display the table of the fixed effects coefficients.

**Random effects coefficients** (only with mixed models): Activate this option to display the table of the random effects coefficients.

**Type III tests of fixed effect**: Activate this option to display the results of the type III tests of the fixed effect.

**Type I tests of fixed effect**: Activate this option to display the results of the type I tests of the fixed effect.

**Type II tests of fixed effect**: Activate this option to display the results of the type II tests of the fixed effect.
R matrix: Activate this option to display the error covariance matrix R for the first subject.

G matrix (only with mixed models): Activate this option to display the random effects covariance matrix G.

Residuals:

Predictions and residuals: Activate this option to display the predictions and residuals for all the observations.

  • Raw residuals: Activate this option to display the raw residuals in the predictions and residuals table.
  
  • Studentized residuals: Activate this option to display the studentized residuals in the predictions and residuals table.
  
  • Pearson residuals: Activate this option to display the Pearson residuals in the predictions and residuals table.

Comparisons (only for repeated measures ANOVA):

Multiple comparisons:

Information on the multiple comparison tests is available in the description section.

Apply to all factors: Activate this option to compute the selected tests for all factors.

Use least squares means: Activate this option to compare the means using their least squares estimators (obtained from the parameters of the model). If this option is not activated, the means are computed using their estimation based on the data.

Sort up: Activate this option to sort the compared categories in increasing order, the sort criterion being their respective means. If this option is not activated, the sort is decreasing.

Pairwise comparisons: Activate this option then choose the comparison methods.

Factors and interactions dialog box

Once the first dialog box disappears, a second one appears, to allow you to what type of factor each factor corresponds. The layout and the aim of the dialog box depends the type of ANOVA you want to run:
- If repeated measures were selected, it is necessary to select the fixed factors (fixed effects), a repeated factor, and a subject factor.

- If random effects have been selected, it is necessary to specify which factors are fixed and which are random.

- If both repeated measures and random effects have been selected, it is necessary to specify which factors are fixed, which are random, and to define which is the repeated factor and which is subject factor.

Each factor must be selected only once. Repeated and subject factors must be qualitative.

Results

Summary statistics: The tables of descriptive statistics show the simple statistics for all the variables selected. The number of observations, missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the dependent variables (in blue) and the quantitative explanatory variables. For qualitative explanatory variables the names of the various categories are displayed together with their respective frequencies.

Correlation matrix: This table is displayed to give you a view of the correlations between the various variables selected.

Goodness of fit statistics: The statistics relating to the fitting of the regression model are shown in this table:

- **Observations**: The number of observations used in the calculations. In the formulas shown below, \( n \) is the number of observations.

- **Sum of weights**: The sum of the weights of the observations used in the calculations. In the formulas shown below, \( W \) is the sum of the weights.

- **AIC**: the Akaike’s Information Criterion (AIC) is defined by:

\[
AIC = -2l(\theta) + 2d
\]

where \( l \) is the likelihood function and \( d \) equals the number of parameters to be estimated. This criterion, proposed by Akaike (1973) is derived from the information theory and uses Kullback and Leibler's measurement (1951). It is a model selection criterion which penalizes models for which adding new explanatory variables does not supply sufficient information to the model, the information being measured through the MSE. The aim is to minimize the AIC criterion.

- **AICC**: This criterion derived from the AIC is defined by:
\[ AICC = -2l(\theta) + 2dn/(n-d-1) \]

- **SBC**: Schwarz's Bayesian Criterion is defined by:
  \[ SBC = -2l(\theta) + d \ln(n) \]
  This criterion, proposed by Schwarz (1978) is similar to the AIC and the aim is to minimize it.

- **CAIC**: This criterion (Bodzogan, 1987) is defined by:
  \[ CAIC = -2l(\theta) + d(\ln(n)+1) \]

- **Iterations**: This value gives the number of iteration necessary to reach the convergence of the Newton-Raphson algorithm.

- **Covariance parameters**: This value gives the number of parameters to be estimated in the covariance matrix \( V \).

- **Number of fixed effects**: This value gives the number of selected fixed effects.

- **Number of random effects**: This value gives the number of selected random effects.

**Covariance parameters – Repeated factors**: This table displays the covariance parameters associated to the repeated factor. For each parameter, the corresponding standard error, the Z statistic, the corresponding probability, as well as the confidence interval are presented.

**Covariance parameters – Random factors** (only with mixed models): This table displays covariance parameters associated to the random. For each parameter, the corresponding standard error, the Z statistic, the corresponding probability, as well as the confidence interval are presented.

The **null model likelihood ratio test table** compares the likelihood of the null model and the likelihood of the selected model. The likelihood ration, the Chi-square statistic and the corresponding probability are displayed.

The **model parameters** table displays the estimate of the parameters, the corresponding standard error, the Student’s t, the corresponding probability, as well as the confidence interval

The **random effects coefficients** (only with mixed models) table displays the estimate of the random effects parameters, the corresponding standard error, the number of degrees of freedom, the Student’s t, the corresponding probability and confidence interval

If the Type I tests and Type III tests of fixed effects have been requested, the corresponding tables are displayed.
The table of **Type I tests of fixed effects** values is used to evaluate the influence of sequentially adding explanatory variables on the fit of the model, through the Fisher's F or its corresponding p-value. The lower the probability, the larger the contribution of the variable to the model (given that all the previously added variables are in the model).

Note: the order in which the variables are selected in the model influences the values obtained.

The table of **Type III tests of fixed effects** values is used to evaluate the impact of removing an explanatory variable, all other variables being retained, in terms of Fisher's F and its corresponding p-value. The lower the probability, the larger the contribution of the variable to the model, all other variables already being in the model.

Note: unlike Type I tests of fixed effects, the order in which the variables are selected in the model does not have any influence on the values obtained.

The **predictions and residuals** table shows, for each observation, its weight, the observed value of the dependent variable, the model's prediction, the residuals, the confidence intervals. Several types of residuals are displayed:

- Raw residuals: $r_i = y_i - x_i' \hat{\beta}$.

- Studentized residuals: $r_i^{\text{stud}} = \frac{r_i}{\sqrt{\text{var}(r_i)}}$

- Pearson's residuals: $r_i^{\text{pearson}} = \frac{r_i}{\sqrt{\text{var}(y_i)}}$

- If one or more random effects are selected, we have:
  - Conditional raw residuals: $r_i^{\text{cond}} = r_i - z_i' \hat{\gamma}$
  - Studentized conditional residuals: $r_i^{\text{cond/stud}} = \frac{r_i^{\text{cond}}}{\sqrt{\text{var}(r_i^{\text{cond}})}}$
  - Pearson's conditional residuals: $r_i^{\text{cond/pearson}} = \frac{r_i^{\text{cond}}}{\sqrt{\text{var}(y_i)}}$

If multiple comparison tests have been requested, the corresponding results are then displayed.
Example

A tutorial on repeated measures ANOVA is available on the Addinsoft website:
http://www.xlstat.com/demo-anorep.htm

A tutorial on random component model is available on the Addinsoft website:
http://www.xlstat.com/demo-mixed.htm

References


MANOVA

Use this model to carry out a MANOVA (Multivariate ANalysis Of VAriance) for two or more balanced or unbalanced factors. The advanced options enable you to choose the confidence level on the model and to take into account interactions between the factors. Multivariate tests can be calculated.

Description

The MANOVA uses the same conceptual framework as the ANOVA. The main difference comes from the nature of the dependent variables: instead of one, we can study many of them. With the MANOVA, explanatory variables are often called factors. Effects of factors are estimated on a combination of several response variables.

The advantage of the MANOVA as opposed to several simultaneous ANOVA lies in the fact that it takes into account correlations between response variables which results in a richer use of the information contained in the data.

The MANOVA tests the presence of significant differences among combinations of levels of factors on several response variables. MANOVA also enables the simultaneous tests of all hypotheses tested by an ANOVA and is more likely to detect differences between levels of factors.

Furthermore, the computation of several ANOVA instead of one MANOVA increases the Type I error which is the probability to reject wrongly the null hypothesis.

The potential covariation between response variables is not taken into account with several ANOVA. Instead, the MANOVA is sensitive to both the difference of averages between levels of factors and the covariation between explanatory variables. And a potential correlation between response variables is more likely to be detected when these variables are studied together as it is the case with a MANOVA.

Let’s consider as an illustrative example a two-way MANOVA. The model is written as followed:

\[
Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk} \tag{1}
\]

where \(Y_{ijk}\) is the kth observation of the ith level of A and jth level of B, \(\epsilon\) is the error of the model.

The hypotheses used in a MANOVA are identical to those used in linear regression: errors \(\epsilon\) follow the same normal distribution \(N(0,\sigma)\) and are independent.
To use the various tests proposed in the results of linear regression, it is recommended to check retrospectively that the underlying hypotheses have been correctly verified. The normality of the residuals can be checked by analyzing certain charts or by using a normality test. The independence of the residuals can be checked by analyzing certain charts or by using the Durbin Watson test.

**Interactions**

By interaction is meant an artificial factor (not measured) which reflects the interaction between at least two measured factors. For example, if we carry out treatment on a plant, and tests are carried out under two different light intensities, we will be able to include in the model an interaction factor treatment*light which will be used to identify a possible interaction between the two factors. If there is an interaction between the two factors, we will observe a significantly larger effect on the plants when the light is strong and the treatment is of type 2 while the effect is average for weak light, treatment 2 and strong light, treatment 1 combinations.

To make a parallel with linear regression, the interactions are equivalent to the products between the continuous explanatory values although here obtaining interactions requires nothing more than simple multiplication between two variables. However, the notation used to represent the interaction between factor A and factor B is A*B.

The interactions to be used in the model can be easily defined in XLSTAT.

**Balanced and unbalanced MANOVA**

We talk of balanced MANOVA when the numbers of categories are equal for all combination of factors. When the numbers of all categories for one of the combination of factors are not equal, then the MANOVA is said to be unbalanced. XLSTAT can handle both cases.

**Nested effects**

When constraints prevent us from crossing every level of one factor with every level of the other factor, nested factors can be used. We say we have a nested effect when fewer than all levels of one factor occur within each level of the other factor. An example of this might be if we want to study the effects of different machines and different operators on some output characteristic, but we can't have the operators change the machines they run. In this case, each operator is not crossed with each machine but rather only runs one machine.

Nested effects are automatically treated in the XLSTAT MANOVA.
Constraints

During the calculations, each factor is broken down into a sub-matrix containing as many columns as there are categories in the factor. Typically, this is a full disjunctive table. Nevertheless, the breakdown poses a problem: if there are g categories, the rank of this sub-matrix is not g but g-1. This leads to the requirement to delete one of the columns of the sub-matrix and possibly to transform the other columns. The strategy taken in XLSTAT is the following:

\[ a_1 = 0 \]: the parameter for the first category is null. This choice allows us to force the effect of the first category as a standard. In this case, the constant of the model is equal to the mean of the dependent variable for group 1.

Moreover, the number of observations should be equal at least to the sum of the number of dependent variables and the number of factors and interactions included in the model (+1).

Multivariate tests

One of the main applications of the MANOVA is multivariate comparison testing where parameters for the various categories of a factor are tested to be significantly different or not. For example, in the case where four treatments are applied to plants, we want to know if treatments have a significant effect and also if treatments have different effects.

Numerous tests have been proposed to compare means of each category. Most of them rely on the relationships that exist between the error matrix E and the matrix symbolizing the tested hypotheses H that is the eigenvalues of the matrix \( S^{-1} H \). XLSTAT provides the main tests including:

**Wilks Lambda test**: the likelihood ratio test statistic also known as Wilks Lambda (1932) is given by:

\[
\Lambda = \prod_{i=1}^{g} \frac{1}{1 + \lambda_i}
\]

The null hypothesis is rejected for small values of Lambda, indicating that the error E is small compared to the total SSCP matrix E+H. This test is the most frequently used.

**Hotelling-Lawley’s Trace test**:

\[
T_{HL} = \sum_{i=2}^{g} \lambda_i = \text{Trace}(E^{-1} H)
\]

A large H compared to E indicates a larger trace. Hence, the null hypothesis of no effects rejected for large values of \( T_{HL} \). This test is efficient if all factors have exactly two levels.
Pillai's Trace test:

\[ T_P = \sum_{i=1}^{k} \frac{\lambda_i}{1 + \lambda_i} = \text{Trace} ((E + H)^{-1} H) \]

As with Hotelling-Lawley's trace, the null hypothesis is rejected for large values of \( T_P \), indicating a large \( H \) relative to \( E \). This test is efficient if all samples have the same number of observations.

Roy's greatest root test:

\[ \lambda_{\text{max}} = \max_{1 \leq i \leq n} \lambda_i \]

The computed p-value for this test is always smaller than from other tests. Roy's test is a powerful but not robust test. For this reason, it's not recommended.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload default options**: Click this button to reload the default options.
- **Delete data selections**: Click this button to delete the data selections.
- **Change data handling**: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.
**General tab:**

**Y / Dependent variables:**
Select the response variable(s) you want to model. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

**X / Explanatory variables:**
Select the qualitative explanatory variables (the factors) in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Variable labels" option has been activated.

**Range:** Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Activate this option to display the results in a new worksheet of the active workbook.

**Workbook:** Activate this option to display the results in a new workbook.

**Variable labels:** Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

**Observation weights:** Activate this option if the observations are weighted. If you do not activate this option, the weights will all be taken as 1. Weights must be greater than or equal to 0. A weight of 2 is equivalent to repeating the same observation twice. If a column header has been selected, check that the "Variable labels" option has been activated.

**Options tab:**

**Interactions / Level:** Activate this option to include interactions in the model then enter the maximum interaction level (value between 1 and 4).

**Confidence interval (%):** Enter the percentage range of the confidence interval to use for the various tests and for calculating the confidence intervals around the parameters and predictions. Default value: 95.

**Wilks:** Activate this option if you want to run Wilks Lambda test.

**Hotelling-Lawley:** Activate this option if you want to run Hotelling-Lawley’s trace test.

**Pillai:** Activate this option if you want to run Pillai’s trace test.
Roy: Activate this option if you want to run Roy’s greatest root test.

Missing data tab:
Remove observations: Activate this option to remove the observations with missing data.
Do not allow missing values observations: Activate this option to avoid missing values.

Outputs tab:
Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.
SSCP matrices: Activate this option to display the SSCP matrices for factors and interactions.

Results
Summary statistics: The tables of descriptive statistics show the simple statistics for all the variables selected. The number of observations, missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the dependent variables (in blue) and the quantitative explanatory variables. For qualitative explanatory variables the names of the various categories are displayed together with their respective frequencies.
SSCP matrices: These tables are displayed to give a general view of the effects of the factors and interactions between factors.

When a k-way MANOVA is applied and the corresponding option is enabled, the results of the multivariate tests are displayed. The associated statistics, the degrees of freedom and the p-values are displayed.

Example
A tutorial on two-way MANOVA and multiple comparisons tests is available on the Addinsoft website:
http://www.xlstat.com/demo-mano.htm

A tutorial on three-way MANOVA is available on the Addinsoft website:
http://www.xlstat.com/demo-mano2.htm
References


Zetterberg, P. (2013). Effects of unbalancedness and heteroscedasticity on two way MANOVA., Department of statistics, Stockholm University.
Logistic regression

Use logistic regression to model a binary or polytomous variable using quantitative and/or qualitative explanatory variables.

Description

Logistic regression is a frequently-used method as it enables binary variables, the sum of binary variables, polytomous variables (variables with more than two categories) or ordinal variables (polytomous variables with ordered values) to be modeled. It is frequently used in the medical domain (whether a patient will get well or not), in sociology (survey analysis), epidemiology and medicine, in quantitative marketing (whether or not products are purchased following an action) and in finance for modeling risks (scoring).

The principle of the logistic regression model is to link the occurrence or non-occurrence of an event to explanatory variables. For example, in the phytosanitary domain, we are seeking to find out from which dosage of a chemical agent an insect will be neutralized.

Models

Logistic and linear regression belong to the same family of models called GLM (Generalized Linear Models): in both cases, an event is linked to a linear combination of explanatory variables.

For linear regression, the dependent variable follows a normal distribution \( N (\mu, \sigma) \) where \( \mu \) is a linear function of the explanatory variables. For logistic regression, the dependent variable, also called the response variable, follows a Bernoulli distribution for parameter \( p \) (\( p \) is the mean probability that an event will occur) when the experiment is repeated once, or a Binomial \((n, p)\) distribution if the experiment is repeated \( n \) times (for example the same dose tried on \( n \) insects). The probability parameter \( p \) is here a linear combination of explanatory variables.

The most common functions used to link probability \( p \) to the explanatory variables are the logistic function (we refer to the Logit model) and the standard normal distribution function (the Probit model). Both these functions are perfectly symmetric and sigmoid: XLSTAT provides two other functions: the complementary Log-log function is closer to the upper asymptote. The Gompertz function is on the contrary closer the axis of abscissa.

The analytical expression of the models is as follows:

\[
p = \frac{\exp(\beta X)}{1 + \exp(\beta X)}
\]
Probit: \[ p = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[ -\frac{x^2}{2} \right] dx \]

Complementary Log-log: \[ p = 1 - \exp\left[ -\exp\left( \beta X \right) \right] \]

Gompertz: \[ p = \exp\left[ -\exp\left( -\beta X \right) \right] \]

Where \( \beta X \) represents the linear combination of variables (including constants).

The knowledge of the distribution of the event being studied gives the likelihood of the sample. To estimate the \( \beta \) parameters of the model (the coefficients of the linear function), we try to maximize the likelihood function. Contrary to linear regression, an exact analytical solution does not exist. So an iterative algorithm has to be used. XLSTAT uses a Newton-Raphson algorithm. The user can change the maximum number of iterations and the convergence threshold if desired.

**Separation problem**

In the example above, the treatment variable is used to make a clear distinction between the positive and negative cases.

<table>
<thead>
<tr>
<th>Treatment 1</th>
<th>Treatment 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response +</td>
<td>121</td>
</tr>
<tr>
<td>Response -</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>85</td>
</tr>
</tbody>
</table>

In such cases, there is an indeterminacy on one or more parameters for which the variance is as high as the convergence threshold is low which prevents a confidence interval around the parameter from being given. To resolve this problem and obtain a stable solution, Firth (1993) proposed the use of a *penalized likelihood* function. XLSTAT offers this solution as an option and uses the results provided by Heinze (2002). If the standard deviation of one of the parameters is very high compared with the estimate of the parameter, it is recommended to restart the calculations with the "Firth" option activated.

**Confidence interval**

In most software, the calculation of confidence intervals for parameters is as for linear regression assuming that the parameters are normally distributed. XLSTAT also offers the alternative "profile likelihood" method (Venzon and Moolgavkar, 1988). This method is more reliable as it does not require the assumption that the parameters are normally distributed. Being iterative, however, it can slow down the calculations.
The multinomial logit model

The multinomial logit model, that correspond to the case where the dependent variable has more than two categories, has a different parameterization from the logit model because the response variable has more than two categories. It focuses on the probability to choose one of the J categories knowing some explanatory variables.

The analytical expression of the model is as follows:

\[
\log \left( \frac{p(y = j | x_i)}{p(y = 1 | x_i)} \right) = \alpha_j + \beta_j x_i
\]

where the category 1 is called the reference or control category. All obtained parameters have to be interpreted relatively to this reference category.

The probability to choose category j is:

\[
p(y = j | x_i) = \frac{\exp(\alpha_j + \beta_j x_i)}{1 + \sum_{k=2}^{J} \exp(\alpha_k + \beta_k x_i)}
\]

For the reference category, we have:

\[
p(y = 1 | x_i) = \frac{1}{1 + \sum_{k=2}^{J} \exp(\alpha_k + \beta_k x_i)}
\]

The model is estimated using a maximum likelihood method; the log-likelihood is as follows:

\[
l(\alpha, \beta) = \sum_{i=1}^{n} \sum_{j=1}^{J} y_{ij} \log \left( p(y = j | x_i) \right)
\]

To estimate the \( \beta \) parameters of the model (the coefficients of the linear function), we try to maximize the likelihood function. Contrary to linear regression, an exact analytical solution does not exist. XLSTAT uses the Newton-Raphson algorithm to iteratively find a solution.

Some results that are displayed for the logistic regression are not applicable in the case of the multinomial case.

The ordinal logit model

The ordinal logit model, that corresponds to the case where the dependent variable has more than two categories and where these categories are ordered in a specific way, has a different parameterization from the logit model. It focuses on the probability to choose one of the J categories knowing some explanatory variables. It is based on the cumulative probabilities and the cumulative logit method.
The analytical expression of the model is as follows (when the logit link function is used):

$$\log\left( \frac{p(y \leq j|x_i)}{p(y > j|x_i)} \right) = \alpha_j + \beta X_i$$

We can see that there is one intercept for each category and only one set of beta coefficients.

The probability to choose category $j$ or a lower category is:

$$p(y \leq j|x_i) = \frac{\exp(\alpha_j + \beta X_i)}{1 + \exp(\alpha_j + \beta X_i)}$$

This probability is equal to 1 when $j=J$. We can obtain the probability to choose category $j$:

$$p(y = j|x_i) = p(y \leq j|x_i) - p(y \leq j-1|x_i)$$

The model is estimated using a maximum likelihood method; the log-likelihood is as follows:

$$l(\alpha, \beta) = \sum_{i=1}^{n} \sum_{j=1}^{J} y_{ij} \log(p(y \leq j|x_i) - p(y \leq j-1|x_i))$$

To estimate the $\beta$ parameters and the $(J-1)\alpha$ parameters of the model (the coefficients of the linear function), we try to maximize the likelihood function. Contrary to linear regression, an exact analytical solution does not exist. XLSTAT uses the Newton-Raphson algorithm to iteratively find a solution.

The logit and probit link functions are available in XLSTAT.

Some results that are displayed for the logistic regression are not applicable in the case of the ordinal case.

**Marginal effects**

The calculation of marginal effects after a binomial logistic regression allows calculating the influence on the probability of occurrence of the event of interest, of an explanatory variable at a given point in the space of explanatory variables. XLSTAT provides the value of the marginal effects at the point corresponding to the means of the explanatory variables. The marginal effects are mainly of interest when compared to each other. By comparing them, one can measure the relative impact of each variable at the given point. The impact can be interpreted as the influence of a small variation. A confidence interval calculated using the Delta method is displayed. XLSTAT provides these results for both quantitative and qualitative variables, whether simple factors or interactions. For qualitative variables, the marginal effect indicates the impact of a change in category (from the first category to the category of interest).
**Percentage of well-classified observations and ROC curve**

XLSTAT can display the classification table (also called the confusion matrix) used to calculate the percentage of well-classified observations for a given cutoff point. Typically, for a cutoff value of 0.5, if the probability is less than 0.5, the observation is considered as being assigned to class 0, otherwise it is assigned to class 1.

The ROC curve can also be displayed. The ROC curve (*Receiver Operating Characteristics*) displays the performance of a model and enables a comparison to be made with other models. The terms used come from signal detection theory.

The proportion of well-classified positive events is called the *sensitivity*. The *specificity* is the proportion of well-classified negative events. If you vary the threshold probability from which an event is to be considered positive, the sensitivity and specificity will also vary. The curve of points (1-specificity, sensitivity) is the ROC curve.

Let's consider a binary dependent variable which indicates, for example, if a customer has responded favorably to a mail shot. In the diagram below, the blue curve corresponds to an ideal case where the n% of people responding favorably corresponds to the n% highest probabilities. The green curve corresponds to a well-discriminating model. The red curve (first bisector) corresponds to what is obtained with a random Bernoulli model with a response probability equal to that observed in the sample studied. A model close to the red curve is therefore inefficient since it is no better than random generation. A model below this curve would be disastrous since it would be less even than random.

![ROC Curve (AUC=0.836)](image)

The area under the curve (or *AUC*) is a synthetic index calculated for ROC curves. The AUC corresponds to the probability such that a positive event has a higher probability given to it by
the model than a negative event. For an ideal model, \( AUC = 1 \) and for a random model, \( AUC = 0.5 \). A model is usually considered good when the AUC value is greater than 0.7. A well-discriminating model must have an AUC of between 0.87 and 0.9. A model with an AUC greater than 0.9 is excellent.

**Hosmer-Lemeshow Test**

The Hosmer-Lemeshow test is a goodness of fit test for a binary logit model. It uses a statistic that follows a Chi-square distribution.

The calculation of this statistic is separated into several steps:

- The sample is ordered according to the probabilities calculated from the model in a decreasing way.
- The sample is divided into \( k \) parts of equal size.
- The Hosmer-Lemeshow statistic is calculated using the following formula:

\[
S_{HL} = \sum_{i=1}^{k} \frac{(O(i) - n_i P(i))}{n_i P(i) (1 - P(i))}
\]

with \( n_i \) being the group size, \( O(i) \) the number of times \( y = 1 \) in group \( i \) and \( P(i) \) the mean probability obtained from the model for group \( i \).

This statistic follows a Chi-square distribution with \( k-2 \) degrees of freedom. XLSTAT uses \( k=10 \).

When this statistic is large and the p-value is small, then this shows a lackness of fit of the model (poor fit).

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

**OK**: Click this button to start the computations.

**Cancel**: Click this button to close the dialog box without doing any computation.
General tab:

Dependent variables:

**Response variable(s):** Select the response variable(s) you want to model. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

**Response type:** Choose the type of response variable you have selected:

- **Binary variable:** If you select this option, you must select a variable containing exactly two distinct values. If the variable has value 0 and 1, XLSTAT will see to it that the high probabilities of the model correspond to category 1 and that the low probabilities correspond to category 0. If the variable has two values other than 0 or 1 (for example Yes/No), the lower probabilities correspond to the first category and the higher probabilities to the second.

- **Sum of binary variables:** If your response variable is a sum of binary variables, it must be of type numeric and contain the number of positive events (event 1) amongst those observed. The variable corresponding to the total number of events observed for this observation (events 1 and 0 combined) must then be selected in the "Observation weights" field. This case corresponds, for example, to an experiment where a dose D (D is the explanatory variable) of a medicament is administered to 50 patients (50 is the value of the observation weights) and where it is observed that 40 get better under the effects of the dose (40 is the response variable).

- **Multinomial:** if your response variable has more than two categories, a multinomial logit model is estimated. A new field called "control category" appears. You can select the reference category.

- **Ordinal:** if your response variable has ordered categories, an ordinal logit model is estimated. The reference category is the lower category. The type of data has to be numeric with a limited number of categories.
Explanatory variables:

**Quantitative**: Activate this option if you want to include one or more quantitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The data selected may be of the numerical type. If the variable header has been selected, check that the "Variable labels" option has been activated.

**Qualitative**: Activate this option if you want to include one or more qualitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Variable labels" option has been activated.

Method: Choose the logistic regression method to be used:

- **Classic**: Activate this option to calculate a logistic regression on the variables selected in the previous operations.
- **PCR**: Activate this option to calculate a logistic regression on the principal components extracted from the selected explanatory variables.

Model: Choose the type of function to use (see description).

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 ...).
**Observation weights:** This field must be entered if the "sum of binary variables" option has been chosen. Otherwise, this field is not active. If a column header has been selected, check that the "Variable labels" option has been activated.

**Regression weights:** Activate this option if you want to weight the influence of observations to adjust the model. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

**Control category:** In the multinomial case, you need to choose which category is the control.

**Options** tab:

**Tolerance:** Enter the value of the tolerance threshold below which a variable will automatically be ignored.

**Firth's method:** Activate this option to use Firth's penalized likelihood (see description). This option is only available for binary logit model.

**Interactions / Level:** Activate this option to include interactions in the model then enter the maximum interaction level (value between 1 and 4).

**Confidence interval (%%):** Enter the percentage range of the confidence interval to use for the various tests and for calculating the confidence intervals around the parameters and predictions. Default value: 95.

**Stop conditions:**

- **Iterations:** Enter the maximum number of iterations for the Newton-Raphson algorithm. The calculations are stopped when the maximum number of iterations has been exceeded. Default value: 100.

- **Convergence:** Enter the maximum value of the evolution of the log of the likelihood from one iteration to another which, when reached, means that the algorithm is considered to have converged. Default value: 0.000001.

**Options specific to the PCR regression**

**Filter factors:** You can activate one of the following two options in order to reduce the number of factors for which results are displayed.

- **Minimum %:** Activate this option then enter the minimum percentage of the total variability that the chosen factors must represent.

- **Maximum Number:** Activate this option to set the number of factors to take into account.
Options specific to the logistic regression

Model selection: Activate this option if you want to use one of the five selection methods provided:

- **Best model**: This method lets you choose the best model from amongst all the models which can handle a number of variables varying from "Min variables" to "Max Variables". Furthermore, the user can choose several "criteria" to determine the best model.
  - **Criterion**: Choose the criterion from the following list: Likelihood, LR (likelihood ratio), Score, Wald, Akaike's AIC, Schwarz's SBC.
  - **Min variables**: Enter the minimum number of variables to be used in the model.
  - **Max variables**: Enter the maximum number of variables to be used in the model.

Note: although XLSTAT uses a very powerful algorithm to reduce the number of calculations required as much as possible, this method can require a long calculation time. The method is only available for binary logit model.

- **Stepwise (Forward)**: The selection process starts by adding the variable with the largest contribution to the model. If a second variable is such that its entry probability is greater than the **entry threshold value**, then it is added to the model. After the third variable is added, the impact of removing each variable present in the model after it has been added is evaluated. If the probability of the calculated statistic is greater than the **removal threshold value**, the variable is removed from the model.

- **Stepwise (Backward)**: This method is similar to the previous one but starts from a complete model.

- **Forward**: The procedure is the same as for stepwise selection except that variables are only added and never removed.

- **Backward**: The procedure starts by simultaneously adding all variables. The variables are then removed from the model following the procedure used for stepwise selection.

Validation tab:

**Validation**: Activate this option if you want to use a sub-sample of the data to validate the model.

**Validation set**: Choose one of the following options to define how to obtain the observations used for the validation:

- **Random**: The observations are randomly selected. The "Number of observations" N must then be specified.
• **N last rows**: The N last observations are selected for the validation. The “Number of observations” N must then be specified.

• **N first rows**: The N first observations are selected for the validation. The “Number of observations” N must then be specified.

• **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to use for the validation.

**Prediction** tab:

**Prediction**: Activate this option if you want to select data to use them in prediction mode. If activate this option, you need to make sure that the prediction dataset is structured as the estimation dataset: same variables with the same order in the selections. On the other hand, variable labels must not be selected: the first row of the selections listed below must correspond to data.

**Quantitative**: Activate this option to select the quantitative explanatory variables. The first row must not include variable labels.

**Qualitative**: Activate this option to select the qualitative explanatory variables. The first row must not include variable labels.

**Observations labels**: Activate this option if observations labels are available. Then select the corresponding data. If this option is not activated, the observations labels are automatically generated by XLSTAT (PredObs1, PredObs2 …).

**Missing data** tab:

**Remove observations**: Activate this option to remove the observations with missing data.

**Estimate missing data**: Activate this option to estimate missing data before starting the computations.

• **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

• **Nearest neighbour**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.

**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the variables selected.
**Correlations**: Activate this option to display the explanatory variables correlation matrix.

**Goodness of fit statistics**: Activate this option to display the table of goodness of fit statistics for the model.

**Hosmer-Lemeshow test**: Activate this option to display the results of the Hosmer-Lemeshow test.

**Type III analysis**: Activate this option to display the type III analysis of variance table.

**Model coefficients**: Activate this option to display the table of coefficients for the model. Optionally, **confidence intervals** of type "profile likelihood" can be calculated (see description).

**Standardized coefficients**: Activate this option if you want the standardized coefficients (beta coefficients) for the model to be displayed.

**Marginal effects**: Activate this option if you want the marginal effects at the means to be displayed.

**Equation**: Activate this option to display the equation for the model explicitly.

**Predictions and residuals**: Activate this option to display the predictions and residuals for all the observations.

**Multiple comparisons**: This option is only active if qualitative explanatory variables have been selected. Activate this option to display the results of the comparison tests.

**Probability analysis**: If only one explanatory variable has been selected, activate this option so that XLSTAT calculates the value of the explanatory variable corresponding to various probability levels.

**Classification table**: Activate this option to display the posterior observation classification table using a **cutoff point** to be defined (default value 0.5).

**Options specific to the classical PCR logistic regression**:

**Factor loadings**: Activate this option to display the coordinates of the variables (*factor loadings*). The coordinates are equal to the correlations between the principal components and the initial variables for normalized PCA.

**Components/Variables correlations**: Activate this option to display correlations between the principal components and the initial variables.
**Factor scores:** Activate to display the coordinates of the observations (factor scores) in the new space created by PCA. The principal components are afterwards used as explanatory variables in the regression.

**Charts tab:**

**Regression charts:** Activate this option to display regression chart:

- **Standardized coefficients:** Activate this option to display the standardized parameters for the model with their confidence interval on a chart.

- **Predictions:** Activate this option to display the regression curve.
  - **Confidence intervals:** Activate this option to have confidence intervals displayed on charts (1) and (4).

**Options specific to the PCR logistic regression:**

**Correlations charts:** Activate this option to display charts showing the correlations between the components and initial variables.

- **Vectors:** Activate this option to display the input variables in the form of vectors.

**Observations charts:** Activate this option to display charts representing the observations in the new space.

- **Labels:** Activate this option to have observation labels displayed on the charts. The number of labels displayed can be changed using the filtering option.

**Biplots:** Activate this option to display charts representing the observations and variables simultaneously in the new space.

- **Vectors:** Activate this option to display the initial variables in the form of vectors.

- **Labels:** Activate this option to have observation labels displayed on the biplots. The number of labels displayed can be changed using the filtering option.

**Colored labels:** Activate this option to show variable and observation labels in the same color as the corresponding points. If this option is not activated the labels are displayed in black color.

**Filter:** Activate this option to modulate the number of observations displayed:

- **Random:** The observations to display are randomly selected. The “Number of observations” N to display must then be specified.
• **N first rows**: The N first observations are displayed on the chart. The “Number of observations” N to display must then be specified.

• **N last rows**: The N last observations are displayed on the chart. The “Number of observations” N to display must then be specified.

• **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to display.

**Results**

XLSTAT displays a large number of tables and charts to help in analyzing and interpreting the results.

**Summary statistics**: This table displays descriptive statistics for all the variables selected. For the quantitative variables, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed. For qualitative variables, including the dependent variable, the categories with their respective frequencies and percentages are displayed.

**Correlation matrix**: This table displays the correlations between the explanatory variables.

**Correspondence between the categories of the response variable and the probabilities**: This table shows which categories of the dependent variable have been assigned probabilities 0 and 1. It is only available for binary dependent variables.

**Summary of the variables selection**: Where a selection method has been chosen, XLSTAT displays the selection summary. For a stepwise selection, the statistics corresponding to the different steps are displayed. Where the best model for a number of variables varying from p to q has been selected, the best model for each number of variables is displayed with the corresponding statistics and the best model for the criterion chosen is displayed in bold.

**Goodness of fit coefficients**: This table displays a series of statistics for the independent model (corresponding to the case where the linear combination of explanatory variables reduces to a constant) and for the adjusted model.

• **Observations**: The total number of observations taken into account (sum of the weights of the observations);

• **Sum of weights**: The total number of observations taken into account (sum of the weights of the observations multiplied by the weights in the regression);

• **DF**: Degrees of freedom;
• **-2 Log(Like.):** The logarithm of the likelihood function associated with the model;

• **R² (McFadden):** Coefficient, like the R², between 0 and 1 which measures how well the model is adjusted. This coefficient is equal to 1 minus the ratio of the likelihood of the adjusted model to the likelihood of the independent model;

• **R²(Cox and Snell):** Coefficient, like the R², between 0 and 1 which measures how well the model is adjusted. This coefficient is equal to 1 minus the ratio of the likelihood of the adjusted model to the likelihood of the independent model raised to the power 2/Sw, where Sw is the sum of weights.

• **R²(Nagelkerke):** Coefficient, like the R², between 0 and 1 which measures how well the model is adjusted. This coefficient is equal to ratio of the R² of Cox and Snell, divided by 1 minus the likelihood of the independent model raised to the power 2/Sw;

• **AIC:** Akaike’s Information Criterion;

• **SBC:** Schwarz’s Bayesian Criterion.

• **Iterations:** Number of iterations before convergence.

**Test of the null hypothesis H₀: Y=p₀:** The H₀ hypothesis corresponds to the independent model which gives probability p₀ whatever the values of the explanatory variables. We seek to check if the adjusted model is significantly more powerful than this model. Three tests are available: the likelihood ratio test (-2 Log(Like.)), the Score test and the Wald test. The three statistics follow a Chi² distribution whose degrees of freedom are shown.

**Type III analysis:** This table is only useful if there is more than one explanatory variable. Here, the adjusted model is tested against a test model where the variable in the row of the table in question has been removed. If the probability Pr > LR is less than a significance threshold which has been set (typically 0.05), then the contribution of the variable to the adjustment of the model is significant. Otherwise, it can be removed from the model.

For PCR logistic regression, the first table of the model parameters corresponds to the parameters of the model which use the principal components which have been selected. This table is difficult to interpret. For this reason, a transformation is carried out to obtain model parameters which correspond to the initial variables.

Model parameters:

• **Binary case:** The parameter estimate, corresponding standard deviation, Wald’s Chi², the corresponding p-value and the confidence interval are displayed for the constant and each variable of the model. If the corresponding option has been activated, the "profile likelihood" intervals are also displayed.

• **Multinomial case:** In the multinomial case, (J-1)*(p+1) parameters are obtained, where J is the number of categories and p is the number of variables in the model. Thus, for
each explanatory variable and for each category of the response variable (except for the reference category), the parameter estimate, corresponding standard deviation, Wald's Chi^2, the corresponding p-value and the confidence interval are displayed. The odds-ratios with corresponding confidence interval are also displayed.

- **Ordinal case**: In the ordinal case, \((J-1)+p\) parameters are obtained, where \(J\) is the number of categories and \(p\) is the number of variables in the model. Thus, for each explanatory variable and for each category of the response variable, the parameter estimate, corresponding standard deviation, Wald's Chi^2, the corresponding p-value and the confidence interval are displayed.

The **equation of the model** is then displayed to make it easier to read or re-use the model. It is only displayed for the binary case.

The table of **standardized coefficients** (also called beta coefficients) are used to compare the relative weights of the variables. The higher the absolute value of a coefficient, the more important the weight of the corresponding variable. When the confidence interval around standardized coefficients has value 0 (this can easily be seen on the chart of standardized coefficients), the weight of a variable in the model is not significant.

The **marginal effects** at the point corresponding to the means of the explanatory variables are then displayed. The marginal effects are mainly of interest when compared to each other. By comparing them, one can measure the relative impact of each variable at the given point. The impact can be interpreted as the influence of a small variation. A confidence interval calculated using the Delta method is displayed. XLSTAT provides these results for both quantitative and qualitative variables, whether simple factors or interactions. For qualitative variables, the marginal effect indicates the impact of a change in category (from the first category to the category of interest).

The **predictions and residuals** table shows, for each observation, its weight, the value of the qualitative explanatory variable, if there is only one, the observed value of the dependent variable, the model's prediction, the same values divided by the weights, the standardized residuals and a confidence interval. In the ordinal case, the probability associated to each category is displayed.

This **classification table** displays the table showing the number of well-classified and misclassified observations for both categories. The sensitivity, specificity and the overall percentage of well-classified observations are also displayed. If a validation sample has been extracted, this table is also displayed for the validation data.

**ROC curve**: The ROC curve is used to evaluate the performance of the model by means of the area under the curve (AUC) and to compare several models together (see the description section for more details).
Comparison of the categories of the qualitative variables: If one or more explanatory qualitative variables have been selected, the results of the equality tests for the parameters taken in pairs from the different qualitative variable categories are displayed.

If only one quantitative variable has been selected, the probability analysis table allows to see to which value of the explanatory variable corresponds a given probability of success.

Example

Tutorials on how to use logistic regression and the multinomial logit model are available on the Addinsoft website:


References


Log-linear regression

Use this tool to fit a log-linear regression model with three possible probability distributions (Poisson, Gamma, and Exponential).

Description

The log-linear regression is used to model data by a log-linear combination of the model parameters and the covariates (qualitative or quantitative). Furthermore, we assume that the data (response variable) are distributed either according to a Poisson, Gamma or exponential distribution.

The log-linear regression model

Denote by $Y$ the response variable vector, $\beta$ the vector of model parameters and $X$ the matrix of the $p$ covariates. The first column of $X$ is composed by a vector of 1s that deals with the intercept of the model. The log-linear model is given by:

$$E(Y \mid X) = e^{\beta'X}$$

According to the previous equation, we directly obtain:

$$\log[E(Y \mid X)] = \beta'X$$

Inference of the model parameters

If we assume that the variables $Y_i$ are independent from the vector of covariates $X_i$, the model parameters can be estimated by maximizing the likelihood. Whatever the probability distribution (Poisson, Gamma, Exponential), the likelihood function is convex and can be maximized using a Newton-Raphson algorithm.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.
OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

Reload: Click this button to reload the default options.

Delete: Click this button to delete the data selections.

Direction: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab:**

**Dependent variables:**

**Response variable(s):** Select the response variable(s) you want to model. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

**Explanatory variables:**

**Quantitative:** Activate this option if you want to include one or more quantitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The data selected may be of the numerical type. If the variable header has been selected, check that the "Variable labels" option has been activated.

**Qualitative:** Activate this option if you want to include one or more qualitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Variable labels" option has been activated.

**Offset:** Activate this option if you want to include an offset. This option is only available for the Poisson distribution.
**Distribution**: Select the probability distribution (Poisson, Gamma or Exponential).

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Variable labels**: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

**Observation labels**: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

**Regression weights**: Activate this option if you want to weight the influence of observations to adjust the model. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

**Options** tab:

**Tolerance**: Enter the value of the tolerance threshold below which a variable will automatically be ignored.

**Confidence interval (%)**: Enter the percentage range of the confidence interval to use for the various tests and for calculating the confidence intervals around the parameters and predictions. Default value: 95.

**Fixed intercept**: Activate this option to set the intercept (or constant) of the model to a given value. Then enter the value in the corresponding field (0 by default).

**Interactions / Level**: Activate this option to include interactions in the model then enter the maximum interaction level (value between 1 and 4).

**Stop conditions**
• **Iterations**: Enter the maximum number of iterations for the Newton-Raphson algorithm. The calculations are stopped when the maximum number of iterations has been exceeded. Default value: 100.

• **Convergence**: Enter the maximum value of the evolution of the log of the likelihood from one iteration to another which, when reached, means that the algorithm is considered to have converged. Default value: 0.000001.

**Model selection**: Activate this option if you want to use one of the four selection methods provided:

• **Stepwise (Forward)**: The selection process starts by adding the variable with the largest contribution to the model. If a second variable is such that its entry probability is greater than the **entry threshold value**, then it is added to the model. After the third variable is added, the impact of removing each variable present in the model after it has been added is evaluated. If the probability of the calculated statistic is greater than the **removal threshold value**, the variable is removed from the model.

• **Stepwise (Backward)**: This method is similar to the previous one but starts from a complete model.

• **Forward**: The procedure is the same as for stepwise selection except that variables are only added and never removed.

• **Backward**: The procedure starts by simultaneously adding all variables. The variables are then removed from the model following the procedure used for stepwise selection.

• **Criterion**: Choose the criterion from the following list: Likelihood, LR (likelihood ratio), Score, Wald, Akaike's AIC, Schwarz's SBC.

**Validation** tab:

**Validation**: Activate this option if you want to use a sub-sample of the data to validate the model.

**Validation set**: Choose one of the following options to define how to obtain the observations used for the validation:

• **Random**: The observations are randomly selected. The “Number of observations” N must then be specified.

• **N last rows**: The N last observations are selected for the validation. The “Number of observations” N must then be specified.

• **N first rows**: The N first observations are selected for the validation. The “Number of observations” N must then be specified.
• **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to use for the validation.

**Prediction** tab:

**Prediction**: Activate this option if you want to select data to use them in prediction mode. If activate this option, you need to make sure that the prediction dataset is structured as the estimation dataset: same variables with the same order in the selections. On the other hand, variable labels must not be selected: the first row of the selections listed below must correspond to data.

**Quantitative**: Activate this option to select the quantitative explanatory variables. The first row must not include variable labels.

**Qualitative**: Activate this option to select the qualitative explanatory variables. The first row must not include variable labels.

**Observations labels**: Activate this option if observations labels are available. Then select the corresponding data. If this option is not activated, the observations labels are automatically generated by XLSTAT (PredObs1, PredObs2 …).

**Missing data** tab:

**Remove observations**: Activate this option to remove the observations with missing data.

**Estimate missing data**: Activate this option to estimate missing data before starting the computations.

  - **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
  - **Nearest neighbour**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.

**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the variables selected.

**Correlations**: Activate this option to display the explanatory variables correlation matrix.

**Goodness of fit statistics**: Activate this option to display the table of goodness of fit statistics for the model.

**Type III analysis**: Activate this option to display the type III analysis of variance table.
**Model coefficients**: Activate this option to display the table of coefficients for the model.

**Equation**: Activate this option to display the equation for the model explicitly.

**Predictions and residuals**: Activate this option to display the predictions and residuals for all the observations.

**Overdispersion test**: Activate this option to test the overdispersion (only for the Poisson regression).

**Charts tab**:  
**Regression charts**: Activate this option to display regression charts.  
- **Confidence intervals**: Activate this option to display confidence intervals.

**Prediction chart**: Activate this option to display the prediction chart.  
- **Confidence intervals**: Activate this option to display confidence intervals.

**Results**

**Summary statistics**: This table displays descriptive statistics for all the variables selected. For the quantitative variables, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed. For qualitative variables, including the dependent variable, the categories with their respective frequencies and percentages are displayed.

**Correlation matrix**: This table displays the correlations between the explanatory variables.

**Correspondence between the categories of the response variable and the probabilities**:  
This table shows which categories of the dependent variable have been assigned probabilities 0 and 1. It is only available for binary dependent variables.

**Summary of the variables selection**: Where a selection method has been chosen, XLSTAT displays the selection summary. For a stepwise selection, the statistics corresponding to the different steps are displayed. Where the best model for a number of variables varying from p to q has been selected, the best model for each number or variables is displayed with the corresponding statistics and the best model for the criterion chosen is displayed in bold.
**Goodness of fit coefficients:** This table displays a series of statistics for the independent model (corresponding to the case where the linear combination of explanatory variables reduces to a constant) and for the adjusted model.

- **Observations:** The total number of observations taken into account (sum of the weights of the observations);

- **Sum of weights:** The total number of observations taken into account (sum of the weights of the observations multiplied by the weights in the regression);

- **DF:** Degrees of freedom;

- **-2 Log(Like.):** The logarithm of the likelihood function associated with the model;

- **R² (McFadden):** Coefficient, like the $R^2$, between 0 and 1 which measures how well the model is adjusted. This coefficient is equal to 1 minus the ratio of the likelihood of the adjusted model to the likelihood of the independent model;

- **R²(Cox and Snell):** Coefficient, like the $R^2$, between 0 and 1 which measures how well the model is adjusted. This coefficient is equal to 1 minus the ratio of the likelihood of the adjusted model to the likelihood of the independent model raised to the power $2/S_w$, where $S_w$ is the sum of weights.

- **R²(Nagelkerke):** Coefficient, like the $R^2$, between 0 and 1 which measures how well the model is adjusted. This coefficient is equal to ratio of the $R^2$ of Cox and Snell, divided by 1 minus the likelihood of the independent model raised to the power $2/S_w$;

- **Deviance:** Value of the deviance criterion for the adjusted model and the independent model;

- **Pearson Chi-square:** Value of the Pearson Chi-square for the adjusted model and the independent model;

- **AIC:** Akaike’s Information Criterion;

- **SBC:** Schwarz’s Bayesian Criterion.

- **Iterations:** Number of iterations before convergence.

**Nullity test:** These results allow checking whether fitted model is significantly more powerful than the independent model. Three tests are available: the likelihood ratio test (-2 Log(Like.)), the Score test and the Wald test. These three statistics follow a Chi-square distribution which degrees of freedom are shown.

**Type III analysis:** This table is only useful if there is more than one explanatory variable. Here, the adjusted model is tested against a test model where the variable in the row of the table in question has been removed. If the probability $Pr > LR$ is less than a significance threshold
which has been set (typically 0.05), then the contribution of the variable to the adjustment of the model is significant. Otherwise, it can be removed from the model.

**Model parameters:** For the constant and each variable in the model, the parameter estimate, its corresponding standard deviation, the Wald's Chi-square and the corresponding p-value and the confidence interval are displayed in this table.

**Example**

A tutorial on how to use log-linear regression is available on the Addinsoft website at:


**References**

Quantile Regression

Use quantile regression to model a quantitative response variable depending on quantitative or qualitative explanatory variables. Furthermore, quantile regression makes it possible to look beyond classical regression or ANCOVA, by extending the analysis limited to expected values to the entire distribution using quantiles.

Description

Quantile regression keeps growing in importance and interest since it was introduced by Koenker and Basset in 1978. The method popularity among the practitioners and also researchers’ community is without doubt due to its peculiarity to provide them a realistic framework to perform their studies.

Indeed, by nature, quantile regression enables to work with a wide range of distributions, without being subject to any restrictions such as normality assumption, thus contrasting with usual regression.

As a consequence of that flexibility, many fields find great interest in quantile regression including Economics, Social Sciences, Environments, Biometrics and Behavioral Sciences, among others.

The main contributions on the subject may be found in the References section.

Model

As in the ANCOVA framework, the dependent variable $Y$ is quantitative while the set of predictors $X$ can be composed not only with quantitative variables (including interactions between quantitative variables) but also with factors (qualitative variables, interactions between qualitative variables and interactions between quantitative and qualitative variables).

Nevertheless, it’s essential to keep in mind that, unlike ANCOVA, no hypotheses on the errors distribution is required.

Problem

The $\alpha$-th quantile, $\alpha \in [0, 1]$, is defined as the value $y$ s.t. $P(Y \leq y) = \alpha$. Introducing the cumulative distribution function $F$, the quantile function $Q$ is its inverse:

$$Q(\alpha) = F^{-1}(\alpha) = \inf \{ y : F(y) > \alpha \}$$
The mean \( \mu \) of the random variable \( Y \) can be characterized as the value:

\[
\mu = \arg\min \{ c : \mathbb{E} (Y - c)^2 \},
\]  

(1)

that minimizes the squared sum of deviations.

In the same way, the following assertion on the \( \alpha \)-th quantile, \( q\alpha \), holds:

\[
q\alpha = \arg\min\{ c : \mathbb{E} \rho_{\alpha} (Y - c) \},
\]

(2)

where \( q\alpha \) denotes the function:

\[
\rho_{\alpha} = \left[ \alpha I_{\{y \leq 0\}} \right] y = \left[ (1 - \alpha) I_{\{y \leq 0\}} + \alpha I_{\{y > 0\}} \right] y
\]

\( q\alpha \) thus minimizes a weighted sum of absolute deviations.

Coming back to our context in which \( Y \) is a dependent variable and \( X \) a set of explanatory variables and considering the linear framework, the minimization problem (1) becomes:

\[
\hat{\beta}_{\alpha} = \arg\min \{ \beta_{\alpha} : \mathbb{E} \left[ (Y - X^T \beta_{\alpha})^2 \right] \}
\]

In the same manner, (2) turns to:

\[
\hat{\beta}_{\alpha} = \arg\min \{ \beta_{\alpha} : \mathbb{E} \left[ \rho_{\alpha} (Y - X \beta_{\alpha}) \right] \}
\]

where the parameters and the associated estimators depend on \( \alpha \).

Instead of considering the conditional mean of the classical regression problem, the quantile regression problem consists in estimating conditional quantiles.
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload options**: Click this button to reload the default options.
- **Delete selections**: Click this button to delete the data selections.
- **Change data handling**: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab:**

**Y / Dependent variables:**

**Quantitative**: Select the response variable(s) you want to model. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

**X / Explanatory variables:**

**Quantitative**: Select the quantitative explanatory variables in the Excel worksheet. The data selected must be of type numeric. If the variable header has been selected, check that the "Variable labels" option has been activated.

**Qualitative**: Select the qualitative explanatory variables (the factors) in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Variable labels" option has been activated.
**Range:** Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Activate this option to display the results in a new worksheet of the active workbook.

**Workbook:** Activate this option to display the results in a new workbook.

**Variable labels:** Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

**Observation labels:** Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 ...).

**Observation weights:** Activate this option if the observations are weighted. If you do not activate this option, the weights will all be taken as 1. Weights must be greater than or equal to 0. A weight of 2 is equivalent to repeating the same observation twice. If a column header has been selected, check that the "Variable labels" option has been activated.

**Quantile(s):**

**Selection:** Activate this option to work with a quantile selection. Then, select the cells containing the order \( \alpha, \alpha \in [0,1] \), of the \( \alpha \)-th quantiles of interest in the Excel worksheet.

**Process:** Activate this option to get the entire quantile process. The number of calculated quantiles is specified thanks to an heuristic formula depending (in an increasing way) on the number of observations and regressors. The resulting quantile orders \( \alpha \) are then uniformly distributed on \([0,1]\).

**Options** tab:

**Algorithm:** 3 algorithms are available to compute the quantile regression coefficients:

- **Simplex:** Select this option to compute the Barrodale and Roberts algorithm based on simplex methods.
- **Interior point:** Select this option to compute the predictor-corrector Mehrotra algorithm based on interior point methods.
- **Smoothing function:** Select this option to compute the Clark and Osborne algorithm based on an approximation of the objective function with a smooth function whose minimization provides asymptotically the same results than the initial function. This algorithm is very competitive, especially when \( p/n > 0.05 \).
**Stop criterion**: The selected algorithm stops as soon as one of these events occurs:

- End of the algorithm  OR
- The maximum number of iterations specified in **Iterations** has been exceeded. Default value: 100.  OR
- The evolution of the results from one iteration to another is inferior to the **Convergence** value, the algorithm is considered to have converged. Default value: 0.000001.

**Confidence Interval (%)**: Enter the percentage range of the confidence interval to use for the various tests and for calculating the confidence intervals around the parameters and predictions. Default value: 95.

**Interactions / Level**: Activate this option to include interactions in the model then enter the maximum interaction level (value between 1 and 4).

**Constraints**: When qualitative explanatory variables have been selected, you can choose the constraints used on these variables:

- **a1 = 0**: Choose this option so that the parameter of the first category of each factor is set to 0.
- **an = 0**: Choose this option so that the parameter of the last category of each factor is set to 0.

**A priori error type**: Activate this option to precise, if possible, the error. Then, select the type: homogeneous error (i.i.d.), heterogeneous (i.n.i.d.) or dependent (n.i.i.d.) (for instance autocorrelated ones). This option will have an impact on the computation of the covariance matrix of the coefficients and their confidence intervals.

**Validation tab**:

**Validation**: Activate this option if you want to use a sub-sample of the data to validate the model.

**Validation set**: Choose one of the following options to define how to obtain the observations used for the validation:

- **Random**: The observations are randomly selected. The "Number of observations" N must then be specified.
- **N last rows**: The N last observations are selected for the validation. The “Number of observations” N must then be specified.

- **N first rows**: The N first observations are selected for the validation. The “Number of observations” N must then be specified.

- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to use for the validation.

**Prediction** tab:

**Prediction**: activate this option if you want to select data to use them in prediction mode. If activate this option, you need to make sure that the prediction dataset is structured as the estimation dataset: same variables with the same order in the selections. On the other hand, variable labels must not be selected: the first row of the selections listed below must correspond to data.

**Quantitative**: activate this option to select the quantitative explanatory variables. The first row must not include variable labels.

**Qualitative**: activate this option to select the qualitative explanatory variables. The first row must not include variable labels.

**Observations labels**: activate this option if observations labels are available. Then select the corresponding data. If this option is not activated, the observations labels are automatically generated by XLSTAT (PredObs1, PredObs2 …).

**Missing data** tab:

**Remove observations**: Activate this option to remove the observations with missing data.

**Estimate missing data**: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

- **Nearest neighbour**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.

**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the variables selected.
Quantile correlations: Activate this option to display the variables quantile correlation matrix.

Covariance matrix: Activate this option to display the covariance matrix of the quantile regression coefficients.

Goodness of fit statistics: Activate this option to display the table of goodness of fit statistics for the model.

Model significance test: Activate this option to test the significance of the model. More precisely, the hypothesis of the complete model is tested against the hypothesis of the model made up of the intercept. 3 tests are available:

- LR: Activate this option to compute the Likelihood Ratio test,
- LM: Activate this option to compute the Lagrange Multiplier test,
- Wald: Activate this option to compute the Wald test.

Model equation: Activate this option to display the model equation.

Predictions and residuals: Activate this option to display the predictions and the residuals for all the observations.

Computations based on:

- Asymptotic distribution: Activate this option to compute the covariance matrix and the confidence intervals thanks to the theoretical asymptotic distribution of the coefficients. This computation will take into account the A priori error type in the Options tab if informed.

- Resampling (Bootstrap): Activate this option to compute the empirical covariance matrix and the confidence intervals thanks to resampling (Bootstrap). Then, inform B = with an integer value to indicate how many samples will be simulated to compute the estimations.

- Hall and Sheather bandwidth: Activate this option to compute the covariance matrix and the confidence intervals using Hall and Sheather bandwidth ( $O(n^{-1/2})$ ).

- Bofinger bandwidth: Activate this option to compute the covariance matrix and the confidence intervals using Bofinger bandwidth ( $O(n^{-1/2})$ ).

Charts tab:
Regression charts: Activate this option to display the regression charts:

- Predictions and residuals: Activate this option to display the following charts:
  - Explanatory variable versus standardized residuals: this chart is displayed only if there is one explanatory variable and if that variable is quantitative.
  - Dependent variable versus standardized residuals.
  - Predictions versus observed values.

Results

If the quantile process is selected in the General tab, then a global table is displayed, summing up for each computed q-th quantile, the associated coefficients value.

Charts representing the behavior of these coefficients with respect to the value of \( \alpha \) are displayed for a better visualization of the results.

If a quantile selection is chosen in the General tab, then the following results are displayed:

Summary statistics: The tables of descriptive statistics show the simple statistics for all the variables selected. The number of observations, missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the dependent variables (in blue) and the quantitative explanatory variables. For qualitative explanatory variables the names of the various categories are displayed together with their respective frequencies.

Then for each quantile, the following results are displayed:

Goodness of fit statistics: Activate this option to display the table of goodness of fit statistics for the model:

- Observations: The number of observations used in the calculations. In the formulas shown below, \( n \) is the number of observations.
- Sum of weights: The sum of the weights of the observations used in the calculations. In the formulas shown below, \( W \) is the sum of the weights.
- DF: The number of degrees of freedom for the chosen model (corresponding to the error part).
• \( R^2_a \): The determination coefficient for the model. This coefficient, whose value is between 0 and 1. Its value is defined by:

\[
R^2_a = 1 - \frac{SAR_a}{SAT_a} = 1 - \frac{\sum_{i=1}^{n} \rho_\alpha (y_i - x_i^T \hat{\beta}_a)}{\sum_{i=1}^{n} \rho_\alpha (y_i - \hat{\beta}_0, a)}
\]

where \( \mathcal{F}_a \) is the \( \alpha \)-th empirical quantile of the observations of the dependent variable \( Y \).

\( RAS_a \) is the acronym for Residual Absolute Sum of weighted differences and \( TAS_a \) is the acronym for Total Absolute Sum of weighted differences.

The \( R^2_a \) is interpreted as the proportion of the variability of the dependent variable explained by the model. The nearer \( R^2_a \) is to 1, the better is the model. The problem with the \( R^2_a \) is that it does not take into account the number of variables used to fit the model.

• Adjusted \( R^2_a \): The adjusted determination coefficient for the model. The adjusted \( R^2_a \) can be negative if the \( R^2_a \) is near to zero. Its value is defined by:

\[
\text{Adjusted } R^2_a = 1 - (1-R^2_a) \frac{W-1}{W-p-1}
\]

The adjusted \( R^2_a \) is a correction to the \( R^2_a \) which takes into account the number of variables used in the model.

• \( MRAS_a \): The Mean Residual Absolute Sum (MRAS) is defined by:

\[
MRAS_a = \frac{1}{W-p} RAS_a
\]

• \( RMRAS_a \): the square Root of the Mean Residual Absolute Sum (RMRAS).

• \( MAPE_a \): The Mean Absolute Percentage Error is calculated as follows:

\[
MAPE_a = \frac{100}{W} \sum_{i=1}^{n} W_i \left| \frac{y_i - \hat{y}_{i,a}}{y_i} \right|
\]
• \( Cp_\alpha \) : Mallows \( CP_\alpha \) coefficient is defined by:

\[
Cp_\alpha = \frac{RAS_\alpha}{\hat{\sigma}_\alpha} + 2p - W
\]

where \( RAS_\alpha \) is the Residual Absolute Sum of weighted differences and \( \hat{\sigma}_\alpha \) denotes the variance estimator of the residuals. The nearer \( Cp_\alpha \) is to \( p \), the less the model is biased.

• \( AIC_\alpha \) : Akaike's Information Criterion is defined by:

\[
AIC_\alpha = W \ln\left(\frac{RAS_\alpha}{W}\right) + 2p
\]

This criterion, proposed by Akaike (1973) is derived from the information theory and uses Kullback and Leibler's measurement (1951). It is a model selection criterion which penalizes models for which adding new explanatory variables does not supply sufficient information to the model, the information being measured through the MSE. The aim is to minimize the AIC criterion.

• \( SBC_\alpha \) : Schwarz's Bayesian Criterion is defined by:

\[
SBC_\alpha = W \ln\left(\frac{RAS_\alpha}{W}\right) + \ln(W)p
\]

This criterion, proposed by Schwarz (1978) is similar to the AIC and, like this, the aim is to minimize it.

• \( PC_\alpha \) : Amemiya's Prediction Criterion is defined by:

\[
PC_\alpha = \frac{(1 - R_\alpha^2)(W + p)}{W - p}
\]

This criterion, proposed by Amemiya (1980) is used, like the adjusted \( R^2 \) to take account of the parsimony of the model.

The table of model parameters displays the estimate of the parameters, the corresponding standard error, as well as the confidence interval.

The model significance table is used to evaluate the explanatory power of the explanatory variables. The explanatory power is evaluated by comparing the fit of the final model with the
fit of the rudimentary model including only a constant equal to the quantile of the dependent variable.

The predictions and residuals table shows, for each observation, its weight, the value of the qualitative explanatory variable, if there is only one, the observed value of the dependent variable, the model's prediction, the residuals, the confidence. Two types of confidence interval are displayed: a confidence interval around the mean (corresponding to the case where the prediction would be made for an infinite number of observations with a set of given values for the explanatory variables) and an interval around the isolated prediction (corresponding to the case of an isolated prediction for the values given for the explanatory variables). The second interval is always greater than the first, the random values being larger. If the validation data have been selected, they are displayed at the end of the table.

The charts which follow show the results mentioned above. If there is only one explanatory variable in the model, the first chart displayed shows the observed values, the regression line and both types of confidence interval around the predictions. The second chart shows the standardized residuals as a function of the explanatory variable. In principle, the residuals should be distributed randomly around the X-axis. If there is a trend or a shape, this shows a problem with the model.

Example

A tutorial on how to run a quantile regression is available on the Addinsoft website at:

http://www.xlstat.com/demo-quantilereg.htm

References


Cubic splines

This tool allows to fit a cubic spline using a set of nodes defined by the user.

Description

A cubic spline is defined as a piecewise function of polynomials of degree 3. Cubic splines are used in interpolation problems where they are preferred to usual polynomial interpolation methods, because they allow a compromise between the smoothness of the curve and the degree of the polynomials.

Cubic splines

A cubic spline $S$ is a piecewise function defined on an interval $[a,b]$ divided into $K$ intervals $[x_{i-1},x_i]$ such that

$$a = x_0 < x_1 < \ldots < x_{K-1} < x_k = b$$

and it is defined by $P_i$ the polynomial of degree 3 on the interval $[x_{i-1},x_i]$. The spline $S$ is given by:

$$
\begin{align*}
S(t) &= P_1(t) \text{ if } t \in [x_0,x_1] \\
& \vdots \\
S(t) &= P_k(t) \text{ if } t \in [x_{K-1},x_k]
\end{align*}
$$

The calculation of the coefficients of the cubic spline involves the derivatives of the polynomials (for further details see Guillod, 2008).

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

OK: Click this button to start the computations.
Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

Reload: Click this button to reload the default options.

Delete: Click this button to delete the data selections.

Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

Y: Select the data corresponding to the ordinates. If the name of the variable is available in the first position of the data, make sure you activate the “Variable labels” option.

X: Select the data that correspond to the abscissa. If the name of the variable is available in the first position of the data, make sure you activate the “Variable labels” option.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Variable labels**: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

**Observation labels**: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels” option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

**Options** tab:

**Data as nodes**: Activate this option to use the data as nodes for the cubic spline.
**Number of nodes**: Activate this option to select the number of nodes. This nodes are then equally distributed.

**Select the nodes coordinates**: If the option is enabled, you have to select the range containing the coordinates of the nodes.

**Prediction** tab:

**Prediction**: Activate this option if you want to select data to use them in prediction mode. If you activate this option, you need to make sure that the prediction dataset is structured as the estimation dataset: same variables with the same order in the selections. On the other hand, variable labels must not be selected: the first row of the selections listed below must correspond to data.

**Observations**: Select the variables for prediction. The first row must not include variable labels.

**Observations labels**: Activate this option if observations labels are available. Then select the corresponding data. If this option is not activated, the observations labels are automatically generated by XLSTAT (PredObs1, PredObs2 …).

**Missing data** tab:

**Do not accept missing data**: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

**Remove observations**: Activate this option to remove the observations with missing data.

**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the variables selected.

**Correlations**: Activate this option to display the correlation matrix.

**Predictions and residuals**: Activate this option to display the predictions and residuals for all the observations.

**Charts** tab:

**Spline curve**: Activate this option to display the spline curve

- **Predictions and residuals**: Activate this option to display the following charts.
(1) Explanatory variable versus standardized residuals: This chart is only displayed if there is only one explanatory variable and this variable is quantitative.

(2) Dependent variable versus standardized residuals.

(3) Predictions for the dependent variable versus the dependent variable.

(4) Bar chart of standardized residuals.

Results

Summary statistics: This table displays the descriptive statistics for each element

Coefficients of the cubic spline: for each interval the coefficients of the cubic spline are given in a table.

Example

An example of the use of cubic splines is available on Addinsoft website:

http://www.xlstat.com/demo-splines.htm

References

Nonparametric regression

This tool carries out two types of nonparametric regression: *Kernel regression* and LOWESS regression.

Description

Parametric regression can be used when the hypotheses about the more classical regression methods cannot be verified or when we are mainly interested in only the predictive quality of the model and not its structure.

**Kernel regression**

*Kernel regression* is a modeling tool which belongs to the family of smoothing methods. Unlike linear regression which is both used to explain phenomena and for prediction (understanding a phenomenon to be able to predict it afterwards), *Kernel regression* is mostly used for prediction. The structure of the model is variable and complex, the latter working like a filter or black box. There are many variations of *Kernel regression* in existence.

As with any modeling method, a learning sample of size $n_{\text{learn}}$ is used to estimate the parameters of the model. A sample of size $n_{\text{valid}}$ can then be used to evaluate the quality of the model. Lastly, the model can be applied to a prediction sample of size $n_{\text{pred}}$, for which the values of the dependent variable $Y$ are unknown.

The first characteristic of *Kernel Regression* is the use of a **kernel** function, to weigh the observations of the learning sample, depending on their "distance" from the predicted observation. The closer the values of the explanatory variables for a given observation of the learning sample are to the values observed for the observation being predicted, the higher the weight. Many kernel functions have been suggested. XLSTAT includes the following kernel functions: Uniform, Triangle, Epanechnikov, Quartic, Triweight, Tricube, Gaussian, and Cosine.

The second characteristic of *Kernel regression* is the **bandwidth** associated to each variable. It is involved in calculating the kernel and the weights of the observations, and differentiates or rescales the relative weights of the variables while at the same time reducing or augmenting the impact of observations of the learning sample, depending on how far they are from the observation to predict. The term *bandwidth* refers to the filtering methods. The lower a given variable and kernel function, the fewer will be the number of observations to influence the prediction.

Example: let $Y$ be the dependent variable, and $(X_1, X_2, \ldots, X_k)$ the $k$ explanatory variables. For the prediction of $y_i$ from observation $i$ ($1 \leq i \leq n_{\text{valid}}$), given the observation $j$ ($1 \leq j \leq n_{\text{learn}}$), the
weight determined using a Gaussian kernel, with a bandwidth fixed to $h_i$ for each of the $X_i$ variables ($l=1...k$), is given by:

$$w_{ij} = \frac{1}{\left(\sqrt{2\pi}\right)^k \prod_{l=1}^k h_l} \exp\left(-\sum_{l=1}^k \left(\frac{x_{jl} - x_{il}}{h_l}\right)^2\right)$$

The third characteristic is the polynomial degree used when fitting the model to the observations of the learning sample. In the case where the polynomial degree is 0 (constant polynomial), the Nadaraya-Watson formula is used to compute the $i$'th prediction:

$$y_i = \frac{\sum_{j=1}^{n_{app}} w_{ij} y_j}{\sum_{j=1}^{n_{app}} w_{ij}}$$

For the constant polynomial, the explanatory variables are only taken into account for computing of the weight of the observations in the learning sample. For higher polynomial degrees (experience shows that higher orders are not necessary and XLSTAT works with polynomials of degrees 0 to 2), the variables are used in calculating a polynomial model. Once the model has been fitted, it is applied to the validation or prediction sample in order to estimate the values of the dependent variable.

Once the parameters of the model have been estimated, the prediction value is calculated using the following formulae:

- **Degree 1**: $y_i = a_0 + \sum_{l=1}^k a_l x_{il}$
- **Degree 2**: $y_i = a_0 + \sum_{l=1}^k a_l x_{il} + \sum_{l=1}^k \sum_{m=1}^k b_{lm} x_{il} x_{im}$

**Notes:**

- Before we estimate the parameters of the polynomial model, the observations of the learning sample are previously weighted using the Nadaraya-Watson formula.
- For a 1st or 2nd order model, for each observation of the validation and prediction samples, the polynomial parameters are estimated. This makes Kernel Regression a numerically intensive method.

Two strategies are suggested in order to restrict the size of the learning sample taken into account for the estimation of the parameters of the polynomial:

- **Moving window**: to estimate $y_i$, we take into account a fixed number of observations previously observed. Consequently, with this strategy, the learning sample evolves at each step.
• k nearest neighbours: this method, complementary to the previous, restricts the size of
the learning sample to a given value k.

Details of the kernel functions:
The weight \( w_{ij} \) computed for observation j, for the estimation of prediction \( y_i \), is defined as:

\[
w_{ij} = \prod_{l=1}^{k} \frac{K(u_{ijl})}{h_i} \text{ where } u_{ijl} = \frac{x_{il} - x_{jl}}{h_l}
\]

and K is a kernel function. The kernel functions available in XLSTAT are:

- Uniform: the kernel function is defined as:
  \[ K(u) = \frac{1}{2} \cdot 1_{|u| \leq 1} \]

- Triangle: the kernel function is defined as:
  \[ K(u) = (1 - |u|)1_{|u| \leq 1} \]

- Epanechnikov: the kernel function is defined as:
  \[ K(u) = \frac{3}{4} (1 - u^2)1_{|u| \leq 1} \]

- Quartic: the kernel function is defined as:
  \[ K(u) = \frac{15}{16} (1 - u^2)^2 1_{|u| \leq 1} \]

- Triweight: the kernel function is defined as:
  \[ K(u) = \frac{35}{32} (1 - u^2)^3 1_{|u| \leq 1} \]

- Tricube: the kernel function is defined as:
  \[ K(u) = (1 - |u|^3)^3 1_{|u| \leq 1} \]

- Gaussian: the kernel function is defined as:
  \[ K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \]

- Cosine: the kernel function is defined as:
  \[ K(u) = \frac{\pi}{4} \cos\left(\frac{\pi}{2} u\right) 1_{|u| \leq 1} \]
LOWESS regression

LOWESS regression (Locally weighted regression and smoothing scatter plots) was introduced by Cleveland (1979) in order to create smooth curves through scattergrams. New versions have since been perfected to increase the robustness of the models. LOWESS regression is very similar to Kernel regression as it is also based on polynomial regression and requires a kernel function to weight the observations.

The LOWESS algorithm can be described as follows: for each point i to predict:

1 - First, the Euclidean distances \( d(i,j) \) between the observations i and j are computed. The fraction \( f \) of the N closest observations to observation i are selected. The weight of the selected points are selected using the Tricube kernel and the following distance:

\[
D(i, j) = \frac{d(i, j)}{\text{Max}_j(d(i, j))}
\]

\[
Poids(j) = \text{Tricube}(D(i, j))
\]

2 - A regression model is then fitted, and a prediction is computed for observation i.

For the robust LOWESS regression, additional computations are performed:

3 - The weights are computed again using the following distance:

\[
D'(i, j) = \frac{|r(j)|}{6\text{Median}_j(|r(j)|)}
\]

where \( r(j) \) is the residual corresponding to observation j after the previous step.

Using the Quartic kernel:

\[
Poids(j) = \text{Quartic}(D'(i, j))
\]

4 - The regression is then fitted again using the new weights.

5 - Steps 3 and 4 are performed a second time. A final prediction is then computed for observation i.

Notes:

- The only input parameters apart from the observations for the method are the \( f \) fraction of nearest individuals (in \% in XLSTAT) and the polynomial degree.
• Robust LOWESS regression is about three times more time consuming than LOWESS regression.

Dialog box
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

: Click this button to reload the default options.

: Click this button to delete the data selections.

: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Y / Dependent variables:

Quantitative: Select the response variable(s) you want to model. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

X / Explanatory variables:

Quantitative: Activate this option if you want to include one or more quantitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The
data selected must be of type numeric. If the variable header has been selected, check that the "Variable labels" option has been activated.

**Qualitative**: Activate this option if you want to include one or more qualitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Variable labels" option has been activated.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Variable labels**: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

**Observation labels**: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

**Method**: Choose the type of nonparametric regression to use (see description).

**Polynomial degree**: enter the order of the polynomial if the LOWESS regression or a polynomial method is chosen.

**Options** tab:

Learning samples:

- **Moving window**: choose this option if you want the size of the learning sample to be constant. You need to enter the size S of the window. In that case, to estimate Y(i+1), the observations i-S-1 to i will be used, and the first observation XLSTAT will be able to compute a prediction for, is the S+1\textsuperscript{th} observation.

- **Expanding window**: choose this option if you want the size of the learning sample to be expanding step by step. You need to enter the initial size S of the window. In that case, to estimate Y(i+1), observations 1 to i will be used, and the first observation XLSTAT will be able to compute a prediction for, is the S+1\textsuperscript{th} observation.
- **All**: the learning and validation samples are identical. This method has no interest for prediction, but it is a way to evaluate the method in case of perfect information.

K nearest neighbours:

- **Rows**: the k points retained for the analysis are k points which are the closest to the point to predict, for a given bandwidth and a given kernel function. k is the value to enter here.
- **%**: the points retained for the analysis are the closest to the point to predict and represent x% of the total learning sample available, where x is the value to enter.

**Tolerance**: Enter the value of the tolerance threshold below which a variable will automatically be ignored.

**Interactions / Level**: Activate this option to include interactions in the model then enter the maximum interaction level (value between 1 and 4).

**Kernel**: the kernel function that will be used. The possible options are: Uniform, Triangle, Epanechnikov, Quartic, Triweight, Tricube, Gaussian, Cosine. A description of these functions is available in the description section.

**Bandwidth**: XLSTAT allows you to choose a method for automatically computing the bandwidths (one per variable), or you can fix them. The different options are:

- **Constant**: the bandwidth is constant and equal to the fixed value. Enter the value of the bandwidth.
- **Fixed**: the bandwidth is defined in a vertical range of cells in an Excel sheet, which you need to select. The cells must be equal to the number of explanatory variables, and in the same order as the variables.
- **Range**: the value hl of the bandwidth for each variable Xi is determined by the following formula:

\[ h_l = \text{Max}(x_{il})_{i=1,n_{app}} - \text{Min}(x_{il})_{i=1,n_{app}} \]

- **Standard deviation**: the value hl of the bandwidth for each explanatory variable is equal to the standard deviation of the variable computed on the learning sample.

**Validation** tab:

**Validation**: Activate this option if you want to use a sub-sample of the data to validate the model.
**Validation set**: Choose one of the following options to define how to obtain the observations used for the validation:

- **Random**: The observations are randomly selected. The “Number of observations” N must then be specified.
- **N last rows**: The N last observations are selected for the validation. The “Number of observations” N must then be specified.
- **N first rows**: The N first observations are selected for the validation. The “Number of observations” N must then be specified.
- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to use for the validation.

**Prediction** tab:

**Prediction**: Activate this option if you want to select data to use them in prediction mode. If activate this option, you need to make sure that the prediction dataset is structured as the estimation dataset: same variables with the same order in the selections. On the other hand, variable labels must not be selected: the first row of the selections listed below must correspond to data.

**Quantitative**: Activate this option to select the quantitative explanatory variables. The first row must not include variable labels.

**Qualitative**: Activate this option to select the qualitative explanatory variables. The first row must not include variable labels.

**Observations labels**: activate this option if observations labels are available. Then select the corresponding data. If this option is not activated, the observations labels are automatically generated by XLSTAT (PredObs1, PredObs2 …).

**Missing data** tab:

These options are available only for PCR and OLS regression. With PLS regression, the missing data are automatically handled by the algorithm.

**Remove observations**: Activate this option to remove the observations with missing data.

**Estimate missing data**: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
• **Nearest neighbour**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.

**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the variables selected.

**Correlations**: Activate this option to display the explanatory variables correlation matrix.

**Goodness of fit statistics**: Activate this option to display the table of goodness of fit statistics for the model.

**Predictions and residuals**: Activate this option to display the predictions and residuals for all the observations.

**Charts** tab:

**Data and predictions**: Activate this option to display the chart of observations and predictions:

- **As a function of X1**: Activate this option to display the observed and predicted observations as a function of the values of the X1 variable.

- **As a function of time**: Activate this option to select the data giving the date of each observation to display the results as a function of time.

**Residuals**: Activate this option to display the residuals as a bar chart.

**Results**

**Descriptive statistics**: The table of descriptive statistics shows the simple statistics for all the variables selected. The number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the quantitative variables. For qualitative variables, including the dependent variable, the categories with their respective frequencies and percentages are displayed.

**Correlation matrix**: This table displays the correlations between the selected variables.

**Goodness of fit coefficients**: This table shows the following statistics:

- The determination coefficient $R^2$;

- The sum of squares of the errors (or residuals) of the model (SSE or SSR respectively);
• The means of the squares of the errors (or residuals) of the model (MSE or MSR);
• The root mean squares of the errors (or residuals) of the model (RMSE or RMSR).

**Predictions and residuals:** Table giving for each observation the input data, the value predicted by the model and the residuals.

**Charts:**
If only one quantitative explanatory variable or temporal variable has been selected (“As a function of time” option in the "Charts" tab in the dialog box), the first chart shows the data and the curve for the predictions made by the model. If the "As a function of X1" option has been selected, the first chart shows the observed data and predictions as a function of the first explanatory variable selected. The second chart displayed is the bar chart of the residuals.

**Example**
A tutorial on Kernel regression is available on the Addinsoft website:

http://www.xlstat.com/demo-kernel.htm

**References**


Nonlinear regression

Use this tool to fit data to any linear or non-linear function. The method used is least squares. Either pre-programmed functions or functions added by the user may be used.

Description

Nonlinear regression is used to model complex phenomena which cannot be handled by the linear model. XLSTAT provides preprogrammed functions from which the user may be able to select the model which describes the phenomenon to be modeled.

When the model required is not available, the user can define a new model and add it to their personal library. To improve the speed and reliability of the calculations, it is recommended to add derivatives of the function for each of the parameters of the model.

When this is possible (preprogrammed functions or user defined functions where the first derivatives have been entered by the user) the Levenberg-Marquardt algorithm is used. When the derivatives are not available, a more complex and slower but efficient algorithm is used. This algorithm does not, however, enable the standard deviations of the parameter estimators to be obtained.

Adding a function to the library of user-defined functions

Syntax:

The parameters of the function must be written pr1, pr2, etc..

The explanatory variables must be represented as X1, X2, etc..

Excel functions can be used: Exp(), Sin(), Pi(), Max(), etc.

Example of a function: pr1 * Exp( pr2 + pr3 * X1 + pr4 * X2 )

File containing function definitions:

The library of user functions is held in the file Models.txt in the user directory defined during installation or by using the XLSTAT options dialog box. The library is built as follows:

Row 1: number of functions defined by user

Row 2: N1= number of parameters in function 1
Row 3: function 1 definition

Rows 4 to \((3 + N1)\): derivatives definition for function 1

Row 4+N1: \(N2=\) number of parameters in function 2

Row 5+N1: function 2 definition

…

When the derivatives have not been supplied by the user, "Unknown" replaces the derivatives of the function.

You can modify manually the items of this file but you should be cautious not to make an error.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- Click this button to start the computations.
- Click this button to close the dialog box without doing any computation.
- Click this button to display the help.
- Click this button to reload the default options.
- Click this button to delete the data selections.
- Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

Y / Dependent variables:
**Quantitative**: Select the response variable(s) you want to model. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

**X / Explanatory variables:**

**Quantitative**: Activate this option if you want to include one or more quantitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The data selected must be of type numeric. If the variable header has been selected, check that the "Variable labels" option has been activated.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Variable labels**: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

**Observation labels**: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

**Weights**: Activate this option if the observations are weighted. If you do not activate this option, the weights will all be taken as 1. Weights must be greater than or equal to 0. If the "Variable labels" option is activated you need to include a header in the selection.

**Functions** tab:

**Built-in function**: Activate this option to fit one of the functions available from the list of built-in functions to the data. Select a function from the list.

**Edit**: Click this button to display the active built-in function in the "Function: Y=" field. You can then copy the function to afterwards change it to create a new function or the derivatives of a new function.
**User defined functions**: Activate this option to fit one of the functions available from the list of user-defined functions to the data, or to add a new function.

**Delete**: Click this button to delete the active function from the list of user-defined functions.

**Add**: Click this button to add a function to the list of user-defined functions. You must then enter the function in the "**Function: Y=**" field, then, if you want and given that it will speed up the calculations and enable the standard deviations of the parameters to be obtained, you can select the derivatives of the function for each of the parameters. To do this, activate the "**Derivatives**" option, then select the derivatives in an Excel worksheet.

**Derivatives**: These will speed up the calculations and enable the standard deviations of the parameters to be obtained,

Note: the description section contains information on defining user functions.

**Options** tab:

**Initial values**: Activate this option to give XLSTAT a starting point. Select the cells which correspond to the initial values of the parameters. The number of rows selected must be the same as the number of parameters.

**Parameters bounds**: Activate this option to give XLSTAT a possible region for all the parameters of the model selected. You must then select a two-column range, the one on the left being the lower bounds and the one on the right the upper bounds. The number of rows selected must be the same as the number of parameters.

**Parameters labels**: Activate this option if you want to specify the names of the parameters. XLSTAT will display the results using the selected labels instead of using generic labels pr1, pr2, etc. The number of rows selected must be the same as the number of parameters.

**Stop conditions**:

- **Iterations**: Enter the maximum number of iterations for the algorithm. The calculations are stopped when the maximum number if iterations has been exceeded. Default value: 200.

- **Convergence**: Enter the maximum value of the evolution in the Sum of Squares of Errors (SSE) from one iteration to another which, when reached, means that the algorithm is considered to have converged. Default value: 0.00001.

**Validation** tab:
**Validation**: Activate this option if you want to use a sub-sample of the data to validate the model.

**Validation set**: Choose one of the following options to define how to obtain the observations used for the validation:

- **Random**: The observations are randomly selected. The “Number of observations” N must then be specified.
- **N last rows**: The N last observations are selected for the validation. The “Number of observations” N must then be specified.
- **N first rows**: The N first observations are selected for the validation. The “Number of observations” N must then be specified.
- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to use for the validation.

**Prediction** tab:

**Prediction**: Activate this option if you want to select data to use them in prediction mode. If activate this option, you need to make sure that the prediction dataset is structured as the estimation dataset: same variables with the same order in the selections. On the other hand, variable labels must not be selected: the first row of the selections listed below must correspond to data.

**Quantitative**: Activate this option to select the quantitative explanatory variables. The first row must not include variable labels.

**Qualitative**: Activate this option to select the qualitative explanatory variables. The first row must not include variable labels.

**Observations labels**: activate this option if observations labels are available. Then select the corresponding data. If this option is not activated, the observations labels are automatically generated by XLSTAT (PredObs1, PredObs2 …).

**Missing data** tab:

**Remove observations**: Activate this option to remove the observations with missing data.

**Estimate missing data**: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
• **Nearest neighbour:** Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.

**Outputs** tab:

**Descriptive statistics:** Activate this option to display descriptive statistics for the variables selected.

**Correlations:** Activate this option to display the explanatory variables correlation matrix.

**Goodness of fit statistics:** Activate this option to display the table of goodness of fit statistics for the model.

**Model parameters:** Activate this option to display the values of the parameters for the model after fitting.

**Equation of the model:** Activate this option to display the equation of the model once fitted.

**Predictions and residuals:** Activate this option to display the predictions and residuals for all the observations.

Charts:

- **Data and predictions:** Activate this option to display the chart of observations and the curve for the fitted function.

- **Residuals:** Activate this option to display the residuals as a bar chart.

**Results**

**Descriptive statistics:** The table of descriptive statistics shows the simple statistics for all the variables selected: the number of observations, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased).

**Correlation matrix:** This table displays the correlations between the selected variables.

**Goodness of fit coefficients:** This table shows the following statistics:

- The number of observations;

- The degrees of freedom (DF);

- The determination coefficient $R^2$.
• The sum of squares of the errors (or residuals) of the model (SSE or SSR respectively);
• The means of the squares of the errors (or residuals) of the model (MSE or MSR);
• The root mean squares of the errors (or residuals) of the model (RMSE or RMSR);

Model parameters: This table gives the value of each parameter after fitting to the model. For built-in functions, or user-defined functions when derivatives for the parameters have been entered, the standard deviations of the estimators are calculated.

Predictions and residuals: This table gives for each observation the input data, the value predicted by the model and the residuals. It is followed by the equation of the model.

Charts:

If only one quantitative explanatory variable has been selected, the first chart represents the data and the curve for the function chosen. The second chart displayed is the bar chart of the residuals.

Example

Tutorials showing how to run a nonlinear regression are available on the Addinsoft website on the following pages:

http://www.xlstat.com/demo-nonlin.htm

http://www.xlstat.com/demo-nonlin2.htm

References


Two-stage least squares regression

Use this tool to analyze your data with a two-stage least squares regression.

Description

The two-stage least squares method is used to handle model with endogenous explanatory variables in a linear regression framework. An endogenous variable is a variable which is correlated with the error term in the regression model. Using endogenous variable is in contradiction with the linear regression assumptions. This kind of variable can be encountered when variable are measured with error.

The general principle of the two-stage least squares approach is to use instrumental variables uncorrelated with the error term to estimate the model parameters. These instrumental variables are correlated to the endogenous variables but not with the error term of the model.

Denote by $y$ the quantitative dependent variable, $X_1$ the matrix of $p_1$ endogenous explanatory variables, $X_2$ the matrix of $p_2$ exogenous explanatory variables (not correlated to the error term) ($p = p_1 + p_2$) and $Z$ the matrix of $q$ instrumental variables. The structural equations of the model are given by:

$$
\begin{align*}
    y &= X_1 \beta_1 + X_2 \beta_2 + \varepsilon \\
    X_1 &= Z \gamma + \delta
\end{align*}
$$

where $\beta_1$ and $\beta_2$ are the parameters respectively associated to $X_1$ and $X_2$. The variables $\varepsilon$ and $\delta$ are the disturbances with zero means.

According to the estimation technique developed by Theil (1953a, 1953b), the estimate of the parameter $\beta = (\beta_1, \beta_2)$ is given by:

$$
\hat{\beta} = (X' \Omega X)^{-1} X' \Omega y
$$

where $\Omega = Z(Z' Z)^{-1} Z'$ is the projection matrix.

XLSTAT enables you to take into account endogenous, exogenous and instrumental variables. Endogenous and exogenous variables should be selected as explanatory variables and instrumental and exogenous variables should be selected as instrumental variables (exogenous variables should be selected in both selections).
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- 🔄: Click this button to reload the default options.
- 🗑️: Click this button to delete the data selections.
- 🛠️: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

**Y / Dependent variables**:

Quantitative: Select the response variable(s) you want to model. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

**X / Explanatory variables**:

Quantitative: Select the quantitative explanatory variables in the Excel worksheet. The data selected must be of type numeric. If the variable header has been selected, check that the "Variable labels” option has been activated. Exogenous and endogenous variables should be selected here.

**Z / Instrumental variables**:

Quantitative: Select the quantitative instrumental variables in the Excel worksheet. The data selected must be of type numeric. If the variable header has been selected, check that the
"Variable labels" option has been activated. All the exogenous variables must be selected as instrumental variables.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Variable labels**: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

**Observation labels**: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 ...).

**Observation weights**: Activate this option if the observations are weighted. If you do not activate this option, the weights will all be taken as 1. Weights must be greater than or equal to 0. A weight of 2 is equivalent to repeating the same observation twice. If a column header has been selected, check that the "Variable labels" option has been activated.

**Regression weights**: Activate this option if you want to carry out a weighted least squares regression. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

**Options** tab:

**Null intercept**: Activate this option to set the constant of the model to 0.

**Tolerance**: Activate this option to prevent the OLS regression calculation algorithm taking into account variables which might be either constant or too correlated with other variables already used in the model (0.0001 by default).

**Confidence interval (%)**: Enter the percentage range of the confidence interval to use for the various tests and for calculating the confidence intervals around the parameters and predictions. Default value: 95.

**Validation** tab:
**Validation**: Activate this option if you want to use a sub-sample of the data to validate the model.

**Validation set**: Choose one of the following options to define how to obtain the observations used for the validation:

- **Random**: The observations are randomly selected. The “Number of observations” N must then be specified.
- **N last rows**: The N last observations are selected for the validation. The “Number of observations” N must then be specified.
- **N first rows**: The N first observations are selected for the validation. The “Number of observations” N must then be specified.
- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to use for the validation.

**Prediction** tab:

**Prediction**: Activate this option if you want to select data to use them in prediction mode. If activate this option, you need to make sure that the prediction dataset is structured as the estimation dataset: same variables with the same order in the selections. On the other hand, variable labels must not be selected: the first row of the selections listed below must correspond to data.

**X / Explanatory variables**: Select the quantitative explanatory variables. The first row must not include variable labels. Only exogenous and endogenous variables should be selected here.

**Observations labels**: activate this option if observations labels are available. Then select the corresponding data. If this option is not activated, the observations labels are automatically generated by XLSTAT (PredObs1, PredObs2 …).

**Missing data** tab:

**Remove observations**: Activate this option to remove the observations with missing data.

- **Check for each Y separately**: Choose this option to remove the observations with missing data in the select Y (dependent) variables, only if the Y of interest has a missing data.
- **Across all Ys**: Choose this option to remove the observations with missing data in the Y (dependent) variables, even if the Y of interest has no missing data.
**Estimate missing data**: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

- **Nearest neighbour**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.

**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the variables selected.

**Correlations**: Activate this option to display the correlation matrix for quantitative variables (dependent or explanatory).

**Analysis of variance**: Activate this option to display the analysis of variance table.

**Standardized coefficients**: Activate this option if you want the standardized coefficients (beta coefficients) for the model to be displayed.

**Predictions and residuals**: Activate this option to display the predictions and residuals for all the observations.

**Charts** tab:

**Regression charts**: Activate this option to display regression chart:

- **Standardized coefficients**: Activate this option to display the standardized parameters for the model with their confidence interval on a chart.

- **Predictions and residuals**: Activate this option to display the following charts.

  1. Line of regression: This chart is only displayed if there is only one explanatory variable and this variable is quantitative.

  2. Explanatory variable versus standardized residuals: This chart is only displayed if there is only one explanatory variable and this variable is quantitative.

  3. Dependent variable versus standardized residuals.

  4. Predictions for the dependent variable versus the dependent variable.

  5. Bar chart of standardized residuals.

  - **Confidence intervals**: Activate this option to have confidence intervals displayed on charts (1) and (4).
Results

Summary statistics: The tables of descriptive statistics show the simple statistics for all the variables selected. The number of observations, missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the dependent variables (in blue) and the quantitative explanatory variables. For qualitative explanatory variables the names of the various categories are displayed together with their respective frequencies.

Correlation matrix: This table is displayed to give you a view of the correlations between the various variables selected.

Goodness of fit statistics: The statistics relating to the fitting of the regression model are shown in this table:

- **Observations**: The number of observations used in the calculations. In the formulas shown below, \( n \) is the number of observations.
- **Sum of weights**: The sum of the weights of the observations used in the calculations. In the formulas shown below, \( W \) is the sum of the weights.
- **DF**: The number of degrees of freedom for the chosen model (corresponding to the error part).
- **R²**: The determination coefficient for the model. This coefficient, whose value is between 0 and 1, is only displayed if the constant of the model has not been fixed by the user. Its value is defined by:

\[
R^2 = 1 - \frac{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} w_i (y_i - \bar{y})^2}, \quad \text{where} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} w_i y_i,
\]

The \( R^2 \) is interpreted as the proportion of the variability of the dependent variable explained by the model. The nearer \( R^2 \) is to 1, the better is the model. The problem with the \( R^2 \) is that it does not take into account the number of variables used to fit the model.

- **Adjusted R²**: The adjusted determination coefficient for the model. The adjusted \( R^2 \) can be negative if the \( R^2 \) is near to zero. This coefficient is only calculated if the constant of the model has not been fixed by the user. Its value is defined by:

\[
\hat{R}^2 = 1 - (1 - R^2) \frac{W - 1}{W - p - 1}
\]
The adjusted $R^2$ is a correction to the $R^2$ which takes into account the number of variables used in the model.

- **MSE**: The mean squared error (MSE) is defined by:
  \[
  MSE = \frac{1}{W - p^*} \sum_{i=1}^{n} w_i \left( y_i - \hat{y}_i \right)^2
  \]

- **RMSE**: The root mean square of the errors (RMSE) is the square root of the MSE.

- **MAPE**: The *Mean Absolute Percentage Error* is calculated as follows:
  \[
  MAPE = \frac{100}{W} \sum_{i=1}^{n} w_i \left| \frac{y_i - \hat{y}_i}{y_i} \right|
  \]

- **DW**: The Durbin-Watson statistic is defined by:
  \[
  DW = \frac{\sum_{i=2}^{n} \left[ (y_i - \hat{y}_i) - (y_{i-1} - \hat{y}_{i-1}) \right]^2}{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}
  \]

  This coefficient is the order 1 autocorrelation coefficient and is used to check that the residuals of the model are not autocorrelated, given that the independence of the residuals is one of the basic hypotheses of linear regression. The user can refer to a table of Durbin-Watson statistics to check if the independence hypothesis for the residuals is acceptable.

- **Cp**: Mallows Cp coefficient is defined by:
  \[
  Cp = \frac{SSE}{\hat{\sigma}} + 2 p^* - W
  \]

  where $SSE$ is the sum of the squares of the errors for the model with $p$ explanatory variables and $\hat{\sigma}$ is the estimator of the variance of the residuals for the model comprising all the explanatory variables. The nearer the Cp coefficient is to $p^*$, the less the model is biased.

- **AIC**: Akaike’s Information Criterion is defined by:
  \[
  AIC = W \ln \left( \frac{SSE}{W} \right) + 2 p^*
  \]

  This criterion, proposed by Akaike (1973) is derived from the information theory and uses Kullback and Leibler’s measurement (1951). It is a model selection criterion which penalizes models for which adding new explanatory variables does not supply sufficient
information to the model, the information being measured through the MSE. The aim is to minimize the AIC criterion.

- **SBC**: Schwarz’s Bayesian Criterion is defined by:

\[
SBC = W \ln \left( \frac{SSE}{W} \right) + \ln(W) p^* 
\]

This criterion, proposed by Schwarz (1978) is similar to the AIC, and the aim is to minimize it.

- **PC**: Amemiya’s Prediction Criterion is defined by:

\[
PC = \frac{(1 - R^2)(W + p^*)}{W - p^*}
\]

This criterion, proposed by Amemiya (1980) is used, like the adjusted R² to take account of the parsimony of the model.

The **analysis of variance table** is used to evaluate the explanatory power of the explanatory variables. Where the constant of the model is not set to a given value, the explanatory power is evaluated by comparing the fit (as regards least squares) of the final model with the fit of the rudimentary model including only a constant equal to the mean of the dependent variable. Where the constant of the model is set, the comparison is made with respect to the model for which the dependent variable is equal to the constant which has been set.

The **parameters of the model** table displays the estimate of the parameters, the corresponding standard error, the Student’s t, the corresponding probability, as well as the confidence interval.

The **equation of the model** is then displayed to make it easier to read or re-use the model.

The **predictions and residuals** table shows, for each observation, its weight, the value of the qualitative explanatory variable, if there is only one, the observed value of the dependent variable, the model's prediction, the residuals and the confidence intervals with the fitted prediction. Two types of confidence interval are displayed: a confidence interval around the mean (corresponding to the case where the prediction would be made for an infinite number of observations with a set of given values for the explanatory variables) and an interval around the isolated prediction (corresponding to the case of an isolated prediction for the values given...
for the explanatory variables). The second interval is always greater than the first, the random values being larger. If the validation data have been selected, they are displayed at the end of the table.

The charts which follow show the results mentioned above. If there is only one explanatory variable in the model, the first chart displayed shows the observed values, the regression line and both types of confidence interval around the predictions. The second chart shows the normalized residuals as a function of the explanatory variable. In principle, the residuals should be distributed randomly around the X-axis. If there is a trend or a shape, this shows a problem with the model.

The three charts displayed next show respectively the evolution of the standardized residuals as a function of the dependent variable, the distance between the predictions and the observations (for an ideal model, the points would all be on the bisector), and the standardized residuals on a bar chart. The last chart quickly shows if an abnormal number of values are outside the interval \( [-2, 2] \) given that the latter, assuming that the sample is normally distributed, should contain about 95% of the data.

If you have selected the data to be used for calculating predictions on new observations, the corresponding table is displayed next.

**Example**

A tutorial on the two-stage least square approach is available on the Addinsoft website: http://www.xlstat.com/demo-sls.htm

**References**


Theil H. (1953a). Repeated least square applied to complete equation systems. mimeo, Central Planning Bureau, The Hague.

Classification and regression trees

Classification and regression trees are methods that deliver models that meet both explanatory and predictive goals. Two of the strengths of this method are on the one hand the simple graphical representation by trees, and on the other hand the compact format of the natural language rules. We distinguish the following two cases where these modeling techniques should be used:

- use classification trees to explain and predict the belonging of objects (observations, individuals) to a class, on the basis of explanatory quantitative and qualitative variables.

- use regression tree to build an explanatory and predicting model for a dependent quantitative variable based on explanatory quantitative and qualitative variables.

Note: Sometimes the term segmentation tree or decision tree is employed when talking of the abovementioned models.

Description

Classification and regression tree analysis was proposed in different ways. AID trees (Automatic Interaction Detection) have been developed by Morgan and Sonquist (1963). CHAID (CHi-square Automatic Interaction Detection) was proposed by Kass (1980) and later enriched by Biggs (Biggs et al, 1991) when he introduced the exhaustive CHAID procedure. The name of the Classification And Regression Trees (CART) methods, comes from the title of the book of Breiman (1984). The QUEST (QUick, Efficient, Statistical Tree ) method is more recent (Loh and Shih, 1997).

These explanatory and predictive methods can be deployed when one needs to:

- Build of a rules-based model to explain a phenomenon recorded through qualitative or quantitative dependent variables, while identifying the most important explanatory variables.

- Identify groups created by the rules.

- Predict the value of the dependent variable for a new observation.

CHAID, CART and QUEST

XLSTAT offers the choice between four different methods of classification and regression tree analysis: CHAID, exhaustive CHAID, CART and Quest. In most cases CHAID and exhaustive CHAID gives the best results. In special situations the two other methods can be of interest. CHAID is the only algorithm implemented in XLSTAT that can lead to non binary tree.
With all these methods, explanatory quantitative variables are transformed into discrete variables with k categories. The discretization is performed using the Fisher’s method (this method is available in the Univariate partitioning function).

**CHAID and exhaustive CHAID**

These two methods proceed in three steps: splitting, merging and stopping.

Splitting: Starting with the root node that contains all the objects, the best split variable is the one for which the criterion (p-value or Tschuprow’s T) is the lowest. The split is performed if the p-value is lower than the user defined threshold. In the case of a quantitative dependent variable, an ANOVA is used to find the variable that best explains the variation of the dependent variable Y (based on the p-value of the Fisher’s F). In the case of a qualitative dependent variable, the user can choose between the Tschuprow’s T (related to the Pearson’s Chi-square) and maximum likelihood ratio.

Merging: In the case of a qualitative split variable, the procedure tries to merge similar categories of that variable into common sub nodes. This step is repeated in the case of the exhaustive CHAID until only two sub nodes remain. This is the reason, why exhaustive CHAID leads to a binary tree. During the merge, the Tschuprow’s T (related to the Pearson’s Chi-square) or the maximum likelihood ratio is computed. If the maximum value is bigger than the user defined threshold, the two corresponding groups of categories are merged. This step is repeated recursively until the maximum p-value is smaller or equal to the threshold, or until there are only two remaining categories.

Stopping: For every newly created sub-node the stop criteria are checked. If none of the criteria are met, the node is treated in the same way as the root node. The following are the stop criteria:

- Pure node: The node contains only objects of one category or one value of the dependent variable.
- Maximum tree depth: The level of the node has reached the user defined maximum tree depth.
- Minimum size for a parent-node: The node contains fewer objects than the user defined minimum size for a parent-node.
- Minimum size for a son-node: After splitting this node, there is at least one sub-node which size is smaller than the used defined minimum size for a son-node.

**CART**
This method verifies recursively for each node if a splitting is possible using the selected measure. Several measures of impurity are available. In the case of a quantitative dependent variable a measure based on the LSD (Least Square Deviation) is being used. In the case of a qualitative dependent variable, the user has the choice between the Gini and the Twoing indexes. For a quantitative explanatory variable, a univariate partitioning into k clusters is carried out. In the case of a qualitative explanatory variable, every possible grouping of the k categories into 2 subsets is tested (there $2^k - 1$ possibilities). Then all the k-1 possible split points are calculated and tested.

For every newly created sub-node the stop criteria are checked. If none of the criteria are met, the node is treated in the same way as the root node.

- **Pure node**: The node contains only objects of one class or one value of the dependent variable.
- **Maximum tree depth**: The level of the node has reached the user defined maximum tree depth.
- **Minimum size for a parent-node**: The node contains fewer objects than the user defined minimum size for a parent-node.
- **Minimum size for a son-node**: After splitting this node, there is at least one sub-node which size is smaller than the user defined minimum size for a son-node.

**QUEST**

This method can only be applied to qualitative dependent variables. This method carries out a splitting using two separate sub-steps. First, we look for the best splitting variable among the explanatory variables; second, the split point for the split variable is calculated:

Selection of the split variable: For a quantitative explanatory variable, an ANOVA F-test is carried out to compare the mean values of each explanatory variable X for the different categories of the qualitative dependent variable Y. In the case of a qualitative explanatory variable a Chi-square test is performed for each explanatory variable. We define $X^*$ as the explanatory variable for which the p-values is the smallest. If the p-value corresponding to $X^*$ is smaller than $\alpha / p$, where $\alpha$ is the user defined threshold and $p$ is the number of explanatory variables, then $X^*$ is chosen as the split variable. In the case where no $X^*$ is found, Levene's F statistic is calculated for all the quantitative explanatory variables. We define by $X^{**}$ the explanatory variable corresponding to the smallest p-value. If the p-value of $X^{**}$ is smaller than $\alpha / (p + pX)$, $pX$ being the number of quantitative explanatory variables, then $X^{**}$ is chosen as the split variable. In the case where no $X^{**}$ is found, the node is not split.

Choice of the split point: In the case of a qualitative explanatory variable X, the latter variable is first transformed into a qualitative variable X'. The detailed description of the transformation can be found in Loh and Shih (1997). In the case of quantitative variable, similar classes of Y are first grouped together by a k-means clustering of the mean values of X until obtaining two
groups of classes. Then, a discriminant analysis using a quadratic model is carried out on these two groups of classes, in order to determine the optimal split point for that variable.

Stop conditions: For every newly created sub-node the stop criteria are checked. If none of the criteria are met, the node is treated in the same way as the root node.

- Pure node: The node does only contain objects of one class or one value of the dependent variable.
- Maximum tree depth: The level of the node has reached the user defined maximum tree depth.
- Minimal parent node size: The node contains fewer objects than the user defined minimal parent node size.
- Minimal son node size: After splitting this node, a sub node would exist that size would be smaller than the used defined minimal son node size.

**Classification table and ROC curve**

Among the numerous results provided, XLSTAT can display the classification table (also called confusion matrix) used to calculate the percentage of well-classified observations. When only two classes are present in the dependent variable, the ROC curve may also be displayed.

The ROC curve (*Receiver Operating Characteristics*) displays the performance of a model and enables a comparison to be made with other models. The terms used come from signal detection theory.

The proportion of well-classified positive events is called the *sensitivity*. The *specificity* is the proportion of well-classified negative events. If you vary the threshold probability from which an event is to be considered positive, the sensitivity and specificity will also vary. The curve of points (1-specificity, sensitivity) is the ROC curve.

Let's consider a binary dependent variable which indicates, for example, if a customer has responded favorably to a mail shot. In the diagram below, the blue curve corresponds to an ideal case where the n% of people responding favorably corresponds to the n% highest probabilities. The green curve corresponds to a well-discriminating model. The red curve (first bisector) corresponds to what is obtained with a random Bernoulli model with a response probability equal to that observed in the sample studied. A model close to the red curve is therefore inefficient since it is no better than random generation. A model below this curve would be disastrous since it would be less even than random.
The area under the curve (or AUC) is a synthetic index calculated for ROC curves. The AUC corresponds to the probability such that a positive event has a higher probability given to it by the model than a negative event. For an ideal model, AUC=1 and for a random model, AUC = 0.5. A model is usually considered good when the AUC value is greater than 0.7. A well-discriminating model must have an AUC of between 0.87 and 0.9. A model with an AUC greater than 0.9 is excellent.

Lastly, you are advised to validate the model on a validation sample wherever possible. XLSTAT has several options for generating a validation sample automatically.

**Classification and regression trees, discriminant analysis and logistic regression**

Classification and regression trees apply to quantitative and qualitative dependent variables. In the case of a Discriminant analysis or logistic regression, only qualitative dependent variables can be used. In the case of a qualitative depending variable with only two categories, the user will be able to compare the performances of both methods by using ROC curves.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.
OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

Reload: Click this button to reload the default options.

Delete: Click this button to delete the data selections.

Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

**Y / Dependent variables**: Select the dependent variable(s) you want to model. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

**Response type**: Confirm the type of response variable you have selected:

- **Quantitative**: Activate this option if the selected dependent variables are quantitative.
- **Qualitative**: Activate this option if the selected dependent variables are qualitative.

**X / Explanatory variables**:

**Quantitative**: Activate this option if you want to include one or more quantitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The data selected may be of the numerical type. If a variable header has been selected, check that the "Variable labels" option has been activated.

**Qualitative**: Activate this option if you want to include one or more qualitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If a variable header has been selected, check that the "Variable labels" option has been activated.
**Observation weights**: Activate this option if the observations are weighted. If you do not activate this option, the weights will all be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option has been activated.

**Method**: Choose the method to be used. CHAID, exhaustive CHAID, CART and Quest are possible choices. In the case of Quest, the “Response type” is automatically changed to qualitative.

**Measure**: In the case of the CHAID or exhaustive CHAID methods with a qualitative response type, the user can choose between the Pearson Chi-square and the likelihood ratio measures. In the case of the CART method together with a qualitative response type, the user can choose between the Gini and Twoing measures.

**Maximum tree depth**: enter the maximum tree depth.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Variable labels**: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

**Observation labels**: Activate this option if observations labels are available. Then select the corresponding data. If the “Variable labels” option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 ...).

**Options** tab:

**Minimum node size**:

- **Minimum parent size**: Enter the minimum number of objects that a node must contain to be split.

- **Minimum son size**: Enter the minimum number of objects that every newly created node must contain after a possible split in order to allow the splitting.
**Significance level (%):** Enter the significance level. This value is compared to the p-values of the F and Chi-square tests. p-values smaller than this value authorize a split. This option is not active for the CART method.

**CHAID options:** these options are only active with the CHAID methods for the grouping or splitting of qualitative explanatory variables.

- **Merge threshold:** Enter the value of the merge significance threshold. Significance values smaller than this value lead to merge two subgroups of categories. The categories of a qualitative explanatory variable may be merged to simplify the computations and the visualization of results.

- **Authorize redivision:** Activate this option if you want to allow that previously merged categories are split again.
  - **Split threshold:** Enter the value of the split significance threshold. P-values greater than this value lead to split the categories or group of categories into two subgroups of categories.
  - **Bonferroni correction:** Activate this option if you want to use a Bonferroni correction during the computation of the p-value of merged categories.

**Number of intervals:** This option is only active if quantitative explanatory variables have been selected. You can choose the maximum number of intervals generated during the discretization of the quantitative explanatory variables using univariate partitioning by the Fisher’s method. Maximum value is 10.

**Stop conditions:** If the observations are weighted and if the CHAID method is being used, the calculation of the node weights is done by a converging procedure. In that case the convergence criteria can be defined.

- **Iterations:** Enter the maximum number of iterations for the calculus of the node weight.
- **Convergence:** Enter the minimal difference of progress between two iterations. A smaller difference is considered as converged.

**Validation tab:**

**Validation:** Activate this option if you want to use a sub-sample of the data to validate the model.
**Validation set**: Choose one of the following options to define how to obtain the observations used for the validation:

- **Random**: The observations are randomly selected. The “Number of observations” N must then be specified.
- **N last rows**: The N last observations are selected for the validation. The “Number of observations” N must then be specified.
- **N first rows**: The N first observations are selected for the validation. The “Number of observations” N must then be specified.
- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to use for the validation.

**Prediction** tab:

**Prediction**: Activate this option if you want to select data to use them in prediction mode. If activate this option, you need to make sure that the prediction dataset is structured as the estimation dataset: same variables with the same order in the selections. On the other hand, variable labels must not be selected: The first row of the selections listed below must correspond to data.

**Quantitative**: Activate this option to select the quantitative explanatory variables. The first row must not include variable labels.

**Qualitative**: Activate this option to select the qualitative explanatory variables. The first row must not include variable labels.

**Observations labels**: Activate this option if observations labels are available. Then select the corresponding data. If this option is not activated, the observations labels are automatically generated by XLSTAT (PredObs1, PredObs2 …).

**Missing data** tab:

**Remove observations**: Activate this option to remove the observations with missing data.

**Estimate missing data**: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbour**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.
**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the variables selected.

**Correlations**: Activate this option to display the correlation matrix.

**Tree structure**: Activate this option to display a table of the nodes, and information on the number of objects, the p-value of the split, and the first two son-nodes. In the case of a qualitative dependent variable the predicted category is displayed. In the case of a quantitative dependent variable, the expected value of the node is displayed.

**Node frequencies**: Activate this option to display the absolute and relative frequencies of the different nodes.

**Rules**: This table displays the rules in natural language, by default only for the dominant categories of each node. Activate the "All categories" option to display the rules for all the categories of the dependent variable and all nodes.

**Results by object**: Activate this option to display for each observation, the observed category, the predicted category, and, in the case of a qualitative dependent variable, the probabilities corresponding to the various categories of the dependent variable.

**Confusion matrix**: Activate this option to display the table showing the numbers of well- and badly-classified observations for each of the categories.

**Charts** tab:

**Tree chart**: Activate this option to display the classification and regression tree graphical. Pruning can be done by the help of the context menu of the tree chart.

- **Bar charts**: Choose this option so that on the tree, the relative frequencies of the categories are displayed using a bar chart.
  - **Frequencies**: Activate this option to display the frequencies on the bar charts
  - **%**: Activate this option to display the % (of the total population) on the bar charts.

- **Pie charts**: Choose this option so that on the tree, the relative frequencies of the categories are displayed using a pie chart.
**Contextual menu for the trees**

When you click on a node on a classification tree, and then do a right click on the mouse, a contextual menu is displayed with the following commands:

**Show the entire tree:** Select this option to display the entire tree and to undo previous pruning actions.

**Hide the subtree:** Select this option to hide all the nodes below the selected node. Hidden parts of the tree are indicated by a red rectangle of the corresponding parent node.

**Show the subtree:** Select this option to show all the nodes below the selected node.

**Set the pruning level:** Select this option to change the maximum tree depth.

**Reset this Menu:** Select this option to deactivate the context menu of the tree chart and to activate the standard menu of Excel.

**Results**

**Descriptive statistics:** The table of descriptive statistics shows the simple statistics for all the variables selected. The number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the quantitative variables. For qualitative variables, including the dependent variable, the categories with their respective frequencies and percentages are displayed.

**Correlation matrix:** This table displays the correlations between the explanatory variables.

**Tree structure:** This table displays the nodes and information on the number of objects, the significance level of the split, and the two first son-nodes. In the case of a qualitative dependent variable the predicted category is displayed. In the case of a quantitative dependent variable the expected value of the node is displayed.

**Split level chart:** This chart shows the significance level of the split variables for the internal nodes of the tree.

**Tree chart:** A legend is first displayed so that you can identify which color corresponds to which category (qualitative dependent variable) or interval (quantitative dependent variable) of the dependent variable. The graphical visualization of the tree allows to quickly see how it has been iteratively built, in order to obtain rules that are as pure as possible, which means that the leaves of the tree should ideally correspond to only one category (or interval).

Every node is displayed as a bar chart or a pie chart. For the pie chars, the inner circle of the pie corresponds to the relative frequencies of the categories (or intervals) to which the objects contained in the node correspond. The outer ring shows the relative frequencies of the categories of the objects contained in the parent node.
The node identifier, the number of objects, their relative frequency, and the purity (if the dependent variable is qualitative), or the predicted value (if the dependent variable is quantitative) are displayed beside each node. Between a parent and a son node, the split variable is displayed with a grey background. Arrows point from this split variable to the son nodes. The values (categories in the case of a qualitative explanatory variable, intervals in the case of a quantitative explanatory variable) corresponding to each son node are displayed in the top left box displayed next to the son node.

Pruning can be done using the contextual menu of the tree chart. Select a node of the chart and click on the right button of the mouse to activate the context menu. The available options are described in the contextual menu section.

**Node frequencies:** This table displays the frequencies of the categories of the dependent variable.

**Rules:** The rules are displayed in natural language for the dominant categories of each node. If the option "all categories" is checked in the dialog box, then the rules for all categories and every node are displayed.

**Results by object:** This table displays for each observation, the observed category, the predicted category, and, in the case of a qualitative dependent variable, the probabilities corresponding to the various categories of the dependent variable.

**Confusion matrix:** This table displays the numbers of well- and badly-classified observations for each of the categories (see the description section for more details).

**Example**

A tutorial on how to use classification and regression trees is available on the Addinsoft website:

http://www.xlstat.com/demo-dtr.htm

**References**


K Nearest Neighbors

Use this tool to predict the category to which belongs an observation described by a set of variables, based on the categories of its k nearest neighbors, which are k observations for which the category is known, described by the same set of variables.

Description

The K Nearest Neighbors (KNN) method aims to categorize observations of a prediction set whose class is unknown given their respective distances to points in a learning set (i.e. whose class is known a priori).

A simple version of KNN is an intuitive supervised classification approach, it can be regarded as an extension of the nearest neighbor method (NN method is a special case of KNN where k = 1).

The KNN classification approach assumes that each example in the learning set is a random vector in $\mathbb{R}^n$. Each point is described as $x = < a_1(x), a_2(x), a_3(x), ..., a_n(x) >$ where $a_r(x)$ denotes the value of the $r^{th}$ attribute. $a_r(x)$ can be either a quantitative or a qualitative variable.

- To determine the class of the query point $x_q$, each of the $k$ nearest points $x_1, ..., x_k$ to $x_q$ proceed to voting. The class of $x_q$ corresponds to the majority class.

The following algorithm describes the basic KNN method:

Given a set $L$ of size $N$ of pre-classified samples (examples in a learning set):

$L = \{(x_1, f(x_1)), (x_2, f(x_2)), ..., (x_N, f(x_N))\}$

Where $f(x_i)$ is a real value function which denotes the class of $x_i$

$f(x_i) \in V$ Where $V = \{v_1, v_2, ..., v_s\}$

- Given a query point or a sample to be classified $x_q$.
- Let $x_1, x_2, x_3, ..., x_k$ be the nearest pre-classified points with a specific distance function to $x_q$.

$\text{Return}$

$$f(x_q) = \underset{v \in V}{\text{argmax}} \left( \sum_{i=1}^{k} g(v, f(x_i)) \right)$$
Where \( \delta(a, b) = \begin{cases} 
1 & \text{if } a = b \\
0 & \text{else} 
\end{cases} \)

**Origins**

Nearest neighbor rules have traditionally been used in survey pattern recognition (Nilsson, 1965).

This method has also been used in several areas such as:

- Bioinformatics
- Image processing
- Computer vision
- Pattern recognition (such as handwritten character recognition)
- GIS (Geographical Information System): finding the closest cities to given positions.
- Generally, in learning systems, when the problem involves finding the nearest point or the k nearest points to a given query point.

**Quantifying the similarity / dissimilarity between the query point and points in the Learning set:**

The measure of dissimilarity between a given query point and the learning set is computed using a distance function. We recall that a distance function \( d \) on a set \( X \) needs to satisfy the metric conditions:

- \( d(x, y) = d(y, x) \) Symmetry property.
- \( d(x, y) \geq 0 \) Non-negativity property.
- \( d(x, y) \leq d(x, z) + d(z, y) \) triangular inequality.

**Asymptotic result regarding the convergence of the basic KNN**

The result established by Cover and Hart (1966) guarantees the existence of the k nearest neighbors. Let \( x \) and \( x_1, x_2, \ldots, x_N \) be independent identically distributed random variables taking values in a separable metric space \( X \). Let \( x'_n \) denote the nearest neighbor to \( x \) from the set \{\( x_1, x_2, x_N \)\}.

Then \( x'_n \to x \) with probability one (Cover and Hart 1966).
**Complexity Of the basic KNN Method**

In order to find the K nearest neighbors to a given query point, the algorithm needs to compute all the distances separating the query point to each point in the learning set. As a result, the algorithm computes N distances where N is the number of the points in the learning set. Finding the K nearest neighbors requires sorting these N distances. Consequently, the real bottleneck in the basic KNN algorithm resides in the sorting step. Therefore, the complexity of the basic KNN is in the order of $N \log(N)$.

**Quantitative metrics (distances):**

Each point is considered as a quantitative vector whose components are quantitative random variables.

Many quantitative distance can be used such as:

- **Euclidean**: $d(x, y) = \sum_{i=1}^{n} (x_i - y_i)^2$

- **Minkowski**: $d(x, y) = \left( \sum_{i=1}^{n} |x_i - y_i|^q \right)^{1/q}$

- **Manhattan**: $d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$

- **Tchebychev**: $d(x, y) = \max_{i=1..n} (|x_i - y_i|)$

- **Canberra**: $d(x, y) = \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|}$

**Qualitative distances:**

Each point is regarded as a vector whose components are all qualitative variables. When dealing with qualitative vectors, quantitative distances cannot be used. Therefore, several qualitative distances have been introduced:

**The Overlap Metric OM**

The Overlap Metric can be considered as basic qualitative distance:
Two vectors $x$ and $y$ are closer if their attributes are similar (their attributes take the same category/value).

The distance between two vectors $x$, $y$ can be defined as:

$$d(x, y) = \sum_{i=1}^{N} \delta(a_i(x), a_i(y))$$

Where $a_i(x)$, $a_i(y)$ correspond to the $i^{th}$ attributes of the vectors $x$, $y$.

**The Value difference distance (VDM)**

VDM was introduced by Graig Stanfil and David Waltz (1986).

In VDM, two attributes are closer if they have the same classification class.

The VDM distance between vectors $x$ and $y$ is given by:

$$N1: \text{normalized}_{vdm}(x, y) = \sum_{c=1}^{C} \left| P(c|a_i(x)) - P(c|a_i(y)) \right|^q$$

$$N11 \text{ normalized}_{vdm}(x, y) = \sum_{c=1}^{C} \left| P(c|a_i(x)) - P(c|a_i(y)) \right|^q$$

Where $C$ is the total number of classes.

$P(c|a_i(x))$: Provided $a_i(x)$, the probability that $a_i(x)$ is classified into $c$.

$P(c|a_i(y))$: Provided $a_i(y)$, the probability that $a_i(y)$ is classified into $c$.

$q$ generally equals 1 or 2.

$P(c|a_i(x))$ and $P(c|a_i(y))$ are computed as follows:

$$P(c|a_i(x)) = \frac{N(ai, x, c)}{N(ai, x)}$$

$$P(c|a_i(y)) = \frac{N(ai, y, c)}{N(ai, y)}$$

$N(ai, x, c)$: number of instances of $x$ for $a_i$ in $c$

$N(ai, x)$: number of instances of $x$ in the data set.

$N(ai, y, c)$: number of instances of $y$ for $a_i$ in $c$

$N(ai, y)$: number of instances of $y$ in the data set.
**Remark:** Although defined for nominal attributes, the VDM distance can also be used to evaluate the distance between numeric attributes.

### Computing similarity using kernels or kernel trick

Kernels can be regarded as a generalization of distance measures. They can be represented using a Hilbert space (Scholkopf 2001).

The complexity behind the computation of kernels is almost as similar as, but sometimes slightly higher than the computation involved in quantitative metrics.

#### Gaussian Kernel:

\[ k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\delta^2}\right) \]

#### Laplacian Kernel:

\[ k(x, y) = \exp\left(-\frac{\|x - y\|}{\delta}\right) \]

#### Logarithmic Kernel:

\[ k(x, y) = -\log(\|x - y\|^d + 1) \]

#### Power kernel:

\[ k(x, y) = -\|x - y\|^d \]

#### Sigmoid kernel:

\[ k(x, y) = \tanh(\alpha x^T y + c) \]

\( x^T y \) is a dot product

#### Linear kernel

\[ k(x, y) = \alpha x^T y + c \]

\( x^T y \) is a dot product
Where $x, y$ are two vectors in $\mathbb{R}^n$; $\delta$ and $d$ are scalars in $\mathbb{R}$.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- ![Reload](image): Click this button to reload the default options.
- ![Delete](image): Click this button to delete the data selections.
- ![Change](image): Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

**Learning set**

**Y / qualitative variables:**

Qualitative: Select the response dependent variable(s) you want to model. These variables must be qualitative. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

**X / explanatory variables (learning set):**
Quantitative: Select the quantitative explanatory variables from the learning set in the Excel worksheet. The data selected must be of type numeric. If the variable header has been selected, check that the "Variable labels" option has been activated.

Qualitative: Select the qualitative explanatory variables from the learning set in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Variable labels" option has been activated.

Prediction set:

Select the quantitative / qualitative explanatory variable data you want to use to do predictions using KNN classification. The number of variables must be equal to the number of explanatory variables in the learning set.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Weights: Check this option if the observations are weighted. If you do not check this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

Number of neighbors: Select the number of neighbors used during the KNN classification process.

Options tab:

Model: Distance or Kernel:

Select the way to compute the similarity between the prediction set and the learning set. The input parameter can be set to either metrics or kernels.
Depending on the nature of the variables of the input set, either qualitative or quantitative distances can be selected:

Quantitative distances:

- Euclidean
- Minkowski
- Manhattan
- Tchebychev
- Canberra

Qualitative distances:

- Overlap distance
- Value difference metric

The Kernel option enables the use of kernel functions to compute the similarity between query points and points in the learning set:

- Gaussian kernel
- Laplacian kernel
- Spherical kernel
- Linear kernel
- Power kernel

In the case of the kernels option, computations are slightly longer due to the projection of points into a higher dimensional space.

**Breaking ties:**

The majority voting procedure leads to the election of the query point classes. Sometimes, more than one points win the majority. This leads to a tie.

There are several ways to break ties for a given query point depending on the KNN implementation. You can break ties by selecting the options below.
- **Random breaker**: chooses the class corresponding to a random point drawn from the set of equidistant points.

- **Smallest Index**: uses the class corresponding to the first point encountered in the set of equidistant points.

**Weighted vote**: Setting the weighted vote option allows you to choose the inverse distance or the squared inverse distance as a weight for each vote of the nearest neighbors.

**Observations to track**: Activate this option if you want to explore which are the k nearest neighbors for all or a subset of the observations of the prediction set.

**Validation** tab:

This tab allows you to quantify the quality of the KNN classifier. The validation technique used to check the consistency of the classification model is the K-fold cross validation technique. Data is partitioned into k equally sub samples of equal size. Among the k subsamples, a single subsample is retained as the validation data to test the model, and the remaining k – 1 subsamples are used as training data. k can be specified in the number of folds field.

**Missing data** tab:

**Remove observations**: Activate this option to remove the observations with missing data.

**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the selected samples before and after imputation.

**Results by class**: Activate this option to display a table giving the statistics and the objects for each of the classes.

**Results by object**: Activate this option to display a table giving the class each object (observation) is assigned to in the initial object order.
Results

Descriptive statistics: The table of descriptive statistics shows the simple statistics for all the variables selected. The number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the quantitative variables. For qualitative variables, including the dependent variable, the categories with their respective frequencies and percentages are displayed.

Example

A tutorial on how to use K nearest neighbors is available on the Addinsoft website:

http://www.xlstat.com/demo-knn.htm

References


Naive Bayes classifier

Use this method to predict the category to which belongs an observation described by a set of quantitative and qualitative variables (predictors).

Description

The Naive Bayes classifier is a supervised machine learning algorithm that allows you to classify a set of observations according to a set of rules determined by the algorithm itself. This classifier has first to be trained on a training dataset that shows which class is expected for a set of inputs. During the training phase, the algorithm elaborates the classification rules on this training dataset that will be used in the prediction phase to classify the observations of the prediction dataset. Naive Bayes implies that classes of the training dataset are known and should be provided hence the supervised aspect of the technique.

Historically, the Naive Bayes classifier has been used in document classification and spam filtering. As of today, it is a renowned classifier that can find applications in numerous areas. It has the advantage of requiring a limited amount of training to estimate the necessary parameters and it can be extremely fast compared to some other techniques. Finally, in spite of its strong simplifying assumption of independence between variables (see description below), the naive Bayes classifier performs quite well in many real-world situations which makes it an algorithm of choice among the supervised Machine Learning methods.

At the root of the Naive Bayes classifier is the Bayes’ theorem with the “naive” assumption of independence between all pairs of variables/features. Given a class variable $y$ and a set of independent variables $x_1$ through $x_n$, the Bayes’ theorem states that:

$$P(y|x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots, x_n|y)}{P(x_1, \ldots, x_n)}$$

From the naive independence assumption, the following relationship can be derived:

$$P(x_i|y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = P(x_i|y)$$

For all $i$, this relationship leads to:

$$P(y|x_1, \ldots, x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i|y)}{P(x_1, \ldots, x_n)}$$

Since $P(x_1, \ldots, x_n)$ is constant given the input, it is regarded as a normalization constant. Thus, we can use the following classification rule:
\[ P(y|x_1,\ldots,x_n) \propto P(y) \prod_{i=1}^{n} P(x_i|y) \]

\[ \hat{y} = \arg \max_y (P(y) \prod_{i=1}^{n} P(x_i|y)) \]

We can use a Maximum A Posteriori (MAP) estimation to estimate \( P(y) \) and \( P(x_i|y) \). Where \( P(y) \) is the relative frequency of class \( y \) in the training set. Several Naive Bayes classifiers might be considered depending on the assumptions made regarding the distribution of \( P(x_i|y) \).

\( P(x_i|y) \) can be assumed to follow a normal distribution, in which case it has the following expression:

\[ P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(x_i - \mu_i)^2}{2\sigma^2} \right) \]

It can also be assumed to follow a Bernoulli distribution or any of the following parametric distributions available in the XLSTAT software: Log-Normal, Gamma, exponential, logistic, Poisson, binomial, uniform. In any of this cases, the distribution parameters are estimated using the moment method.

If the distribution is not known or if you are using qualitative data, XLSTAT offers the possibility to estimate an empirical distribution from the ratio of the given observation to the total number of observation for a given class \( y \).

If an empirical distribution is used, it might be desirable to use a Laplace smoothing in order to avoid the null probability. This might come in handy, for instance, if a qualitative variable from the prediction data set takes a value that hasn’t been met in the training phase of the algorithm. The corresponding conditional probability \( P(x_i|y) \) would then be equal to 0 for every class \( y_i \) of \( y \), leading to the meaningless classification of the observation. In such a case, the Laplace smoothing would have the virtuous property to assign a low, but not null, conditional probability \( P(x_i|y) \) to the corresponding variable. Allowing the remaining variables to be considered nonetheless to affect a class to the observation.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.
OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

Reload default options: Click this button to reload the default options.

Delete data selections: Click this button to delete the data selections.

Change data handling: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Training set

Qualitative: Select the response variable(s) you want to model. These variables must be qualitative. If several variables have been selected, XLSTAT carries out calculations for each variable separately. If a column header has been selected, check that the "Variable labels" option has been activated.

X / explanatory variables (training set):

Quantitative: Select the quantitative explanatory variables from learning set in the Excel worksheet. The data selected must be of type numeric. If the variable header has been selected, check that the "Variable labels" option has been activated.

Qualitative: Select the qualitative explanatory variables from the learning set in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Variable labels" option has been activated.

Prediction set:

Select the quantitative / qualitative explanatory variable data you want to use to do predictions using KNN classification. The number of variables must be equal to the number of explanatory variables in the training set.
**Range:** Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Activate this option to display the results in a new worksheet of the active workbook.

**Workbook:** Activate this option to display the results in a new workbook.

**Variable labels:** Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

**Observation labels:** Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

**Options tab:**

**Same parametric/Empirical Distribution for all quantitative variables:** this option allows you to choose the same parametric/empirical distribution for all quantitative variables.

**Select a specific distribution for each quantitative variable:** this option allows you to select for each quantitative variable a specific parametric distribution or to consider it as an empirical distribution.

The parametric distribution can be selected from the following set of distributions: Normal, log-Normal, Gamma, exponential, logistic, Poisson, Binomial, Bernoulli, Uniform.

The qualitative variables are implicitly drawn from independent empirical distributions.

The parameters of the selected parametric distributions are estimated using the moment method.

**Breaking ties:**

Prediction using the naive Bayes approach can end up in a case where some classes have the same probability $P(y)$. There are several ways to break ties for a given prediction. The following options are available:

- **Random breaker:** chooses a random class in the set of classes having the same $P(y)$.
- **Smallest Index:** chooses the first class encountered in the set of classes having the same $P(y)$. 

Laplace smoothing parameter:

The Laplace smoothing prevents from getting probabilities equal to zero or one.

The Laplace smoothing parameter $\theta$ is a positive real number added to the computation of the probability mass function $P(X_n = k)$ as follows:

$$P(X_n = k) = \frac{n_k + \theta}{\sum_k n_k + \theta |V|}$$

Where $X_n$ is either a qualitative or a discrete quantitative variable.

The support of $X_n$: $V$ is considered to be finite; the size of $V$ is $|V|$.

Validation tab:

The validation technique used to check the consistency of the Naive Bayes classification model is the K-fold cross validation technique. Data is partitioned into $k$ subsamples of equal size. Among the $k$ subsamples, a single subsample is retained as the validation data to test the model, and the remaining $k - 1$ subsamples are used as training data. $k$ can be specified in the number of folds field.

Missing data tab:

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- Mean or mode: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

- Nearest Neighbor: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Results by class: Activate this option to display a table giving the statistics and the objects for each of the classes.

Results by object: Activate this option to display a table giving the class each object (observation) is assigned to in the initial object order.
**Posterior probabilities of each class**: Activate this option to display the table which summarizes the posterior probabilities corresponding to each class \( P(Y = y) \) for all predicted observations.

**Confusion matrix**: Activate this option to display the confusion matrix. The confusion matrix contains information about observed and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. Diagonal values correspond to correct predictions. The higher the sum of the diagonal values according to the total the better the classifier.

**Accuracy of the model**: Activate this option to display model accuracy, which is the proportion of correct predictions.

---

**Results**

**Descriptive statistics**: The table of descriptive statistics shows the simple statistics for all the variables selected. The number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the quantitative variables. For qualitative variables, including the dependent variable, the categories with their respective frequencies and percentages are displayed.

**Results corresponding to the descriptive statistics of the learning set:**

The number of observations corresponding to each variable in the learning set, its mean (in case of quantitative variable or modes in case of qualitative variable) and its standard deviation.

**Results corresponding to the parameters involved in the classification process:**

The kind of probability distribution is reported.

The qualitative variables are considered to follow implicitly an empirical distribution.

The nature of the a priori distribution of the classes (uniform, not uniform) is also reported.

**Results regarding the classifier**

In order to evaluate and to score the naive Bayes classifier, a simple confusion matrix computed using the leave one out method as well as an accuracy index are displayed.

**Results regarding the validation method**

The error rate of the naive Bayes model obtained using the K folded-cross validation is reported. The number of folds is also reported to the user.

The cross validation results enables the selection of the adequate model parameters.
Result corresponding to the predicted classes

The predicted classes obtained using the naive Bayes classifier are displayed. In addition to the predicted classes, the a posteriori probabilities used to predict each observation are also reported.

Example

A tutorial on how to use the naive Bayes classifier is available on the Addinsoft website:

http://www.xlstat.com/demo-naive.htm

References


PLS/PCR/OLS Regression

Use this module to model and predict the values of one or more dependent quantitative variables using a linear combination of one or more explanatory quantitative and/or qualitative variables.

Description

The three regression methods available in this module have the common characteristic of generating models that involve linear combines of explanatory variables. The difference between the three method lies on the way the correlation structures between the variables are handled.

OLS Regression

From the three methods it is the most classical. Ordinary Least Squares regression (OLS) is more commonly named linear regression (simple or multiple depending on the number of explanatory variables).

In the case of a model with $p$ explanatory variables, the OLS regression model writes

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon$$

where $Y$ is the dependent variable, $\beta_0$, is the intercept of the model, $X_j$ corresponds to the $j^{th}$ explanatory variable of the model ($j=1$ to $p$), and $\epsilon$ is the random error with expectation 0 and variance $\sigma^2$.

In the case where there are $n$ observations, the estimation of the predicted value of the dependent variable $Y$ for the $i^{th}$ observation is given by:

$$\hat{y}_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

The OLS method corresponds to minimizing the sum of square differences between the observed and predicted values. This minimization leads to the following estimators of the parameters of the model:

$$\begin{align*}
\hat{\beta} &= \left(X' DX\right)^{-1} X' D Y \\
\hat{\sigma}^2 &= \frac{1}{W-p} \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2
\end{align*}$$
where $\hat{\beta}$ is the vector of the estimators of the $\beta$ parameters, $X$ is the matrix of the explanatory variables preceded by a vector of 1s, $y$ is the vector of the $n$ observed values of the dependent variable, $p^*$ is the number of explanatory variables to which we add 1 if the intercept is not fixed, $w_i$ is the weight of the $i^{th}$ observation, and $W$ is the sum of the $w_i$ weights, and $D$ is a matrix with the $w_i$ weights on its diagonal.

The vector of the predicted values writes:

$$\hat{y} = X (X' DX)^{-1} X' Dy$$

The limitations of the OLS regression come from the constraint of the inversion of the $X'X$ matrix: it is required that the rank of the matrix is $p+1$, and some numerical problems may arise if the matrix is not well behaved. XLSTAT uses algorithms due to Dempster (1969) that allow circumventing these two issues: if the matrix rank equals $q$ where $q$ is strictly lower than $p+1$, some variables are removed from the model, either because they are constant or because they belong to a block of collinear variables.

Furthermore, an automatic selection of the variables is performed if the user selects a too high number of variables compared to the number of observations. The theoretical limit is $n-1$, as with greater values the $X'X$ matrix becomes non-invertible.

The deleting of some of the variables may however not be optimal: in some cases we might not add a variable to the model because it is almost collinear to some other variables or to a block of variables, but it might be that it would be more relevant to remove a variable that is already in the model and to the new variable.

For that reason, and also in order to handle the cases where there a lot of explanatory variables, other methods have been developed.

**PCR Regression**

PCR (Principal Components Regression) can be divided into three steps: we first run a PCA (Principal Components Analysis) on the table of the explanatory variables, then we run an OLS regression on the selected components, then we compute the parameters of the model that correspond to the input variables.

PCA allows to transform an $X$ table with $n$ observations described by variables into an $S$ table with $n$ scores described by $q$ components, where $q$ is lower or equal to $p$ and such that $(S'S)$ is invertible. An additional selection can be applied on the components so that only the $r$ components that are the most correlated with the $Y$ variable are kept for the OLS regression step. We then obtain the $R$ table.

The OLS regression is performed on the $Y$ and $R$ tables. In order to circumvent the interpretation problem with the parameters obtained from the regression, XLSTAT transforms the results back into the initial space to obtain the parameters and the confidence intervals that correspond to the input variables.
**PLS Regression**

This method is quick, efficient and optimal for a criterion based on covariances. It is recommended in cases where the number of variables is high, and where it is likely that the explanatory variables are correlated.

The idea of PLS regression is to create, starting from a table with \( n \) observations described by \( p \) variables, a set of \( h \) components with \( h < p \). The method used to build the components differs from PCA, and presents the advantage of handling missing data. The determination of the number of components to keep is usually based on a criterion that involves a cross-validation. The user may also set the number of components to use.

Some programs differentiate PLS1 from PLS2. PLS1 corresponds to the case where there is only one dependent variable. PLS2 corresponds to the case where there are several dependent variables. The algorithms used by XLSTAT are such that the PLS1 is only a particular case of PLS2.

In the case of the OLS and PCR methods, if models need to be computed for several dependent variables, the computation of the models is simply a loop on the columns of the dependent variables table \( Y \). In the case of PLS regression, the covariance structure of \( Y \) also influences the computations.

The equation of the PLS regression model writes:

\[
Y = T_h C_h' + E_h \\
= X W_h^* C_h' + E_h \\
= X W_h \left( P_h W_h \right)^{-1} C_h' + E_h
\]

where \( Y \) is the matrix of the dependent variables, \( X \) is the matrix of the explanatory variables. \( T_h, C_h, W_h^*, W_h, \) and \( P_h \), are the matrices generated by the PLS algorithm, and \( E_h \) is the matrix of the residuals.

The matrix \( B \) of the regression coefficients of \( Y \) on \( X \), with \( h \) components generated by the PLS regression algorithm is given by:

\[
B = W_h \left( P_h W_h \right)^{-1} C_h'
\]

Note: the PLS regression leads to a linear model as the OLS and PCR do.

Notes:

The three methods give the same results if the number of components obtained from the PCA (in PCR) or from the PLS regression is equal to the number of explanatory variables.
The components obtained from the PLS regression are built so that they explain as well as possible \( Y \), while the components of the PCR are built to describe \( X \) as well as possible. XLSTAT allows partly compensating this drawback of the PCR by allowing the selection of the components that are the most correlated with \( Y \).

**PLS Discriminant Analysis**

PLS regression can be adapted to fit discriminant analysis. The PLS discriminant analysis uses the PLS algorithm to explain and predict the membership of observations to several classes using quantitative or qualitative explanatory variables. XLSTAT-PLS uses the PLS2 algorithm applied on the full disjunctive table obtained from the qualitative dependant variable.

PLS discriminant analysis can be applied in many cases when classical discriminant analysis cannot be applied. For example, when the number of observations is low and when the number of explanatory variables is high. When there are missing values, PLS discriminant analysis can be applied on the data that is available. Finally, as PLS regression, it is adapted when multicolinearity between explanatory variables is high.

As many models as categories of the dependant variable are obtained. An observation is associated to the category that has an equation with the highest value.

Let \( K \) be the number of categories of the dependant variable \( Y \). For each category \( a_1, \ldots, a_K \), an equation of the model is obtained:

\[
F(y_i, a_k) = b_0 + \sum_{j=1}^{p} b_j x_{ij}
\]

With \( a_k \) being a category of the dependant qualitative variable, \( b_0 \) being the intercept of the model associated to \( a_k \), \( p \) being the number of explanatory variables and \( b_i \) being the coefficients of the same model.

Observation \( i \) is associated to class \( k \) if:

\[
k^* = \arg \max_k F(y_i, a_k)
\]

PLS discriminant analysis offers an interesting alternative to classical linear discriminant analysis.

The output mixes the outputs of the PLS regression with classical discriminant analysis outputs such as confusion matrix.
**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- ![Refresh](image): Click this button to reload the default options.
- ![Delete](image): Click this button to delete the data selections.
- ![Arrow Down](image), ![Arrow Right](image): Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab:**

**Y / Dependent variables:**

**Quantitative**: Select the dependent variable(s). The data must be numerical. If the “Variable labels” option is activated make sure that the headers of the variables have also been selected.

**X / Explanatory variables:**

**Quantitative**: Activate this option if you want to include one or more quantitative explanatory variables. Then select the corresponding data. The data must be numerical. If the “Variable labels” option is activated make sure that the headers of the variables have also been selected.

**Qualitative**: Activate this option if you want to include one or more qualitative explanatory variables. Then select the corresponding data. Whatever their Excel format, the data are considered as categorical. If the “Variable labels” option is activated make sure that the headers of the variables have also been selected.
**Method**: Choose the regression method you want to use:

- **PLS**: Activate this option to compute a Partial Least Squares regression.
- **PLS-DA**: Activate this option to compute a Partial Least Squares Discriminant Analysis.
- **PCR**: Activate this option to compute Principal Components Regression.
- **OLS**: Activate this option to compute an Ordinary Least Squares regression.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Variable labels**: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

**Observation labels**: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

**Observation weights**: Activate this option if you want to weight the observations. If you do not activate this option, the weights are considered to be equal to 1. The weights must be greater or equal to 0 and they must be integer values. Setting a case weight to 2 is equivalent to repeating twice the same observation. If the "Variable labels" option is activated, make sure that the header of the selection has also been selected.

**Regression weights**: This option is active only with PCR and OLS regression. Activate this option if you want to run a weighted least squares regression. If you do not activate this option, the regression weights are considered to be equal to 1. The weights must be greater or equal to 0. If the "Variable labels" option is activated make sure that the header of the selection has also been selected.

**Options** tab:

**Common options:**
Confidence interval (%): Enter the size in % of the confidence interval that is used for the various tests, parameters and predictions. Default value: 95.

Interactions / Level: Activate this option to include interactions in the model then enter the maximum interaction level (value between 1 and 4).

X / Explanatory variables:
- Center: Activate this option is you want to center the explanatory variables before starting the calculations.
- Reduce: Activate this option is you want to reduce the explanatory variables before starting the calculations.

Options for PLS regression:

Stop conditions:
- Automatic: Activate this option so that XLSTAT automatically determines the number of components to keep.
- $Q_i^2$ threshold: Activate this option to set the threshold value of the $Q_i^2$ criterion used to determine if the contribution of a component is significant or not. The default value is 0.0975 which corresponds to 1-0.95².
- $Q_i^2$ improvement: Activate this option to set the threshold value of the $Q_i^2$ improvement criterion used to determine if the contribution of a component is significant or not. The default value is 0.05 which corresponds to a 5% improvement. This value is computed as follows:

$$Q^2(h)\text{Imp} = \frac{Q^2(h) - Q^2(h-1)}{Q^2(h-1)}$$

- Minimum Press: Activate this option so that the number of components used in the model corresponds to the model with the minimum Press statistic.
- Max components: Activate this option to set the pour fixer le maximum number of components to take into account in the model.

Stop conditions:
- Automatic: Activate this option so that XLSTAT automatically determines the number of components to keep.
Options for PCR regression:

**Standardized PCA**: Activate this option to run a PCA on the correlation matrix. Inactivate this option to run a PCA on the covariance matrix (unstandardized PCA).

**Filter components**: You can activate one of the two following options in order to reduce the number of components used in the model:

- **Minimum %**: Activate this option and enter the minimum percentage of total variability that the selected components should represent.
- **Maximum number**: Activate this option to set the maximum number of components to take into account.

**Sort components by**: Choose one of the following options to determine “which criterion should be used to select the components on the basis of the “Minimum %”, or of the “Maximum number”:

- **Correlations with Ys**: Activate this option so that the components selection is based on the sorting down of $R^2$ coefficient between the dependent variable $Y$ and the components. This option is recommended.
- **Eigenvalues**: Activate this option so that the selection of the components is based on the sorting down of the eigenvalues corresponding to the components.

Options for PCR and OLS regression:

**Fixed intercept**: Activate this option to set the intercept (or constant) of the model to a given value. Then enter the value in the corresponding field (0 by default).

**Tolerance**: Activate this option to allow the OLS algorithm to automatically remove the variables that would either be constant or highly correlated with other variables or group of variables (Minimum and default value is 0.0001. Maximum value allowed is 1). The higher the tolerance, the more the model tolerates collinearities between the variables.

Options for OLS regression:

**Constraints**: This option is active only if you have selected qualitative explanatory variables. Choose the type of constraint:

- **$a1 = 0$**: For each qualitative variable, the parameter of the model that corresponds to the first category of the variable is set to 0. This type of constraint is useful when you consider that the first category corresponds to a standard, or to a null effect.
- **$\text{Sum}(ai) = 0$**: For each qualitative variable, the sum of the parameters corresponding to the various categories equals 0.
- **$\text{Sum}(ni.ai) = 0$**: For each qualitative variable, the sum of the parameters corresponding to the various categories weighted by their frequency equals 0.
Model selection: Activate this option if you want to use one of the following model selection methods:

- **Best model**: This method allows choosing the best model among all the models that are based on a number of variables that is bounded by "Min variables" and "Max variables". The quality of the model depends on a selection "Criterion".

- **Criterion**: Select the criterion in the following list: adjusted R², Mean Squares of Errors (MSE), Mallows’ Cp, Akaike’s AIC, Schwarz’s SBC, Amemiya’s PC.

- **Min variables**: Enter the minimum number of variables to take into account in the model.

- **Max variables**: Enter the maximum number of variables to take into account in the model.

- **Note**: this method can lead to very long computations because the total number of models explored is the sum of the Cn,k where k varies between "Min variables" and "Max variables", and where Cn,k is n!/[n-k]!k !. It is therefore highly recommended that you increase step by step the value of "Max variables".

- **Stepwise**: The selection process starts with the adding of the variable that contributes the most to the model (the criterion used here is the Student’s t statistic). If a second variable is such that the probability of its t is lower than the "Threshold level", it is added to the model. The procedure is the same for the third variable. Then, starting with the third variable, the algorithm evaluates how the removal of one of the variables would impact the model. If the probability corresponding to the Student’s t of one of the variables is greater than the "Threshold level", the variable is removed. The procedure continues until no variable can be either added or removed from the model.

- **Forward**: The procedure is identical to the stepwise, except that there are no removal steps.

- **Backward**: The procedure starts with the selection of all the available variables. The variables are then removed from the model one by one using the same methodology as for the stepwise selection.

- **Threshold level**: Enter the value of the threshold probability for the Student’s t statistic during the selection process.

**Validation** tab:

**Validation**: Activate this option if you want to use a sub-sample of the data to validate the model.

**Validation set**: Choose one of the following options to define how to obtain the observations used for the validation:
• **Random**: The observations are randomly selected. The “Number of observations” N must then be specified.

• **N last rows**: The N last observations are selected for the validation. The “Number of observations” N must then be specified.

• **N first rows**: The N first observations are selected for the validation. The “Number of observations” N must then be specified.

• **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to use for the validation.

**Prediction** tab:

**Prediction**: Activate this option if you want to select data to use them in prediction mode. If activate this option, you need to make sure that the prediction dataset is structured as the estimation dataset: same variables with the same order in the selections. On the other hand, variable labels must not be selected: the first row of the selections listed below must correspond to data.

**Quantitative**: Activate this option to select the quantitative explanatory variables. The first row must not include variable labels.

**Qualitative**: Activate this option to select the qualitative explanatory variables. The first row must not include variable labels.

**Observations labels**: Activate this option if observations labels are available. Then select the corresponding data. If this option is not activated, the observations labels are automatically generated by XLSTAT (PredObs1, PredObs2 …).

**Missing data** tab:

These options are available only for PCR and OLS regression. With PLS regression, the missing data are automatically handled by the algorithm.

**Remove observations**: Activate this option to remove the observations with missing data.

**Estimate missing data**: Activate this option to estimate missing data before starting the computations.

• **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

• **Nearest neighbour**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.
**Outputs** tab:

**Options common to the three methods:**

**Descriptive statistics:** Activate this option to display the descriptive statistics for all the selected variables.

**Correlations:** Activate this option to display the correlation matrix for the quantitative variables (dependent and explanatory).

**Standardized coefficients:** Activate this option to display the standardized parameters of the model (also called beta coefficients).

**Equation:** activate this option to explicitly display the equation of the model.

**Predictions and residuals:** Activate this option to display the table of predictions and residuals.

**Option for PLS regression:**

**t, u and u~components:** Activate this option to display the tables corresponding to the components. If this option is not activated the corresponding charts are not displayed.

**c, w, w* and p vectors:** Activate this option to display the tables corresponding to the vectors obtained from the PLS algorithm. If this option is not activated the corresponding charts are not displayed.

**VIPs:** Activate this option to display the table and the charts of the Variable Importance for the Projection.

**Confidence intervals:** Activate this option to compute the confidence intervals of the standardized coefficients. The computations involve a jackknife method.

**Outliers analysis:** Activate this option to display the table and the charts of the outliers analysis.

**Option for PLS-DA:**

**t, u and u~components:** Activate this option to display the tables corresponding to the components. If this option is not activated the corresponding charts are not displayed.

**c, w, w* and p vectors:** Activate this option to display the tables corresponding to the vectors obtained from the PLS algorithm. If this option is not activated the corresponding charts are not displayed.
**VIPs**: Activate this option to display the table and the charts of the Variable Importance for the Projection.

**Confidence intervals**: Activate this option to compute the confidence intervals of the standardized coefficients. The computations involve a jackknife method.

**Outliers analysis**: Activate this option to display the table and the charts of the outliers analysis.

**Confusion matrix**: Activate this option to display the table showing the numbers of well- and badly-classified observations for each of the classes.

**Options for PCR regression:**

**Factor loadings**: Activate this option to display the factor loadings. The factor loadings are equal to the correlations between the principal components and the input variables if the PCA is based on the correlation matrix (standardized PCA).

**Correlations Factors/Variables**: Activate this option to display the correlations between the principal component and the input variables.

**Factor scores**: Activate this option to display the factor scores (coordinates of the observations in the new space) generated by the PCA. The scores are used in the regression step of the PCR.

**Options for PCR and OLS regression:**

**Analysis of variance**: Activate this option to display the analysis of variance table.

**Adjusted predictions**: Activate this option to compute and display the adjusted predictions in the predictions and residuals table.

**Cook’s D**: Activate this option to compute and display the Cook’s distances in the predictions and residuals table.

**Press**: Activate this option to compute and display the Press statistic.

**Charts** tab:

**Options common to the three methods:**

**Regression charts**: Activate this option to display the regression charts:

- **Standardized coefficients**: Activate this option to display a chart with the standardized coefficients of the model, and the corresponding confidence intervals.
• **Predictions and residuals**: Activate this option to display the following charts:

  (1) Regression line: this chart is displayed only if there is one explanatory variable and if that variable is quantitative.

  (2) Explanatory variable versus standardized residuals: this chart is displayed only if there is one explanatory variable and if that variable is quantitative.

  (3) Dependent variable versus standardized residuals.

  (4) Predictions versus observed values.

  (5) Bar chart of the standardized residuals.

• **Confidence intervals**: Activate this option to display the confidence intervals on charts (1) and (4).

**Options for the PLS regression and the PCR:**

**Correlation charts**: Activate this option to display the charts involving correlations between the components and input variables. In the case of PCR, activate this option to display the correlation circle.

  • **Vectors**: Activate this option to display the input variables with vectors.

**Observations charts**: activate this option to display the charts that allow visualizing the observations in the new space.

  • **Labels**: Activate this option to display the observations labels on the charts. The number of labels can be modulated using the filtering option.

**Biplots**: Activate this option to display the charts where the input variables and the observations are simultaneously displayed.

  • **Vectors**: Activate this option to display the input variables with vectors.

  • **Labels**: Activate this option to display the observations labels on the biplots. The number of labels can be modulated using the filtering option.

**Colored labels**: Activate this option to display the labels with the same color as the corresponding points. If this option is not activated the labels are displayed in black.

**Filter**: Activate this option to modulate the number of observations displayed:

  • **Random**: The observations to display are randomly selected. The “Number of observations” N to display must then be specified.

  • **N first rows**: The N first observations are displayed on the chart. The “Number of observations” N to display must then be specified.
• **N last rows**: The N last observations are displayed on the chart. The “Number of observations” N to display must then be specified.

• **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to display.

### Results

**Descriptive statistics**: the tables of descriptive statistics display for all the selected variables a set of basic statistics. For the dependent variables (colored in blue), and the quantitative explanatory variables, XLSTAT displays the number of observations, the number of observations with missing data, the number of observations with no missing data, the mean, and the unbiased standard deviation. For the qualitative explanatory variables XLSTAT displays the name and the frequency of the categories.

**Correlation matrix**: this table is displayed to allow you visualizing the correlations among the explanatory variables, among the dependent variables and between both groups.

### Results of the PLS regression:

The first table displays the **model quality** indexes. The quality corresponds here to the cumulated contribution of the components to the indexes:

- The $Q^2_{cum}$ index measures the global contribution of the $h$ first components to the predictive quality of the model (and of the sub-models if there are several dependent variables). The $Q^2_{cum}(h)$ index writes:

$$Q^2_{cum}(h) = 1 - \prod_{j=1}^{h} \frac{\sum_{j}^{q} PRESS_{kj}}{\sum_{k=1}^{q} SCE_{k(j-1)}}$$

The index involves the PRESS statistic (that requires a cross-validation), and the Sum of Squares of Errors (SSE) for a model with one less component. The search for the maximum of the $Q^2_{cum}$ index is equivalent to finding the most stable model.

- The $R^2_{Ycum}$ index is the sum of the coefficients of determination between the dependent variables and the $h$ first components. It is therefore a measure of the explanatory power of the $h$ first components for the dependent variables of the model.

- The $R^2_{Xcum}$ index is the sum of the coefficients of determination between the explanatory variables and the $h$ first components. It is therefore a measure of the explanatory power of the $h$ first components for the explanatory variables of the model.
A bar chart is displayed to allow the visualization of the evolution of the three indexes when the number of components increases. While the $R^2Y_{cum}$ and $R^2X_{cum}$ indexes necessarily increase with the number of components, this is not the case with $Q^2cum$.

The next table corresponds to the correlation matrix of the explanatory and dependent variables with the $t$ and $\bar{u}$ components. A chart displays the correlations with the $t$ components.

The next table displays the $w$ vectors, followed by the $w^*$ vectors and the $c$ vectors, that are directly involved in the model, as it is shown in the “Description” section. If to $h=2$ corresponds a valid model, it is shown that the projection of the $x$ vectors on the $y$ vectors on the variables on the $w^*/c$ axes chart, gives a fair idea of the sign and the relative weight of the corresponding coefficients in the model.

The next table displays the scores of the observations in the space of the $t$ components. The corresponding chart is displayed. If some observations have been selected for the validation, they are displayed on the chart.

The next table displays the standardized scores of the observations in the space of the $t$ components. These scores are equivalent to computing the correlations of each observation (represented by an indicator variable) with the components. This allows displaying the observations on the correlations map that follows where the Xs, the Ys and the observations are simultaneously displayed. An example of an interpretation of this map is available in Tenenhaus (2003).

The next table corresponds to the scores of the observations in the space of the $u$ and then the $u^\sim$ components. The chart based on the $u^\sim$ is displayed. If some observations have been selected for the validation, they are displayed on the chart.

The table with the $Q^2$ quality indexes allows visualizing how the components contribute to the explanation of the dependent variables. The table of the cumulated $Q^2$ quality indexes allows measuring the quality that corresponds to a space with an increasing number of dimensions.

The table of the $R^2$ and redundancies between the input variables (dependent and explanatory) and the components $t$ and $u^\sim$ allow evaluating the explanatory power of the $t$ and $u^\sim$. The redundancy between an $X$ table ($n$ rows and $p$ variables) and a c component is the part of the variance of $X$ explained by $c$. We define it as the mean of the squares of the correlation coefficients between the variables and the component:

$$Rd(X, c) = \frac{1}{p} \sum_{j=1}^{p} R^2(x_j, c)$$

From the redundancies one can deduce the VIPs (Variable Importance for the Projection) that measure the importance of an explanatory variable for the building of the $t$ components. The VIP for the $j$th explanatory variable and the component $h$ is defined by:
On the VIP charts (one bar chart per component), a border line is plotted to identify the VIPs that are greater than 0.8 and above: these thresholds suggested by Wold (1995) and Ericksson (2001) allow identifying the variables that are moderately (0.8<VIP<1) or highly influential (VIP>1).

The next table displays the outliers analysis. The DModX (distances from each observation to the model in the space of the X variables) allow identifying the outliers for the explanatory variables, while the DModY (distances from each observation to the model in the space of the Y variables) allow identifying the outliers for the dependent variables. On the corresponding charts the threshold values DCrit are also displayed to help identifying the outliers: the DMod values that are above the DCrit threshold correspond to outliers. The DCrit are computed using the threshold values classically used in box plots. The value of the DModX for the ith observation writes:

\[ D_{ModX_i} = \sqrt{\frac{n}{n-h-1} \frac{\sum_{j=1}^{p} e(X,t)^2_{ij}}{p-h}} \]

where the \( e(X,t)_{ij} \) (i = 1 … n) are the residuals of the regression of X on the jth component. The value of the DModY for the ith observation writes:

\[ D_{ModY_i} = \sqrt{\frac{\sum_{j=1}^{q} e(Y,t)^2_{ij}}{q-h}} \]

where q is the number of dependent variables and the \( e(Y,t)_{ij} \) (i = 1 … n) are the residuals of the regression of Y on the jth component.

The next table displays the parameters of the models corresponding to the one or more dependent variables. It is followed by the equation corresponding to each model, if the number of explanatory variables does not exceed 20.

For each of the dependent variables a series of tables and charts is displayed.

**Goodness of fit statistics:** This table displays the goodness of fit statistics of the PLS regression model for each dependent variable. The definition of the statistics is as follows:

The table of the standardized coefficients (also named beta coefficients) allows comparing the relative weight of the variables in the model. To compute the confidence intervals, in the case of PLS regression, the classical formulae based on the normality hypotheses used in
OLS regression do not apply. A bootstrap method suggested by Tenenhaus et al. (2004) allows estimating the confidence intervals. The greater the absolute value of a coefficient, the greater the weight of the variable in the model. When the confidence interval around the standardized coefficients includes 0, which can easily be observed on the chart, the weight of the variable in the model is not significant.

In the predictions and residuals table, the weight, the observed value of the dependent variable, the corresponding prediction, the residuals and the confidence intervals are displayed for each observation. Two types of confidence intervals are displayed: an interval around the mean (it corresponds to the case where the prediction is made for an infinite number of observations with a give set of values of the explanatory variables) and an interval around an individual prediction (it corresponds to the case where the prediction is made for only one observation). The second interval is always wider than the first one, as the uncertainty is of course higher. If some observations have been selected for the validation, they are displayed in this table.

The three charts that are displayed afterwards allow visualizing:

- the residuals versus the dependent variable,
- the distance between the predicted and observed values (for an ideal model the all the points would be on the bisecting line),
- the bar chart of the residuals.

If you have selected data to use in prediction mode, a table displays the predictions on the new observations and the corresponding confidence intervals.

**PLS-DA specific results:**

**Classification functions:** The classification functions can be used to determine which class an observation is to be assigned to using values taken for the various explanatory variables. These functions are linear. An observation is assigned to the class with the highest classification function F().

**Prior and posterior classification and scores:** This table shows for each observation its membership class defined by the dependent variable, the membership class as deduced by the membership probabilities and the classification function score for each category of the dependent variable.

**Confusion matrix for the estimation sample:** The confusion matrix is deduced from prior and posterior classifications together with the overall percentage of well-classified observations.

**Results of the PCR regression:**
The PCR regression requires a Principal Component Analysis step. The first results concern the latter.

**Eigenvalues**: the table of the eigenvalues and the corresponding scree plot are displayed. The number of eigenvalues displayed is equal to the number of non null eigenvalues. If a components filtering option has been selected it is applied only before the regression step.

If the corresponding outputs options have been activated, XLSTAT displays the **factor loadings** (the coordinates of the input variables in the new space), then the correlations between the input variables and the components. The **correlations** are equal to the factor loadings if the PCA is performed on the correlation matrix. The next table displays the **factor scores** (the coordinates of the observations in the new space), and are later used for the regression step. If some observations have been selected for the validation, they are displayed in this table. A biplot is displayed if the corresponding option has been activated.

If the filtering option based on the correlations with the dependent variables has been selected, the components used in the regression step are those that have the greatest determination coefficients (R²) with the dependent variables. The matrix of the correlation coefficients between the components and the dependent variables is displayed. The number of components that are kept depends on the number of eigenvalues and on the selected options ("Minimum %" or "Max components").

If the filtering option based on the eigenvalues has been selected, the components used in the regression step are those that have the greatest eigenvalues. The number of components that are kept depends on the number of eigenvalues and on the selected options ("Minimum %" or "Max components").

**Results common to the PCR and OLS regressions:**

**Goodness of fit statistics**: this table displays statistics that are related to the goodness of fit of the regression model:

- **Observations**: the number of observations taken into account for the computations. In the formulae below, \( n \) corresponds to number of observations.
- **Sum of weights**: the sum of weights of the observations taken into account. In the formulae below, \( W \) corresponds to the sum of weights.
- **DF**: the number of degrees of freedom of the selected model (corresponds to the error DF of the analysis of variance table).
- **R²**: the coefficient of determination of the model. This coefficient, which value is between 0 and 1, is displayed only if the intercept of the model has not been fixed by the user. The value of this coefficient is computed as follows:
\[ R^2 = 1 - \frac{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} w_i (y_i - \bar{y})^2}, \text{ with } \bar{y} = \frac{1}{n} \sum_{i=1}^{n} w_i y_i \]

- The \( R^2 \) is interpreted as the proportion of variability of the dependent variable explained by the model. The closer the \( R^2 \) to 1, the better fitted the model. The major drawback of the \( R^2 \) is that it does not take into account the number variables used to fit the model.

- **Adjusted \( R^2 \):** the adjusted coefficient of determination of the model. The adjusted \( R^2 \) can be negative if the \( R^2 \) is close to zero. This coefficient is displayed only if of the model has not been fixed by the user. The value of this coefficient is computed as follows:

\[ \hat{R}^2 = 1 - \left(1 - R^2\right) \frac{W-1}{W - p - 1} \]

The adjusted \( R^2 \) is a correction of the \( R^2 \) that allows taking into account the number of variables used in the model.

- **MSE:** the Mean Squares of Errors (MSE) is defined by:

\[ MSE = \frac{1}{W - p} \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2 \]

- **RMSE:** the Root Mean Squares of Errors (RMSE) is the square root of the MSE.

- **MAPE:** the Mean Absolute Percentage Error (MAPE) is computed as follows:

\[ MAPE = \frac{100}{W} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \]

- **DW:** the Durbin-Watson statistic is defined by

\[ DW = \frac{\sum_{i=2}^{n} \left[ (y_i - \hat{y}_i) - (y_{i-1} - \hat{y}_{i-1}) \right]^2}{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2} \]

This statistic corresponds to the order 1 autocorrelation coefficient and allows verifying if the residuals are not autocorrelated. The independence of the residuals is one of the hypotheses of the linear regression. The user will need to look at a Durbin-Watson table to know if the hypothesis of independence between the residuals is accepted or rejected.

- **Cp:** the Mallows’ Cp is defined by:
\[ Cp = \frac{SCE}{\hat{\sigma}} + 2p^* - W \]

where SSE is the sum of squares of errors for the model with p explanatory variables, and where \( \hat{\sigma} \) corresponds to the estimator of the variance of the residuals for the model that includes all the explanatory variables. The closer the Cp coefficient to p* the less biased the model.

- **AIC**: the Akaike’s Information Criterion (AIC) is defined by:

\[ AIC = W \ln \left( \frac{SCE}{W} \right) + 2p^* \]

This criterion suggested by Akaike (1973) derives from the information theory and is based on the Kullback and Leibler measure (1951). It is a models selection criterion that penalizes models for which the addition of a new explanatory variable does not bring sufficient information. The lower the AIC, the better the model.

- **SBC**: the Schwarz’s Bayesian Criterion writes:

\[ SBC = W \ln \left( \frac{SCE}{W} \right) + \ln(W)p^* \]

This criterion suggested by Schwarz (1978) is close to the AIC, and the goal is to minimize it.

- **PC**: the Amemiya’s Prediction Criterion) writes

\[ PC = \frac{(1 - R^2)(W + p^*)}{W - p^*} \]

This criterion suggested by Amemiya (1980) allows as the adjusted \( R^2 \) to take into account the parsimony of the model.

- **Press RMCE**: the Press RMSE statistic is displayed only if the corresponding option has been activated in the dialog box. The Press statistic is defined by

\[ \text{Press} = \sum_{i=1}^{n} w_i \left( y_i - \hat{y}_{i(-i)} \right)^2 \]

where \( \hat{y}_{i(-i)} \) is the prediction of the \( i^{th} \) observation when it is not included in the data set used for the estimation of the parameters of the model. When obtain:

\[ \text{Press RMCE} = \sqrt{\frac{\text{Press}}{W - p^*}} \]
The Press RMSE can then be compared to the RMSE. A large difference between both indicates that the model is sensitive to the presence or absence of some observations.

The analysis of variance table allows evaluating how much information the explanatory variables bring to the model. In the case where the intercept of the model is not fixed by the user, the explanatory power is measured by comparing the fit of the selected model with the fit of a basic model where the dependent variable equals its mean. When the intercept is fixed to a given value, the selected model is compared to a basic model where the dependent model equals the fixed intercept.

In the case of a PCR regression, the first table of model parameters corresponds to the parameters of the model based on the selected components. This table is not easy to interpret. For that reason a transformation is performed to obtain the parameters of the model corresponding to the input variables. The latter table is directly obtained in the case of an OLS regression. In this table you will find the estimate of the parameters, the corresponding standard error, the Student’s t, the corresponding probability, as well as the confidence interval.

The equation of the model is then displayed to facilitate the visualization or the reuse of the model.

The table of the standardized coefficients (also named beta coefficients) allows comparing the relative weight of the variables in the model. The greater the absolute value of a coefficient, the greater the weight of the variable in the model. When the confidence interval around the standardized coefficients includes 0, which can easily be observed on the chart, the weight of the variable in the model is not significant.

In the predictions and residuals table, the weight, the value of the explanatory variable if there is only one, the observed value of the dependent variable, the corresponding prediction, the residuals and the confidence intervals, the adjusted prediction and the Cook’s D, are displayed for each observation. Two types of confidence intervals are displayed: an interval around the mean (it corresponds to the case where the prediction is made for an infinite number of observations with a give set of values of the explanatory variables) and an interval around an individual prediction (it corresponds to the case where the prediction is made for only one observation). The second interval is always wider than the first one, as the uncertainty is of course higher. If some observations have been selected for the validation, they are displayed in this table.

The charts that follow allow visualizing the results listed above. If there is only one explanatory variable in the model, and if that variable is quantitative, then the first chart allows visualizing the observations, the regression line and the confidence intervals around the prediction. The second chart displays the standardized residuals versus the explanatory variable. The residuals should be randomly distributed around the abscissa axis. If a trend can be observed, that means there is a problem with the model.

The three charts that are displayed afterwards allow visualizing respectively the standardized residuals versus the dependent variable, the distance between the predicted and observed
values (for an ideal model the all the points would be on the bisecting line), and the bar chart of
the standardized residuals. The third chart makes it possible to quickly see if there is an
unexpected number of high residuals: the normality assumption for the residuals is such that
only 5% of the standardized residuals should be out of the \([-2, 2]\) interval.

If you have selected data to use in prediction mode, a table displays the **predictions on the new observations** and the corresponding confidence intervals.

**OLS regression results:**

If the Type I SS and Type III SS (SS: Sum of Squares) options have been activated, the corresponding tables are displayed.

The **Type I SS** table allows visualizing the influence of the progressive addition of new explanatory variables to the model. The influence is given by the Sum of Squares of Errors (SSE), de la Mean Squares of Errors (MSE), the Fisher’s F statistic, and the probability corresponding to the Fisher’s F. The smaller the probability, the more information the variable brings to the model. Note: the order of selection of the variables influences the results obtained here.

The **Type III SS** table allows visualizing the influence of the withdrawal of an explanatory variable on the goodness of fit of the model, all the other variables being included. The influence is measured by the Sum of Squares of Errors (SSE), de la Mean Squares of Errors (MSE), the Fisher’s F statistic, and the probability corresponding to the Fisher’s F. The smaller the probability, the more information the variable brings to the model. Note: the order of the variables in the selection does not influence the results in this table.

**Examples**

A tutorial on how to use PLS regression is available on the Addinsoft website on following page:

[http://www.xlstat.com/demo-pls.htm](http://www.xlstat.com/demo-pls.htm)

**References**


Correlated Component Regression (CCR)

Use this module to model and predict a designated dependent variable Y (continuous or dichotomous) from a set of P correlated explanatory variables (predictors) X = (X1, X2, … , XP).

Description

The four regression methods available in the Correlated Component Regression (CCR) module use fast cross-validation to determine the amount of regularization to produce reliable predictions from data with P correlated explanatory (X) variables, where multicollinearity may exist and P can be greater than the sample size N. The methods are based on Generalized Linear Models (GLM). As an option, the CCR step-down algorithm may be activated to exclude irrelevant Xs.

The linear part of the model is a weighted average of K components S = (S1, S2, … , SK) where each component itself is a linear combination of the predictors. For Y dichotomous, these methods provide an alternative to Logistic regression (CCR-Logistic) and linear discriminant analysis (CCR-LDA). For a continuous Y, these procedures provide an alternative to traditional linear regression methods, where components may be correlated (CCR-LM procedure), or restricted to be uncorrelated with components obtained by PLS regression techniques (CCR-PLS). Typically K<P, resulting in model regularization that reduces prediction error.

Traditional maximum likelihood regression methods, which employ no regularization at all, can be obtained as a special case of these models when K=P (the saturated model). Regularization, inherent in the CCR methods, reduces the variance (instability) of prediction and also lowers the mean squared error of prediction when the predictors have moderate to high correlation. The smaller the value for K, the more regularization is applied. Typically, K will be less than 10 (quite often K = 2, 3 or 4) regardless of P. M-fold cross-validation techniques are used to determine the amount of regularization $K^*$ to apply, and the number of predictors $P^*$ to include in the model when the step-down algorithm is utilized.

When the CCR step-down option is activated with M-fold cross-validation, output includes a table of predictor counts, reporting the number of times each predictor is included in a model estimated with one omitted fold. The counts can be used as an alternative measure of variable importance (Tenenhaus, 2010), as a supplement to the standardized regression coefficients. Additional options can limit the number of predictors to be included in the model.

The regression methods in the XLSTAT-CCR module differ according to the assumptions made about the scale type of the dependent variable Y (continuous vs. dichotomous), and the distributions (if any) assumed about the predictors.
Linear regression (CCR.LM, PLS)

Predictions for the dependent variable Y based on the linear regression model are obtained as follows:

\[ \hat{Y} = S(S'DS)^{-1}S'DY \] (0.1)

where D is a diagonal matrix with case weights as the diagonal entries.

For example, with K=2 components we have:

\[ \hat{Y} = \alpha + b_{1.2}S_1 + b_{2.1}S_2 \] (0.2)

where \( b_{1.2} \) and \( b_{2.1} \) are the component weights, the components defined as:

\[ S_1 = \sum_{g=1}^{p} \lambda_{g.1}X_g \quad \text{and} \quad S_2 = \sum_{g=1}^{p} \lambda_{g.2}X_g \]

where \( \lambda_{g.1} \) and \( \lambda_{g.2} \) are component coefficients (loadings) for the gth predictor on components \( S_1 \) and \( S_2 \) respectively.

The component weights and loadings are obtained from traditional OLS regression. By substitution we get the reduced form expression:

\[ \hat{Y} = \alpha + \sum_{g=1}^{p} (b_{1.2}\lambda_{g.1} + b_{2.1}\lambda_{g.2})X_g \] (0.3)

where

\[ \beta_g = b_{1.2}\lambda_{g.1} + b_{2.1}\lambda_{g.2} \] (0.4)

is the (regularized) regression coefficient associated with predictor \( X_g \).

Regardless which linear regression model (CCR-LM, or PLS) is used to generate the predictions, when the number of components K equals the number of predictors P, the results are identical to those obtained from traditional least squares (OLS or WLS) regression. Traditional least squares regression produces unbiased predictions, but such predictions may have large variance and hence higher mean squared error than regularized solutions (K < P). Thus, predictions obtained from the CCR module are typically more reliable than those obtained from a traditional regression model.

Methods CCR.LM and PLS assume that the dependent variable Y is continuous:

- CCR.LM is invariant to standardization and also allows the components to be correlated (recommended)
- PLS produces different results depending upon whether or not the predictors are standardized to have variance 1. By default, the PLS ‘standardize’ option is activated.
Logistic Regression (CCR.Logistic) and Linear Discriminant Analysis (CCR.LDA)

Logistic regression is the standard regression (classification) approach for predicting a dichotomous dependent variable. Both Linear and Logistic regression are GLM (Generalized Linear Models) in that a linear combination of the explanatory variables ('linear predictor') is used to predict a function of the dependent variable. In the case of linear regression, the mean of Y is predicted as a linear function of the X variables. For logistic regression, the logit of Y is predicted as a linear function of X.

\[
\text{Logit}(Y \mid S) = \alpha + b_{1.2}S_1 + b_{2.1}S_2
\]

which in reduced form yields:

\[
\text{Logit}(Y \mid X) = \alpha + \sum_{g=1}^{P} (b_{1.2}\lambda_{g.1} + b_{2.1}\lambda_{g.2})X_g
\]

Logit(Y), defined as the natural logarithm of the probability of being in dependent variable group 1 (say Y=1) divided by the probability of being in group 2 (say Y=0), can easily be transformed to yield the probability of being in either category. For example, the conditional probability of being in group 1 can be expressed as:

\[
\text{Prob}(Y = 1 \mid X) = \frac{\exp(\text{Logit}(Y \mid X))}{1 + \exp(\text{Logit}(Y \mid X))}
\]

\[
= \frac{1}{1 + \exp(-\text{Logit}(Y \mid X))}
\]

and

\[
\text{Prob}(Y = 0 \mid X) = \frac{1}{1 + \exp(\text{Logit}(Y \mid X))}
\]

Thus, the logistic regression model is a model for predicting the probability of being in a particular group. Predictions are reported for group 1, which is defined as the category of Y associated with the higher of the 2 numeric values taken on by Y.

Linear Discriminant Analysis (LDA) is another model used commonly to obtain predicted probabilities for a dichotomous Y:

- CCR.LDA assumes that the X variables follow a multivariate normal distribution within each Y group, with different group means but common variances and covariances.

- CCR.Logistic makes no distributional assumptions.

CCR.LM

The CCR.LM method applies CCR techniques to obtain a regularized linear regression based on the Correlated Component Regression (CCR) model for a continuous Y (Magidson, 2010; Magidson and Wassmann, 2010). It is recommended especially in cases where several
explanatory variables have moderate to high correlation. Tutorial 1 (see the Example section) illustrates an analysis involving a relatively small number of correlated Xs, and Tutorial 3 (see the Example section) shows how separate models can be estimated for different (latent class) segments using "Observation weights".

Method CCR.LM differs from Method PLS in that the components are allowed to be correlated, there is no need to deflate (and then restore) predictors, and similar to traditional OLS regression, predictions are invariant to linear transformations applied to the predictors. Thus, the explanatory variables do not need to be standardized prior to estimation.

CCR.PLS

The PLS method applies CCR techniques to obtain a regularized linear regression based on the PLS regression (PLS) model for a continuous Y. For an introduction to PLS regression see Tenenhaus (1998). For a comparison of the CCR.LM and PLS methods see Tenenhaus (2011).

Unlike CCR.LM which is invariant with respect to the scale of the predictors, when K < P, PLS regression can yield substantially different predictions depending upon whether the predictors are standardized or not. For a detailed comparison of CCR.LM, PLS with unstandardized Xs and PLS with standardized Xs, see Magidson (2011).

CCR.LDA and CCR.Logistic

The CCR.LDA and CCR.Logistic methods apply CCR techniques to obtain regularized regressions based on the Correlated Component Regression (CCR) model for a dichotomous Y.

Tutorial 2 (see the Example section) illustrates such an analysis involving many Xs, and where P > N (i.e., the number of predictors exceeds the number of cases).

Notes:

Depending upon which method is selected, CCR.LM, PLS, CCR.LDA, or CCR.Logistic, in the case where P < N, setting K = P yields the corresponding (saturated) regression models:

Method CCR.LM (or PLS) is equivalent to OLS regression (for K = P)

Method CCR.Logistic yields traditional Logistic regression (for K = P)

Method CCR.LDA yields traditional Linear Discriminant Analysis (for K = P)

where prior probabilities are computed from group sizes.
M-fold Cross-Validation

R rounds of M-fold Cross-validation (CV) may be used to determine the number of components $K^*$ and number of predictors $P^*$ to include in a model. For $R>1$ rounds, the standard error of the relevant CV statistic is also reported. When multiple records (rows) are associated with the same case ID (in XLSTAT, case IDs are specified using ‘Observation labels’), for each round, the CV procedure assigns all records corresponding to the same case to the same fold.

The Automatic Option in M-fold Cross-Validation

When the CV option is performed in Automatic mode (see ‘Automatic’ option in Options tab) a maximum number $K$ is specified for the number of components, all $K$ models containing between 1 and $K$ components are estimated, and the $K^*$ model selected as the one with the best CV statistic. When the step-down option is also activated, the $K$ models are estimated with all predictors prior to beginning the step-down algorithm.

The CV statistic used to determine $K^*$ depends upon the model type as follows:

For CCR.LM or PLS: The CV-R2 is the default statistic. Alternatively, the Normed Mean Squared Error (NMSE) can be used instead of CV-R2.

For CCR.LDA or CCR.Logistic: The CV-Accuracy (ACC), based on the probability cut-point of .5, is used by default. In the case of two or more values of $K$ yielding identical values for CV-Accuracy, the one with the higher value for the Area Under the ROC Curve (AUC) is selected.

Predictor Selection Using the CCR/Step-Down Algorithm

In step 1 of the step-down option, a model containing all predictors is estimated with $K^*$ components (where $K^*$ is specified by the user or determined by the program if the Automatic option is activated), and the relevant CV statistics are computed. In step 2, the model is then re-estimated after excluding the predictor whose standardized coefficient is smallest in absolute value, and CV statistics are computed again. Note that both steps 1 and 2 are performed within each subsample formed by eliminating one of the folds. This process continues until the user-specified minimum number of predictors remains in the model (by default, $P_{min} = 1$). The number of predictors included in the reported model, $P^*$, is the one with the best CV statistic.

In any step of the algorithm, if the number of predictors remaining in the model falls below $K^*$, the number of components is automatically reduced by 1, so that the model remains saturated. For example, suppose that $K^*=5$, but after a certain number of predictors are eliminated $P=4$ predictors remain. Then, the $K^*$ is reduced to 4 and the step-down algorithm continues.
If a maximum number of predictors to be included in a model, Pmax, is specified, the step-down algorithm still begins with all predictors included in the model, but results are reported only for P less than or equal to Pmax, and the CV statistics are only examined for P in the range [Pmin, Pmax].

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

Reload: Click this button to reload the default options.

Delete selections: Click this button to delete the data selections.

Changing data handling: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Y / Dependent variables:

Quantitative: Select the dependent variable(s) for the CCR linear or PLS model. The data must be numerical. If the “Variable labels” option is activated make sure that the headers of the variables have also been selected.

Qualitative: Select the dependent variable(s) for the logistic or discriminant CCR model. The data will be considered categorical but it can be numerical data (0/1 for example). If the...
“Variable labels” option is activated make sure that the headers of the variables have also been selected.

**X / Explanatory variables:**

**Quantitative:** Activate this option if you want to include one or more quantitative explanatory variables. Then select the corresponding data. The data must be numerical. If the “Variable labels” option is activated make sure that the headers of the variables have also been selected.

**Method:** Choose the regression method you want to use:

- **CCR.LM:** Activate this option to compute a Correlated Component Linear Regression model with a continuous dependent variable. Predictors assumed to be numeric (continuous, dichotomous, or discrete).

- **PLS:** Activate this option to compute a Partial Least Squares Regression with a continuous dependent variable. Predictors assumed to be numeric (continuous, dichotomous, or discrete).

- **CCR.LDA:** Activate this option to compute a Correlated Component Regression with a dichotomous (binary) dependent variable Y. According to assumptions of Linear Discriminant Analysis (LDA), predictors assumed to be multivariate normal with differing means but constant variances and correlations within each dependent variable group).

- **CCR.Logistic:** Activate this option to compute a Correlated Component Logistic Regression model with a dichotomous (binary) dependent variable. Predictors assumed to be numeric (continuous, dichotomous, or discrete).

**Range:** Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Activate this option to display the results in a new worksheet of the active workbook.

**Workbook:** Activate this option to display the results in a new workbook.

**Variable labels:** Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

**Observation labels:** Activate this option if labels are available for the N observations. Then select the corresponding data. If the ‘Variable labels’ option is activated you need to include a header in the selection.
With repeated measures data (multiple records per case) the Observation labels variable serves as a case ID variable, which groups the records from a given case together so that they are assigned to the same fold during cross-validation. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …), so that each case contains a single record.

**Observation weights:** Activate this option if you want to weight the observations. If you do not activate this option, all weights are set to 1. The weights must be non-negative values. Setting a case weight to 2 is equivalent to repeating the same observation twice. If the ‘Variable labels’ option is activated, make sure that the header (first row) has also been selected.

**Options** tab:

**Component options:**

**Automatic:** When the ‘Automatic’ option is activated, XLSTAT-CCR estimates K-component models for all values of K less than or equal to the number specified in the ‘Number of components’ text box, and produces the ‘Cross-validation Component Plot’ (see Charts tab). This chart plots the CV-R^2 (or NMSE) if the CCR.LM or PLS method is activated, or for CV-ACC (accuracy) and CV-AUC (Area Under ROC Curve) if the CCR.logistic or CCR.LDA method is activated. Coefficients are provided for the model with the best CV result.

Note: Activating the ‘Automatic’ option will have no effect if ‘Cross-validation’ option is not also activated.

**Number of components / Max Components:** When Automatic is activated, separate K-component models are estimated for each value K=1, 2, …, K_{MAX} where the number K_{MAX} is specified in the ‘Max Components’ field. If Automatic is not activated, enter the desired number of components K (positive integer) in the ‘Number of Components’ field. If the number entered exceeds the number of selected predictors P or N-1, K will automatically be reduced to the minimum of P and N-1.

**Step-Down options:**

**Perform Step Down:** Activate this option to estimate a K*-component model containing the subset of candidate predictors selected according to the chosen option settings:

**Min variables:** Enter the minimum number of predictors to be included in the model. The default value is 1.

**Max variables:** Enter the maximum number of predictors to be included in the model. The default value is 20.
**Remove by percent:** Activate this option to specify the percentage of predictors to be removed at each step. If not activated, the step-down algorithm removes 1 predictor at a time, which might take a considerable amount of time to run when the number of predictors is large.

**Percent:** Enter the percentage of predictors to be removed at each step. The specified percentage of predictors will be removed at each step, until 100 predictors remain, at which time the step-down algorithm removes 1 predictor at a time. By default, the percentage is set to 1%, meaning that if you had say 10,000 predictors to begin with, after 460 steps you have fewer than 100 predictors. Or, if you used 2%, after 229 steps you would be under 100 predictors.

**Note:** If the ‘Automatic’ option is also activated, $K^*$ is the value for $K$ having the best cross-validation (CV) statistic. Otherwise, $K^* = \text{the number entered in the ‘Number of Components’ field.}$

**Additional Options for CCR.Logistic method**

The following additional options apply to the Iteratively Re-weighted Least Squares (IRLS) algorithm that is used repeatedly to estimate parameters for the CCR.logistic model.

**Iterations:** Enter the number of iterations for IRLS. The default (recommended) number is 4.

**Ridge:** Enter the Ridge penalty number for CCR.logistic models. The default number is 0.001. With no penalty (Ridge parameter = 0), the separation problems may cause nonconvergence, in which case increasing the number of iterations will yield larger and larger estimates for at least one regression coefficient.

**Additional Options for CCR.LM and PLS methods:**

**NMSE:** Activate this option to use Normed Mean Squared Error (NMSE) as an alternative to the default criterion, CV-$R^2$, for determining the tuning parameters $K^*$ (if the ‘Automatic’ option is activated) and/or the number of predictors to be included in the model, $P^*$, if the ‘Perform Step-down’ option is activated.

NMSE is defined as the Mean Squared Error divided by the Variance of $Y$. It should provide values that are greater than 0, and usually less than 1. Values greater than 1 indicate a poor fit in that the predictions (when applied to cases in the omitted folds) tend to be further from the observed $Y$ than the baseline prediction provided by the observed mean of $Y$ (a constant). If the NMSE option is not activated, the default criterion CV-$R^2$ will be used. These two criteria should give the same or close to the same solutions in most cases. CV-$R^2$ is computed as the square of the correlation between the predicted and observed dependent variable.
**Additional Options for PLS method:**

**Standardize:** Activated by default, this option standardizes the explanatory variables to have variance 1. Unlike the other methods which are invariant with respect to linear transformations on the variables, the PLS regression method produces different results depending upon whether or not the explanatory variables are standardized. Deactivate this option to use the PLS method with unstandardized predictors.

**Validation tab:**

**Validation options:**

**Validation:** Activate this option if you want to use a sub-sample of the data to validate the model.

**Validation set:** Choose one of the following options to define how to obtain the observations used for the validation:

- **Random:** The observations are randomly selected. The “Number of observations” N must then be specified.
- **N last rows:** The N last observations are selected for the validation. The “Number of observations” N must then be specified.
- **N first rows:** The N first observations are selected for the validation. The “Number of observations” N must then be specified.
- **Group variable:** If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to use for the validation.

**Cross-validation options:**

**Cross-Validation:** Activate this option to use cross-validation.

**Number of Rounds:** The default number is 1. Enter the number of rounds (positive integer) of cross-validation to be performed. When a value greater than 1 is entered, the standard error for the relevant CV statistic is calculated. This option does not apply when a Fold variable is specified.

**Number of Folds:** The default number is 10. Enter the number of cross-validation folds (positive integer greater than 1). Typically, a value between 5-10 is specified that divides evenly (when possible) into the number of observations in the estimation sample. This option does not apply when a Fold variable is specified.
**Stratify:** Activate this option to use the 2 categories of dependent variable Y as a stratifier for fold assignment (applies only to CCR.LDA and CCR.Logistic).

**Fold variable:** Activate this option to use a variable to specify to which fold each case is assigned. If no grouping variable is specified, each case is assigned to 1 of the folds $M$ folds randomly. A fold variable contains positive integer values 1, 2, …, $M$ where $M = \#$ folds.

Note: When Observation labels are specified with the same label for multiple records, all records with the same observation label are grouped together and assigned to the same fold. This assures that in the case of repeated measures data (multiple records per case) the records associated with a given case are all allocated to the same fold during cross-validation.

**Missing data tab:**

**Remove observations:** Activate this option to remove the observations with missing data.

**Estimate missing data:** Activate this option to estimate missing data before starting the computations.

- **Mean or mode:** Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

**Outputs tab:**

**Tab [1]**

**Descriptive statistics:** Activate this option to display the descriptive statistics for all the selected variables.

**Correlations:** Activate this option to display the correlation matrix for the quantitative variables (dependent and explanatory).

**Coefficients:**

- **Unstandardized:** Activate this option to display the unstandardized parameters of the model.

- **Standardized:** Activate this option to display the standardized parameters of the model (also called beta coefficients).

- **Predictions and residuals:** Activate this option to display the predictions and residuals associated with the dependent variable. For methods CCR.LM and PLS, predictions

**Equation:** activate this option to explicitly display the equation of the model.
For model types CCR.LM and PLS, the equation predicts the mean of the dependent variable for given values of the predictors. For model types CCR.LDA and CCR.Logistic, the equation predicts the probability of being in dependent variable group 1 (group 1 is the group that is coded with the higher value).

**Tab [2]**

The following parameters can be included in the output by activating the associated output options.

**Component weights:**

- **Unstandardized:** Activate this option to display the unstandardized component weights table.
- **Standardized:** Activate this option to display the standardized component weights table.

**Loadings:**

- **Unstandardized:** Activate this option to display the unstandardized loadings table.
- **Standardized:** Activate this option to display the standardized loadings table.

**Cross-validation predictor count table:** Activate this option to display the predictor count table. This option can only be activated if ‘Step-down’ option is activated in the Options tab and the ‘Cross-Validation’ option is activated in the Validation tab.

**Cross-validation step-down table:** Activate this option to display the table corresponding to cross-validation step-down. This option can only be activated if the ‘Step-down’ option is activated in the Options tab and the ‘Cross-Validation’ option is activated in the Validation tab.

**Tab [3]**

This tab is only available for model types CCR.LDA and CCR.Logistic.

**Classification table:** Activate this option to display the posterior observation classification table (confusion table) using a specified probability cutpoint (default probability cutpoint = 0.5).

**Charts tab:**

**Cross-Validation Component Plot:** Activate this option to display the chart produced when both the Automatic option and Cross-validation are activated. This chart plots the relevant CV statistic as a function of the number of components \( K=1, 2, \ldots, K_{\text{MAX}} \).
For model types CCR.LDA and CCR.Logistic:

The Cross-Validation Component Plot corresponds to the cross-validation AUC and model accuracy (ACC) based on the number of components K ranging from 1 to the specified Number of components.

For model types CCR.LM and PLS:

The $R^2$ plot corresponds to the cross-validation $R^2$ (or NMSE if this option is activated in the Options tab) based on the number of components K ranging from 1 to the specified Number of components.

Cross-Validation Step-down Plot: Activate this option to display the chart associated with the Step-down option and Cross-validation.

For CCR.LDA and CCR.Logistic options:

The Cross-Validation Step-down Plot corresponds to the cross-validation AUC and model accuracy based on the specified K-component model for numbers of predictors $P$ ranging from the specified ‘Max variables’ down to the specified ‘Min variables’.

For CCR.LM and PLS:

The $R^2$ graph corresponds to the cross-validation $R^2$ (or NMSE if this option is activated in the Options tab) based on the specified K-component model for numbers of predictors $P$ ranging from the specified ‘Max variables’ down to the specified ‘Min variables’.

Results

Summary (descriptive) statistics: the tables of descriptive statistics display for all the selected variables a set of basic statistics. For the dependent variables (colored in blue), and the quantitative explanatory variables, XLSTAT displays the number of observations, the number of observations with missing data, the number of observations with no missing data, the mean, and the unbiased standard deviation.

Correlation Matrix: this table is displayed to allow your visualizing the correlations among the explanatory variables, among the dependent variables and between both groups.

Goodness of Fit Statistics:

For all models:

The number of observations in the training set and in the validation set (if any), as well as the sum of weights are first displayed.
For model types CCR.LM and PLS:

The table displays the model quality indices.

- The R² is shown for the estimation sample. If a validation is specified, the Validation-R2 will be included in the table. If the cross-validation option is activated, the CV-R2 will be included in the table. The CV- R2 reported in the table is the average of the CV- R2(P*.r) across the rounds. For round r, the OPTIMAL NUMBER OF PREDICTORS P*.r, is determined for that round, and an average is computed of these CV- R2(P*.r).

- If the NMSE option is activated, the normed mean squared error (NMSE) is reported in addition to R2. For the NMSE reported in the ‘Validation’ column, the variance of the dependent variable is computed based on the validation sample. For the NMSE reported in the ‘Training’ and ‘Cross-validation’ columns, the variance of the dependent variable is computed based on the estimation sample.

For model types CCR.LDA and CCR.Logistic:

The table displays the model quality indices.

- The Area Under the Curve (AUC) is shown for the estimation sample. If a validation is specified, the Validation-AUC will be included in the table. If the cross-validation option is activated, the CV-AUC will be included in the table.

- The accuracy (ACC) is shown for the estimation sample. If a validation is specified, the Validation-ACC will be included in the table. If the cross-validation option is activated, the CV-ACC will be included in the table.

Predictors retained in the model: A list of the names of the predictors retained in the model.

Number of components: The number of components in the model.

Unstandardized component weights table: The unstandardized component weights for each component.

Standardized component weights table: The standardized component weights for each component.

Unstandardized loadings table: The unstandardized predictor loadings for each component.

Standardized loadings table: The standardized predictor loadings for each component.
Cross-Validation component table (and associated plot): This output appears only if the 'Automatic' option is activated in the Options tab and the 'Cross-Validation' option is activated in the Validation tab. If more than 1 round of M-folds are used, the relevant CV statistics are computed as the average over all Rounds, and the associated standard error is also reported. Coefficients and other output are provided for the model containing K* components where K* is the value of K shown in this table associated with the best CV statistic.

Results for model types CCR.LM and PLS:

The relevant CV statistic is the CV-R2. The NMSE statistic is also reported if requested in the Options tab.

Results for model types CCR.LDA and CCR.Logistic:

The relevant CV statistics are the Cross-Validated Accuracy (CV-ACC) and the CV-AUC.

Cross-Validated step-down table (and associated plot): The Cross-Validation step-down table appears only if the Step-Down option and the Cross-Validation option are activated. If more than 1 round of M-folds are used, the relevant CV statistics are computed as the average over all Rounds, and the associated standard error is also reported. Coefficients and other output are provided for the model containing P* predictors where P* is the value of P shown in this table associated with the best CV statistic.

Results for model types CCR.Lm and PLS:

For each number of predictors in the model, the table reports the CV-R2. If more than 1 round of M-folds are used, the reported CV-R2 is the average over all Rounds, and the associated standard error is also reported.

Results for model types CCR.LDA and CCR.Logistic:

For each number of predictors in the model, the table reports the CV-AUC (and associated standard error) and the CV-ACC.

Note: The value for the CV statistic provided in this table for P* predictors, along with the associated standard error, may differ from the CV statistic provided in the Goodness of Fit Table. For example, suppose that P* = 4 predictors and R = 10 rounds of M-folds are used. Then the value of the CV statistic reported in this table is computed as the average over all 10 rounds of the corresponding CV statistic within each of the 10 rounds, where all CV statistics are based on P* predictors. On the other hand, as mentioned above, the CV statistic (and
associated standard error) reported in the Goodness of Fit Table is computed as the average across all 10 rounds where in each round r the CV statistic is used based on P*r predictors.

**Cross-Validation predictor count table:**

The Cross-Validation step-down table is available only if the Step-Down option and the Cross-Validation option are activated.

In the table, the first column lists the number of times each candidate predictor showed up in the final model for each round. The last column (Total) reports the sum of counts for each round. The last row (Total) reports the sum of the totals for a given round (= M * Pr).

Optimal number of predictors for each round table: Reports the optimal # of predictors selected in each round (Pr).

**Unstandardized coefficients table:**

Unstandardized regression coefficients are used to predict the dependent variable Y.

For CCR.LDA and CCR.Logistic, Y is dichotomous and predictions are for the probability of being in the dependent variable group associated with the higher of the 2 numeric values taken on by Y.

For PLS with the Standardize option activated in the Options tab, predictors are standardized by dividing by their standard deviation. The unstandardized regression coefficient reported is for the standardized predictor.

The equation of the model is displayed if the corresponding option has been activated. For model types CCR.LM and PLS, the equation computes conditional mean of the dependent variable, while for model types CCR.LDA and CCR.Logistic the equation computes the predicted probability of the dependent variable group coded with the highest value.

**Standardized coefficients table (and associated column chart):**

**Standardized regression coefficients** are used to assess the importance of the predictors, predictors with the highest magnitude are the most important. Each standardized regression coefficient equals the corresponding unstandardized coefficient multiplied by the ratio std(Xg)/std(Y), where ‘std’ denotes standard deviation.

For PLS with the Standardize option activated in the Options tab, predictors are standardized by dividing by their standard deviation, so that std(Xg) = 1 for each predictor g =1,2,….P. The
standardized regression coefficient in this case equals the corresponding unstandardized coefficient reported divided by std(Y).

Predictions and residuals:

This table reports the predictions for the dependent variable, residuals and standardized residuals.

Additional output for model types CCR.LDA and CCR.Logistic:

Classification table for the estimation sample (and associated ROC Curve): The table reports the correct classification rates for each of the 2 dependent variable groups. This classification table is based on the cutpoint specified in Output Tab 3 (default probability = .5).

Classification table for the validation sample (and associated ROC Curve): The table reports the correct classification rates for each of the 2 dependent variable groups. This classification table is based on the cutpoint specified in Output Tab 3 (default probability = .5).

Copyright ©2011 Statistical Innovations Inc. All rights reserved.

Examples

The following tutorials on how to use XLSTAT-CCR are available:

Tutorial 1: Getting Started with Correlated Component Regression (CCR) in XLSTAT-CCR

http://www.xlstat.com/demo-ccr1.htm

Tutorial 2: Using Correlated Component Regression with a Dichotomous Y and Many Correlated Predictors in XLSTAT-CCR

http://www.xlstat.com/demo-ccr2.htm

Tutorial 3: Developing a Separate CCR Model for Each Segment in XLSTAT-CCR

http://www.xlstat.com/demo-ccr3.htm
References


Correlation tests

Use this tool to compute the correlation coefficients of Pearson, Spearman or Kendall, between two or more variables, and to determine if the correlations are significant or not. Several visualizations of the correlation matrices are proposed.

Description

Three correlation coefficients are proposed to compute the correlation between a set of quantitative variables, whether continuous, discrete or ordinal (in the latter case, the classes must be represented by values that respect the order):

Pearson correlation coefficient: this coefficient corresponds to the classical linear correlation coefficient. This coefficient is well suited for continuous data. Its value ranges from -1 to 1, and it measures the degree of linear correlation between two variables. Note: the squared Pearson correlation coefficient gives an idea of how much of the variability of a variable is explained by the other variable. The p-values that are computed for each coefficient allow testing the null hypothesis that the coefficients are not significantly different from 0. However, one needs to be cautious when interpreting these results, as if two variables are independent, their correlation coefficient is zero, but the reciprocal is not true.

Spearman correlation coefficient (rho): this coefficient is based on the ranks of the observations and not on their value. This coefficient is adapted to ordinal data. As for the Pearson correlation, one can interpret this coefficient in terms of variability explained, but here we mean the variability of the ranks.

Kendall correlation coefficient (tau): as for the Spearman coefficient, it is well suited for ordinal variables as it is also based on ranks. However, this coefficient is conceptually very different. It can be interpreted in terms of probability: it is the difference between the probabilities that the variables vary in the same direction and the probabilities that the variables vary in the opposite direction. When the number of observations is lower than 50 and when there are no ties, XLSTAT gives the exact p-value. If not, an approximation is used. The latter is known as being reliable when there are more than 8 observations.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.
Click this button to start the computations.

Click this button to close the dialog box without doing any computation.

Click this button to display the help.

Click this button to reload the default options.

Click this button to delete the data selections.

Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

**Observations/variables table:** Select a table comprising N observations described by P variables. If column headers have been selected, check that the "Variable labels" option has been activated.

**Weights:** Activate this option if the observations are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

**Type of correlation:** Choose the type of correlation to use for the computations (see the description section for more details).

**Subsamples:** Check this option to select a column showing the names or indexes of the subsamples for each of the observations. All computations are then performed subsample by subsample.

- **Variable-Category labels:** Activate this option to use variable-category labels when displaying outputs for the quantitative variables. Variable-Category labels include the variable name as a prefix and the category name as a suffix.

**Range:** Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Activate this option to display the results in a new worksheet of the active workbook.
Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (row and column variables, weights) includes a header.

Significance level (%): Enter the significance level for the test of on the correlations (default value: 5%).

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove the observations: Activate this option to remove observations with missing data.

Pairwise deletion: Activate this option to remove observations with missing data only when the variables involved in the calculations have missing data. For example, when calculating the correlation between two variables, an observation will only be ignored if the data corresponding to one of the two variables is missing.

Estimate missing data: Activate this option to estimate the missing data before the calculation starts.

- Mean or mode: Activate this option to estimate the missing data by using the mean (quantitative variables) or the mode (qualitative variables) for the corresponding variables.

- Nearest neighbour: Activate this option to estimate the missing data for an observation by searching for the nearest neighbour to the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the selected variables.

Correlations: Activate this option to display the correlation matrix that corresponds to the correlation type selected in the "General" tab. If the "significant correlations in bold" option is activated, the correlations that are significant at the selected significance level are displayed in bold.

p-values: Activate this option to display the p-values that correspond to each correlation coefficient.

Coefficients of determination: Activate this option to display the coefficients of determination. These correspond to squared correlations coefficients. When the using the Pearson correlation
coefficient, the coefficients of determination are equal to the $R^2$ of the regression of one variable by the other.

**Sort the variables:** Activate this option to sort and group variables that are highly correlated.

**Charts** tab:

**Correlation maps:** Several visualizations of a correlation matrix are proposed.

- The “blue-red” option allows to represent low correlations with cold colors (blue is used for the correlations that are close to -1) and the high correlations are with hot colors (correlations close to 1 are displayed in red color).

- The “Black and white” option allows to either display in black the positive correlations and in white the negative correlations (the diagonal of 1s is display in grey color), or to display in black the significant correlations, and in white the correlations that are not significantly different from 0.

- The “Patterns” option allows to represent positive correlations by lines that rise from left to right, and the negative correlations by lines that rise from right to left. The higher the absolute value of the correlation, the large the space between the lines.

**Scatter plots:** Activate this option to display the scatter plots for all two by two combinations of variables.

- **Matrix of plots:** Check this option to display all possible combinations of variables in pairs in the form of a two-entry table with the various variables displayed in rows and in columns.

- **Histograms:** Activate this option so that XLSTAT displays a histogram when the X and Y variables are identical.

- **Q-Q plots:** Activate this option so that XLSTAT displays a Q-Q plot when the X and Y variables are identical.

- **Confidence ellipses:** Activate this option to display confidence ellipses. The confidence ellipses correspond to a $x\%$ confidence interval (where $x$ is determined using the significance level entered in the General tab) for a bivariate normal distribution with the same means and the same covariance matrix as the variables represented in abscissa and ordinates.
Results

The correlation matrix and the table of the p-values are displayed. The correlation maps allow identifying potential structures in the matrix, or to quickly identify interesting correlations.

Example

A tutorial on how to compute a Spearman correlation coefficient and the corresponding significance test is available on the Addinsoft website:

http://www.xlstat.com/demo-corrsp.htm

References


RV coefficient

Use this tool to compute the similarity between two matrices of quantitative variables recorded from the same observations or two configurations resulting from multivariate analyses for the same set of observations.

Description

This tool allows computing the RV coefficient between two matrices of quantitative variables recorded from the same observations. The RV coefficient is defined as (Robert and Escoufier, 1976; Schlich, 1996):

$$RV(W_i, W_j) = \frac{\text{trace}(W_i, W_j)}{\sqrt{\text{trace}(W_i, W_i) \cdot \text{trace}(W_j, W_j)}}$$

where $\text{trace}(W_i, W_j) = \sum_{l,m} w_{i,l,m} w_{j,l,m}^{*}$ is a generalized covariance coefficient between matrices $W_i$ and $W_j$, $\text{trace}(W_i, W_i) = \sum_{l,m} w_{i,l,m}^{2}$ is a generalized variance of matrix $W_i$, and $w_{i,l,m}^{*}$ is the $(l,m)$ element of matrix $W_i$.

The RV coefficient is a generalization of the squared Pearson correlation coefficient. The RV coefficient is between 0 and 1. The closer to 1 the RV is, the more similar the two matrices $W_i$ and $W_j$ are.

XLSTAT offers the possibility:

- to compute the RV coefficient between two matrices, including all variables from both matrices;
- to choose the $k$ first variables from both matrices and compute the RV coefficient between the two resulting matrices.

XLSTAT allows testing if the obtained RV coefficient is significantly different from 0 or not.

Two methods to compute the p-value are proposed by XLSTAT. The user can choose between a p-value computed using an approximation of the exact distribution of the RV statistic with the Pearson type III approximation (Kazi-Aoual et al., 1995), and a p-value computed using Monte Carlo resamplings.
Note: the XLSTAT_RVcoefficient spreadsheet function can be used to compute the RV coefficient between two matrices of quantitative variables.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload options**: Click this button to reload the default options.
- **Delete data**: Click this button to delete the data selections.
- **Change data handling**: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab:**

**Matrix A**: Select the data that correspond to N observations described by P quantitative variables. If a column header has been selected, check that the "Column labels" option is activated.

**Matrix B**: Select the data that correspond to N observations described by Q quantitative variables. If a column header has been selected for Matrix B, a column header should be selected for matrix B.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.
Workbook: Activate this option to display the results in a new workbook.

Column labels: Activate this option if the first row of the data selections (variables, observations labels) includes a header.

Row labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Column labels" option is activated you need to include a header in the selection.

Options tab:

Selected variables:

All: Choose this option to compute the RV coefficient between Matrix A and Matrix B using all variables from both matrices.

User defined: Choose this option to compute the RV coefficient between sub-matrices of Matrix A and Matrix B with the same number of variables. Then, enter the number of variables to be selected. For example to compute the RV coefficient on the first two variables (or the first two dimensions when comparing results from multidimensional analysis), enter 2 for both From and To. To compute RV coefficients for a series of number of variables, enter a for From and b for To where a<b. For example, to compute, RV coefficients for the first 2, 3 and 4 variables, enter 2 for From and 4 for To.

p-value computation:

Extrapolation: Choose this option to use an approximation of the exact distribution of the RV statistic with the Pearson type III approximation to compute the p-value associated with the RV coefficient.

permutation: Choose this option to calculate the p-value based on Monte Carlo permutations, and select the number of random permutations to perform or the maximum time to spend.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove the observations: Activate this option to remove observations with missing data.

Outputs tab:
RV coefficients: Activate this option to display the RV coefficient(s), standardized RV coefficient(s), and mean(s) and variance(s) of the RV coefficient distribution.

Adjusted RV coefficients: Activate this option to display the adjusted RV coefficient(s).

p-value: Activate this option to display the p-value(s) associated with the RV coefficient(s).

Charts tab:

Chart of RV coefficients: Activate this option to display a bar chart of the RV coefficient(s) (with colors corresponding to the significance of the associated p-value(s) if the option p-value has been selected).

Results

RV coefficients: A table including the RV coefficient(s), standardized RV coefficient(s), and mean(s) and variance(s) of the RV coefficient distribution; and the adjusted RV coefficient(s) and p-value(s) if requested by the user.

Bar chart: A bar chart showing the RV coefficient(s) (with color codes to show significance of the associated p-value(s) if requested).

Example

An example showing how to compute the RV coefficient is available on the Addinsoft web site. To download this data, go to:

http://www.xlstat.com/demo-rv.htm

References


Tests on contingency tables (chi-square, ...)

Use this tool to study the association between the rows and the columns of a contingency table, and to test the independence between the rows and the columns.

Note: to build a contingency table from two qualitative variables you can use the “Build a contingency table” feature of XLSTAT.

Description

Many association measures and several tests have been proposed to evaluate the relationship between the R rows and the C columns of a contingency table.

Some association measures have been specifically developed for the 2x2 tables. Others can only be used with ordinal variables.

XLSTAT always displays all the measures. However, measures that concern ordinal variables should only be interpreted if the variables are ordinal and sorted in increasing order in the contingency table.

Tests of independence between the rows and the columns of a contingency table

- The Pearson chi-square statistic allows to test the independence between the rows and the columns of the table, by measuring to which extent the observed table is far (in the chi-square sense) from the expected table computed using the same marginal sums. The statistic writes:

\[ \chi^2 = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(n_{ij} - f_{ij})^2}{f_{ij}}, \quad \text{with} \quad f_{ij} = \frac{n_i n_j}{n}, \quad n = \sum_{i=1}^{R} \sum_{j=1}^{C} n_{ij}, \quad n_i = \sum_{j=1}^{C} n_{ij}, \quad n_j = \sum_{i=1}^{R} n_{ij}, \]

One shows that this statistic follows a Chi-square distribution with (R-1)(C-1) degrees of freedom. However, this result is asymptotical and, before using the test, it is recommended to make sure that:

- That n is greater or equal to 20;
- That no marginal sum (n_i or n_j) is less than 5;
- That at least 80% of the expected values (f_{ij}) is above 5.

- In the case where R=2 and C=2, a continuity correction has been suggested by Yates (1934). The modified statistic writes:
A test based on the likelihood ratio and on the Wilks’ $G^2$ statistic has been developed as an alternative to the Pearson chi-square test. It consists in comparing the likelihood of the observed table to the likelihood of the expected table defined as for the Pearson chi-square test. $G^2$ is defined by:

$$G^2 = 2 \sum_{i=1}^{R} \sum_{j=1}^{C} n_{ij} \log \left( \frac{n_{ij}}{f_{ij}} \right)$$

As for the Pearson statistic, $G^2$ follows asymptotically a Chi-square distribution with $(R-1)(C-1)$ degrees of freedom.

The Fisher’s exact test allows to compute the probability that a table showing a stronger association between the rows and the columns would be observed, the marginal sums being fixed, and under the null hypothesis of independence between rows and columns. In the case of a 2x2 table, the independence is measured through the odds ratio (see below for further details) given by $\theta = \frac{n_{11}n_{22}}{n_{12}n_{21}}$. The independence is corresponds to the case where $\theta = 1$. There are three possible alternative hypotheses: the two-sided test corresponds to $\theta \neq 1$, the lower one-sided test to $\theta < 1$ and the upper one-sided test to $\theta > 1$.

XLSTAT allows to compute the Fisher’s exact two-sided test when $R \geq 2$ and $C \geq 2$. The computing method is based on the network algorithm developed by Mehta (1986) and Clarkson (1993). It may fail in some cases. The user is prompted when this happens.

Monte Carlo test: A nonparametric test based on simulations has been developed to test the independence between rows and columns. A number of Monte Carlo simulations defined by the user are performed in order to generate contingency tables that have the same marginal sums as the observed table. The chi-square statistic is computed for each of the simulated tables. The p-value is then determined by suing the distribution obtained from the simulations.

### Association measures (1)

A first series of association coefficients between the rows and the columns of a contingency table is proposed:

- **The Pearson’s Phi coefficient** allows to measure the association between the rows and the columns of an RxC table. In the case of a 2x2 table, its value ranges from -1 to 1 and writes:

$$\phi_p = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{11}n_{22}n_{12}n_{21}}}$$
When \( R>2 \) and/or \( C>2 \), it ranges between 0 and the minimum of the square roots of \( R-1 \) and \( C-1 \). In that case, the Pearson’s Phi writes:

\[
\phi_p = \sqrt{\frac{\chi_p^2}{n}}
\]

- **Contingency coefficient**: This coefficient, also derived from the Pearson’s chi-square statistic, writes:

\[
C = \sqrt{\frac{\chi_p^2}{\left(\chi_p^2 + n\right)}}
\]

- **Cramer’s V**: This coefficient is also derived from the Pearson chi-square statistic. In the case of \( 2x2 \) table, its value has the \([-1; 1]\) range. It writes:

\[
V = \phi_p
\]

When \( R>2 \) and/or \( C>2 \), it ranges between 0 and 1 and its value is given by:

\[
V = \sqrt{\frac{\chi_p^2}{\min(R-1, C-1)}}
\]

The closer \( V \) is to 0, the more the rows and the columns are independent.

- **Tschuprow’s T**: This coefficient is also derived from the Pearson chi-square statistic. Its value ranges from 0 to 1 and is given by:

\[
T = \sqrt{\frac{\chi_p^2}{\left(\chi_p^2 + n\right)}}
\]

The closer \( T \) is to 0, the more the rows and the columns are independent.

- **Goodman and Kruskal tau (R/C) and (C/R)**: This coefficient, unlike the Pearson coefficient is asymmetric. It allows to measure the degree of dependence of the rows on the columns (R/C) or vice versa (C/R).

- **Cohen’s kappa**: This coefficient is computed on \( RxR \) tables. It is useful in the case of paired qualitative samples. For example, we ask the same question to the same individuals at two different times. The results are summarized in a contingency table. The Cohen’s kappa, which value ranges from 0 to 1, allows to measure to which extent the answer are identical. The close the kappa is to 1, the higher the association between the two variables.

- **Yule’s Q**: This coefficient is used on \( 2x2 \) tables only. It is computed using the product of the concordant data \((n11.n22)\) and the product of the discordant data \((n12.n21)\). It ranges from -1 to 1. A negative value corresponds to a discordance between the two variables, a value close to 0 corresponds to the independence, and a value close to 1 to
the concordance. The Yule’s Q is equal to the Goodman and Kruskal Gamma when the latter is computed on a 2x2 table.

- **Yule’s Y**: This coefficient is used on 2x2 tables only. It is similar to the Yule’s Q and ranges from -1 to 1.

**Association measures (2)**

A second series of association coefficients between the rows and the columns of a contingency table is proposed. Confidence ranges around the estimated values are available. As the confidence ranges are computed using asymptotical results, their reliability increased with the number of the data.

- **Goodman and Kruskal Gamma**: This coefficient allows to measure on a -1 to 1 scale the degree of concordance between two ordinal variables.

- **Kendall’s tau**: This coefficient, also referred to as tau-b, allows to measure on a -1 to 1 scale the degree of concordance between two ordinal variables. Unlike the Gamma coefficient, the Kendall’s tau allows to take ties into account.

- **Stuart’s tau**: This coefficient, also referred to as tau-c, allows to measure on a -1 to 1 scale the degree of concordance between two ordinal variables. As the Kendall’s tau, the tau-c allows to take ties into account. In addition, it allows to adjust for the size of the table.

- **Somers’ D (R/C) and (C/R)**: This coefficient is an asymmetrical alternative to the Kendall’s tau. In the (R/C) case, the rows are assumed to depend on the columns and reciprocally in the (C/R) case; the correction for ties applies only to the “explanatory” variable.

- **Theil’s U (R/C) and (C/R)**: The asymmetric coefficient U of uncertainty of Theil (R/C) allows to measure the proportion of uncertainty of the row variable that is explained by the column variable, and reciprocally in the C/R case. These coefficients range from 0 to 1. The symmetric version of the coefficient that ranges from 0 to 1 is computed using the two asymmetric (R/C) and (C/R) coefficients.

- **Odds ratio and Log(Odds ratio)**: The odds ratio is given in the case of 2x2 by $\theta=(n_{11}.n_{22})/(n_{12}.n_{21})$. $\theta$ varies from 0 to infinity. $\theta$ can be interpreted as the increase in chances of being in column 1, when being in row 1 compared to when being in row 2. The case $\theta=1$ corresponds to no advantage. When $\theta>1$, the probability is $\theta$ times higher for row 1 than for row 2. We compute the logarithm of the odds because its variance is easier to compute, and because it is symmetric around 0, which allows to obtain a confidence interval. The confidence of the odds ration itself is computed by taking the exponential of the confidence interval on the log(odds ratio).
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload options**: Click this button to reload the default options.
- **Delete selection**: Click this button to delete the data selections.
- **Change data format**: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab:**

**Contingency table**: If the data format selected is "contingency table", select the data that correspond to the contingency table. If row and column labels are included, make sure that the "Labels included" option is checked.

**Row variable(s)**: If the data format selected is "contingency table", select the data that correspond to the variable(s) that will be used to construct the rows of the contingency table(s).

**Column variable(s)**: If the data format selected is "contingency table", select the data that correspond to the variable(s) that will be used to construct the columns of the contingency table(s).

**Data format**: Select the data format.

- **Contingency table**: Activate this option if your data are correspond to a contingency table.
• **Qualitative variables**: Activate this option if your data are available as two qualitative variables to be used to create a contingency table.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Labels included**: Activate this option if the row and column labels of the contingency table are selected.

**Variable labels**: Activate this option if the first row of the data selections (data and observations labels) includes a header.

**Options** tab:

**Chi-square test**: Activate this option to display the statistics and the interpretation of the Chi-square test of independence between rows and columns.

**Likelihood ratio test**: Activate this option to perform the Wilks G² likelihood ratio test.

**Monte Carlo method**: Activate this option to compute the p-value using Monte Carlo simulations.

**Significance level (%)**: Enter the significance level for the test.

**Fisher’s exact test**: Activate this option to compute the Fisher’s exact test. In the case of a 2x2 table, you can choose the alternative hypothesis. In the other cases, the two-sided is automatically used (see the description section for more details).

**Missing data** tab:

**Do not accept missing data**: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

**Replace missing data by 0**: Activate this option if you consider that missing data are equivalent to 0.

**Replace missing data by their expected value**: Activate this option if you want to replace the missing data by the expected value. The expectation is given by:
\[ E(n_{ij}) = \frac{n_i \cdot n_j}{n} \]

where \( n_i \) is the row sum, \( n_j \) is the column sum, and \( n \) is the grand total of the table before replacement of the missing data.

**Outputs** tab:

**List of combines**: Activate this option to display the table that lists all the possible combines between the two variables that are used to create a contingency table, and the corresponding frequencies.

**Contingency table**: Activate this option to display the contingency table.

**Inertia by cell**: Activate this option to display the inertia for each cell of the contingency table.

**Chi-square by cell**: Activate this option to display the contribution to the chi-square of each cell of the contingency table.

**Significance by cell**: Activate this option to display a table indicating, for each cell, if the actual value is equal (=), lower (<) or higher (>), than the theoretical value, and to run a test (Fisher's exact test of on a 2x2 table having the same total frequency as the complete table, and the same marginal sums for the cell of interest), in order to determine if the difference with the theoretical value is significant or not.

**Association coefficients**: Activate this option pour display the various association coefficients.

**Observed frequencies**: Activate this option to display the table of the observed frequencies. This table is almost identical to the contingency table, except that the marginal sums are also displayed.

**Theoretical frequencies**: Activate this option to display the table of the theoretical frequencies computed using the marginal sums of the contingency table.

**Proportions or percentages / Row**: Activate this option to display the table of proportions or percentages computed by dividing the values of the contingency table by the marginal sums of each row.

**Proportions or percentages / Column**: Activate this option to display the table of proportions or percentages computed by dividing the values of the contingency table by the marginal sums of each column.

**Proportions or percentages / Total**: Activate this option to display the table of proportions or percentages computed by dividing the values of the contingency table by the sum of all the cells of the contingency table.
Charts tab:

3D view of the contingency table: Activate this option to display the 3D bar chart corresponding to the contingency table.

Results

The results that are displayed correspond to the various statistics, tests and association coefficients described in the description section.

References


**Cochran-Armitage trend test**

Use this tool to test if a series of proportions, possibly computed from a contingency table, can be considered as varying linearly with an ordinal or continuous variable.

**Description**

The Cochran-Armitage test allows to test if a series of proportions, can be considered as varying linearly with an ordinal or continuous score variable.

If $X$ is the score variable, the statistic that is computed to test for the linearity is given by:

$$z = \frac{\sum_{i=1}^{r} n_{i+}(X_i - \bar{X})}{\sqrt{p_{+1}(1-p_{+1})} s^2}$$

with $s^2 = \sum_{i=1}^{r} n_{i+}(X_i - \bar{X})^2$

Note: if $X$ is an ordinal variable, the minimum value of $X$ has no influence on the value of $z$.

In the case of the two-tailed (or two-sided) test, the null (H0) and alternative (Ha) hypotheses are:

- **H0**: $z = 0$
- **Ha**: $z \neq 0$

Note: $z$ is asymptotically distributed as a standard Normal variable. Some statistical programs use $z^2$ to test the linearity. $z^2$ follows a Chi-square distribution with one degree of freedom.

In the one-tailed case, you need to distinguish the left-tailed (or lower-tailed or lower one-sided) test and the right-tailed (or upper-tailed or upper one-sided) test. In the left-tailed test, the following hypotheses are used:

- **H0**: $z = 0$
- **Ha**: $z < 0$

If Ha is chosen, one concludes that the proportions decrease when the score variable increases.

In the right-tailed test, the following hypotheses are used:

- **H0**: $z = 0$
- **Ha**: $z > 0$
If Ha is chosen, one concludes that the proportions increase when the score variable increases.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload default options**: Click this button to reload the default options.
- **Delete data selections**: Click this button to delete the data selections.
- **Change data handling**: Click these buttons to change the way XLSTAT handles the data. If the arrow points down (column mode), XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right (row mode), XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

**Contingency table**: Select a contingency table. If the column labels of the table have been selected, make sure the “Column labels” option is checked.

**Proportions**: Select the column (or row if in row mode) that contains the proportions. If a column has been selected, make sure the “Column labels” option is checked.

**Sample sizes**: If you selected proportions, you must select the corresponding sample sizes. If a column has been selected, make sure the “Column labels” option is checked.

**Row labels**: Activate this option to select the labels of the rows.
Data format:

- **Contingency table**: Activate this option if your data are contained in a contingency table.

- **Proportions**: Activate this option if your data are available as proportions and sample sizes.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Column labels**: Activate this option if column headers have been selected within the selections.

**Scores**: You can choose between ordinal scores (1, 2, 3, ...) or user defined scores.

- **Ordinal**: Activate this option to use ordinal scores.

- **User defined**: Activate this option to select the scores. If a column has been selected, make sure the “Column labels” option is checked.

**Options** tab:

**Alternative hypothesis**: Choose the alternative hypothesis to be used for the test (see the description section for more information).

**Significance level (%)**: Enter the significance level for the test (default value: 5%).

**Asymptotic p-value**: Activate this option to compute the p-value based on the asymptotic distribution of the z statistic.

**Monte Carlo method**: Activate this option to compute the p-value using Monte Carlo simulations. Enter the number of simulations to perform.

**Missing data** tab:

**Do not accept missing data**: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.
**Outputs** tab:

**Descriptive statistics:** Activate this option to display the descriptive statistics.

**Charts** tab:

**Proportions:** Activate this option to display a scatter plot with the scores as abscissa and the proportions as ordinates.

**Results**

The results include a summary table with the input data, a chart showing the proportions as a function of the scores. The next results correspond to the test itself, and its interpretation.

**References**


Mantel test

Use this test to compute the linear correlation between two proximity matrices (simple Mantel test), or to compute the linear correlation between two matrices knowing their correlation with a third matrix (partial Mantel test).

Description

Mantel (1967) proposed a first statistic to measure the correlation between two proximity (similarity or dissimilarity) and symmetric A and B matrices of size n:

\[ z(AB) = \sum_{i=j}^{n} a_{ij}b_{ij} \]

The standardized Mantel statistic, easier to use because it varies between -1 and 1, is the Pearson correlation coefficient between the two matrices:

\[ r(AB) = \frac{1}{n(n-1)/2} \sum_{i=a}^{n-1} \sum_{j=a+1}^{n} \left( \frac{a_{ij} - \bar{a}}{s_a} \right) \left( \frac{b_{ij} - \bar{b}}{s_b} \right) \]

Notes:

In the case where the similarities or dissimilarities would be ordinal, one can use the Spearman or Kendall correlation coefficients.

In the case where the matrices are not symmetric, the computations are possible.

While it is not a problem to compute the correlation coefficient between two sets of proximity coefficients, testing their significance can not be done using the usual approach that is used to test correlations: to use the latter tests, one needs to assume the independence of the data, which is not the case here. A permutation test has been proposed to determine if the correlation coefficient can be considered as showing a significant correlation between the matrices or not.

In the case of the two-tailed (or two-sided) test, the null (H0) and alternative (Ha) hypotheses are:

- H0: \( r(AB) = 0 \)
- Ha: \( r(AB) \neq 0 \)
In the one-tailed case, you need to distinguish the left-tailed (or lower-tailed or lower one-sided) test and the right-tailed (or upper-tailed or upper one-sided) test. In the left-tailed test, the following hypotheses are used:

- H0: \( r(AB) = 0 \)
- Ha: \( r(AB) < 0 \)

In the right-tailed test, the following hypotheses are used:

- H0: \( r(AB) = 0 \)
- Ha: \( r(AB) > 0 \)

The Mantel test consists of computing the correlation coefficient that would be obtained after permuting the rows and columns of one of the matrices. The p-value is calculated using the distribution of the \( r(AB) \) coefficients obtained from S permutations. In the case where \( n \), the number of rows and columns of the matrices, is lower than 10, all the possible permutations can easily be computed. If \( n \) is greater than 10, one needs to randomly generate a set of S permutations in order to estimated the distribution of \( r(AB) \).

A Mantel test for more than two matrices has been proposed (Smouse et al., 1986): when we have three proximity matrices A, B and C, the partial Mantel statistic \( r(AB.C) \) for the A and B matrices knowing the C matrix is computed as a partial correlation coefficient. In order to determine if the coefficient is significantly different from 0, a p-value is computed using random permutations as described by Smouse et al (1986).

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
Click this button to reload the default options.

Click this button to delete the data selections.

Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab:**

**Matrix A:** Select the first proximity matrix. If the row and column labels are included, make sure the "labels included" option is checked.

**Matrix B:** Select the second proximity matrix. If the row and column labels are included, make sure the "labels included" option is checked.

**Matrix C:** Activate this option if you want to compute the partial Mantel test. Then select the third proximity matrix. If the row and column labels are included, make sure the "labels included" option is checked.

**Range:** Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Activate this option to display the results in a new worksheet of the active workbook.

**Workbook:** Activate this option to display the results in a new workbook.

**Labels included:** Activate this option if the row and column labels have been selected.

**Options tab:**

**Alternative hypothesis:** Choose the alternative hypothesis to be used for the test (see description).

**Significance level (%):** Enter the significance level for the test.

**Exact p-values:** Activate this option so that XLSTAT tries to compute all the possible permutations when possible, to obtain an exact distribution of the Mantel statistic.
**Number of permutations**: Enter the number of permutations to perform in the case where it is not possible to generate all the possible permutations.

**Type of correlation**: Select the type of correlation to use to compute the standardized Mantel statistic.

**Missing data** tab:

**Do not accept missing data**: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

**Charts** tab:

**Scatter plot**: Activate this option to display a scatter plot using the values of matrix A on the X axis and the values of the matrix B on the Y axis.

**Histogram**: Activate this option to display the histogram computed from the distribution of $r(AB)$ based on the permutations.

**Results**

The displayed results correspond to the standardized Mantel statistic, to the corresponding p-value for the selected alternative hypothesis. A first level interpretation of the test is provided. The histogram of the $r(AB)$ distribution is displayed if the corresponding option has been checked. The observed value of $r(AB)$ is displayed on the histogram.

**Example**

An example showing how to use the Mantel test is displayed on the Addinsoft website:

[http://www.xlstat.com/demo-mantel.htm](http://www.xlstat.com/demo-mantel.htm)

**References**


**One-sample t and z tests**

Use this tool to compare the mean of a normally-distributed sample with a given value.

**Description**

Let the average of a sample be represented by $\mu$. To compare this mean with a reference value, two parametric tests are possible:

- **Student's t test** if the true variance of the population from which the sample has been extracted is not known; the variance of sample $s^2$ is used as variance estimator.

- **The z test** if the true variance $s^2$ of the population is known.

These two tests are said to be parametric as their use requires the assumption that the samples are distributed normally. Moreover, it also assumed that the observations are independent and identically distributed. The normality of the distribution can be tested beforehand using the normality tests.

Three types of test are possible depending on the alternative hypothesis chosen:

For the two-tailed test, the null H0 and alternative Ha hypotheses are as follows:

- $H_0$: $\mu = \mu_0$
- $H_a$: $\mu \neq \mu_0$

In the left one-tailed test, the following hypotheses are used:

- $H_0$: $\mu = \mu_0$
- $H_a$: $\mu < \mu_0$

In the right one-tailed test, the following hypotheses are used:

- $H_0$: $\mu = \mu_0$
- $H_a$: $\mu > \mu_0$
**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload**: Click this button to reload the default options.
- **Delete**: Click this button to delete the data selections.
- **Data Handling**: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab:**

- **Data**: Select the data in the Excel worksheet.

**Data format**: choose the data format.

- **One column/row per sample**: Activate this option for XLSTAT to consider that each column (column mode) or row (row mode) corresponds to a sample. You can then test the hypothesis on several samples at the same time.

- **One sample**: Activate this for XLSTAT to consider that all the selected values, whatever the number of rows or columns belong to the same sample.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.
Column/row labels: Activate this option if the first row (column mode) or first column (rows mode) of the selected data contain labels.

z Test: Activate this option to carry out a z test.

Student's t Test: Activate this option to carry out Student's t test.

Options tab:

Alternative hypothesis: Choose the alternative hypothesis to be used for the test (see description).

Theoretical mean: Enter the value of the theoretical mean with which the mean of the sample is to be compared.

Significance level (%): Enter the significance level for the tests (default value: 5%).

Where a z test has been requested, the population variance value must be entered.

Variance for the z test:

- Estimated using samples: Activate this option for XLSTAT to estimate the variance of the population from the sample data. This should, in principle, lead to a t test, but this option is offered for teaching purposes only.

- User defined: enter the value of the known variance of the population.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove the observations: Activate this option to remove observations with missing data.

Outputs tab:
**Descriptive statistics**: Activate this option to display descriptive statistics for the selected samples.

**Results**

The results displayed by XLSTAT relate to the various statistics of the tests selected and the interpretation arising from these.

**References**


Two-sample t and z tests

Use this tool to compare the means of two normally distributed independent or paired samples.

Description

Parametric t and z tests are used to compare the means of two samples. The calculation method differs according to the nature of the samples. A distinction is made between independent samples (for example a comparison of annual sales by shop between two regions for a chain of supermarkets), or paired samples (for example if comparing the annual sales within the same region over two years).

The t and z tests are known as parametric because the assumption is made that the samples are normally distributed. This hypothesis could be tested using normality tests.

Comparison of the means of two independent samples

Take a sample S1 comprising n1 observations, of mean μ1 and variance s1². Take a second sample S2, independent of S1 comprising n2 observations, of mean μ2 and variance s2². Let D be the assumed difference between the means (D is 0 when equality is assumed).

As for the z and t tests on a sample, we use:

Student's t test if the true variance of the populations from which the samples are extracted is not known;

The z test if the true variance s² of the population is known.

Student's t Test

The use of Student's t test requires a decision to be taken beforehand on whether variances of the samples are to be considered equal or not. XLSTAT gives the option of using Fisher's F test to test the hypothesis of equality of the variances and to use the result of the test in the subsequent calculations.

If we consider that the two samples have the same variance, the common variance is estimated by:

\[ s² = \frac{[(n1-1)s1² + (n2-1)s2²]}{(n1 + n2 - 2)} \]

The test statistic is therefore given by:
\[ t = \frac{(\mu_1 - \mu_2 - D)}{s \sqrt{1/n_1 + 1/n_2}} \]

The t statistic follows a Student distribution with \( n_1+n_2-2 \) degrees of freedom.

If we consider that the variances are different, the statistic is given by:
\[ t = \frac{(\mu_1 - \mu_2 - D)}{\sqrt{s_1^2 / n_1 + s_2^2 / n_2}} \]

A change in the number of degrees of freedom was proposed by Satterthwaite:
\[ df = \frac{(s_1^2 / n_1 + s_2^2 / n_2)^2}{\frac{(s_1^2 / n_1)^2}{n_1-1} + \frac{(s_2^2 / n_2)^2}{n_2-1}} \]

Note: when \( n_1=n_2 \), we simply have \( df = 2(n_1-1) \).

Cochran and Cox (1950) proposed an approximation to determine the p-value. It is given as an option in XLSTAT.

**z-Test**

For the z-test, the variance \( s^2 \) of the population is presumed to be known. The user can enter this value or estimate it from the data (this is offered for teaching purposes only). The test statistic is given by:
\[ z = \frac{(\mu_1 - \mu_2 - D)}{\sigma \sqrt{1/n_1 + 1/n_2}} \]

The z statistic follows a normal distribution.

**Comparison of the means of two paired samples**

If two samples are paired, they have to be of the same size. Where values are missing from certain observations, either the observation is removed from both samples or the missing values are estimated.

We study the mean of the calculated differences for the \( n \) observations. If \( d \) is the mean of the differences, \( s^2 \) the variance of the differences and \( D \) the supposed difference, the statistic of the t test is given by:
\[
t = \frac{(d - D)}{s / \sqrt{n}}
\]

The t statistic follows a Student distribution with n-1 degrees of freedom.

For the z test, the statistic is as follows where \( \sigma^2 \) is the variance

\[
z = \frac{(d - D)}{\sigma / \sqrt{n}}
\]

The z statistic follows a normal distribution.

Alternative hypotheses

Three types of test are possible depending on the alternative hypothesis chosen:

For the two-tailed test, the null H0 and alternative Ha hypotheses are as follows:
- H0: \( \mu_1 - \mu_2 = D \)
- Ha: \( \mu_1 - \mu_2 \neq D \)

In the left-tailed test, the following hypotheses are used:
- H0: \( \mu_1 - \mu_2 = D \)
- Ha: \( \mu_1 - \mu_2 < D \)

In the right-tailed test, the following hypotheses are used:
- H0: \( \mu_1 - \mu_2 = D \)
- Ha: \( \mu_1 - \mu_2 > D \)

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

[OK] Click this button to start the computations.
Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

Reload: Click this button to reload the default options.

Delete: Click this button to delete the data selections.

Arrow buttons: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

**Data / Sample 1:** If the format of the selected data is "one column per variable", select the data for the various samples in the Excel worksheet. If the format of the selected data is "one column per sample" or "paired samples", select a column of data corresponding to the first sample.

**Sample identifiers / Sample 2:** If the format of the selected data is "one column per variable", select the data identifying the two samples to which the selected data values correspond. If the format of the selected data is "one column per sample" or "paired samples", select a column of data corresponding to the second sample.

**Data format:** choose the data format.

- **One column/row per sample:** Activate this option to select one column (or row in row mode) per sample.

- **One column/row per variable:** Activate this option for XLSTAT to carry out as many tests as there are columns/rows, given that each column/row must contain the same number of rows/columns and that a sample identifier which enables each observation to be assigned to a sample must also be selected.

- **Paired samples:** Activate this option to carry out tests on paired samples. You must then select a column (or row in row mode) per sample, all the time ensuring that the samples are of the same size.

**Range:** Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.
**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Column/row labels**: Activate this option if the first row (column mode) or first column (rows mode) of the selected data contain labels.

**z-test**: Activate this option to carry out a z test.

**Student's t test**: Activate this option to carry out Student's t test.

**Options** tab:

**Alternative hypotheses**: Choose the alternative hypothesis to be used for the test (see description).

**Hypothesized difference (D)**: Enter the value of the supposed difference between the samples.

**Significance level (%)**: Enter the significance level for the tests (default value: 5%).

**Weights**: This option is only available if the data format is “One column/row per variable” or if the data are paired. Check this option if the observations are weighted. If you do not check this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the “Column/row labels” option is activated.

Where a z test has been requested, the value of the known variance of the populations, or, for a test on paired samples, the variance of the difference must be entered.

**Variances for the z test**: 

**Estimated using samples**: Activate this option for XLSTAT to estimate the variance of the population from the sample data. This should, in principle, lead to a t test, but this option is offered for teaching purposes only.

**User defined**: Enter the values of the known variances of the populations.

**Sample variances for the t-test**:
**Assume equality:** Activate this option to consider that the variances of the samples are equal.

**Cochran-Cox:** Activate this option to calculate the p-value by using the Cochran and Cox method where the variances are assumed to be unequal.

**Use an F test:** Activate this option to use Fisher's F test to determine whether the variances of both samples can be considered to be equal or not.

**Missing data tab:**

**Do not accept missing data:** Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

**Remove the observations:** Activate this option to remove observations with missing data.

**Outputs tab:**

**Descriptive statistics:** Activate this option to display descriptive statistics for the selected samples.

**Charts tab:**

**Dominance diagram:** Activate this option to display a dominance diagram in order to make a visual comparison of the samples.

**Example**

An example showing how to run a two sample Student’s t test is available at:

[http://www.xlstat.com/demo-ttest.htm](http://www.xlstat.com/demo-ttest.htm)

**Results**

The results displayed by XLSTAT relate to the various statistics of the tests selected and the interpretation arising from these.

The dominance diagram enables a visual comparison of the samples to be made. The first sample is represented on the x-axis and the second on the y-axis. To build this diagram, the
data from the samples is sorted first of all. When an observation in the second sample is
greater than an observation in the first sample, a "+" is displayed. When an observation in the
second sample is less than an observation in the first sample, a "-" is displayed. In the case of
a tie, a "o" is displayed.

References


River.


Comparison of the means of k samples

If you want to compare the means of k samples, you have to use the ANOVA tool which enables multiple comparison tests to be used.
One-sample variance test

Use this tool to compare the variance of a normally-distributed sample with a given value.

Description

Let us consider a sample of n independent normally distributed observations. One shows that the sample variance, \( s^2 \) follows a scale chi-squared distribution with \( n-1 \) degrees of freedom.

\[
\frac{s^2}{\sigma^2} \sim \frac{\chi^2}{n-1}
\]

where \( \sigma^2 \) is the theoretical sample variance. This allows us to compute a confidence interval around the variance.

To compare this variance to a reference value, a parametric test is proposed. It is based on the following statistic:

\[
\chi^2_0 = (n-1) \frac{s^2}{\sigma^0^2}
\]

which follows a chi-square distribution with \( n-1 \) degrees of freedom.

This test is said to be parametric as its use requires the assumption that the samples are distributed normally. Moreover, it also assumed that the observations are independent and identically distributed. The normality of the distribution can be tested beforehand using the normality tests.

Three types of test are possible depending on the alternative hypothesis chosen:

For the two-tailed test, the null H0 and alternative Ha hypotheses are as follows:

- H0 : \( \sigma^2 = \sigma_0^2 \)
- Ha : \( \sigma^2 \neq \sigma_0^2 \)

In the left one-tailed test, the following hypotheses are used:

- H0 : \( \sigma^2 = \sigma_0^2 \)
- Ha : \( \sigma^2 < \sigma_0^2 \)

In the right one-tailed test, the following hypotheses are used:

- H0 : \( \sigma^2 = \sigma_0^2 \)
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- Click this button to start the computations.
- Click this button to close the dialog box without doing any computation.
- Click this button to display the help.
- Click this button to reload the default options.
- Click this button to delete the data selections.

Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Data: Select the data in the Excel worksheet.

Data format: choose the data format.

- One column/row per sample: Activate this option for XLSTAT to consider that each column (column mode) or row (row mode) corresponds to a sample. You can then test the hypothesis on several samples at the same time.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.
Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column/row labels: Activate this option if the first row (column mode) or first column (rows mode) of the selected data contain labels.

Options tab:

Alternative hypothesis: Choose the alternative hypothesis to be used for the test (see description).

Theoretical variance: Enter the value of the theoretical mean with which the mean of the sample is to be compared.

Significance level (%): Enter the significance level for the tests (default value: 5%).

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove the observations: Activate this option to remove observations with missing data.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the selected samples.

Results

The results displayed by XLSTAT relate to the confidence interval around the variance and to the test comparing the observed variance to the theoretical variance.

Example

An example showing how to run a one sample variance test is available at:
References

Two-sample comparison of variances

Use this tool to compare the variances of two samples.

Description

Three parametric tests are offered for the comparison of the variances of two samples. Take a sample S1 comprising n1 observations with variance s1². Take a second sample S2 comprising n2 observations with variance s2². XLSTAT offers three tests for comparing the variances of the two samples.

Fisher’s F test

Let \( R \) be the assumed ratio of the variances (\( R = 1 \) when equality is assumed).

The test statistic \( F \) is given by:

\[
F = \frac{s1^2}{R.s2^2}
\]

This statistic follows a Fisher distribution with (n1-1) and (n2-1) degrees of freedom if both samples follow a normal distribution.

Three types of test are possible depending on the alternative hypothesis chosen:

For the two-tailed test, the null H0 and alternative Ha hypotheses are as follows:

- H0: \( s1^2 = s2^2.R \)
- Ha: \( s1^2 \neq s2^2.R \)

In the left-tailed test, the following hypotheses are used:

- H0: \( s1^2 = s2^2.R \)
- Ha: \( s1^2 < s2^2.R \)

In the right-tailed test, the following hypotheses are used:

- H0: \( s1^2 = s2^2.R \)
- Ha: \( s1^2 > s2^2.R \)
**Levene’s test**

Levene’s test can be used to compare two or more variances. It is a two-tailed test for which the null and alternative hypotheses are given by the following for the case where two variances are being compared:

- H0: $s_1^2 = s_2^2$
- Ha: $s_1^2 \neq s_2^2$

The statistic from this test is more complex than that from the Fisher test and involves absolute deviations at the mean (original article by Levene, 1960) or at the median (Brown and Forsythe, 1974). The use of the mean is recommended for symmetrical distributions with averagely thick tails. The use of the median is recommended for asymmetric distributions.

The Levene statistic follows a Fisher’s $F$ distribution with 1 and $n_1+n_2-2$ degrees of freedom.

**Bartlett’s homogeneity of variances test**

Bartlett’s test can be used to compare two or more variances. This test is sensitive to the normality of the data. In other words, if the hypothesis of normality of the data seems fragile, it is better to use Levene's or Fisher's test. On the other hand, Bartlett's test is more powerful if the samples follow a normal distribution.

This also is a two-tailed test which can be used with two or more variances. Where two variances are compared, the hypotheses are:

- H0: $s_1^2 = s_2^2$
- Ha: $s_1^2 \neq s_2^2$

Bartlett's statistic follows a Chi-square distribution with one degree of freedom.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

[Click this button to start the computations.]
Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

Reload: Click this button to reload the default options.

Delete: Click this button to delete the data selections.

General: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Data / Sample 1: If the format of the selected data is "one column per variable", select the data for the various samples in the Excel worksheet. If the format of the selected data is "one column per sample", select a column of data corresponding to the first sample.

Sample identifiers / Sample 2: If the format of the selected data is "one column per variable", select the data identifying the two samples to which the selected data values correspond. If the format of the selected data is "one column per sample", select a column of data corresponding to the second sample.

Data format: choose the data format.

- One column/row per sample: Activate this option to select one column (or row in row mode) per sample.

- One column/row per variable: Activate this option for XLSTAT to carry out as many tests as there are columns/rows, given that each column/row must contain the same number of rows/columns and that a sample identifier which enables each observation to be assigned to a sample must also be selected.

Column/row labels: Activate this option if the first row (column mode) or first column (rows mode) of the selected data contain labels.

Fisher's F test: Activate this option to use Fisher's F test (see description).

Levene's test: Activate this option to use Levene's test (see description).
- **Mean**: Activate this option to use Levene's test based on the mean.
- **Median**: Activate this option to use Levene's test based on the median.

**Bartlett's test**: Activate this option to use Bartlett's test (see description).

**Options** tab:

**Alternative hypothesis**: Choose the alternative hypothesis to be used for the test (see description).

**Hypothesized ratio (R)**: Enter the value of the supposed ratio between the variances of the samples.

**Significance level (%)**: Enter the significance level for the tests (default value: 5%).

**Missing data** tab:

**Do not accept missing data**: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

**Remove the observations**: Activate this option to remove observations with missing data.

**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the selected samples.

**Results**

The results displayed by XLSTAT relate to the various statistics of the tests selected and the interpretation arising from these.

**References**


k-sample comparison of variances

Use this tool to compare the variances of k samples.

Description

Two parametric tests are offered for the comparison of the variances of k samples (k ≥ 2). Take k samples S1, S2, ..., Sk, comprising n1, n2, ..., nk observations with variances s1², s2², ..., sk².

Levene's test

Levene's test can be used to compare two or more variances. This is a two-tailed test for which the null and alternative hypotheses are:

- H0: s¹² = s²² = ... = sₖ²
- Ha: There is at least one pair (i, j) such that sᵢ² ≠ sᵣ²

The statistic from this test involves absolute deviations at the mean (original article by Levene, 1960) or at the median (Brown and Forsythe, 1974). The use of the mean is recommended for symmetrical distributions with averagely thick tails. The use of the median is recommended for asymmetric distributions.

The Levene statistic follows a Fisher distribution with k-1 and n₁+n₂-2 degrees of freedom.

Bartlett's homogeneity of variances test

Bartlett's test can be used to compare two or more variances. This test is sensitive to the normality of the data. In other words, if the hypothesis of normality of the data seems fragile, it is better to use Levene's or Fisher's test. On the other hand, Bartlett's test is more powerful if the samples follow a normal distribution.

This also is a two-tailed test which can be used with two or more variances. Where two variances are compared, the hypotheses are:

- H0: s¹² = s²² = ... = sₖ²
- Ha: There is at least one pair (i, j) such that sᵢ² ≠ sᵣ²

Bartlett's statistic follows a Chi² distribution with k-1 degree of freedom.
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload**: Click this button to reload the default options.
- **Delete**: Click this button to delete the data selections.
- **Change orientation**: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

**Data / Sample 1**: If the format of the selected data is "one column per variable", select the data for the various samples in the Excel worksheet. If the format of the selected data is "one column per sample", select a column of data corresponding to the first sample.

**Sample identifiers / Sample 2**: If the format of the selected data is "one column per variable", select the data identifying the k samples to which the selected data values correspond. If the format of the selected data is "one column per sample", select a column of data corresponding to the second sample.

**Data format**: choose the data format.

- **One column/row per sample**: Activate this option to select one column (or row in row mode) per sample.
- **One column/row per variable**: Activate this option for XLSTAT to carry out as many tests as there are columns/rows, given that each column/row must contain the same
number of rows/columns and that a sample identifier which enables each observation to be assigned to a sample must also be selected.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Column/row labels**: Activate this option if the first row (column mode) or first column (rows mode) of the selected data contain labels.

**Levene's test**: Activate this option to use Levene's test (see description).

- **Mean**: Activate this option to use Levene's test based on the mean.
- **Median**: Activate this option to use Levene's test based on the median.

**Bartlett's test**: Activate this option to use Bartlett's test (see description).

**Options** tab:

**Significance level (%)**: Enter the significance level for the tests (default value: 5%).

**Missing data** tab:

**Do not accept missing data**: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

**Remove the observations**: Activate this option to remove observations with missing data.

**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the selected samples.
Results

The results displayed by XLSTAT relate to the various statistics of the tests selected and the interpretation arising from these.

References


Multidimensional tests (Mahalanobis, ...)

Use this tool to compare two or more samples simultaneously on several variables.

Description

The tests implemented in this tool are used to compare samples described by several variables. For example, instead of comparing the average of two samples as with the Student t test, we compare here simultaneously for the same samples averages measured for several variables.

Compared to a procedure that would involve as many Student t tests as there are variables, the method proposed here has the advantage of using the structure of covariance of the variables and of obtaining an overall conclusion. It may be that two samples are different for a variable with a Student t test, but that overall it is impossible to reject the hypothesis that they are similar.

Mahalanobis distance

The Mahalanobis distance, from the name if the Indian statistician Prasanta Chandra Mahalanobis (1893-1972), allows computing the distance between two points in a p-dimensional space, while taking into account the covariance structure across the p dimensions. The square of the Mahalanobis distance writes:

$$d_M^2 = (\bar{x}_1 - \bar{x}_2)' \Sigma^{-1} (\bar{x}_1 - \bar{x}_2)$$

In other words, it is the transposed of the vector of the difference of coordinates for p dimensions between the two points, multiplied by the inverse of the covariance matrix multiplied by the vector of differences. The Euclidean distance corresponds to the Mahalanobis distance where the covariance matrix is the identity matrix, which means that the variables are standardized and independent.

The Mahalanobis distance can be used to compare two groups (or samples) because the Hotelling $T^2$ statistic defined by:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} d_M^2$$

follows a Hotelling distribution, if the samples are normally distributed for all variables. The F statistic that is used for the comparison test where the null hypothesis $H_0$ is that the means of the two samples are equal, is defined by:

$$F = \frac{n_1 + n_2 - (p + 1)}{(n_1 + n_2 - 2) p} T^2$$
This statistic follows a Fisher’s F distribution with p and n1+n2-p-1 degrees of freedom if the samples are normally distributed for all the variables.

Note: This test can only be used if we assume that the samples are normally distributed and have identical covariance matrices. The second hypothesis can be tested with the Box or Kullback tests available in this tool.

If we want to compare more than two samples, the test based on the Mahalanobis distance can be used to identify possible sources of the difference observed at the global level. It is then recommended to use the Bonferroni correction for the alpha significance level. For k samples, we use the following significance level should be used:

\[ \alpha^* = \frac{2\alpha}{k(k - 1)} \]

**Wilks' lambda**

The Wilks’ lambda statistic follows the three parameters Wilks’ distribution defined by:

\[ \Lambda(p, m, n) = \frac{|A|}{|A + B|} \]

where A and B are two semi-defined positive matrices that respectively follow Wishart Wp(I, m) and Wp(I, n) distributions, where I is the identity matrix.

When we want to compare the means of p variables for k independent groups (or samples or classes), testing as null hypothesis H0 that the p averages are equal, if we assume that the covariance matrices are the same for the k groups, is equivalent to calculate the following statistic:

\[ \Lambda(p, n-k, k-1) = \frac{|W|}{|W + B|} \]

where

- W is the pooled within-group covariance matrix,
- B is the pooled between-groups covariance matrix,
- n is the total number of observations.

The distribution of the Wilks lambda is complex, so we use instead the Rao’s F statistic given by:

\[ F = \left(1 - \Lambda^{1/s}\right) \frac{m_2}{\Lambda^{1/s} m_1} \]
with
\[
\begin{align*}
    s &= \sqrt{\frac{p^2 (k - 1)^2 - 4}{p^2 + (k - 1)^2 - 5}} \\
    m_1 &= p(k - 1) \\
    m_2 &= s\left[n - (p + k + 2)/2\right] - p(k - 1)/2 + 1
\end{align*}
\]

One can show that if the sample size is large, then \(F\) follows a Fisher’s F distribution with \(m_1\) and \(m_2\) degrees of freedom. When \(p \leq 2\) or \(k = 2\), the F statistic is exactly distributed as \(F(m_1, m_2)\).

Note: This test can only be used if we assume that the samples are normally distributed and have identical covariance matrices. The second hypothesis can be tested with the Box or Kullback tests available in this tool.

**Testing the equality of the within-groups covariance matrices**

**Box test:** The Box test is used to test the assumption of equality for intra-class covariance matrices. Two approximations are available, one based on the Chi\(^2\) distribution, and the other on the Fisher distribution.

**Kullback’s test:** The Kullback’s test is used to test the assumption of equality for intra-class covariance matrices. The statistic calculated is approximately distributed according to a Chi\(^2\) distribution.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

**OK**: Click this button to start the computations.

**Cancel**: Click this button to close the dialog box without doing any computation.

**Help**: Click this button to display the help.

**Reloading button**: Click this button to reload the default options.
Click this button to delete the data selections.

Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

**Observations/variables table**: Select a table comprising N objects described by P descriptors. If column headers have been selected, check that the "Variable labels" option has been activated.

**Groups**: Check this option to select the values which correspond to the identifier of the group to which each observation belongs.

**Weights**: Activate this option if the observations are weighted. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Column labels" option is activated.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Variable labels**: Activate this option if the first row of the data selections includes a header.

**Options** tab:

**Wilks’ Lambda test**: Activate this option to compute the Wilks’ lambda test.

**Mahalanobis test**: Activate this option to compute the Mahalanobis distances as well as the corresponding F statistics and p-values.

- **Bonferroni correction**: Activate this option if you want to use a Bonferroni correction during the computation of the p-values corresponding to the Mahalanobis distances.

**Box test**: Activate this option to compute the Box test using the two available approximations.
**Kullback's test**: Activate this option to compute the Kullback's test.

**Significance level (%)**: Enter the significance level for the tests (default value: 5%).

**Missing data** tab:

**Do not accept missing data**: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

**Remove the observations**: Activate this option to remove observations with missing data.

**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the selected variables.

**Correlations**: Activate this option to display the correlation matrix.

**Covariance matrices**: Activate this option to display the inter-class, intra-class, intra-class total, and total covariance matrices.

**Results**

The results displayed by XLSTAT correspond to the various tests that have been selected.

**Example**

An example showing how to compare multidimensional samples is available on the Addinsoft website:

[http://www.xlstat.com/demo-maha.htm](http://www.xlstat.com/demo-maha.htm)

**References**

z-test for one proportion

Use this test to compare a proportion calculated from a sample with a given proportion.

Description

Let \( n \) be the number of observations verifying a certain property among a sample of size \( N \). The proportion of the sample verifying the property is defined by \( p = n / N \). Let \( p0 \) be a known proportion with which we wish to compare \( p \). Let \( D \) be the assumed difference (exact, minimum or maximum) between the two proportions \( p \) and \( p0 \). \( D \) is usually 0.

The two-tailed (or two-sided) test corresponds to testing the difference between \( p – p0 \) and \( D \), using the null (H0) and alternative (Ha) hypotheses shown below:

- H0: \( p – p0 = D \)
- Ha: \( p – p0 \neq D \)

In the one-tailed case, you need to distinguish the left-tailed (or lower-tailed or lower one-sided) test and the right-tailed (or right-sided or upper one-sided) test. In the left-tailed test, the following hypotheses are used:

- H0: \( p – p0 = D \)
- Ha: \( p – p0 < D \)

In the right-tailed test the following hypotheses are used:

- H0: \( p – p0 = D \)
- Ha: \( p – p0 > D \)

This z-test is based on the following assumptions:

- The observations are mutually independent,
- The probability \( p \) of having the property in question is identical for all observations,
- The number of observations is large enough, and the proportions are neither too close to 0 nor to 1.

Note: to determine whether \( N \) is sufficiently large one should make sure that:
\[
\begin{align*}
0 < p - 2\sqrt{p \frac{1-p}{N}} &< 1 \\
p + 2\sqrt{p \frac{1-p}{N}} &< 1
\end{align*}
\]

**z statistic**

One can find several ways to compute the z statistic in the statistical literature. The most used version is:

\[
z = \frac{p - p_0 - D}{\sigma}
\]

The large sample approximation leads to the following estimate for its standard deviation \( \sigma \):

\[
\hat{\sigma} = \sqrt{\frac{p(1-p)}{N}}
\]

However if one think that proportion we are comparing our sample proportion with might be a better estimate, one can use.

\[
\hat{\sigma} = \sqrt{\frac{p_0(1-p_0)}{N}}
\]

This version of the statistic should not be used when \( D \) is not null.

The z statistic is asymptotically normally distributed. The larger \( N \), the better the approximation. The p-value is computed using the normal approximation.

**Confidence intervals**

Many methods exist to compute confidence intervals on a proportion. XLSTAT offers the choice between four different versions: Wald, Wilson score, Clopper-Pearson, Agresti Coull.

**Dialog box**

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
General tab:

Frequency / Proportion: Enter the number of observations \( n \) for which the property is observed, or the corresponding proportion (see "data format" below).

Sample size: Enter the number of observations in the sample.

Test proportion: Enter the value of the test proportion with which the proportion observed is to be compared.

Data format: Choose here if you would prefer to enter the value of the **number of observations** for which the property is observed, or the **proportion** observed.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

z-Test: Activate this option to use a z-test.

Options tab:

Alternative hypotheses: Choose the alternative hypothesis to be used for the test.

Hypothesized difference (D): Enter the value of the supposed difference between the proportions.

Significance level (%): Enter the significance level for the test (default value: 5%).
**Variance:** Select the method used to estimate the variance of the proportion (used only for the confidence interval with the Wald interval).

- **Sample:** Activate this option to compute the variance using the proportion obtained for the sample.

- **Test proportion:** Activate this option to compute the variance using the test proportion and the size of the sample.

**Confidence interval:** Select the method used to compute the confidence interval (Wald, Wilson score, Clopper-Pearson, Agresti Coull).

**Results**

The results displayed by XLSTAT relate to the various statistics of the tests selected and the interpretation arising from these.

**Example**

An example showing how to compare proportions is available on the Addinsoft website:

http://www.xlstat.com/demo-prop.htm

**References**


**z-test for two proportions**

Use this tool to compare two proportions calculated for two samples.

**Description**

Let $n_1$ be the number of observations verifying a certain property for sample S1 of size $N_1$, and $n_2$ the number of observations verifying the same property for sample S2 of size $N_2$. The proportion of sample S1 verifying the property is defined by $p_1 = n_1 / N_1$, and the proportion for S2 is defined by $p_2 = n_2 / N_2$. Let D be the assumed difference (exact, minimum or maximum) between the two proportions $p_1$ and $p_2$. D is usually set to 0.

The two-tailed (or two-sided) test corresponds to testing the difference between $p_1 - p_2$ and D, using the null (H0) and alternative (Ha) hypotheses shown below:

- H0: $p_1 - p_2 = D$
- Ha: $p_1 - p_2 \neq D$

In the one-tailed case, you need to distinguish the left-tailed (or lower-tailed or lower one-sided) test and the right-tailed (or right-sided or upper one-sided) test. In the left-tailed test, the following hypotheses are used:

- H0: $p_1 - p_2 = D$
- Ha: $p_1 - p_2 < D$

In the right-tailed test the following hypotheses are used:

- H0: $p_1 - p_2 = D$
- Ha: $p_1 - p_2 > D$

This test is based on the following assumptions:

- The observations are mutually independent,
- The probability $p_1$ of having the property in question is identical for all observations in sample S1,
- The probability $p_2$ of having the property in question is identical for all observations in sample S2,
- The number of observations $N_1$ and $N_2$ are large enough, and the proportions are neither too close to 0 nor to 1.
Note: to determine whether N1 and N2 are sufficiently large one should make sure that:

\[
\begin{align*}
0 < p_i - 2\sqrt{p_i (1 - p_i) / N_i} & \quad \text{and} \quad 0 < p_2 - 2\sqrt{p_2 (1 - p_2) / N_2} \\
p_i + 2\sqrt{p_i (1 - p_i) / N_i} < 1 & \quad \text{and} \quad p_2 + 2\sqrt{p_2 (1 - p_2) / N_2} < 1
\end{align*}
\]

Dialog box

- Click this button to start the computations.
- Click this button to close the dialog box without doing any computation.
- Click this button to display the help.
- Click this button to reload the default options.
- Click this button to delete the data selections.

**General** tab:

**Frequency 1 / Proportion 1**: Enter the number of observations n1 for which the property is observed (see the description section), or the corresponding proportion (see "data format" below).

**Sample size 1**: Enter the number of observations in sample 1.

**Frequency 2 / Proportion 2**: Enter the number of observations n2 for which the property is observed (see the description section), or the corresponding proportion (see "data format" below).

**Sample size 2**: Enter the number of observations in sample 2.

**Data format**: Choose here if you would prefer to enter the values of the number of observations for which the property is observed, or the proportions observed.
Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

**z-Test:** Activate this option to use a z-test.

**Monte Carlo method:** Activate this option to compute the p-value using Monte Carlo simulations. Enter the number of simulations to perform.

**Options** tab:

**Alternative hypotheses:** Choose the alternative hypothesis to be used for the test (see the description section).

**Hypothesized difference (D):** Enter the value of the supposed difference between the proportions.

**Significance level (%):** Enter the significance level for the test (default value: 5%).

**Variance:** Select the method used to estimate the variance of the difference between the proportions.

- **p1q1/n1+ p2q2/n2:** Activate this option to compute the variance using this formula.
- **pq(1/n1+ 1/n2):** Activate this option to compute the variance using this formula.

**Results**

The results displayed by XLSTAT relate to the various statistics of the tests selected and the interpretation arising from these.

**Example**

An example showing how to compare proportions is available on the Addinsoft website:

References


Comparison of k proportions

Use this tool to compare k proportions, and to determine if they can be considered as equal, or if at least one pair of proportions shows a significant difference.

Description

XLSTAT offers three different approaches to compare proportions and to determine whether they can be considered as equal (null hypothesis H0) or if at least two proportions are significantly different (alternative hypothesis Ha):

Chi-square test: This test is identical to that used for the contingency tables;

Monte Carlo method: The Monte Carlo method is used to calculate a distribution of the Chi2 distance based on simulations with the constraint of complying with the total number of observations for the k groups. This results in an empirical distribution which gives a more reliable critical value (on condition that the number of simulations is large) than that given by the Chi2 theoretical distribution which corresponds to the asymptotic case.

Marascuilo procedure: It is advised to use the Marascuilo procedure only if the Chi-square test or the equivalent test based on Monte Carlo simulations reject H0. The Marascuilo procedure compares all pairs of proportions, which enables the proportions possibly responsible for rejecting H0 to be identified.

Dialog box

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload**: Click this button to reload the default options.
- **Delete**: Click this button to delete the data selections.
**Frequencies / Proportions**: Select the data in the Excel worksheet.

**Sample sizes**: Select the data corresponding to the sizes of the samples.

**Sample labels**: Activate this option if sample labels are available. Then select the corresponding data. If the “Column labels” option is activated you need to include a header in the selection. If this option is not activated, the row labels are automatically generated by XLSTAT (Sample1, Sample2 …).

**Data format**: Choose here if you would prefer to enter the value of the number of observations for which the property is observed, or the proportions observed.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Column labels**: Activate this option if the first line of the data selected (frequencies/proportions, sample size and sample labels) contain a label.

**Chi-square test**: Activate this option to use the $\chi^2$ test.

**Monte Carlo method**: Activate this option to use the simulation method and enter the number of simulations.

**Marascuilo procedure**: Activate this option to use the Marascuilo procedure.

**Significance level (%)**: Enter the significance level for the three tests (default value: 5%).

**Results**

The results of the Chi2 test are displayed first if the corresponding option has been activated. For the Chi2 test and the Monte Carlo method, the p-value is compared with the significance level in order to validate the null hypothesis.
The results obtained from Monte Carlo simulations are all the more close to the Chi-square results the higher the total number of observations and number of simulations. The difference relates to the critical value and the p-value.

The Marascuilo procedure identifies which proportions are responsible for rejecting the null hypothesis. It is possible to identify which pairs of proportions are significantly different by looking at the results in the "Significant" column.

Note: it might be that the Marascuilo procedure does not identify significant differences among the pairs of proportions, while the Chi-square test rejects the null hypothesis. In general, this happens when the two proportions are significantly different as identified by the Marascuilo procedure. More in-depth analysis might be necessary before making a decision.

Example

An example showing how to compare k proportions is available on the Addinsoft website:

http://www.xlstat.com/demo-kprop.htm

References


Multinomial goodness of fit test

Use this tool to check whether the observed frequencies of the values (categories) of a qualitative variable correspond to the expected frequencies or proportions.

Description

The multinomial goodness of fit test allows verifying whether the distribution of a sample corresponding to a qualitative variable (or discretized quantitative) is consistent with what is expected. The test is based on the multinomial distribution which is the extension of the binomial distribution if there are more than two possible outcomes.

Let k be the number of possible values (categories) for variable X. We write p1, p2, …, pk the probabilities (or densities) corresponding to each value.

Let n1, n2, n3, …, nk be the frequencies of each value for a sample.

The null hypothesis of the test writes:

- H0: The distribution of the values in the sample is consistent with what is expected, meaning the distribution of the sample is not different from the distribution of X.

The alternative hypothesis of the test writes:

- Ha: The distribution of the values in the sample is not consistent with what is expected, meaning the distribution of the sample is different from the distribution of X.

Several methods and statistics have been proposed for this test. XLSTAT offers the following two possibilities:

1. Chi-square test:

We compute the following statistic:

\[ \chi^2 = \sum_{i=1}^{k} \frac{(n_i - Np_i)^2}{Np_i} \]

This statistic is asymptotically distributed as Chi-square with k-1 degrees of freedom.

2. Monte Carlo test:
This version of the test overcomes some heavy calculations of the exact method based on the multinomial distribution, and avoids the approximation by the Chi-square distribution that may be of poor quality with small samples. This test consists of a random resampling of \( N \) observations in a distribution having the expected properties. For each resampling, we compute the \( \chi^2 \) statistic, then once the resampling process is finished, we evaluate how many times the value observed on the sample is exceeded, from what we deduce the \( p \)-value.

**Dialog box**

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- ![Reload](reload.png): Click this button to reload the default options.
- ![Delete](delete.png): Click this button to delete the data selections.

**Frequencies**: Select the data corresponding to the observed frequencies in the Excel worksheet.

**Expected frequencies / Expected proportions**: Select the data corresponding to the expected frequencies or to the expected proportions. If you select expected frequencies, they must sum to the same value as the sum of the observed frequencies.

**Data format**: Choose here if you would prefer to select expected frequencies or expected proportions.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Column labels**: Activate this option if the first line of the data selected (frequencies/proportions, sample size and sample labels) contain a label.
**Chi-square test**: Activate this option to use the Chi-square test.

**Monte Carlo method**: Activate this option to use the simulation method and enter the number of simulations.

**Significance level (%)**: Enter the significance level for the two tests (default value: 5%).

**Results**

The results of the Chi-square test are displayed first if the corresponding option has been activated. For the Chi-square test and the Monte Carlo method, the p-value is compared with the significance level in order to validate the null hypothesis.

The results obtained from Monte Carlo simulations are all the more close to the Chi-square results the higher the total number of observations and number of simulations. The difference relates to the critical value and the p-value.

For the Monte Carlo test, a confidence interval on the p-value is displayed.

**Example**

An example showing how to run a multinomial goodness of fit test is available on the Addinsoft website:

[http://www.xlstat.com/demo-goodness.htm](http://www.xlstat.com/demo-goodness.htm)

**References**

Equivalence test (TOST)

Use this tool to test the equivalence of the two normally distributed independent samples.

Description

Unlike classical hypothesis testing, equivalence tests are used to validate the fact that a difference is in a given interval.

This type of test is used primarily to validate bioequivalence. When we want to show the equivalence of two drugs, classical hypothesis testing does not apply, we will use equivalence testing which will validate the equivalence between the two drugs.

In a classical hypothesis test, we try to reject the null hypothesis of equality. As part of an equivalence test, we try to validate the equivalence between two samples. The TOST (two one-sided test) is a test of equivalence that is based on the classical t test used to test the hypothesis of equality between two means.

So we will have two samples, a theoretical difference between the means as well as a range within which we can say that the sample means are equivalent.

The test is known as parametric because the assumption is made that the samples are normally distributed. This hypothesis could be tested using normality tests.

The TOST test uses Student's test to check the equivalence between the means of two samples. A detailed description of such tests can be found in the chapter dedicated to t tests.

XLSTAT offers two equivalent methods to test equivalence using the TOST test.

- Using the 100 * (1-2 * alpha)% confidence interval around the mean. By comparing this interval to the user-defined interval of equivalence, we can conclude the equivalence or non equivalence. Thus, if the confidence interval is within the interval defined by the user, we conclude the equivalence between the two samples. If one of the bounds of the confidence interval is outside the interval defined by the user, then the two samples are not equivalent.

- Using two one-sided tests, one on the right and one on the left. We apply a right one-sided t-test on the lower bound of the interval defined by the user and a left one-sided t-test on the upper bound of the interval defined by the user. We obtain p-values for both tests. We take the greatest of these p-values as p-value of the equivalence test.

These two tests are similar and should give similar results. They were introduced by Schuirman's (1987).
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **reload the default options**: Click this button to reload the default options.
- **Delete data selections**: Click this button to delete the data selections.
- **Change handling of data**: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab:**

**Sample 1**: Select a column of data corresponding to the first sample.

**Sample 2**: Select a column of data corresponding to the second sample.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Column/row labels**: Activate this option if the first row (column mode) or first column (rows mode) of the selected data contain labels.

**Options** tab:
**Hypothesized difference (D):** Enter the value of the supposed difference between the samples.

**Lower bound:** Enter the value of the supposed lower bound for equivalence testing.

**Upper bound:** Enter the value of the supposed upper bound for equivalence testing.

**Significance level (%):** Enter the significance level for the tests (default value: 5%).

**Weights:** This option is only available if the data format is “One column/row per variable” or if the data are paired. Check this option if the observations are weighted. If you do not check this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the “Column/row labels” option is activated.

Sample variances for the t-test:

**Assume equality:** Activate this option to consider that the variances of the samples are equal.

**Cochran-Cox:** Activate this option to calculate the p-value by using the Cochran and Cox method where the variances are assumed to be unequal.

**Use an F test:** Activate this option to use Fisher's F test to determine whether the variances of both samples can be considered to be equal or not.

**Missing data** tab:

**Do not accept missing data:** Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

**Remove the observations:** Activate this option to remove observations with missing data.

**Outputs** tab:

**Descriptive statistics:** Activate this option to display descriptive statistics for the selected samples.

**Results**

The first table displays the descriptive statistics associated with the two samples.
The following table of results can be used to validate the hypothesis of equivalence for two means. If the confidence interval around the difference with a confidence level of (1-2 *alpha)*100% is included in the interval defined by the user in the dialog box, then the samples are equivalent. You have to check if the four values in this table are ordered increasingly. The last line gives an interpretation (equivalence or non equivalence).

The following table allows you to view two one-sided tests based on the bounds defined by the user. The p-value for the test of equivalence is the largest p-value obtained with the one-sided t tests.

Example

An example showing how to run an equivalence test for two samples is available at:

http://www.xlstat.com/demo-tost.htm

References


Comparison of two distributions (Kolmogorov-Smirnov)

Use this tool to compare the distributions of two samples and to determine whether they can be considered identical.

Description

The Kolmogorov-Smirnov test compares two distributions. This test is used for distribution fitting tests for comparing an empirical distribution determined from a sample with a known distribution. It can also be used for comparing two empirical distributions.

Note: this test enables the similarity of the distributions to be tested at the same time as their shape and position.

Take a sample S1 comprising n1 observations, with F1 the corresponding empirical distribution function. Take a second sample S2 comprising n2 observations, with F2 the corresponding empirical distribution function.

The null hypothesis of the Kolmogorov-Smirnov test is defined by:

\[ H_0: F_1(x) = F_2(x) \]

The Kolmogorov statistic is given by:

\[ D_1 = \sup_x \left| F_1(x) - F_2(x) \right| \]

D1 is the maximum absolute difference between the two empirical distributions. Its value therefore lies between 0 (distributions perfectly identical) and 1 (separations perfectly separated). The alternative hypothesis associated with this statistic is:

\[ H_a: F_1(x) \neq F_2(x) \]

The Smirnov statistics are defined by:

\[ D_2 = \sup_x \left( F_1(x) - F_2(x) \right) \]

\[ D_3 = \sup_x \left( F_2(x) - F_1(x) \right) \]

The alternative hypothesis associated with D2 is:

\[ H_a: F_1(x) < F_2(x) \]
The alternative hypothesis associated with D3 is:

\[ \text{Ha: } F_1(x) > F_2(x) \]

Nikoforov (1994) proposed an exact test method for the Kolmogorov-Smirnov on two samples. This method is used by XLSTAT for the three alternative hypotheses. XLSTAT also enables the supposed difference D between the distributions to be introduced. The value must be between 0 and 1.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- ![Reload Default Options](Reload Default Options.png): Click this button to reload the default options.
- ![Delete Data Selections](Delete Data Selections.png): Click this button to delete the data selections.
- ![Change Data Handling](Change Data Handling.png): Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab:**

**Data / Sample 1**: If the format of the selected data is "one column per variable", select the data for the various samples in the Excel worksheet. If the format of the selected data is "one column per sample", select a column of data corresponding to the first sample.
**Sample identifiers / Sample 2:** If the format of the selected data is "one column per variable", select the data identifying the two samples to which the selected data values correspond. If the format of the selected data is "one column per sample", select a column of data corresponding to the second sample.

**Data format:** choose the data format.

- **One column/row per sample:** Activate this option to select one column (or row in row mode) per sample.

- **One column/row per variable:** Activate this option for XLSTAT to carry out as many tests as there are columns/rows, given that each column/row must contain the same number of rows/columns and that a sample identifier which enables each observation to be assigned to a sample must also be selected.

**Range:** Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Activate this option to display the results in a new worksheet of the active workbook.

**Workbook:** Activate this option to display the results in a new workbook.

**Column/Row labels:** Activate this option if the first row (column mode) or first column (rows mode) of the selected data contain labels.

**Kolmogorov-Smirnov test:** Activate this option to run the Kolmogorov-Smirnov test (see description).

**Options** tab:

**Alternative hypothesis:** Choose the alternative hypothesis to be used for the test (see description).

**Hypothesized difference (D):** Enter the value of the maximum supposed difference between the empirical distribution functions of the samples. The value must be between 0 and 1.

**Significance level (%):** Enter the significance level for the test (default value: 5%).

**Missing data** tab:
**Do not accept missing data:** Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

**Remove the observations:** Activate this option to remove observations with missing data.

**Outputs tab:**

**Descriptive statistics:** Activate this option to display descriptive statistics for the selected samples.

**Charts tab:**

**Dominance diagram:** Activate this option to display a dominance diagram in order to make a visual comparison of the samples.

**Cumulative histograms:** Activate this option to display the chart showing the empirical distribution functions for the samples.

**Results**

The results displayed by XLSTAT relate to the various statistics of the tests selected and the interpretation arising from these.

**References**


**Smirnov N. V. (1939).** On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bulletin Moscow University, 2*, 3-14.
Comparison of two samples (Wilcoxon, Mann-Whitney, ...)

Use this tool to compare two samples described by ordinal or discrete quantitative data whether independent or paired.

Description

To get round the assumption that a sample is normally distributed required for using the parametric tests (z test, Student's t test, Fisher's F test, Levene's test and Bartlett's test), non-parametric tests have been put forward.

As for parametric tests, a distinction is made between independent samples (for example a comparison of annual sales by shop between two regions for a chain of supermarkets), or paired samples (for example if comparing the annual sales within the same region over two years).

If we designate $D$ to be the assumed difference in position between the samples (in general we test for equality, and $D$ is therefore 0), and $P_1 - P_2$ to be the difference of position between the samples, three tests are possible depending on the alternative hypothesis chosen:

For the two-tailed test, the null $H_0$ and alternative $H_a$ hypotheses are as follows:

- $H_0$: $P_1 - P_2 = D$
- $H_a$: $P_1 - P_2 \neq D$

In the left-tailed test, the following hypotheses are used:

- $H_0$: $P_1 - P_2 = D$
- $H_a$: $P_1 - P_2 < D$

In the right-tailed test, the following hypotheses are used:

- $H_0$: $P_1 - P_2 = D$
- $H_a$: $P_1 - P_2 > D$

Comparison of two independent samples

Three researchers, Mann, Whitney, and Wilcoxon, separately perfected a very similar non-parametric test which can determine if the samples may be considered identical or not on the basis of their ranks. This test is often called the Mann-Whitney test, sometimes the Wilcoxon-Mann-Whitney test or the Wilcoxon Rank-Sum test (Lehmann, 1975).
We sometimes read that this test can determine if the samples come from identical populations or distributions. This is completely untrue. It can only be used to study the relative positions of the samples. For example, if we generate a sample of 500 observations taken from an N(0,1) distribution and a sample from a distribution of 500 observations from an N(0,4) distribution, the Mann-Whitney test will find no difference between the samples.

Let S1 be a sample made up of n1 observations \((x_1, x_2, \ldots, x_{n_1})\) and S2 a second sample made up of n2 observations \((y_1, y_2, \ldots, y_{n_1})\) independent of S1. Let N be the sum of n1 and n2.

To calculate the Wilcoxon \(W_s\) statistic which measures the difference in position between the first sample S1 and sample S2 from which \(D\) has been subtracted, we combine the values obtained for both samples, then put them in order. The \(W_s\) statistic is the sum of the ranks of one of the samples. For XLSTAT, the sum is calculated on the first sample.

For the expectation and variance of \(W_s\) we therefore have:

\[
E(W_s) = \frac{1}{2} n_1 (N + 1) \quad \text{and} \quad V(W_s) = \frac{1}{12} n_1 n_2 (N + 1)
\]

The Mann-Whitney U statistic is the sum of the number of pairs \((x_i, y_i)\) where \(x_i > y_i\), from among all the possible pairs. We show that

\[
E(U) = \frac{n_1 n_2}{2} \quad \text{and} \quad V(U) = \frac{1}{12} n_1 n_2 (N + 1)
\]

We may observe that the variances of \(W_s\) and \(U\) are identical. In fact, the relationship between \(U\) and \(W_s\) is:

\[
W_s = U + \frac{n_1 (n_1 + 1)}{2}
\]

The results offered by XLSTAT are those relating to Mann-Whitney’s \(U\) statistic.

When there are ties between the values in the two samples, the rank assigned to the tied values is the mean of their rank before processing (for example, for two samples of respective size 3 and 3, if the ordered list of values is \(\{1, 1.2, 1.2, 1.4, 1.5, 1.5\}\), the ranks are initially \(\{1, 2, 3, 4, 5, 6\}\) then after inclusion \(\{1, 2.5, 2.5, 4, 4.5, 5.5\}\). Although this does not change the expectation of \(W_s\) and \(U\), the variance is, on the other hand, modified.

\[
V(W_s) = V(U) = \frac{1}{12} n_1 n_2 (N + 1) - \frac{n_1 n_2 \sum_{i=1}^{nd} (d_i^3 - d_i)}{12N(N - 1)}
\]

where \(nd\) is the number of distinct values and \(d_i\) the number of observations for each of the values.
For the calculation of the p-values associated with the statistic, XLSTAT can use an exact method if the user wants for the following cases:

\[ U*n1*n2 \leq 10e7, \] if there are no ties

\[ U*nd \leq 5000 \text{ if there are ties.} \]

The calculations may be appreciably slowed down where there are ties. A normal approximation has been proposed to get round this problem. We have:

\[
P(U \leq u) \approx \Phi\left(\frac{u - E(U) + c}{\sqrt{V(U)}}\right)
\]

where \(F\) is the distribution function for the standardized normal distribution, and \(c\) is a continuity correction used to increase the quality of the approximation (\(c\) is \(\frac{1}{2}\) or \(-\frac{1}{2}\) depending on the nature of the test). The approximation is more reliable the higher \(n1\) and \(n2\) are.

If the user requests that an exact test be used and this is not possible because of the constraints given below, XLSTAT indicates in the results report that an approximation has been used.

A Monte Carlo approximation of the p-value is also possible for this test.

**Comparison of two paired samples**

Two tests have been proposed for the cases where samples are paired: the **sign test** and the **Wilcoxon signed rank test**.

Let \(S1\) be a sample made up of \(n\) observations \((x1, x2, \ldots, xn)\) and \(S2\) a second sample paired with \(S1\), also comprising \(n\) observations \((y1, y2, \ldots, yn)\). Let \((p1, p2, \ldots, pn)\) be the \(n\) pairs of values \((xi, yi)\).

**Sign test**

Let \(N+\) be the number of pairs where \(yi > xi\), \(N0\) the number of pairs where \(yi = xi\), and \(N-\) the number of pairs where \(yi < xi\). We can show that \(N+\) follows a binomial distribution with parameters \((n-N0)\) and probability \(\frac{1}{2}\). The expectation and the variance of \(N+\) are therefore:

\[
E(N+) = \frac{n - N0}{2} \quad \text{and} \quad V(N+) = \frac{n - N0}{4}
\]

The p-value associated with \(N+\) and the type of test chosen (two-tailed, right or left one-tailed) can therefore be determined exactly.
Note: This test is called the sign test as it constructs the differences within the n pairs from the sign. This test is therefore used to compare evolutions evaluated on an ordinal scale. For example, this test would be used to determine if the effect of a medicine is positive from a survey where the patient simply declares if he feels less well, not better, or better after taking it.

The disadvantage of the sign test is that it does not take into account the size of the difference between each pair, data which is often available.

Wilcoxon signed-rank test

Wilcoxon proposed a test which takes into account the size of the difference within pairs. This test is called the Wilcoxon signed rank test, as the sign of the differences is also involved.

As for the sign test, the differences for all the pairs is calculated, then they are ordered and finally the positive differences $S_1, S_2, \ldots, S_p$ and the negative differences $R_1, R_2, \ldots, R_m$ ($p+m=n$) are separated.

The statistic used to show whether both samples have the same position is defined as the sum of the $S_i$'s:

$$V_s = \sum_{i=1}^{p} S_i$$

The expectation and the variance of $V_s$ are:

$$E(V_s) = \frac{n(n+1)}{4} \quad \text{and} \quad V(V_s) = \frac{n(n+1)(2n+1)}{24}$$

Where they might be ties among the differences, or null differences for certain pairs, we have:

$$E(V_s) = \frac{n(n+1) - d_0 (d_0 + 1)}{4}$$

$$V(V_s) = \frac{\left[n(n+1)(2n+1) - d_0 (d_0 + 1)(2d_0 + 1)\right]}{24} - \frac{\sum_{j=1}^{nd} (d_i^3 - d_i)}{48}$$

where $d_0$ is the number of null differences, $nd$ the number of distinct differences, and $d_i$ the number of values corresponding to the i'th distinct difference value (it is the same as considering that the d_i's are the number of ties for the i'th distinct difference value).

Where there are no null differences or ties among the differences, if n is less than or equal to 100, XLSTAT calculates an exact p-value (Lehmann, 1975). Where there are ties, a normal approximation is used. We have:
\[ P(V_s \leq v) \approx \Phi \left( \frac{v - E(V_s) + c}{\sqrt{V(V_s)}} \right) \]

where \( \Phi \) is the distribution function for the standardized normal distribution, and \( c \) is a continuity correction used to increase the quality of the approximation (\( c \) is \( \frac{1}{2} \) or \(-\frac{1}{2}\) depending on the nature of the test). The approximation is more reliable the higher \( n \) is.

A Monte Carlo approximation of the p-value is also possible for this test.

**Computation of the p-values**

To compute the p-values corresponding to the various statistics, XLSTAT offers several alternatives:

- **Asymptotic method**: The p-value is obtained using the asymptotic approximation of the distribution of the statistic. The reliability of the approximation depends on the number of samples and on the number of measures per sample.

- **Exact method**: The computation of the p-value is based on the exact distribution of the statistic.

- **Monte Carlo method**: The computation of the p-value is based on random resamplings. The user must set the number of resamplings. A confidence interval on the p-value is provided. The more resamplings are performed, the better the estimation of the p-value.

In order to avoid freezing Excel because of too long computations, it is possible with the two latter methods to set the maximum time that should be spent computing the p-value.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.

- **Cancel**: Click this button to close the dialog box without doing any computation.

- **Help**: Click this button to display the help.
Click this button to reload the default options.

Click this button to delete the data selections.

Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

**Data / Sample 1**: If the format of the selected data is "one column per variable", select the data for the various samples in the Excel worksheet. If the format of the selected data is "one column per sample" or "paired samples", select a column of data corresponding to the first sample.

**Sample identifiers / Sample 2**: If the format of the selected data is "one column per variable", select the data identifying the two samples to which the selected data values correspond. If the format of the selected data is "one column per sample" or "paired samples", select a column of data corresponding to the second sample.

**Data format**: choose the data format.

- **One column/row per sample**: Activate this option to select one column (or row in row mode) per sample.

- **One column/row per variable**: Activate this option for XLSTAT to carry out as many tests as there are columns/rows, given that each column/row must contain the same number of rows/columns and that a sample identifier which enables each observation to be assigned to a sample must also be selected.

- **Paired samples**: Activate this option to carry out tests on paired samples. You must then select a column (or row in row mode) per sample, all the time ensuring that the samples are of the same size.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.
**Column/Row labels**: Activate this option if the first row (column mode) or first column (rows mode) of the selected data contain labels.

**Mann-Whitney test**: Activate this option to run the Mann-Whitney test (see description).

**Sign test**: Activate this option to use sign test (see description).

**Wilcoxon signed rank test**: Activate this option to use Wilcoxon signed rank test (see description).

**Options** tab:

**Alternative hypothesis**: Choose the alternative hypothesis to be used for the test (see description).

**Hypothesized difference (D)**: Enter the value of the supposed difference between the samples.

**Significance level (%)**: Enter the significance level for the test (default value: 5%).

Depending on the test that is being used, several methods can be available to compute the p-value. Choose among the **asymptotic**, **exact** or **Monte Carlo** methods (see the description section for more information). In the case of the exact and Monte Carlo method you can set the maximum time you want to spend computing the p-value.

**Continuity correction**: Activate this option if you want XLSTAT to use the continuity correction when computing the asymptotic p-value (see description).

**Missing data** tab:

**Do not accept missing data**: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

**Remove the observations**: Activate this option to remove observations with missing data.

**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the selected samples.
Charts tab:

**Dominance diagram:** Activate this option to display a dominance diagram in order to make a visual comparison of the samples.

Results

The results displayed by XLSTAT relate to the various statistics of the tests selected and the interpretation arising from these.

Example

References


Comparison of k samples (Kruskal-Wallis, Friedman, ...)

Use this tool to compare k independent samples (Kruskal-Wallis test and Dunn's procedure) or paired samples (Friedman's test and Nemenyi's procedure, GPU accelerated).

Description

To get round the assumption that a sample is normally distributed required for using multiple comparison tests (offered in XLSTAT after an ANOVA), non-parametric tests were proposed.

As for parametric tests, a distinction is made between independent samples (for example a comparison of crop yields from fields with similar properties but treated with three different types of fertilizer), from cases where they are paired (for example if comparing the scores given by 10 judges to 3 different products).

Comparison of k independent samples

The Kruskal-Wallis test is often used as an alternative to the ANOVA where the assumption of normality is not acceptable. It is used to test if k samples (k≥2) come from the same population or populations with identical properties as regards a position parameter (the position parameter is conceptually close to the median, but the Kruskal-Wallis test takes into account more information than just the position given by the median).

If Mi is the position parameter for sample i, the null H0 and alternative Ha hypotheses for the Kruskal-Wallis test are as follows:

- H0: M1 = M2 = … = Mk
- Ha: There is at least one pair (i, j) such that Mi ≠ Mj

The calculation of the K statistic from the Kruskal-Wallis test involves, as for the Mann-Whitney test, the rank of the observations once the k samples (or groups) have been mixed. K is defined by:

\[ K = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1) \]

where \( n_i \) is the size of sample i, \( N \) is the sum of the \( n_i \)'s, and \( R_i \) is the sum of the ranks for sample i.

When k=2, the Kruskal-Wallis test is equivalent to the Mann-Whitney test and K is equivalent to \( Ws \).
When there are ties, the mean ranks are used for the corresponding observations as in the case of the Mann-Whitney test. $K$ is then given by:

$$K = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N + 1)$$

$$1 - \sum_{i=1}^{nd} \left( d_i^3 - d_i \right) / \left( N^3 - N \right)$$

where $nd$ is the number of distinct values and $d_i$ the number of observations for each of the values.

The distribution of the $K$ statistic can be approximated by a Chi-square distribution with $(k-1)$ degrees of freedom. This approximation is reliable, except when $N$ is small. The p-values associated with $K$, which for the exact case depends on the statistic $K$ and the $k$ sizes of the samples, have been tabulated for the case where $k = 3$ (Lehmann 1975, Hollander and Wolfe 1999).

**Comparison of $k$ paired samples**

The **Friedman test** is a non-parametric alternative to the two-way ANOVA where the assumption of normality is not acceptable. It is used to test if $k$ paired samples ($k \geq 2$) of size $n$, come from the same population or from populations having identical properties as regards the position parameter. As the context is often that of the two-way ANOVA factors, we sometimes speak of the Friedman test with $k$ treatments and $n$ blocks.

If $M_i$ is the position parameter for sample $i$, the null $H_0$ and alternative $H_a$ hypotheses for the Friedman test are as follows:

- $H_0$: $M_1 = M_2 = \ldots = M_k$
- $H_a$: There is at least one pair $(i, j)$ such that $M_i \neq M_j$

Let $n$ be the size of $k$ paired samples. The $Q$ statistic from the Friedman test is given by:

$$Q = \frac{12}{nk(k+1)} \sum_{i=1}^{k} R_i^2 - 3n(k + 1)$$

where $R_i$ is the sum of the ranks for sample $i$.

Where there are ties, the average ranks are used for the corresponding observations. $Q$ is then given by:
\[ Q = \frac{12}{nk(k+1)} \sum_{i=1}^{k} R_i^2 - 3n(k+1) \\
1 - \sum_{j=1}^{N} \sum_{i=1}^{nd(j)} \frac{d_{ij}^3 - d_{ij}}{n(k^3 - k)} \]

where \( nd(j) \) is the number of distinct values for block \( j \), and \( d_{ij} \) the number of observations for each of the values.

As for the Kruskal-Wallis test, the p-value associated with a given value of \( Q \) can be approximated by a Chi-square distribution with \((k-1)\) degrees of freedom. This approximation is reliable when \( kn \) is greater than 30, the quality also depending on the number of ties. The p-values associated with \( Q \) have been tabulated for \((k = 3, n \leq 15)\) and \((k = 4, n \leq 8)\) (Lehmann 1975, Hollander and Wolfe 1999).

**Computation of the p-values**

To compute the p-values corresponding to the various statistics, XLSTAT offers several alternatives:

- **Asymptotic method**: The p-value is obtained using the asymptotic approximation of the distribution of the statistic. The reliability of the approximation depends on the number of samples and on the number of measures per sample.

- **Exact method**: The computation of the p-value is based on the exact distribution of the statistic. This method is available for the Kruskal-Wallis test when there are no ties.

- **Monte Carlo method**: The computation of the p-value is based on random resamplings. The user must set the number of resamplings. A confidence interval on the p-value is provided. The more resamplings are performed, the better the estimation of the p-value.

In order to avoid freezing Excel because of too long computations, it is possible with the two latter methods to set the maximum time that should be spent computing the p-value.

**Multiple pairwise comparisons**

Whether for the Kruskal-Wallis or the Friedman test, if the p-value is such that the H0 hypothesis has to be rejected, then at least one sample (or group) is different from another. To identify which samples are responsible for rejecting H0, multiple comparison procedures can be used.

For the Kruskal-Wallis test, three multiple comparison methods are available:

- **Dunn (1963)**: the method based on the comparison of the mean of the ranks of each treatment, the ranks being those used for the computation of \( K \). The normal distribution is
used as the asymptotic distribution of the standardized difference of the mean of the ranks.

- Conover and Iman (1999): close to Dunn’s method, this method uses a Student distribution. It corresponds to a t test performed on the ranks.

- Steel-Dwass-Critchlow-Fligner (1984): This more complex method is recommended by Hollander (1999). It requires the recalculation of the ranks for each combination of treatments. The Wij statistic is calculated for each combination. XLSTAT then calculates the corresponding p-value using the asymptotic distribution.

For the Friedman test, one multiple comparison method is available:

- Nemenyi (1963). This method is close to the Dunn’s procedure, but it takes into account the fact that data are paired.

For the Dunn, the Conover-Iman and the Nemenyi methods, to take into account the fact that there are \( k(k-1)/2 \) possible comparisons, the correction of the significance level proposed by Bonferroni can be applied. The significance level used for pairwise comparisons is:

\[
\alpha' = \frac{2\alpha}{k(k-1)}
\]

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **Ok**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload**: Click this button to reload the default options.
- **Delete**: Click this button to delete the data selections.
- **Change**: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If
the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab:**

**Data:** If the format of the selected data is "one column per variable", select the data for the various samples in the Excel worksheet. If the format of the selected data is "one column per sample" or "paired samples", select the columns of data corresponding to the various samples.

**Sample identifiers:** If the format of the selected data is "one column per variable", select the data identifying the k samples to which the selected data values correspond.

**Data format:** choose the data format.

- **One column/row per sample:** Activate this option to select one column (or row in row mode) per sample.

- **One column/row per variable:** Activate this option for XLSTAT to carry out as many tests as there are columns/rows, given that each column/row must contain the same number of rows/columns and that a sample identifier which enables each observation to be assigned to a sample must also be selected.

- **Paired samples:** Activate this option to carry out tests on paired samples. You must then select a column (or row in row mode) per sample, all the time ensuring that the samples are of the same size.

**Range:** Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Activate this option to display the results in a new worksheet of the active workbook.

**Workbook:** Activate this option to display the results in a new workbook.

**Column/Row labels:** Activate this option if the first row (column mode) or first column (rows mode) of the selected data contain labels.

**Kruskal-Wallis test:** Activate this option to run the Kruskal-Wallis test (see description).

**Friedman test:** Activate this option to run a Friedman test (see description).
Multiple pairwise comparisons: Activate this option to compute multiple pairwise comparisons (see description).

- Bonferroni correction: Activate this option to use the Bonferroni corrected significance level for the multiple comparisons.

Options tab:

Alternative hypothesis: Choose the alternative hypothesis to be used for the test (see description).

Significance level (%): Enter the significance level for the test (default value: 5%).

Depending on the test that is being used, several methods can be available to compute the p-value. Choose among the asymptotic, exact or Monte Carlo methods (see the description section for more information). In the case of the exact and Monte Carlo method you can set the maximum time you want to spend computing the p-value.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove the observations: Activate this option to remove observations with missing data.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the selected samples.

Results

The results displayed by XLSTAT relate to the various statistics of the tests selected and the interpretation arising from these. Results of multiple comparisons are then displayed to identify the treatments responsible for rejecting the null hypothesis, if it has been rejected.
Example

A tutorial showing how to use the Friedman’s test is available on the Addinsoft website:

http://www.xlstat.com/demo-friedman.htm

A tutorial showing how to use the Kruskal-Wallis test is available on the Addinsoft website:

http://www.xlstat.com/demo-kruskal.htm

References


**Durbin-Skillings-Mack test**

Use this tool to test if $k$ treatments being measured within a (balanced or not) incomplete block design are identical or different (GPU accelerated).

**Description**

The goal of the test proposed by Durbin (1951) is to allow analyzing rigorously the results of a study carried out within the framework of a balanced incomplete block design (BIBD), using a nonparametric procedure – thus not making any assumption on the distribution of the measurements. Skillings and Mack (1981) suggested an extension of this approach for more general incomplete block designs.

**Block designs**

A block design is a design in which we study the influence of two factors on one or more phenomena. We know that one factor has an impact that we cannot control, but that is not of interest. So we want to ensure that this factor does not disturb the analysis that we perform once the data have been collected. For this we make sure that the various levels of other factors are well represented in each block.

The blocking factor can correspond to judges evaluating products, and the factor of interest would then be the products being studied.

A complete block design is a design in which all levels of the factors of interest are present once within each block. For a sensory design, this corresponds to a design where all products are seen once by each judge.

In an incomplete block design, all levels of the factors of interest are not present for all levels of the blocking factor. It is balanced if each level of the factor of interest is present a same number of times $r$ in the design, and if each pair of levels of each factor is present the same number of times $\lambda$.

If $t$ is the number of treatments, $b$ the number of blocks, $k$ the number of treatments measured within each block, we show that the following conditions are necessary (but not sufficient) to have a balanced incomplete block design:

\[ bk = tr \]
\[ r(k-1) = \lambda(t-1) \]

**The Durbin and Skillings-Mack tests**
The Durbin and Skillings-Marck tests are an extension of the Friedman test (1937) which can only be used in the case of complete block designs.

If T1, T2, …, Tt correspond to the t treatments, as for the Friedman test, the null and alternative hypotheses used in the test are:

- H0 : The t treatments are not different.
- Ha : At least one of the treatments is different from another.

The Durbin statistic is given by

$$Q = \frac{12(t-1)}{rt(k-1)(k+1)} \sum_{j=1}^{t} \left( R_j - \frac{r(k+1)}{2} \right)^2$$

where $R_j$ is the sum over the b blocs for treatment j on ranks $R_{ij}$ of bloc i.

In the case where there are ties in one or more blocks, the variance must be corrected. We then have:

$$Q = \frac{t-1}{A-C} \left( \left( \sum_{j=1}^{t} R_j^2 \right) - rC \right)$$

with

$$A = \sum_{i=1}^{b} \sum_{j=1}^{t} R_{ij}^2$$

$$C = bk \left( k + 1 \right) / 4$$

This statistic has the property to be asymptotically distributed according to a Khi-2 law at a t-1 degree of freedom. Alvo and Cabilio (1995) propose a modified statistic with, according to Conover (1999), better asymptotic properties

$$F = \frac{Q / (t-1)}{b(k-1)-Q(b(k-1)-t+1)}$$

This statistic has the property to be asymptotically distributed according to a Fisher F at t-1 and b(k-1)-t+1 degrees of freedom.
The computation of Skillings and Mack $T$ statistic which allows to treat unbalanced incomplete block designs is more complex. The missing values are replaced by an average of the ranks, and a compensatory weight is applied to blocks with missing values. The $T$ statistic asymptotically follows a Chi-square with t-1 degree of freedom.

**Computation of the p-values**

To compute the p-values corresponding to the various statistics, XLSTAT offers two alternative methods:

- Asymptotic method: The p-value is obtained using the asymptotic approximation of the distribution of the $Q$ and $F$ statistics. The reliability of the approximation depends on the number of treatments and on the number of blocks.

- Monte Carlo method: The computation of the p-value is based on random resamplings. The user must set the number of resamplings. A confidence interval on the p-value is provided. The more resamplings are performed, the better the estimation of the p-value.

In order to avoid freezing Excel because of too long computations, it is possible with the two latter methods to set the maximum time that should be spent computing the p-value.

**Multiple pairwise comparisons**

If the p-value is such that the H0 hypothesis has to be rejected, then at least one treatment is different from another. To identify which treatments are responsible for rejecting H0, a multiple comparison procedure can be used. XLSTAT allows to use for the Durbin test the procedure suggested by Conover (1999). In the case of non balanced incomplete block designs no procedure is available, but the average ranks are displayed so that treatments can be ranked.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
Click this button to display the help.

Click this button to reload the default options.

Click this button to delete the data selections.

Click these buttons to change the way XLSTAT handles the data. If the arrow points down (column mode), XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right (row mode), XLSTAT considers that rows correspond to variables and columns to observations.

**General tab:**

**Subjects/Treatments table:** Select a table where each row (or column if in column mode) corresponds to a block, and each column (or row in row mode) corresponds to a treatment. If headers have been selected with the data, make sure the “Treatment labels” option is checked.

**Range:** Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Activate this option to display the results in a new worksheet of the active workbook.

**Workbook:** Activate this option to display the results in a new workbook.

**Treatment labels:** Activate this option if headers have been selected with the input data.

**Multiple pairwise comparisons:** Activate this option to compute multiple pairwise comparisons.

**Options tab:**

**Significance level (%)** Enter the significance level for the test (default value: 5%).

To compute the p-value, you can either choose the asymptotical approximation method or the Monte Carlo resamplings based method (see the description section). In the latter case, you can set the number of resamplings you want to make, and the maximum time you want XLSTAT to spend on making the resamplings.
Outputs tab:

**Descriptive statistics**: Activate this option to compute and display the statistics that correspond to each treatment.

Results

**Descriptive statistics**: This table displays the descriptive statistics that correspond to the k treatments.

The results that correspond to the Durbin test (in the case of balanced incomplete block design) or to the Skillings-Mack test (in the case of incomplete block designs) are then displayed, followed by a short interpretation of the test. Results of multiple comparisons are then displayed to identify the treatments responsible for rejecting the null hypothesis, if it has been rejected.

Example

An example showing how to run a Durbin test is available on the Addinsoft website:

http://www.xlstat.com/demo-durbin.htm

References


Page test

Use this tool to test if k treatments being measured within a (balanced or not) incomplete block design are identical or if a pretended sorting of the treatments is preferable (GPU accelerated).

Description

The goal of the test proposed by Page (1963) is to allow analyzing rigorously the results of a study carried out within the framework of a complete design, to verify if a series of several treatments should be considered as not different, or if alternatively a ranking of the treatment makes sense. The Page test is a nonparametric method, thus not making any assumption on the distribution of the measurements. This test differs from the Friedman test by the fact that the alternative hypothesis is a ranking of the treatments and not only a difference. This test has been extented to the case of incomplete blocks by Alvo and Cabilio (2005).

Block designs

A block design is a design in which we study the influence of two factors on one or more phenomena. We know that one factor has an impact that we cannot control, but that is not of interest. So we want to ensure that this factor does not disturb the analysis that we perform once the data are collected. For this we make sure that the various levels of other factors are well represented in each block.

The blocking factor can correspond to judges evaluating products, and the factor of interest would then be the products being studied.

A complete block design is a design in which all levels of the factors of interest are present once within each block. For a sensory design, this corresponds to a design where all products are seen once by each judge.

In an incomplete block design, all levels of the factors of interest are not present for all levels of the blocking factor. It is balanced if each level of the factor of interest is present a same number of times r in the design, and if each pair of levels of each factor is present the same number of times $\lambda$.

If t is the number of treatments, b the number of blocks, k the number of treatments measured within each block, we show that the following conditions are necessary (but not sufficient) to have a balanced incomplete block design:

$bk=tr$

$r(k-1)=\lambda(t-1)$
The Page test

If $t_1, t_2, \ldots, t_k$ correspond to the $k$ treatments, the null and alternative hypotheses used in the test are:

- $H_0$: The $k$ treatments are not significantly different.
- $H_a$: $t_1 \leq t_2 \leq \ldots \leq t_k$

or

- $H_a$: $t_1 \geq t_2 \geq \ldots \geq t_k$

Where, for the alternative hypotheses, at least one inequality is strict.

The statistic suggested by is given by:

$$L = \sum_{j=1}^{k} (jR_j)$$

Page tabulated that statistic, and also gave an asymptotic approximation with 1 degree Chi-square distribution. Conover (1999) uses the square root of the same statistic that follows a standard normal distribution:

$$z = \frac{12L - 3bk(k+1)^2}{\sqrt{bk^2(k^2-1)(k+1)}}$$

Where $b$ is the number of blocs and $k$ the number of treatments.

In the case of incomplete blocks, Alvo and Cabilio (2005) suggest an alternative statistic which value is identical in the case of complet blocks and that has the same asymptotical properties.

Computation of the p-values

To compute the p-values corresponding to the various statistics, XLSTAT offers two alternative methods:

- Asymptotic method: The p-value is obtained using the asymptotic approximation of the distribution of the $z$ statistics. The reliability of the approximation depends on the number of treatments and on the number of blocks.
- Monte Carlo method: The computation of the p-value is based on random resamplings. The user must set the number of resamplings. A confidence interval on the p-value is provided. The more resamplings are performed, the better the estimation of the p-value.

In order to avoid freezing Excel because of too long computations, it is possible with the two latter methods to set the maximum time that should be spent computing the p-value.

Multiple pairwise comparisons

If the p-value is such that the H0 hypothesis has to be rejected, then at least one treatment is different from another. To identify which treatment(s) is/are responsible for rejecting H0, a multiple comparison procedure can be used, XLSTAT allows using the procedure suggested by Cabilio and Peng (2008), with two alternative ways to compute the p-value of the paired comparisons. It can either use the normal approximation of a Monte Carlo based p-value.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- OK: Click this button to start the computations.
- Cancel: Click this button to close the dialog box without doing any computation.
- Help: Click this button to display the help.
- : Click this button to reload the default options.
- : Click this button to delete the data selections.
- : Click these buttons to change the way XLSTAT handles the data. If the arrow points down (column mode), XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right (row mode), XLSTAT considers that rows correspond to variables and columns to observations.

General tab:
Subjects/Treatments table: Select a table where each row (or column if in column mode) corresponds to a block, and each column (or row in row mode) corresponds to a treatment. If headers have been selected with the data, make sure the “Treatment labels” option is checked.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Treatment labels: Activate this option if headers have been selected with the input data.

Multiple pairwise comparisons: Activate this option to compute multiple pairwise comparisons. Two methods are proposed: Cabilio et Peng using an asymptotical p-value or the same procedure with p-values computed through Monte Carlo resamplings (see the description section for more information).

Options tab:

Significance level (%): Enter the significance level for the test (default value: 5%).

To compute the p-value, you can either choose the asymptotical approximation method or the Monte Carlo resamplings based method (see the description section). In the latter case, you can set the number of resamplings you want to make, and the maximum time you want XLSTAT to spend on making the resamplings.

Outputs tab:

Descriptive statistics: Activate this option to compute and display the statistics that correspond to each treatment.

Results

Descriptive statistics: This table displays the descriptive statistics that correspond to the k treatments.
The results that correspond to the Page test (complete block design) or to the Alvo and Cabilio variant of the test (incomplete block designs) are then displayed, followed by a short interpretation of the test. Results of multiple comparisons are then displayed to identify the treatments responsible for rejecting the null hypothesis, if it has been rejected.

Example

An example showing how to run a Page test is available on the Addinsoft website:

http://www.xlstat.com/demo-page.htm

References


Cochran's Q test

Use this tool to compare k≥2 paired samples which values are binary (GPU accelerated).

Description

The Cochran's Q test is presented using two different approaches. Some authors present it as a particular case of the Friedman's test (comparison a k paired samples) when the variable is binary (Lehmann, 1975), while other present it as a marginal homogeneity test for a k-dimensional contingency table (Agresti, 1990).

As a consequence, the null H0 and alternative Ha hypotheses for the Cochran's Q test can either be,

- H0: The k treatments are not different.
- Ha: At least on of the treatment is different from another.

or,

- H0: the k distributions are marginally homogeneous.
- Ha: the k distributions are marginally inhomogeneous.

XLSTAT uses the first approach, as it is the most used. The term “treatment” has been chosen for the k samples that are being compared.

Two possible formats are available for the input data:

- You can select data in a “raw” format. In this case, each column corresponds to a treatment and each row to a subject (or individual, or bloc).

- You can also select the data in a “grouped” format. Here, each column corresponds to a treatment, and each row corresponds to a unique combine of the k treatments. You then need to select the frequencies corresponding to each combine (field “Frequencies” in the dialog box).

Computation of the p-values

To compute the p-values corresponding to the various statistics, XLSTAT offers several alternatives:
- Asymptotic method: The p-value is obtained using the asymptotic approximation of the distribution of the statistic. The reliability of the approximation depends on the number of samples and on the number of measures per sample.

- Exact method: The computation of the p-value is based on the exact distribution of the statistic.

- Monte Carlo method: The computation of the p-value is based on random resamplings. The user must set the number of resamplings. A confidence interval on the p-value is provided. The more resamplings are performed, the better the estimation of the p-value.

In order to avoid freezing Excel because of too long computations, it is possible with the two latter methods to set the maximum time that should be spent computing the p-value.

**Multiple pairwise comparisons**

If the p-value is such that the H0 hypothesis has to be rejected, then at least one treatment is different from another. To identify which samples are responsible for rejecting H0, multiple comparison procedures can be used. XLSTAT suggests using the Marascuilo and McSweeney (1977) option, while a McNemar test with the Bonferroni correction is also available.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- ![Reload Options](image): Click this button to reload the default options.
- ![Delete Selections](image): Click this button to delete the data selections.
Click these buttons to change the way XLSTAT handles the data. If the arrow points down (column mode), XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right (row mode), XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

**Subjects/Treatments table**: Select a table where each row (or column if in column mode) corresponds to a subject, and each column (or row in row mode) corresponds to a treatment. If headers have been selected with the data, make sure the “Treatment labels” or “Labels included” is checked.

**Data format**:

- **Subjects/Treatments table**:
  - **Raw**: Choose that option if the input data are in a raw format (as opposed to grouped).
  - **Grouped**: Choose that option if your data correspond to a summary table where each row corresponds to a unique combine of treatments. You then need to select the frequencies that correspond to each combine (see “Frequencies” below).

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Treatment labels**: Activate this option if headers have been selected with the input data.

**Weights**: Select the weights that correspond to the combines of treatments. Weights must be greater than or equal to 0. If a column header has been selected, check that the “Treatment labels” option is activated.

**Multiple pairwise comparisons**: Activate this option to compute multiple pairwise comparisons. Two methods are proposed: McNemar (Bonferroni) for testing each pair with a McNemar test where the significance level is amended using the Bonferroni approach, or Marascuilo to use the procedure suggested by Marascuilo and McSweeney (1977).
Options tab:

Significance level (%): Enter the significance level for the test (default value: 5%).

Depending on the test that is being used, several methods can be available to compute the p-value. Choose among the asymptotic, exact or Monte Carlo methods (see the description section for more information). In the case of the exact and Monte Carlo method you can set the maximum time you want to spend computing the p-value.

Outputs tab:

Descriptive statistics: Activate this option to compute and display the statistics that correspond to each treatment.

Results

Descriptive statistics: This table displays the descriptive statistics that correspond to the k treatments.

The results that correspond to the Cochran's Q test are then displayed, followed by a short interpretation of the test.

Example

An example showing how to run a Cochran's Q test is available on the Addinsoft website:

http://www.xlstat.com/demo-cochranq.htm

References


McNemar’s test

Use this tool to compare 2 paired samples which values are binary. The data can be summarized in a 2x2 contingency table.

Description

The McNemar’s test is a special case of the Cochran’s Q test when there are only two treatments. As for the Cochran’s Q test, the variable of interest is binary. However, the McNemar’s test has two advantages:

- Obtaining an exact p-value is possible (Lehmann, 1975);
- The data can be summarized in a 2x2 contingency table.

In the case of the two-tailed (or two-sided) test, the null (H0) and alternative (Ha) hypotheses are:

- H0: Treatment 1 = Treatment 2
- Ha: Treatment 1 ≠ Treatment 2

In the one-tailed case, you need to distinguish the left-tailed (or lower-tailed or lower one-sided) test and the right-tailed (or upper-tailed or upper one-sided) test. In the left-tailed test, the following hypotheses are used:

- H0: Treatment 1 = Treatment 2
- Ha: Treatment 1 < Treatment 2

In the right-tailed test, the following hypotheses are used:

- H0: Treatment 1 = Treatment 2
- Ha: Treatment 1 > Treatment 2

Three possible formats are available for the input data:

- You can select data in a “raw” format. In this case, each column corresponds to a treatment and each row to a subject (or individual, or bloc).

- You can also select the data in a “grouped” format. Here, each column corresponds to a treatment, and each row corresponds to a unique combine of the k treatments. You then need
to select the frequencies corresponding to each combine (field “Frequencies” in the dialog box).

- You can also select a contingency table with two rows and two columns. In the case where you choose this, the first and second treatments are respectively considered as corresponding to the rows and the columns. The positive response cases (or successes) are considered as corresponding to the first row of the contingency table for the first treatment, and to the first column for the second treatment.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- ![Reload](reload.png): Click this button to reload the default options.
- ![Delete](delete.png): Click this button to delete the data selections.
- ![Mode](mode.png): Click these buttons to change the way XLSTAT handles the data. If the arrow points down (column mode), XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right (row mode), XLSTAT considers that rows correspond to variables and columns to observations.

**General tab:**

**Subjects/Treatments table / Contingency table (2x2):** In the case of a “Subjects/Treatments table), select a table where each row (or column if in column mode) corresponds to a subject, and each column (or row in row mode) corresponds to a treatment. In the case of a “Contingency table”, select the contingency table. If headers have been selected with the data, make sure the “Treatment labels” or “Labels included” is checked.
Data format:

- **Subjects/Treatments table**: Choose this option if the data correspond to a Subjects/Treatments table.
  - **Raw**: Choose that option if the input data are in a raw format (as opposed to grouped).
  - **Grouped**: Choose that option if your data correspond to a summary table where each row corresponds to a unique combine of treatments. You then need to select the frequencies that correspond to each combine (see “Frequencies” below).

- **Contingency table (2x2)**: Activate this option if your data are available in a 2x2 contingency table.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Treatment labels/Labels included**: Activate this option if headers have been selected with the input data. In the case of a contingency table, the row and column labels must be selected if this option is checked.

**Weights**: Select the weights that correspond to the combines of treatments. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Treatment labels" option is activated.

**Positive response code**: Enter the value that corresponds to a positive response in your experiment.

**Options** tab:

**Alternative hypothesis**: Choose the alternative hypothesis to be used for the test (see description).

**Significance level (%)**: Enter the significance level for the test (default value: 5%).

**Exact p-value**: Activate this option to compute the exact p-value.
**Outputs** tab:

This tab is only visible if the “Subjects/Treatments table” format has been chosen.

**Descriptive statistics**: Activate this option to compute and display the statistics that correspond to each treatment.

**Contingency table**: Activate this option to display the 2x2 contingency table.

**Results**

**Descriptive statistics**: This table displays the descriptive statistics that correspond to the two treatments.

**Contingency table**: The 2x2 contingency table built from the input data is displayed.

The results that correspond to the McNemar’s test are then displayed, followed by a short interpretation of the test.

**Example**

An example showing how to run a McNemar test is available on the Addinsoft website:

http://www.xlstat.com/demo-mcnemar.htm

**References**


Cochran-Mantel-Haenszel Test

Use this tool to test the hypothesis of independence on a series of contingency tables corresponding to an experiment crossing two categorical variables, with a control variable taking multiple values.

Description

Imagine the case of a laboratory working on a new antifungal agent. In order to define the appropriate dose and dosage form, an experiment is conducted with four dose levels and two different dosage forms (ointment or shower gel). For each dose level, the test is performed on about twenty patients, divided equally for each presentation. The experimenters record for each patient whether the treatment is effective or not. The results are thus in the form of a contingency table with three dimensions, or more simply in the form of 4 two-way contingency tables. The variable corresponding to the dose is the control variable.

One could be willing to do a test of independence on the table resulting from the sum of 4 contingency tables, however in this case one could conclude that there is independence for the sole reason that the sub-contingency table with the largest number of respondents corresponds a case of independence, while the other tables do not at all.

Cochran (1954) then Mantel and Haenszel (1959) developed a test that allows to test whether there is independence or not between the rows and columns of the contingency tables, taking into account the fact that the tables are independent of each other (for each dose the patients are different), and by conditioning on the marginal sums of each table, as in the standard test of independence on contingency tables.

The test commonly named the Cochran-Mantel-Haenszel (CMH) test is based on the $M^2$ statistic defined by:

$$M^2 = \left( \frac{\sum_{i=1}^{k} (n_{i1k} - n_{1+k}n_{+1k}/n_{++k})^2}{\sum_{i=1}^{k} n_{i+k}n_{2+k}n_{1+k}n_{2+k}/(n_{++k}^2(n_{++k}-1))} - \frac{1}{2} \right)^2$$

This statistic follows asymptotically a chi-square distribution with 1 degree of freedom. Knowing $M^2$, we can therefore compute the p-value, and knowing the risk of Type I, alpha, we can determine the critical value. It is also possible, as for the test of independence on a contingency table, to calculate the exact p-value, if the contingency tables are of size 2x2. The use of absolute value and the subtraction of -1 / 2 and the division by $(n_{++k}^2-1)$ instead of $n_{++k}^2$. 
corresponds to a continuity correction proposed by Mantel and Haenszel. Its use is strongly recommended. With XLSTAT you have the choice to use it (default) or not.

It may be noted that the numerator measures for the upper left cell the difference between the actual value and the expected value corresponding to independence, and that then sum these differences. If the differences are in opposite directions from one table to another we could therefore conclude that there is independence while there is dependence in each table (Type II error). This situation happens when there is a three-way interaction between the three variables. This test is to be used with caution.

The Cochran-Mantel-Haenszel test has been generalized by Birch (1965), Landis et al. (1978) and Mantel and Byar (1978) to the case of RxC contingency tables where R and C can be greater than 2. The computation of $M^2$ is more complex, but it still leads to a statistic that asymptotically follows a chi-square with $(L-1)(C-1)$ degrees of freedom.

It is recommended to perform separately from the CMH test, the analysis of the Cramer's V for the individual contingency tables to get an idea of their contribution to independence. XLSTAT displays automatically for each contingency table, a table with the Cramer's V, the chi-square and the corresponding p-values (exact for 2x2 tables and asymptotic for higher dimensional tables) where possible, that is, when there are no null marginal sums.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload default options**: Click this button to reload the default options.
- **Delete selections**: Click this button to delete the data selections.
- **Column mode**: Click these buttons to change the way XLSTAT handles the data. If the arrow points down (column mode), XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right (row mode), XLSTAT considers that rows correspond to variables and columns to observations.
General tab:

Contingency tables: If the data format selected is "contingency tables", select the k contingency tables, and then specify the value of k by entering the number of strata.

Variable 1: If the data format selected is "variables", select the data corresponding to the first qualitative variable used to construct contingency tables.

Variable 2: If the data format selected is "variables", select the data corresponding to the second qualitative variable used to construct contingency tables.

Strata: If the selected data format is "variables", select the data corresponding to the various strata.

Data format: Select the data format.

- Contingency tables: Activate this option if your data are available as a set of k contingency tables one under the other.

- Variables: Activate this option if your data are available as two qualitative variables with one row for each observation and one variable corresponding to the various strata (control variable).

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column/Row labels: Activate this option if the first row (column mode) or first column (rows mode) of the selected data contain labels.

Options tab:

Significance level (%): Enter the significance level for the test (default value: 5%).

Exact p-values: Activate this option to compute the exact p-values when possible (see the description section).
**Alternative hypothesis**: Choose the alternative hypothesis to be used for the test in the case of an exact p-value computed on a set of 2x2 tables (see the description section).

**Common odds ratio**: Enter the value of the assumed common odds-ratio.

**Continuity correction**: Activate this option if you want XLSTAT to use the continuity correction if the exact p-values calculation has not been requested or is not possible (see the description section).

**Missing data** tab:

**Do not accept missing data**: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

**Results**

The results that correspond to the Cochran-Mantel-Haenszel test are displayed, followed by a short interpretation of the test.

**Example**

A tutorial showing how to use the Cochran-Mantel-Haenszel test is available on the Addinsoft website:

[http://www.xlstat.com/demo-cmh.htm](http://www.xlstat.com/demo-cmh.htm)

**References**


(1978).

Mantel N. and Haenszel W. (1959) Statistical aspects of the analysis of data from
retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.


common odds ratio in several 2 x 2 contingency tables. Journal of the American Statistical
Association, 80, 969-973.
One-sample runs test

Use this tool to test whether a series of binary events is randomly distributed or not.

Description

The first version of this nonparametric test was presented by Mood (1940) and is based on the same runs statistic as the two-sample test by Wald and Wolfowitz (1940), which is why this test is sometimes mistakenly referred to as the Wald and Wolfowitz runs test. However, the article by Mood makes reference to the article by Wald and Wolfowitz and the asymptotic distribution of the statistic uses also the results given by these authors.

A run is a sequence of identical events, preceded and succeeded by different or no events. The runs test used here applies to binomial variables only. For example, in ABBABBB, we have 4 runs (A, BB, A, BBB).

XLSTAT accepts as input, continuous data or binary categorical data. For continuous data, a cut-point must be chosen by the user so that the data are transformed into a binary sample.

A sample will be considered as randomly distributed if no particular structure can be identified. Extreme cases are repulsion, where you have all observations of one kind on the left, and all the remaining observations on the right, and alternation where the elements of the two kinds are alternating as much as possible. With the previous case, repulsion would give “AABBBBB” or “BBBBBBAA”, and alternation “BABABB” or “BABBBB” or “BBABABB” or “BBABBAB” or “BBBBABAB”.

In the case of the two-tailed (or two-sided) test, the null (H0) and alternative (Ha) hypotheses are:

- H0: data are randomly distributed.
- Ha: data are not randomly distributed.

In the one-tailed case, you need to distinguish the left-tailed (or lower-tailed or lower one-sided) test and the right-tailed (or upper-tailed or upper one-sided) test. In the left-tailed test, the following hypotheses are used:

- H0: data are randomly distributed.
- Ha: there is repulsion between the two types of events.

In the right-tailed test, the following hypotheses are used:

- H0: data are randomly distributed.
- Ha: The two types of events are alternating.
The expectation of the number of runs R is given by:

$$E(R) = \frac{2mn}{N}$$

where m is the number of events of type 1, and n the number of events of type 2, and N is the total sample size.

The variance of the number of runs R is given by:

$$V(R) = \frac{2mn(2mn - N)}{[N^2(N-1)]}$$

The minimum value of R is always 2. The maximum value is given by $2\text{Min}(m, n) - t$, where t is 1 if $m=n$, and 0 if not.

If r is the number of runs measured on the sample, it was shown by Wald and Wolfowitz that asymptotically, when m or n tend to infinity,

$$\frac{(r - E(R))}{\sqrt{V(R)}} \rightarrow N(0,1)$$

where $N(0,1)$ is the standard normal distribution.

XLSTAT offers three ways to compute the p-values. You can compute the p-value based on:

- The exact distribution of R,
- The asymptotic distribution of R,
- An approximated distribution based on P Monte Carlo permutations. As the number of possible permutations is high (it is equal to $N!$), P must be set to a high value so that the approximation is fine.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

[OK]: Click this button to start the computations.
General tab:

Data: Select a column (or row in row mode) of data corresponding to the series of data to analyze.

Data type: Select the data type.

- **Quantitative**: Activate this option to select one column (or row in row mode) of quantitative data. The data will then be transformed on the basis of the cut point (see below).

- **Qualitative**: Activate this option to select one column (or row in row mode) of binary data.

Cut point: Choose the type of value that will be used to discretize the continuous data into a binary sample.

- **Mean**: Observations are split into two groups depending on whether there are lower or greater than the mean.

- **Median**: Observations are split into two groups depending on whether there are lower or greater than the median.

- **User defined**: Select this option to enter the value used to transform the data and enter that value. The observations are split into two groups depending on whether there are lower or greater than the given value.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.
Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column/Row labels: Activate this option if the first row (column mode) or first column (rows mode) of the selected data contain labels.

Options tab:

Alternative hypothesis: Choose the alternative hypothesis to be used for the test (see description).

Significance level (%): Enter the significance level for the test (default value: 5%).

Exact p-value: Activate this option if you want XLSTAT to calculate the exact p-value (see description).

Asymptotic p-value: Activate this option if you want XLSTAT to calculate the p-value based on the asymptotic approximation (see description).
  
  • Continuity correction: Activate this option if you want XLSTAT to use the continuity correction when computing the asymptotic p-value.

Monte Carlo method: Activate this option if you want XLSTAT to calculate the p-value based on Monte Carlo permutations, and enter the number of random permutations to perform.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove the observations: Activate this option to remove observations with missing data.

Results

The results that correspond to the one-sample runs test are displayed, followed by a short interpretation of the test.
References


Grubbs test

Use this tool to test whether one or two outliers are present in a sample for which we assume that it is extracted from a population that follows a normal distribution.

Description

Grubbs (1950, 1969, 1972) developed several tests in order to determine whether the greatest value or the lowest value (Grubbs test) are outliers, or, for the double Grubbs test, whether the two greatest values or the two lowest ones are outliers. This test assumes that the data corresponds to a sample extracted from a population that follows a normal distribution.

Detecting outliers

In statistics, an outlier is a value recorded for a given variable, that seems unusual and suspiciously lower or greater than the other observed values. One can distinguish two types of outliers:

- An outlier can simply be related to a reading error (on an measuring instrument), a keyboarding error, or a special event that disrupted the observed phenomenon to the point of making it incomparable to others. In such cases, you must either correct the outlier, if possible, or otherwise remove the observation to avoid that it disturbs the analyses that are planed (descriptive analysis, modeling, predicting).

- An outlier can also be due to an atypical event, but nevertheless known or interesting to study. For example, if we study the presence of certain bacteria in river water, you can have samples without bacteria, and other with aggregates with many bacteria. These data are of course important to keep. The models used should reflect that potential dispersion.

When there are outliers in the data, depending on the stage of the study, we must identify them, possibly with the aid of tests, flag them in the reports (in tables or on graphical representations), delete or use methods able to treat them as such.

To identify outliers, there are different approaches. For example, in classical linear regression, we can use the value of Cook's d values, or submit the standardized residuals to a Grubbs test to see if one or two values are abnormal. The classical Grubbs test can help identifying one outlier, while the double Grubbs test allows identifying two. It is not recommended to use these methods repeatedly on the same sample. However, it may be appropriate if you really suspect that there are more than two outliers.

Definitions
Let $x_1, x_2, \ldots, x_i, \ldots, x_n$, be a sample that is extracted from a population that we assume is following a normal distribution $N(\mu, \sigma^2)$. Parameters $\mu$ and $\sigma^2$ are respectively estimated by:

$$
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \\
\hat{s}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2
$$

We define:

$$
x_{\text{max}} = \arg \max_{i=1, n} (x_i)
$$

and

$$
x_{\text{min}} = \arg \min_{i=1, n} (x_i)
$$

**The Grubbs test (for one outlier)**

The statistics that used for the Grubbs test for one outlier are:

- For the one-sided left-tailed case: $G_{\text{min}} = \frac{\bar{x} - x_{\text{min}}}{s}$

- For the one-sided right-tailed case: $G_{\text{max}} = \frac{x_{\text{max}} - \bar{x}}{s}$

- For the two-sided case: $G = \max \left( G_{\text{min}}, G_{\text{max}} \right)$

For the two-sided test, the null (H0) and alternative (Ha) hypotheses are given by:

H0: The sample does not contain any outlier.

Ha: The lowest or the greatest value is an outlier.

For the one-sided left-tailed test, the null (H0) and alternative (Ha) hypotheses are given by:

H0: The sample does not contain any outlier.

Ha: The lowest value is an outlier.

For the one-sided right-tailed test, the null (H0) and alternative (Ha) hypotheses are given by:

H0: The sample does not contain any outlier.
Ha: The greatest value is an outlier.

An approximation of the critical value \(G_{\text{crit}}\) giving the threshold above which, for a given significance level \(\alpha\) (typically 5\%), one must reject the null hypothesis is given by:

\[
G_{\text{crit}}(n, \alpha) \approx \frac{(n-1)t_{n-2, 1-\alpha/k}}{\sqrt{n(n-2 + t_{n-2, 1-\alpha/k}^2)}}
\]

where \(t(n-2, 1-\alpha/k)\) is the value of the inverse of the Student cumulative distribution function at \(1-\alpha/k\) with \(n-2\) degrees of freedom, and where \(k\) equals \(n\) for one-sided tests and \(2n\) for the two-sided test. We can compare this value to the \(G\) statistic computed for the sample, and deduce that one can keep \(H_0\) if \(G_{\text{crit}}\) is greater than \(G\) (or \(G_{\min}\) or \(G_{\max}\)) and reject it otherwise. From the \(G_{\text{crit}}\) approximation we can also deduce an approximation of the p-value that corresponds to \(G\). XLSTAT displays all these results as well as the conclusion based on the significance level given by the user.

Double Grubbs test

For this test, we first sort up the \(x_i\) observations. The statistics used for the double Grubbs test are given by:

- One-sided left-tailed test: \(G_{2, \min} = \frac{Q_{\min}}{(n-1)s}\)

  With \(Q_{\min} = \sum_{i=3}^{n} (x_i - \overline{x}_3)^2\), \(\overline{x}_3 = \frac{1}{n-2} \sum_{i=3}^{n} x_i\)

- One-sided right-tailed test: \(G_{2, \max} = \frac{Q_{\max}}{(n-1)s}\)

  With \(Q_{\max} = \sum_{i=1}^{n-2} (x_i - \overline{x}_{n-2})^2\), \(\overline{x}_{n-2} = \frac{1}{n-2} \sum_{i=1}^{n-2} x_i\)

- Two-sided test: \(G_{2, \min, \max} = \text{Max}(G_{2, \min}, G_{2, \max})\)

For the two-sided test, the null (\(H_0\)) and alternative (\(H_a\)) hypotheses are given by:

\(H_0\): The sample does not contain any outlier.
Ha: The two lowest or two greatest values are outliers.

For the one-sided left-tailed test, the null (H0) and alternative (Ha) hypotheses are given by:

H0: The sample does not contain any outlier.
Ha: The two lowest values are outliers.

For the one-sided right-tailed test, the null (H0) and alternative (Ha) hypotheses are given by:

H0: The sample does not contain any outlier.
Ha: The two greatest values are outliers.

Wilrich (2013) gives an approximation of the $G_{2 \text{crit}}$ critical value above which one should reject $H_0$ for a given significance level $\alpha$. However XLSTAT gives an approximation based on Monte Carlo simulations. The default number of simulations is set to 1000000, which allows to obtain an accuracy that is higher than the ones available in original papers of Grubbs, and sufficient for any operational problem. Using the same set of simulations, XLSTAT gives the p-value that corresponds to the computed $G_2$ statistic as well as the conclusion of the test taking into account the significance level given by the user.

**Z-scores**

Z-scores are displayed by XLSTAT to help you identify potential outliers. Z-scores correspond to the standardised sample:

$$z_i = \frac{x_i - \bar{x}}{s} \ (i=1,...,n)$$

The problem with these scores is that once the acceptance interval is set (typically -1.96 and 1.96 for a 95% interval), any value that is outside is considered suspicious. However we know if we have 100 values, it is statistically normal to have 5 outside this interval. Furthermore, one can show that for a given $n$, the highest z-score is at most given by:

$$\arg \max_{i=1,...,n} z_i \leq \frac{n-1}{\sqrt{n}}$$

Iglewicz and Hoaglin (1993) recommend using a modified z-score in order to better identify outliers:

$$z_i = 0.6745 \frac{x_i - \bar{x}}{MAD} \ (i=1,...,n)$$

where the MAD is the *Median Absolute Deviation*. The acceptant interval is given by $]-3.5 ; 3.5[$ whatever $n$. 

677
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **_reload**: Click this button to reload the default options.
- **delete**: Click this button to delete the data selections.
- **

: Click these buttons to change the way XLSTAT handles the data. If the arrow points down (column mode), XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right (row mode), XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

**Data**: Select the data on the Excel sheet. If you select several columns, XLSTAT considers column (or row in row mode) corresponds to a sample. If headers have been selected with the data, make sure the “Column labels” option is checked.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Column/Row labels**: Activate this option if the first row (column mode) or first column (rows mode) of the selected data contain labels.
You can choose the test to apply on your data:

- **Grubbs test**: Select this test to run a Grubbs test to identify one outlier.
- **Double Grubbs test**: Select this test to run a Grubbs test to identify two outliers.

**Options** tab:

**Alternative hypothesis**: Choose the alternative hypothesis to be used for the test (see description).

**Significance level (%)**: Enter the significance level for the test (default value: 5%).

**Iterations**: Choose whether you want to apply the selected test data a limited number of times (default is 1), or if you want to let XLSTAT iterate until no more outlier is found.

**Critical value / p-value**: Enter the number of Monte Carlo simulations to perform to compute the critical value and the p-value. This option is only available for the double Grubbs test.

**Missing data** tab:

**Do not accept missing data**: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

**Remove the observations**: Activate this option to remove observations with missing data.

**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the selected samples.

**Z-scores**: Activate this option to calculate and display the z-scores and the corresponding graph. You can choose between the **modified z-scores** or standard **z-scores**. For z-scores you can choose which limits to display on the charts.
Results

**Descriptive statistics**: This table displays the descriptive statistics that correspond to the k samples.

The results correspond to the **Grubbs test** are then displayed. An interpretation of the test is provided if a single iteration of the test was requested, or if no observation was identified as being an outlier.

In case several iterations were required, also display a table showing, for each observation, the iteration in which it was removed from the sample.

The z-scores are then displayed if they have been requested.

Example

A tutorial showing how to use the Grubbs test is available on the Addinsoft website:


References


**Iglewicz B. and Hoaglin D. (1993)**. "Volume 16: How to Detect and Handle Outliers", The ASQC Basic References in Quality Control: Statistical Techniques, Edward F. Mykytka, Ph.D., Editor.


Dixon test

Use this tool to test whether one or two outliers are present in a sample for which we assume that it is extracted from a population that follows a normal distribution.

Description

The Dixon test (1950, 1951, 1953), which is actually divided into six tests depending on the chosen statistic and on the number of outliers to identify, was developed to help determine if the greatest value or lowest value of a sample, or the two largest values, or the two smallest ones can be considered as outliers. This test assumes that the data correspond to a sample extracted from a population that follows a normal distribution.

Detecting outliers

In statistics, an outlier is a value recorded for a given variable, that seems unusual and suspiciously lower or greater than the other observed values. One can distinguish two types of outliers:

- An outlier can simply be related to a reading error (on an measuring instrument), a keyboarding error, or a special event that disrupted the observed phenomenon to the point of making it incomparable to others. In such cases, you must either correct the outlier, if possible, or otherwise remove the observation to avoid that it disturbs the analyses that are planed (descriptive analysis, modeling, predicting).

- An outlier can also be due to an atypical event, but nevertheless known or interesting to study. For example, if we study the presence of certain bacteria in river water, you can have samples without bacteria, and other with aggregates with many bacteria. These data are of course important to keep. The models used should reflect that potential dispersion.

When there are outliers in the data, depending on the stage of the study, we must identify them, possibly with the aid of tests, flag them in the reports (in tables or on graphical representations), delete or use methods able to treat them as such.

To identify outliers, there are different approaches. For example, in classical linear regression, we can use the value of Cook’s d values, or submit the standardized residuals to a Grubbs test to see if one or two values are abnormal. The classical Grubbs test can help identifying one outlier, while the double Grubbs test allows identifying two. It is not recommended to use these methods repeatedly on the same sample. However, it may be appropriate if you really suspect that there are more than two outliers.
Definitions

Let \( x_1, x_2, \ldots, x_i, \ldots, x_n \), be a sample that is extracted from a population that we assume is following a normal distribution \( N(\mu, \sigma^2) \). Parameters \( \mu \) and \( \sigma^2 \) are respectively estimated by:

We assume that the \( x_i \) are sorted up.

Dixon test for one outlier

This test is used to determine if the largest or smallest value can be considered as being an outlier. This test assumes that the data corresponds to a sample coming from a population that follows a normal distribution.

The statistics used for the Dixon test and the corresponding ranges of number of observations they should be used with (Barnett and Lewis 1994 and Verma and Quiroz-Ruiz 2006) are:

- \( R_{10} = \frac{x_n - x_{n-1}}{x_n - x_1} \), recommended for \( 3 \leq n \leq 100 \), also named N7
- \( R_{11} = \frac{x_n - x_{n-1}}{x_n - x_2} \), recommended for \( 4 \leq n \leq 100 \), also named N9
- \( R_{12} = \frac{x_n - x_{n-1}}{x_n - x_3} \), recommended for \( 5 \leq n \leq 100 \), also named N10

These statistics are valid for testing whether the maximum value is an outlier. To identify if the minimum value is an outlier, simply sort the data in descending order and use the same statistics. If we want to identify if the minimum or the maximum value is an outlier, we calculate the statistics for the two alternatives (sort ascending or descending) and keep the largest value for the statistic.

For the two-sided test, the null (H0) and alternative (Ha) hypotheses are given by:

- H0: The sample does not contain any outlier.
- Ha: The lowest or the greatest value is an outlier.

For the one-sided left-tailed test, the null (H0) and alternative (Ha) hypotheses are given by:

- H0: The sample does not contain any outlier.
- Ha: The lowest value is an outlier.

For the one-sided right-tailed test, the null (H0) and alternative (Ha) hypotheses are given by:
H0: The sample does not contain any outlier.

Ha: The greatest value is an outlier.

**Dixon test for two outliers**

This test is used to determine if the two largest or two smallest values can be considered as being an outlier. This test assumes that the data corresponds to a sample coming from a population that follows a normal distribution.

The statistics used for the Dixon test and the corresponding ranges of number of observations they should be used with (Barnett and Lewis 1994 and Verma and Quiroz-Ruiz 2006) are:

- \( R_{21} = \frac{x_n - x_{n-2}}{x_n - x_2} \), recommended for \( 4 \leq n \leq 100 \), also named N11
- \( R_{21} = \frac{x_n - x_{n-2}}{x_n - x_2} \), recommended for \( 5 \leq n \leq 100 \), also named N12
- \( R_{22} = \frac{x_n - x_{n-2}}{x_n - x_3} \), recommended for \( 6 \leq n \leq 100 \), also named N13

These statistics are valid for testing whether the maximum value is an outlier. To identify if the minimum value is an outlier, simply sort the data in descending order and use the same statistics. If we want to identify if the minimum or the maximum value is an outlier, we calculate the statistics for the two alternatives (sort ascending or descending) and keep the largest value for the statistic.

For the two-sided test, the null (H0) and alternative (Ha) hypotheses are given by:

H0: The sample does not contain any outlier.

Ha: The two lowest or two greatest values are outliers.

For the one-sided left-tailed test, the null (H0) and alternative (Ha) hypotheses are given by:

H0: The sample does not contain any outlier.

Ha: The two lowest values are outliers.

For the one-sided right-tailed test, the null (H0) and alternative (Ha) hypotheses are given by:

H0: The sample does not contain any outlier.
Ha: The two greatest values are outliers.

**Critical value and p-value for the Dixon test**

Literature provides more or less accurate approximations of the critical value beyond which, for a given significance level $\alpha$, we cannot keep the null hypothesis. However XLSTAT provides an approximation of the critical values based on Monte Carlo simulations. The number of these approximations is by default set to 1000000, which provides more reliable than those provided in the historical articles. XLSTAT also provides on the basis of these simulations, a p-value and the conclusion of the test based on the significance level chosen by the user.

**Z-scores**

Z-scores are displayed by XLSTAT to help you identify potential outliers. Z-scores correspond to the standardised sample:

$$z_i = \frac{x_i - \bar{x}}{s} \quad (i=1,\ldots,n)$$

The problem with these scores is that once the acceptance interval is set (typically -1.96 and 1.96 for a 95% interval), any value that is outside is considered suspicious. However we know if we have 100 values, it is statistically normal to have 5 outside this interval. Furthermore, one can show that for a given $n$, the highest z-score is at most given by:

$$\arg \max_{i=1\ldots n} z_i \leq \frac{n - 1}{\sqrt{n}}$$

Iglewicz and Hoaglin (1993) recommend using a modified z-score in order to better identify outliers:

$$z_i = 0.6745 \frac{x_i - \bar{x}}{MAD} \quad (i=1,\ldots,n)$$

where the MAD is the Median Absolute Deviation. The acceptant interval is given by $]-3.5 ; 3.5[$ whatever $n$.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.
OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

Reload: Click this button to reload the default options.

Delete: Click this button to delete the data selections.

Click these buttons to change the way XLSTAT handles the data. If the arrow points down (column mode), XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right (row mode), XLSTAT considers that rows correspond to variables and columns to observations.

**General tab:**

**Data:** Select the data on the Excel sheet. If you select several columns, XLSTAT considers column (or row in row mode) corresponds to a sample. If headers have been selected with the data, make sure the “Column labels” option is checked.

**Range:** Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Activate this option to display the results in a new worksheet of the active workbook.

**Workbook:** Activate this option to display the results in a new workbook.

**Column/Row labels:** Activate this option if the first row (column mode) or first column (rows mode) of the selected data contain labels.

You can choose the test to apply on your data:

- **User defined:** Choose this option to be able to select the statistic you want to use to identify outliers.

- **Automatic:** Choose this option to let XLSTAT choose the appropriate statistic, based on what is recommended in the literature (Böhrer, 2008).
**Options** tab:

**Alternative hypothesis**: Choose the alternative hypothesis to be used for the test (see description).

**Significance level (%)**: Enter the significance level for the test (default value: 5%).

**Iterations**: Choose whether you want to apply the selected test data a limited number of times (default is 1), or if you want to let XLSTAT iterate until no more outlier is found.

**Critical value / p-value**: Enter the number of Monte Carlo simulations to perform to compute the critical value and the p-value.

**Missing data** tab:

**Do not accept missing data**: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

**Remove the observations**: Activate this option to remove observations with missing data.

**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the selected samples.

**Z-scores**: Activate this option to calculate and display the z-scores and the corresponding graph. You can choose between the modified z-scores or standard z-scores. For z-scores you can choose which limits to display on the charts.

**Results**

**Descriptive statistics**: This table displays the descriptive statistics that correspond to the k samples.

The results correspond to the Dixon test are then displayed. An interpretation of the test is provided if a single iteration of the test was requested, or if no observation was identified as being an outlier.

In case several iterations were required, also display a table showing, for each observation, the iteration in which it was removed from the sample.
The z-scores are then displayed if they have been requested.

**Example**

A tutorial showing how to use the Dixon test is available on the Addinsoft website:

http://www.xlstat.com/demo-dixon.htm

**References**


Cochran’s C test

Use this tool to test whether there is an outlying variance among a series of k variances.

Description

The Cochran’s C test (Cochran 1941) is one of the tests developed to identify and study the homogeneity of a series of variances (Bartlett's test, Brown-Forsythe, Levene or Hartley in particular). Cochran's test was developed to answer a specific question: Are the variances homogeneous or is the highest variance different from others. XLSTAT also offers two alternatives and uses the results of 't Lam (2010) for an extension of the balanced case to unbalanced cases.

Detecting outliers

In statistics, an outlier is a value recorded for a given variable, that seems unusual and suspiciously lower or greater than the other observed values. One can distinguish two types of outliers:

- An outlier can simply be related to a reading error (on an measuring instrument), a keyboarding error, or a special event that disrupted the observed phenomenon to the point of making it incomparable to others. In such cases, you must either correct the outlier, if possible, or otherwise remove the observation to avoid that it disturbs the analyses that are planned (descriptive analysis, modeling, predicting).

- An outlier can also be due to an atypical event, but nevertheless known or interesting to study. For example, if we study the presence of certain bacteria in river water, you can have samples without bacteria, and other with aggregates with many bacteria. These data are of course important to keep. The models used should reflect that potential dispersion.

When there are outliers in the data, depending on the stage of the study, we must identify them, possibly with the aid of tests, flag them in the reports (in tables or on graphical representations), delete or use methods able to treat them as such.

To identify outliers, there are different approaches. For example, in classical linear regression, we can use the value of Cook’s d values, or submit the standardized residuals to a Grubbs test to see if one or two values are abnormal. The classical Grubbs test can help identifying one outlier, while the double Grubbs test allows identifying two. It is not recommended to use these methods repeatedly on the same sample. However, it may be appropriate if you really suspect that there are more than two outliers.
If the sample can be divided into sub-samples, we can look for changes from a sub-sample to another. The test Cochran’s C test and the Mandel’s h and k statistics are part of the methods suitable for such studies.

**Definitions**

Let \( x_{11}, x_{12}, \ldots, x_{1n1}, x_{22}, \ldots, x_{2n2}, \ldots, x_{p1}, x_{p2}, \ldots, x_{pnp} \) be a sample of that we distinguish for their belonging to \( p \) groups (for example laboratories) of respective size \( n_i \) (\( i=1\ldots p \)). Let \( \bar{x}_i \) be the estimated mean for the \( i \) group, and let \( s_i^2 \) be the group \( i \) variance. We have:

\[
\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}
\]

\[
s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2
\]

It is assumed that the observations are identically distributed and follow a normal distribution.

**Cochran’s C test**

The \( C_i \) statistic corresponding to group (or sub-sample) \( i \) (\( i=1\ldots p \)) given by Cochran (1941) writes:

\[
C_i = \frac{s_i^2}{\sum_{j=1}^{k} s_j^2}
\]

and

\[
C = \underset{i=1\ldots k}{\text{arg max}} \left( C_i \right)
\]

is the statistic used for the test. The critical value corresponding to that statistic has abundantly been tabulated and various authors have given approximations (Wilrich, 2013). However, as noticed by ‘t Lam (2010), this statistic has several drawbacks:

- This test requires that the groups have identical sizes (balanced design),
- Only the maximum variance is being studied, the minimum variance is ignored even if it is the true outlier (right-tailed test only),
- Tables of critical values are limited and sometimes contain errors,
- The use of tables is not convenient.
For that reason, 't Lam proposes a generalization of the Cochran statistic for unbalanced designs, and a generalization test where the alternative hypothesis may be one or two-sided. The statistics for the group $i$ is given by:

$$G_i = \frac{V_iS_i^2}{\sum_{j=1}^{p}V_jS_j^2} \quad \text{with} \quad V_j = n_j - 1$$

For a significance level $\alpha$, 't Lam gives the lower and upper critical values for this statistic:

$$G_{LL}(i) = \left[ 1 + \frac{\left( \frac{v_{total}}{v_i} \right) - 1}{F^{-1}(1 - \delta / k, v_i, v_{total} - v_i)} \right]^{-1} \quad (1)$$

$$G_{UL}(i) = \left[ 1 + \frac{\left( \frac{v_{total}}{v_i} \right) - 1}{F^{-1}(\delta / k, v_i, v_{total} - v_i)} \right]^{-1} \quad (2)$$

with $v_{total} = \left( \sum_{i=1}^{p}v_i \right) - p$, and $\delta = \alpha$ for a one-sided test and $\delta = \alpha/2$ for a two-sided test, and $F^{-1}$ is the inverse Fisher cumulative distribution function.

For the two-sided test, the null (H0) and alternative (Ha) hypotheses are given by:

- H0: The variances are homogeneous.
- Ha: One of the variances is lower than the others.

For the one-sided left-tailed test, the null (H0) and alternative (Ha) hypotheses are given by:

- H0: The variances are homogeneous.
- Ha: At least one of the variances is lower than the others.

For the one-sided right-tailed test, the null (H0) and alternative (Ha) hypotheses are given by:

- H0: The variances are homogeneous.
- Ha: At least one of the variances is greater than the others.

Under a two-sided test, to identify the potentially outlying variance we compute:

$$G_{min} = \arg \min_{i=1 \ldots k} (G_i) \quad \text{and} \quad G_{max} = \arg \max_{i=1 \ldots k} (G_i)$$
Then, if one or two statistics are not within the critical range given by (1) and (2), the p-values associated with the two statistics are calculated. We identify the abnormal variance as the one that corresponds to the lowest p-value.

**Z-scores**

Z-scores are displayed by XLSTAT to help you identify potential outliers. Z-scores correspond to the standardised sample:

\[
z_i = \frac{x_i - \bar{x}}{s} \quad (i=1,...,n)
\]

The problem with these scores is that once the acceptance interval is set (typically -1.96 and 1.96 for a 95% interval), any value that is outside is considered suspicious. However we know if we have 100 values, it is statistically normal to have 5 outside this interval. Furthermore, one can show that for a given n, the highest z-score is at most given by:

\[
\arg \max_{i=1,...,n} z_i \leq \frac{n - 1}{\sqrt{n}}
\]

Iglewicz and Hoaglin (1993) recommend using a modified z-score in order to better identify outliers:

\[
z_i = 0.6745 \frac{x_i - \bar{x}}{MAD} \quad (i=1,...,n)
\]

where the MAD is the *Median Absolute Deviation*. The acceptant interval is given by \([-3.5; 3.5]\) whatever n.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
Click this button to reload the default options.

Click this button to delete the data selections.

Click these buttons to change the way XLSTAT handles the data. If the arrow points down (column mode), XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right (row mode), XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

**Data:** If the format of the selected data is "one column per variable", select the data for the various samples in the Excel worksheet. If the format of the selected data is "one column per group", select the columns of data corresponding to the various groups.

**Group identifiers / Group size:** If the format of the selected data is "one column per variable", select the data identifying the groups to which the selected data values correspond. If the format of the selected data is “Variances” you need to enter the group size (balanced design) or select the group sizes (unbalanced design).

**Data format:** Select the data format.

- **One column/row per group:** Activate this option to select one column (or row in row mode) per group.

- **One column/row per variable:** Activate this option for XLSTAT to carry out as many tests as there are columns/rows, given that each column/row must contain the same number of rows/columns and that a sample identifier which enables each observation to be assigned to a sample must also be selected.

- **Variances:** Activate this option if your data correspond to variances. In that case you need to define the sample size (balanced design) or select the sample sizes (unbalanced design).

**Range:** Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Activate this option to display the results in a new worksheet of the active workbook.

**Workbook:** Activate this option to display the results in a new workbook.
Column/Row labels: Activate this option if the first row (column mode) or first column (rows mode) of the selected data contain labels.

You can choose the test to apply on your data:

- **Cochran’s C (balanced)**: choose this option if the design is balanced.
- **‘t Lam’s G (unbalanced)**: choose this option if the design is unbalanced.

**Options** tab:

**Alternative hypothesis**: Choose the alternative hypothesis to be used for the test (see description).

**Significance level (%)**: Enter the significance level for the test (default value: 5%).

**Iterations**: Choose whether you want to apply the selected test data a limited number of times (default is 1), or if you want to let XLSTAT iterate until no more outlier is found.

**Missing data** tab:

**Do not accept missing data**: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

**Remove the observations**: Activate this option to remove observations with missing data.

**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the selected samples.

**Z-scores**: Activate this option to calculate and display the z-scores and the corresponding graph. You can choose between the modified z-scores or standard z-scores. For z-scores you can choose which limits to display on the charts.

**Results**

**Descriptive statistics**: This table displays the descriptive statistics that correspond to the groups.
The results correspond to the **Cochran’s C test** are then displayed. An interpretation of the test is provided if a single iteration of the test was requested, or if no observation was identified as being an outlier.

In case several iterations were required, also display a table showing, for each observation, the iteration in which it was removed from the sample.

The z-scores are then displayed if they have been requested.

**Example**

A tutorial showing how to use the Cochran test is available on the Addinsoft website:

http://www.xlstat.com/demo-cochran.htm

**References**


**Wilrich P.-T. (2013).** Critical values of Mandel’s h and k, the Grubbs and the Cochran test statistic. *Advances in Statistical Analysis*, **97**(1), 1-10.
Mandel’s h and k statistics

Use this tool to calculate the h and k Mandel’s statistics to identify potential outliers in a sample.

Description

The Mandel’s h and k statistics (1985, 1991) have been developed to help identifying outliers during inter-laboratories studies.

Detecting outliers

In statistics, an outlier is a value recorded for a given variable, that seems unusual and suspiciously lower or greater than the other observed values. One can distinguish two types of outliers:

- An outlier can simply be related to a reading error (on an measuring instrument), a keyboarding error, or a special event that disrupted the observed phenomenon to the point of making it incomparable to others. In such cases, you must either correct the outlier, if possible, or otherwise remove the observation to avoid that it disturbs the analyses that are planed (descriptive analysis, modeling, predicting).

- An outlier can also be due to an atypical event, but nevertheless known or interesting to study. For example, if we study the presence of certain bacteria in river water, you can have samples without bacteria, and other with aggregates with many bacteria. These data are of course important to keep. The models used should reflect that potential dispersion.

When there are outliers in the data, depending on the stage of the study, we must identify them, possibly with the aid of tests, flag them in the reports (in tables or on graphical representations), delete or use methods able to treat them as such.

To identify outliers, there are different approaches. For example, in classical linear regression, we can use the value of Cook’s d values, or submit the standardized residuals to a Grubbs test to see if one or two values are abnormal. The classical Grubbs test can help identifying one outlier, while the double Grubbs test allows identifying two. It is not recommended to use these methods repeatedly on the same sample. However, it may be appropriate if you really suspect that there are more than two outliers.

If the sample can be divided into sub-samples, we can look for changes from a sub-sample to another. The test Cochran’s C test and the Mandel’s h and k statistics are part of the methods suitable for such studies.
**Definitions**

Let \( x_{11}, x_{12}, \ldots, x_{1n_1}, x_{21}, x_{22}, \ldots, x_{2n_2}, \ldots, x_{p1}, x_{p2}, \ldots, x_{pn_p} \) be a sample of that we distinguish for their belonging to \( p \) groups (for example laboratories) of respective size \( n_i \) (\( i=1 \ldots p \)). Let \( \bar{x}_i \) be the estimated mean for the \( i \) group, and let \( s_i^2 \) be the group \( i \) variance. We have:

\[
\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}
\]

\[
s_i^2 = \frac{1}{n_i-1} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2
\]

It is assumed that the observations are identically distributed and follow a normal distribution.

**Mandel’s h statistic**

Mandel’s \( h_i \) for group \( i (i=1 \ldots p) \) is given by:

\[
h_i = \frac{\bar{x}_i - \bar{x}}{s}
\]

with

\[
\bar{x} = \frac{1}{p} \sum_{j=1}^{p} \bar{x}_j \quad \text{and} \quad s = \frac{1}{p-1} \sum_{j=1}^{p} (\bar{x}_j - \bar{x})^2
\]

XLSTAT provides \( h_i \) statistics for each group. To identify groups for which the mean is potentially abnormal, we can calculate the critical values and confidence intervals for a given level of significance \( \alpha \) around statistic \( h \) (Wilrich, 2013). The critical value is given by:

\[
h_{\text{crit}}(p, \alpha) = \frac{(p-1)t_{p-2,1-\alpha/2}}{\sqrt{p(p-2+t_{p-2,1-\alpha/2}^2)}}
\]

where \( t \) corresponds to the quantile of the Student distribution for \( 1-\alpha/2 \) and \( p-2 \) degree of freedom.

The confidence interval (two-sided) of size \( 100(1-\alpha)% \) around \( h_i \) is given by \(-h_{\text{crit}}; h_{\text{crit}}\].

XLSTAT displays the critical value on the chart of the \( h_i \) if the \( n_i \) are constant.

**Mandel’s k statistic**

Mandel’s \( k_i \) for group \( i (i=1 \ldots p) \) is given by:
\[ k_i = \frac{s_i}{\bar{s}} \]

with

\[ s_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 \quad \text{and} \quad \bar{s} = \sqrt{\frac{1}{p} \sum_{i=1}^{p} s_i^2} \]

XLSTAT provides \( k_i \) statistics for each group. To identify groups for which the variance is potentially abnormal, we can calculate the critical values and confidence intervals for a given level of significance \( \alpha \) around statistic \( h \) (Wilrich, 2013). The critical value is given by:

\[ k_{crit} (n, \alpha) = \sqrt{p \left( 1 + \left( p - 1 \right) F^{-1}_{1-\alpha,(p-1),(n-1)} \right)} \]

where \( F^{-1}(1-\alpha,v_1,v_2) \) is the value of the inverse cumulative distribution function of the Fisher distribution for probability \( 1-\alpha \) with \( v_1 \) and \( v_2 \) degrees of freedom.

The confidence interval (one-sided) of size \( 100(1-\alpha)\% \) around \( k_i \) is given by \([0 ; k_i,crit]\). XLSTAT displays the critical value on the chart of the \( h_i \) if the \( n_i \) are constant.

**Z-scores**

Z-scores are displayed by XLSTAT to help you identify potential outliers. Z-scores correspond to the standardised sample:

\[ z_i = \frac{x_i - \bar{x}}{s} \quad (i=1,...,n) \]

The problem with these scores is that once the acceptance interval is set (typically -1.96 and 1.96 for a 95% interval), any value that is outside is considered suspicious. However we know if we have 100 values, it is statistically normal to have 5 outside this interval. Furthermore, one can show that for a given \( n \), the highest z-score is at most given by:

\[ \arg \max_{i=1,...,n} z_i \leq \frac{n-1}{\sqrt{n}} \]

Iglewicz and Hoaglin (1993) recommend using a modified z-score in order to better identify outliers:

\[ z_i = 0.6745 \frac{x_i - \bar{x}}{MAD} \quad (i=1,...,n) \]

where the MAD is the Median Absolute Deviation. The acceptant interval is given by \([-3.5 ; 3.5]\) whatever \( n \).
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- OK: Click this button to start the computations.
- Cancel: Click this button to close the dialog box without doing any computation.
- Help: Click this button to display the help.
- Reload default options: Click this button to reload the default options.
- Delete data selections: Click this button to delete the data selections.
- Data format: Click these buttons to change the way XLSTAT handles the data. If the arrow points down (column mode), XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right (row mode), XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Data: If the format of the selected data is "one column per variable", select the data for the various samples in the Excel worksheet. If the format of the selected data is "one column per group", select the columns of data corresponding to the various groups.

Group identifiers / Group size: If the format of the selected data is "one column per variable", select the data identifying the groups to which the selected data values correspond. If the format of the selected data is "Variances" you need to enter the group size (balanced design) or select the group sizes (unbalanced design).

Data format: Select the data format.

- One column/row per group: Activate this option to select one column (or row in row mode) per group.
- **One column/row per variable**: Activate this option for XLSTAT to carry out as many tests as there are columns/rows, given that each column/row must contain the same number of rows/columns and that a sample identifier which enables each observation to be assigned to a sample must also be selected.

- **Variances**: Activate this option if your data correspond to variances. In that case you need to define the sample size (balanced design) or select the sample sizes (unbalanced design).

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Column/Row labels**: Activate this option if the first row (column mode) or first column (rows mode) of the selected data contain labels.

You can choose the test to apply on your data:

- **Mandel's h statistic**: choisissez cet option pour calculer la statistique h de Mandel.

- **Mandel's k statistic**: choisissez cet option pour calculer la statistique h de Mandel.

**Options** tab:

**Significance level (%)**: Enter the significance level for the test (default value: 5%).

**Missing data** tab:

**Do not accept missing data**: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

**Remove the observations**: Activate this option to remove observations with missing data.

**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the selected samples.
**Z-scores:** Activate this option to calculate and display the z-scores and the corresponding graph. You can choose between the **modified z-scores** or standard **z-scores.** For z-scores you can choose which limits to display on the charts.

**Results**

**Descriptive statistics:** This table displays the descriptive statistics that correspond to the groups.

The results correspond to the **Mandel's statistics** are then displayed.

The z-scores are then displayed if they have been requested.

**Example**

A tutorial showing how to compute the Mandel's statistics is available on the Addinsoft website:

http://www.xlstat.com/demo-mandel.htm

**References**


DataFlagger

Use DataFlagger to show up the values within or outside a given interval, or which are equal to certain values.

Dialog box

- **OK**: Click this button to start flagging the data.
- **Cancel**: Click this button to close the dialog box without doing any change.
- **Help**: Click this button to display the help.
- ![Reload](image): Click this button to reload the default options.
- ![Delete](image): Click this button to delete the data selections.
- ![Change](image): Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

Data: Select the data in the Excel worksheet.

Flag a value or a text: Activate this option if you want to identify or show up a value or a series of values in the selected range.

- **Value or text**: Choose this option to find and flag a single value or a character string.
- **List values or texts**: Choose this option to find and flag a series of values or texts. You must then select the series of values or texts in question in an Excel worksheet.

Flag an interval: Activate this option if you want to identify or show up values within or outside an interval. You then have to define the interval.

- **Inside**: Choose this option to find and flag values within an interval. Afterwards choose the boundary types (open or closed) for the interval, then enter the values of the boundaries.
• **Outside**: Choose this option to find and flag values outside an interval. Afterwards choose the boundary types (open or closed) for the interval, then enter the values of the boundaries.

**Font**: Use the following options to change the font of the values obeying the flagging rules.

- **Style**: Choose the font style
- **Size**: Choose the font size
- **Color**: Choose the font color

**Cell**: Use the following option to change the background color of the cell.

- **Color**: Choose the cell color
**Min/Max Search**

Use this tool to locate the minimum and/or maximum values in a range of values. If the minimum value is encountered several times, XLSTAT makes a multiple selection of the minimum values enabling you afterwards to browse between them simply using the "Enter" key.

**Dialog box**

- **OK**: Click this button to start the search.
- **Cancel**: Click this button to close the dialog box without doing any search.
- **Help**: Click this button to display the help.

**Data**: Select the data in the Excel worksheet.

**Find the minimum**: Activate this option to make XLSTAT look for the minimum value(s) in the selection. If the "Multiple selection" option is activated and several minimum values are found, they will all be selected and you can navigate between them using the "Enter" key.

**Find the maximum**: Activate this option to make XLSTAT look for the maximum value(s) in the selection. If the "Multiple selection" option is activated and several maximum values are found, they will all be selected and you can navigate between them using the "Enter" key.

**Multiple selection**: Activate this option to enable multiple occurrences of the minimum and/or maximum values to be selected at the same time.
Remove text values in a selection

Use this tool to remove text values in a data set that is expected to contain only numerical data. This tool is useful if you are importing data from a format that generates empty cells with a text format in Excel.

Dialog box

- **OK**: Click this button to start removing the text values.
- **Cancel**: Click this button to close the dialog box without doing any change.
- **Help**: Click this button to display the help.
- [Revert](#): Click this button to reload the default options.
- [Delete](#): Click this button to delete the data selections.

**Data**: Select the data in the Excel worksheet.

**Clean only the cells with empty strings**: Activate this option if you want to only clean the cells that correspond to empty strings.
Sheets management

Use this tool to manage the sheets contained in the open Excel workbooks.

Dialog box

When you start this tool, it displays a dialog box that lists all the sheets contained in all the workbooks, whether they are hidden or not.

Activate: Click this button to go to the first sheet that is selected.

Unhide: Click this button to unhide all the selected sheets.

Hide: Click this button to hide all the selected sheets.

Delete: Click this button to delete all the selected sheets. Warning: deleting hidden sheets is irreversible.

Cancel: Click this button to close the dialog box.

Help: Click this button to display help.
Delete hidden sheets

Use this tool to delete the hidden sheets generated by XLSTAT or other applications. XLSTAT generates hidden sheets to create certain charts. This tool is used to choose which hidden sheets are to be deleted and which kept.

Dialog box

Hidden sheets: The list of hidden sheets is displayed. Select the hidden sheets you want to delete.

All: Click this button to select all the sheets in the list.

None: Click this button to deselect all the sheets in the list.

Delete: Click this button to delete all the selected sheets. Warning: deleting hidden sheets is irreversible.

Cancel: Click this button to close the dialog box.

Help: Click this button to display help.
Unhide hidden sheets

Use this tool to unhide the hidden sheets generated by XLSTAT or other applications. XLSTAT generates hidden sheets to create certain charts.

Dialog box

**Hidden sheets**: The list of hidden sheets is displayed. Select the hidden sheets you want to unhide.

**All**: Click this button to select all the sheets in the list.

**None**: Click this button to deselect all the sheets in the list.

**Unhide**: Click this button to unhide all the selected sheets.

**Cancel**: Click this button to close the dialog box.

**Help**: Click this button to display help.
Export to GIF/JPG/PNG/TIF

Use this tool to export a table, a chart, or any selected object on an Excel sheet to a GIF, JPG, PNG or TIF file.

Dialog box

- **Save**: Click this button to save the selected object to a file.
- **Cancel**: Click this button to close the dialog box.
- **Help**: Click this button to display the help.

- Click this button to reload the default options.

**Format**: Choose the graphic format of the file.

**File name**: Enter the name of the file to which the image should be saved, or select the file in a folder.

**Resize**: Activate this option to modify the size of the graphic before saving it to a file.

- **Width**: Enter the value in points of the graphic’s width;
- **Height**: Enter the value in points of the graphic’s height.

**Display the grid**: Activate this option if you want that while generating the file, XLSTAT keeps the gridlines that separate the cells. This option is only active when cells or tables are selected.
**Display the main bar**

Use this tool to display the main XLSTAT toolbar if it is no longer displayed, or to place the main toolbar on the top left of the Excel worksheet.

**Hide the sub-bars**

Use this tool to hide the XLSTAT sub-bars.
External Preference Mapping (PREFMAP)

Use this method to model and represent graphically the preference of assessors for a series of objects depending on objective criteria or linear combinations of criteria.

Description

External preference mapping (PREFMAP) is used to display on the same chart (in two of three dimensions) objects and indications showing the preference levels of assessors (in general, consumers) in certain points in the representation space. The preference level is represented on the preference map in the form of vectors, ideal or anti-ideal points, or isopreference curves depending on the type of model chosen.

These models are themselves constructed from objective data (for example physico-chemical descriptors, or scores provided by experts on well-determined criteria) which enable the position of the assessors and the products to be interpreted according to objective criteria.

If there are only two or three objective criteria, the axes of the representation space are defined by the criteria themselves (possibly standardized to avoid the effects of scale). On the other hand, if the number of descriptors is higher, a method for reducing the number of dimensions must be used. In general, PCA is used. Nevertheless, it is also possible to use factorial analysis if it is suspected that underlying factors are present, or MDS (multidimensional scaling) if the initial data are the distances between the products. If the descriptors used by the experts are qualitative variables, a PCA can be used to create a 2- or 3-dimensional space.

The PREFMAP can be used to answer the following questions:

- How is the product positioned with respect to competitive products?
- What is the nearest competitive product to a given product?
- What type of consumer prefers a product?
- Why are certain products preferred?
- How can I reposition a product so that it is again more preferred by its core target?
- What new products might it be relevant to create?

Preference models
To model the preferences of assessors depending on objective criteria or a combination of objective criteria (if a PCA has enabled a 2- or 3-dimensional space to be created) four models have been proposed within the framework of PREFMAP. For a given assessor, if we designate $y_i$ to be their preference for product $i$, and $X_1, X_2, \ldots, X_p$ to be the $p$ criteria or combinations of criteria (in general $p=2$) describing product $i$, the models are:

- **Vector:**
  $$y_i = a_0 + \sum_{j=1}^{p} a_j x_{ij}$$

- **Circular:**
  $$y_i = a_0 + \sum_{j=1}^{p} a_j x_{ij} + b \sum_{j=1}^{p} x_{ij}^2$$

- **Elliptic:**
  $$y_i = a_0 + \sum_{j=1}^{p} a_j x_{ij} + \sum_{j=1}^{p} b_j x_{ij}^2$$

- **Quadratic:**
  $$y_i = a_0 + \sum_{j=1}^{p} a_j x_{ij} + \sum_{j=1}^{p} b_j x_{ij}^2 + \sum_{j=1}^{p-1} \sum_{k=j+1}^{p} c_{jk} x_{j} x_{ik}$$

The coefficients $a_j$ are estimated by multiple linear regression. It will be noted that the models are classified from the simplest to the most complex. XLSTAT lets you either chose one model to use for all assessors, or choose a model giving the best result as regards the p-value of Fisher's F for a particular assessor or the p-value of the F-ratio test. In other words, you can choose a model which is both parsimonious and powerful at the same time.

The **vector model** represents individuals on the sensorial map in the form of vectors. The size of the vectors is a function of the $R^2$ of the model: the longer the vector, the better the corresponding model. The preference of the assessor will be stronger the further you are in the direction indicated by the vector. The interpretation of the preference can be done by projecting the different products on the vectors (product preference). The disadvantage of the vector model is that it neglects the fact that for certain criteria, like the saltiness or temperature for example, there can be an increase of preference to an optimum value then a decrease.

The **circular model** takes into account this concept of optimum. If the surface area for the model has a maximum in terms of preference (this happens if the $b$ coefficient is estimated negative), this is known as the ideal point. If the surface area for the model has a minimum in terms of preference (this happens if the $b$ coefficient is estimated positive), this is known as the anti-ideal point. With the circular model, circular lines of isopreference can be drawn around the ideal or anti-ideal points.

The **elliptical model** is more flexible, as it takes the effect of scale into account better. The disadvantage of this model is that there is not always an optimum: as with the circular model, it can generate an ideal point or an anti-ideal point if all the $b_j$ coefficients have the same sign, but we may also obtain a saddle point (in the form of a surface shaped like a horse's saddle) if all the $b_j$ coefficients do not have the same sign. The saddle point cannot easily be interpreted. It corresponds only to an area where the preference is less sensitive to variations.
Lastly, the **quadratic model** takes more complex preference structures into account, as it includes interaction terms. As with the elliptical model we can obtain an ideal, an anti-ideal, or a saddle point.

**Preference map**

The preference map is a summary view of three types of element:

The assessors (or groups of assessors if a classification of assessors has been carried out beforehand) represented in the corresponding model by a vector, an ideal point (labeled +), an anti-ideal point (labeled -), or a saddle point (labeled 0);

The objects whose position on the map is determined by their coordinates;

The descriptors which correspond to the representation axes with which they are associated (when a PCA precedes the PREFMAP, a biplot from the PCA is studied to interpret the position of the objects as a function of the objective criteria).

The PREFMAP, with the interpretation given by the preference map is an aid to interpretation and decision-making which is potentially very powerful since it allows preference data to be linked to objective data. However, the models associated with the assessors must be adjusted correctly in order that the interpretation is reliable.

**Preference scores**

The preference score for each object for a given assessor, whose value is between 0 (minimum) and 1 (maximum), is calculated from the prediction of the model for the assessor. The more the product is preferred, the higher the score. A preference order of objects is deducted from the preference scores for each of the assessors.

**Contour plot**

The contour plot shows the regions corresponding to the various preference consensus levels on a chart whose axes are the same as the preference map. At each point on the chart, the percentage of assessors for whom the preference calculated from the model is greater than their mean preference is calculated. In the regions with cold colors (blue), a low proportion of models give high preferences. On the other hand, the regions with hot colors (red) indicate a high proportion of models with high preferences.
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload**: Click this button to reload the default options.
- **Delete data**: Click this button to delete the data selections.
- **Change data handling**: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab:**

- **Y / Preference data**: Select the preference data. The table contains the various objects (products) studied in the rows and the assessors in the columns. This is reversed in transposed mode. If column headers have been selected, check that the "Variable labels" option has been activated.

  Note: XLSTAT considers that the preferences are the increasing data (the more an assessor likes an object, the higher the preference).

- **Center**: Activate this option if you want to center the preference data before starting the calculations.
- **Reduce**: Activate this option if you want to reduce the preference data before starting the calculations.

**X / Configuration**: Select the data corresponding to the objective descriptors or to a 2- or 3-dimensional configuration if a method has already been used to generate the configuration. If column headers have been selected, check that the "Variable labels" option has been activated.
**Preliminary transformation**: Activate this option if you want to transform the data.

- **Normalization**: Activate this option to standardize the data for the X-configuration before carrying out the PREFMAP.

- **PCA (Pearson)**: Activate this option for XLSTAT to transform the selected descriptors using a normalized Principle Components Analysis (PCA). The number of factors used afterwards used for the calculations is determined by the number of **dimensions** chosen.

- **PCA (Covariance)**: Activate this option for XLSTAT to transform the selected descriptors using a non-normalized Principle Components Analysis (PCA). The number of factors used afterwards used for the calculations is determined by the number of **dimensions** chosen.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Variable labels**: Check this option if the first line of the data selected (Y, X, object labels) contains a label.

**Objects labels**: Activate this option if observation labels are available. Then select the corresponding data. If the “Variable labels” option is activated you need to include a header in the selection. If this option is not activated, the labels are automatically generated by XLSTAT (Obs1, Obs2 …).

**Model**: Choose the type of model to use to link the preferences to the X configuration if the option “Find the best model” (see Options tab) has not been activated.

**Dimensions**: Enter the number of dimensions to use for the PREFMAP model (default value: 2).

**Options** tab:

**Find the best model**: Activate this option to allow XLSTAT to find the best model for each assessor.

- **F-ratio**: Activate this option to use the F-ratio test to select the model that is the best compromise between quality of the fit and parsimony in variables. A more complex
model is accepted if the p-value corresponding to the F is lower than the significance level.

- **F**: Activate this option to select the model that gives the best p-value based computed the Fisher's F.

**Significance level (%)**: enter the significance level. The p-values of the models are displayed in bold when they are less than this level.

**Weights**: If you want to weigh the assessors, activate this option, then select the weight corresponding to each observation.

These options are visible only if a PCA based preliminary transformation has been requested.

**Supplementary variables**: Activate this option if you want to calculate coordinates afterwards for variables which were not used in calculating the factor axes (passive variables as opposed to active variables).

- **Quantitative**: Activate this option if you have supplementary quantitative variables. If column headers were selected for the main table, ensure that a label is also present for the variables in this selection.

**Prediction** tab:

This tab is not visible if a preliminary PCA transformation was requested.

**Prediction**: activate this option if you want to select data to use them in prediction mode. If you activate this option, you need to make sure that the prediction dataset is structured as the estimation dataset: same variables with the same order in the selections. On the other hand, variable labels must not be selected: the first row of the selections listed below must correspond to data.

**X / Configuration**: Activate this option to select the configuration data to use for the predictions. The first row must not include variable labels.

**Object labels**: Activate this option if you want to use object labels for the prediction data. The first row must not include variable labels. If this option is not activated, the labels are automatically generated by XLSTAT (PredObs1, PredObs2, etc.).

**Missing data** tab:
Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- Mean or mode: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- Nearest neighbour: Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Correlations: Activate this option to display the correlation matrix for the different variables selected.

Analysis of variance: Activate this option to display the analysis of variance table for the various models.

Model coefficients: Activate this option to display the parameters of the models.

Model predictions: Activate this option to display the predictions of the models.

Preference scores: Activate this option to display preference scores on a scale of 0 to 1.

Ranks of the preference scores: Activate this option to display the ranks for the preference scores.

Sorted objects: Activate this option to display the objects in decreasing order of preference for each of the assessors.

If a preliminary transformation based on PCA has been requested, the following options are available:

Factor loadings: Activate this option to display the coordinates of the variables (factor loadings). The coordinates are equal to the correlations between the principal components and the initial variables for normalized PCA.

Components/Variables correlations: Activate this option to display correlations between the principal components and the initial variables.

Factor scores: Activate to display the coordinates of the observations (factor scores) in the new space created by PCA. These coordinates are afterwards used for the PREFMAP.
If a preliminary transformation based on PCA has been requested, the following options are available:

**Factor loadings**: Activate this option to display the coordinates of the variables (*factor loadings*). The coordinates are equal to the correlations between the principal components and the initial variables for normalized PCA.

**Components/Variables correlations**: Activate this option to display correlations between the principal components and the initial variables.

**Factor scores**: Activate to display the coordinates of the observations (*factor scores*) in the new space created by PCA. The principal components are afterwards used as explanatory variables in the regression.

**Charts (PCA) tab**: This tab is visible only if a PCA based preliminary transformation has been requested.

**Correlations charts**: Activate this option to display charts showing the correlations between the components and initial variables.

- **Vectors**: Activate this option to display the input variables in the form of vectors.

**Observations charts**: Activate this option to display charts representing the observations in the new space.

- **Labels**: Activate this option to have observation labels displayed on the charts. The number of labels displayed can be changed using the filtering option.

**Biplots**: Activate this option to display charts representing the observations and variables simultaneously in the new space.

- **Vectors**: Activate this option to display the initial variables in the form of vectors.

- **Labels**: Activate this option to have observation labels displayed on the biplots. The number of labels displayed can be changed using the filtering option.

**Type of biplot**: Choose the type of biplot you want to display. See the description section of the PCA for more details.

- **Correlation biplot**: Activate this option to display correlation biplots.

- **Distance biplot**: Activate this option to display distance biplots.

- **Symmetric biplot**: Activate this option to display symmetric biplots.
- **Coefficient**: Choose the coefficient whose square root is to be multiplied by the coordinates of the variables. This coefficient lets you to adjust the position of the variable points in the biplot in order to make it more readable. If set to other than 1, the length of the variable vectors can no longer be interpreted as standard deviation (correlation biplot) or contribution (distance biplot).

**Colored labels**: Activate this option to show variable and observation labels in the same color as the corresponding points. If this option is not activated the labels are displayed in black color.

**Charts** tab:

- **Preference map**: Activate this option to display the preference map.
  - **Display ideal points**: Activate this option to display the ideal points.
  - **Display anti-ideal points**: Activate this option to display the anti-ideal points.
  - **Display saddle points**: Activate this option to display the saddle points.
  - **Domain restriction**: Activate this option to only display the solution points (ideal, anti-ideal, saddle) if they are within a domain to be defined. Then enter the size of the area to be used for the display: this is expressed as a percentage of the area delimited by the X configuration (value between 100 and 500).
  - **Vectors length**: The options below are used to determine the lengths of the vectors on the preference map when a vector model is used.
    - **Coefficients**: Choose this option so that the length of the vectors is only determined by the coefficients of the vector model.
    - **R²**: Choose this option so that the length of the vectors is only determined by the R² value of the model. Thus the better the model is adjusted, the longer is the corresponding vector on the map.
    - **=**: Activate this option to display the vectors with an equal length.
    - **Lengthening factor**: Use this option to multiply the length of all vectors by an arbitrary value (default value: 1)

**Circular model**:

- **Display circles**: Enter the number of isopreference circles to be displayed.
Contour plot: Activate this option to display the contour plot (see the description section). Afterwards, you need to choose which criterion is used to determine the % of assessors that prefer products at a given point of the preference map.

- **Threshold / Mean (%):** Enter the level in % with respect to the preference mean above which an assessor can be considered to like a product (the default value, 100, is the mean).
- **Threshold (value):** Enter the preference value above which an assessor can be considered to like a product (the default value, 100, is the mean).

PREFMAP & Contour plot: Activate this option to display the superposition of the preference map and of the contour plot. Three quality levels are possible. If you notice some defects in the map, you can increase the number of points.

**Results**

**Summary statistics:** This table shows the number of non-missing values, the mean and the standard deviation (unbiased) for all assessors and all dimensions of the X configuration (before transformation if that has been requested).

**Correlation matrix:** This table is displayed to give you a view of the correlations between the various variables selected.

**Model selection:** This table shows which model was used for each assessor. If the model is not a vector model, the solution point type is displayed (ideal, anti-ideal, saddle) with its coordinates.

**Analysis of variance:** This table shows the statistics used to evaluate the goodness of fit of the model ($R^2$, $F$, and $Pr>F$). When the p-value ($Pr>F$) is less than the chosen significance level, it is displayed in bold. If the F-ratio test was chosen in the “Options” tab, the results of the F-ratio test are displayed if it was successful at least once.

**Model coefficients:** This table displays the various coefficients of the chosen model for each assessor.

**Model predictions:** This table shows the preferences estimated by the model for each assessor and each product. Note: if the preferences have been standardized, these results therefore apply to the standardized preferences.

**Preference scores from 0 to 1:** This table shows the predictions on a scale of 0 to 1.

**Ranks of the preference scores:** This table displays the ranks of the preference scores. The higher the rank, the higher the preference.
**Objects sorted by increasing preference order:** This table shows the list of objects in increasing order of preference, for each assessor. In other words, the last line corresponds to objects preferred by the assessors according to the preference models.

The **preference map** and the **contour plot** are then displayed. On the preference map, the ideal points are shown by (+), the anti-ideal points by (-) and saddle points by (o).

If the option is enabled and you use Excel 2003 or higher, you can view the superposition of the preference map and contour plot. This chart can be resized, but so that the overlay is maintained after resizing, you must click in the Excel sheet and then again on the graph.

**Example**

A example of Preference Mapping is available on the Addinsoft website:

http://www.xlstat.com/demo-prefmap.htm

**References**

*Danzart M. and Heyd B. (1996).* Le modèle quadratique en cartographie des préférences. 3ème Congrès Sensometrics, ENITIAA.


Internal Preference Mapping

Use Internal Preference Mapping (IPM) to analyze the ratings given on P products by J assessors (consumers, experts, ...): While External Preference Mapping allows to relate consumers ratings to sensory data (chemical measurements, ratings by experts), for Internal Preference Mapping only preference data is necessary. IPM is based on PCA and adds two options to improve the visualization on the results.

Description

Internal Preference Mapping (IPM) is based on Principle Component Analysis (PCA) to allow identifying which products correspond to groups of consumers.

For more information on PCA, you can read the description available in the chapter dedicated to that method. While PCA does not filter out variables, this tool allows removing (after the PCA step) from the plots the assessors that are not well enough displayed on a given 2-dimensional map. The measure of how well a point is projected from a d-dimensional space to a 2-dimensional map is named communality. It can also be understood as the sum of the squared cosines between the vector and the axes of the sub-space.

The biplot that is then produced is not a true biplot as all the retained assessors are moved on a virtual circle surrounding the product points in order to facilitate the visual interpretation.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- OK: Click this button to start the computations.
- Cancel: Click this button to close the dialog box without doing any computation.
- Help: Click this button to display the help.
- : Click this button to reload the default options.
- : Click this button to delete the data selections.
Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

**Products\Assessors table:** Select the quantitative data corresponding to P products described by J assessors. If column headers have been selected, check that the "Variable labels" option has been activated.

**PCA type:** Choose the type of matrix to be used for PCA. The difference between the Pearson (n) and the Pearson (n-1) options, only influences the way the variables are standardized, and the difference can only be noticed on the coordinates of the observations.

**Range:** Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Activate this option to display the results in a new worksheet of the active workbook.

**Workbook:** Activate this option to display the results in a new workbook.

**Variable labels:** Activate this option if the first row of the data selections (products\assessors table, weights, products labels) includes a header.

**Product labels:** Activate this option if product labels are available. Then select the corresponding data. If the "Assessor labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

**Weights:** Activate this option if the observations are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

**Options** tab:

**Filter factors:** You can activate one of the following two options in order to reduce the number of factors for which results are displayed.

- **Minimum %:** Activate this option then enter the minimum percentage of the total variability that the chosen factors must represent.
• **Maximum Number**: Activate this option to set the number of factors to take into account.

**Rotation**: Activate this option if you want to apply a rotation to the factor coordinate matrix.

- **Number of factors**: Enter the number of factors the rotation is to be applied to.
- **Method**: Choose the rotation method to be used. For certain methods a parameter must be entered (Gamma for Orthomax, Tau for Oblimin, and the power for Promax).
- **Kaiser normalization**: Activate this option to apply Kaiser normalization during the rotation calculation.

**Supplementary data** tab:

**Supplementary observations**: Activate this option if you want to calculate the coordinates and represent additional observations. These observations are not taken into account for the factor axis calculations (passive observations as opposed to active observations). Several methods for selecting supplementary observations are provided:

- **Random**: The observations are randomly selected. The “Number of observations” N to display must then be specified.
- **N last rows**: The last N observations are selected for validation. The “Number of observations” N to display must then be specified.
- **N first rows**: The first N observations are selected for validation. The “Number of observations” N to display must then be specified.
- **Group variable**: If you choose this option, you must then select an indicator variable set to 0 for active observations and 1 for passive observations.

**Supplementary variables**: Activate this option if you want to calculate coordinates afterwards for variables which were not used in calculating the factor axes (passive variables as opposed to active variables).

- **Quantitative**: Activate this option if you have supplementary quantitative variables. If column headers were selected for the main table, ensure that a label is also present for the variables in this selection.
- **Qualitative**: Activate this option if you have supplementary qualitative variables. If column headers were selected for the main table, ensure that a label is also present for the variables in this selection.
  - **Color observations**: Activate this option so that the observations are displayed in different colors depending on the value of the first qualitative variable.
- **Display the centroids**: Activate this option to display the centroids that correspond to the categories of the supplementary qualitative variables.

**Missing data** tab:

**Do not accept missing data**: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

**Remove the observations**: Activate this option to remove observations with missing data.

**Pairwise deletion**: Activate this option to remove observations with missing data only when the variables involved in the calculations have missing data. For example, when calculating the correlation between two variables, an observation will only be ignored if the data corresponding to one of the two variables is missing.

**Estimate missing data**: Activate this option to estimate the missing data before the calculation starts.

- **Mean or mode**: Activate this option to estimate the missing data by using the mean (quantitative variables) or the mode (qualitative variables) for the corresponding variables.

- **Nearest neighbour**: Activate this option to estimate the missing data for an observation by searching for the nearest neighbour to the observation.

**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the variables selected.

**Correlations**: Activate this option to display the correlation or covariance matrix depending on the type of options chosen in the "General" tab.

- **Test significance**: Where a correlation was chosen in the "General" tab in the dialog box, activate this option to test the significance of the correlations.

- **Bartlett’s sphericity test**: Activate this option to perform the Bartlett sphericity test.

- **Significance level (%)**: Enter the significance level for the above tests.

- **Kaiser-Meyer-Olkin**: Activate this option to compute the Kaiser-Meyer-Olkin Measure of Sampling Adequacy.

**Eigenvalues**: Activate this option to display the table and chart (scree plot) of eigenvalues.
Factor loadings: Activate this option to display the coordinates of the variables in the factor space.

Variables/Factors correlations: Activate this option to display correlations between factors and variables.

Factor scores: Activate to display the coordinates of the observations (factor scores) in the new space created by PCA.

Contributions: Activate this option to display the contribution tables for the variables and observations.

Squared cosines: Activate this option to display the tables of squared cosines for the variables and observations.

Filter out assessors: Activate this option if you want to remove in the outputs and on the maps, the assessors for which the communality is below a given threshold.

Charts tab:

Correlations charts: Activate this option to display charts showing the correlations between the components and initial variables.

- Vectors: Activate this option to display the initial variables in the form of vectors.

Observations charts: Activate this option to display charts representing the observations in the new space.

- Labels: Activate this option to have observation labels displayed on the charts. The number of labels displayed can be changed using the filtering option.

Biplots: Activate this option to display charts representing the observations and variables simultaneously in the new space.

- Vectors: Activate this option to display the initial variables in the form of vectors.

- Labels: Activate this option to have observation labels displayed on the biplots. The number of labels displayed can be changed using the filtering option.

- Move to circle: Activate this option to move all the points corresponding to the assessors moved to a circle that surrounds the points corresponding to the products.

Colored labels: Activate this option to show labels in the same color as the points.
Results

Descriptive statistics: The table of descriptive statistics shows the simple statistics for all the variables selected. This includes the number of observations, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased).

Correlation/Covariance matrix: This table shows the data to be used afterwards in the calculations. The type of correlation depends on the option chosen in the "General" tab in the dialog box. For correlations, significant correlations are displayed in bold.

Bartlett's sphericity test: The results of the Bartlett sphericity test are displayed. They are used to confirm or reject the hypothesis according to which the variables do not have significant correlation.

Measure of Sample Adequacy of Kaiser-Meyer-Olkin: this table gives the value of the KMO measure for each individual variable and the overall KMO measure. The KMO measure ranges between 0 and 1. A low value corresponds to the case where it is not possible to extract synthetic factors (or latent variables). In other words, observations do not bring out the model that one could imagine (the sample is "inadequate"). Kaiser (1974) recommends not to accept a factor model if the KMO is less than 0.5. If the KMO is between 0.5 and 0.7 then the quality of the sample is mediocre, it is good for a KMO between 0.7 and 0.8, very good between 0.8 and 0.9 and excellent beyond.

Eigenvalues: The eigenvalues and corresponding chart (scree plot) are displayed. The number of eigenvalues is equal to the number of non-null eigenvalues.

If the corresponding output options have been activated, XLSTAT afterwards displays the factor loadings in the new space, then the correlations between the initial variables and the components in the new space. The correlations are equal to the factor loadings in a normalized PCA (on the correlation matrix).

If supplementary variables have been selected, the corresponding coordinates and correlations are displayed at the end of the table.

Contributions: Contributions are an interpretation aid. The variables which had the highest influence in building the axes are those whose contributions are highest.

Squared cosines: As in other factor methods, squared cosine analysis is used to avoid interpretation errors due to projection effects. If the squared cosines associated with the axes used on a chart are low, the position of the observation or the variable in question should not be interpreted.

The factor scores in the new space are then displayed. If supplementary data have been selected, these are displayed at the end of the table.
Contributions: This table shows the contributions of the observations in building the principal components.

Squared cosines: This table displays the squared cosines between the observation vectors and the factor axes.

Where a rotation has been requested, the results of the rotation are displayed with the rotation matrix first applied to the factor loadings. This is followed by the modified variability percentages associated with each of the axes involved in the rotation. The coordinates, contributions and cosines of the variables and observations after rotation are displayed in the following tables.

Example

A tutorial on how to use Internal Preference Mapping is available on the Addinsoft website:

http://www.xlstat.com/demo-intprefmap.htm

References


Panel analysis

Use this tool to check whether your sensory or consumer panel allows to differentiate a series of products. If it does, measure to what extent and make sure that the ratings given by the assessors are reliable.

Description

This tool enables chaining different analyzes proposed by XLSTAT, to assess the ability of a panel of J consumers, experts, judges, or assessors (the term assessor is used in the XLSTAT interface), to differentiate P products using K descriptors (variables in the statistical sense) and to control if the ratings are reliable (if repeated measurements are available through different evaluation sessions). A classification can also be done to identify homogeneous groups among the assessors.

The first step consists of a series of ANOVA with the aim to verify for each descriptor if there is a product effect or not. For each descriptor, the table of Type III SS of the ANOVA is displayed for the selected model. Then, a summary table allows comparing the p-values of the product effect for the different descriptors. The analyses that follow will only be conducted for the descriptors that allow discriminating the products. Different ANOVA models are possible depending on the presence or absence of sessions (repetitions), the willingness to take into account interactions and one wants to consider the effect of assessors and sessions as fixed or random.

The second step consists of a graphical analysis. For each of the k descriptors that are kept after the ANOVAs, box plots and strip plots are displayed. We can thus see how, for each descriptor, different assessors use the rating scale to evaluate the different products.

The third step starts with the restructuring of the data table, in order to have a table containing one row per product and one column per pair of assessor and descriptor - if there are several sessions, then the table contains averages - followed by a PCA (normalized) on this same table. The number of products P is generally less than the product k * J, so we should have at most P factorial axes. We then display as many PCA correlations plots as there are descriptors, in order to highlight on each plot the points corresponding to the assessors ratings for a given descriptor. This allows to check in one step the extent to which assessors agree or not for each of the k descriptors, once the effect of position and scale is removed (because the PCA is normalized), and to what extent the descriptors are linked or not. To study more precisely the relationship between descriptors, an MFA (multiple factor analysis) plot is displayed.

During the fourth step an ANOVA is performed for each assessor separately, and for each of the k descriptors in order to check whether there is a product effect or not. This allows to assess for each assessor if he is able to distinguish the products using the available descriptors. A summary table is then used to count for each assessor the number of
descriptors for which he was able to differentiate the products. The corresponding percentage is displayed. This percentage is a simple measure of the discriminating power of assessors.

For the **fifth step**, a global table initially presents ratings (averaged over the repetitions if available) for each assessor in rows, and each pair (product,descriptor) in columns. It is followed by a series of P tables and charts to compare, product by product, assessors (averaged over the possible repetitions) for the set of descriptors. These charts can be used to identify strong trends and possible atypical ratings for some assessors.

The **sixth step** allows identifying atypical assessors through the measure for each product of the Euclidean distance of each assessor to an average for all assessors in the space of the k descriptors. A table showing these distances for each product and the minimum and maximum computed over all assessors, allows identifying assessors that are close to or far from the consensus. A chart is displayed to allow visualizing these distances.

If a “session” variable was selected, the **seventh step** checks if for some assessors there is a session effect, typically an order effect. This is assessed using a Friedman test (or Wilcoxon signed rank test if there are only two sessions). The test is calculated on all products, descriptor by descriptor. Then, for each assessor and each descriptor, we calculate which is the maximum observed range between sessions across products. The product corresponding to the maximum range is indicated. This table is used to identify possible anomalies in the ratings given by some assessors and possibly remove some observations for future analyses.

If for each triple (assessor,product,descriptor) there exists at least one rating, the **eighth step** consists of a classification of the assessors. The classification is first performed on the raw data, then on the standardized data to eliminate possible effects of scale and position.

Finally a table preformatted for Generalized Procrustean Analysis (GPA) is displayed in case you want to run such an analysis.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
Click this button to reload the default options.

Click this button to delete the data selections.

Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

**Y / Descriptors:** Select the preference data associated to each descriptor. The table contains the scores given by the assessors for the different descriptors corresponding to a product and to a session. If column headers have been selected, check that the "Variable labels" option has been activated.

**Products:** Select the data corresponding to the tested products. Only one column has to be selected. If column headers have been selected, check that the "Variable labels" option has been activated.

**Assessors:** Select the data corresponding to the assessors. Only one column has to be selected. If column headers have been selected, check that the "Variable labels" option has been activated.

**Sessions:** Activate this option if more than one tasting session has been organized. Select the data corresponding to the sessions. Only one column has to be selected. If column headers have been selected, check that the "Variable labels" option has been activated.

**Range:** Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Activate this option to display the results in a new worksheet of the active workbook.

**Workbook:** Activate this option to display the results in a new workbook.

**Variable labels:** Check this option if the first line of the data selected (Y, X, object labels) contains a label.

**Observation labels:** Activate this option if observation labels are available. Then select the corresponding data. If the “Variable labels” option is activated you need to include a header in the selection. If this option is not activated, the labels are automatically generated by XLSTAT (Obs1, Obs2 …).
Weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

Options tab:

Model: Select the ANOVA model you want to use to identify the non-discriminating descriptors. If the Session option is not active, the two possible models are $Y = \text{Product} + \text{Assessor}$ and $Y = \text{Product} + \text{Assessor} + \text{Product} \times \text{Assessor}$. If the Session option is active, the two possible models are $Y = \text{Product} + \text{Juge} + \text{Session}$ and $Y = \text{Product} + \text{Assessor} + \text{Session} + \text{Product} \times \text{Assessor} + \text{Product} \times \text{Session} + \text{Session} \times \text{Assessor}$.

Random effects (Assessor / Session): Activate this option if you want to consider that the Assessor and Repetition effects as well as the interactions involving them are random effects. If this option is not checked, all effects are considered as fixed.

Significance level (%): Enter the significance level that will be used to determine above which level p-values lead to validate the null hypotheses of the various tests that are computed during the analysis.

Filter out non discriminating descriptors: Activate this option to remove from the analysis all the descriptors for which there is no product effect. You can then specify the threshold p-value above which one can consider there is no product effect.

Missing data tab:

Remove observations: Activate this option to remove the observations with missing data.

- Check each Y separately: Activate this option to remove observations for each descriptor separately (the sample size will vary from one model to another).

- For all Y: Activate this option to remove all observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- Mean or mode: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

- Nearest neighbour: Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.
**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the variables selected.

**ANOVA summaries**: Activate this option to display the summaries of the various ANOVA models that are computed.

**Assessors' ability to discriminate products**: Activate this option to display the tables and charts and allow evaluating the ability of the assessors to differentiate the various products.

**Distance to consensus**: Activate this option to the table of distances to consensus.

**Assessor means by (product, descriptor)**: Activate this option to display the table of the means for each pair (product, descriptor) and, for each product, the table of the means by assessor and descriptor.

**Sessions analysis**: Activate this option to assess the reliability of the assessors using the sessions information.

**GPA table**: Activate this option to display a table that is formatted in a way to allow running a GPA (Generalized Procrustean Analysis).

**Charts** tab:

**Box plots**: Activate this option to display the box plots that allow to compare the various assessors for each descriptor.

**Strip plots**: Activate this option to display the strip plots that allow to compare the various assessors for each descriptor.

**PCA plots**: Activate this option to display the various plots obtained from the PCA and MFA.

**Line plot for each product**: Activate this option to display the line plots that allow for each product to compare the assessors for all descriptors.

**Line plot of distances to consensus**: Activate this option to display the chart that allows to evaluate how far each assessor is from the consensus, product by product.

**Dendrogram**: Activate this option to display the dendrograms obtained from the classification of the assessors.
Results

Summary statistics: The tables of descriptive statistics show the simple statistics for all the variables selected. The number of observations, missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the descriptors. For qualitative explanatory variables the names of the various categories are displayed together with their respective frequencies.

The first step consists of a series of ANOVA with the aim to verify for each descriptor if there is a product effect or not. For each descriptor, the table of Type III SS of the ANOVA is displayed for the selected model. Then, a summary table allows comparing the p-values of the product effect for the different descriptors. The analyses that follow will only be conducted for the descriptors that allow discriminating the products. Different ANOVA models are possible depending on the presence or absence of sessions (repetitions), the willingness to take into account interactions and one wants to consider the effect of assessors and sessions as fixed or random.

The second step consists of a graphical analysis. For each of the k descriptors that are kept after the ANOVAs, box plots and strip plots are displayed. We can thus see how, for each descriptor, different assessors use the rating scale to evaluate the different products.

The third step starts with the restructuring of the data table, in order to have a table containing one row per product and one column per pair of assessor and descriptor - if there are several sessions, then the table contains averages - followed by a PCA (normalized) on this same table. The number of products P is generally less than the product k * J, so we should have at most P factorial axes. We then display as many PCA correlations plots as there are descriptors, in order to highlight on each plot the points corresponding to the assessors ratings for a given descriptor. This allows to check in one step the extent to which assessors agree or not for each of the k descriptors, once the effect of position and scale is removed (because the PCA is normalized), and to what extent the descriptors are linked or not. To study more precisely the relationship between descriptors, an MFA (multiple factor analysis) plot is displayed.

During the fourth step an ANOVA is performed for each assessor separately, and for each of the k descriptors in order to check whether there is a product effect or not. This allows to assess for each assessor if he is able to distinguish the products using the available descriptors. A summary table is then used to count for each assessor the number of descriptors for which he was able to differentiate the products. The corresponding percentage is displayed. This percentage is a simple measure of the discriminating power of assessors.

A summary table also shows the assessors' performance score based on ANOVA model. The first line indicates the number of descriptors for which each assessor was able to differentiate products (Discrimination), the second line shows the repeatability associated with the assessor (no session effect, meaning the assessor is consistent with himself). The third line is only displayed if an interaction effect has been included in the model, and it gives for each assessor the number of descriptors for which the assessor has not contributed to an
interaction effect. The last line is the sum of the previous. The higher the value on this line the better the assessor.

For the fifth step, a global table initially presents ratings (averaged over the repetitions if available) for each assessor in rows, and each pair (product, descriptor) in columns. It is followed by a series of P tables and charts to compare, product by product, assessors (averaged over the possible repetitions) for the set of descriptors. These charts can be used to identify strong trends and possible atypical ratings for some assessors.

The sixth step allows identifying atypical assessors through the measure for each product of a Euclidean distance of each assessor to an average for all assessors in the space of descriptors. A table showing these distances for each product and the minimum and maximum computed over all assessors, allows identifying assessors that are close to or far from the consensus. A chart is displayed to allow visualizing these distances.

If a “session” variable was selected, the seventh step checks if for some assessors there is a session effect, typically an order effect. This is assessed using a Friedman test (or Wilcoxon signed rank test if there are only two sessions). The test is calculated on all products, descriptor by descriptor. Then, for each assessor and each descriptor, we calculate which is the maximum observed range between sessions across products. The product corresponding to the maximum range is indicated. This table is used to identify possible anomalies in the ratings given by some assessors and possibly remove some observations for future analyses.

If for each triple (assessor, product, descriptor) there exists at least one rating, the eighth step consists of a classification of the assessors. The classification is first performed on the raw data, then on the standardized data to eliminate possible effects of scale and position.

Finally a table preformatted for Generalized Procrustean Analysis (GPA) is displayed in case you want to run such an analysis.

Example

A tutorial explaining how to use Panel Analysis is available on the Addinsoft website:

http://www.xlstat.com/demo-panel.htm

References


Product characterization

Use this tool to identify which descriptors best discriminate a set of products and which characteristics of the products are important in a sensory study.

Description

This tool has been developed using the recommendations given by Pr. Jérôme Pagès and Sébastien Lê from the Laboratory for Applied Mathematics at Agrocampus (Rennes, France). It provides the XLSTAT users with a user-friendly tool that helps finding in a sensory study which descriptors are discriminating well a set of products. You can also identify which are the most important characteristics of each product.

All computations are based on the analysis of variance (ANOVA) model. For more details on technical aspects, see the analysis of variance chapter of the XLSTAT help.

The data table must have a given format. Each row should concern a given product, eventually a given session and should gather scores given by an assessor for one or more descriptors associated to the designated product. The dataset must contain the following columns: one identifying the assessor, one identifying the product, eventually one identifying the session, and as many columns as there are descriptors or characteristics.

For each descriptor an ANOVA model is applied to check if the scores given by the assessors are significantly different. The simplest model is:

Score = product effect + judge effect

If different sessions have been organized (each judge has evaluated at least twice each product), the session factor can be added and the model becomes:

Score = product effect + judge effect + session effect

An interaction factor can also be included. We then can test if some combines of the judges and products are giving higher or lower grades on the descriptors. The model is:

Score = product effect + judge effect + product effect * judge effect

The judge effect is always supposed to be random. It means we consider each judge to have its own way of giving scores to the products (on the score scale).

Product characterization is a very efficient tool to characterize products using judges' preferences.
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload default options**: Click this button to reload the default options.
- **Delete data selections**: Click this button to delete the data selections.
- **Change data handling**: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab:**

**Y / Descriptors**: Select the preference data associated to each descriptor. The table contains the scores given by the judges for the different descriptors corresponding to a product and to a session. If column headers have been selected, check that the "Variable labels" option has been activated.

**Products**: Select the data corresponding to the tested products. Only one column has to be selected. If column headers have been selected, check that the "Variable labels" option has been activated.

**Assessors**: Select the data corresponding to the assessors. Only one column has to be selected. If column headers have been selected, check that the "Variable labels" option has been activated.

**Sessions**: Activate this option if more than one tasting session has been organized. Select the data corresponding to the sessions. Only one column has to be selected. If column headers have been selected, check that the "Variable labels" option has been activated.
Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Check this option if the first line of the data selected (Y, X, object labels) contains a label.

Observation labels: Activate this option if observation labels are available. Then select the corresponding data. If the “Variable labels” option is activated you need to include a header in the selection. If this option is not activated, the labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

Options tab:

Model: Select the ANOVA model you want to use to identify the non-discriminating descriptors. If the Session option is not active, the two possible models are \( Y = \text{Product} + \text{Assessor} \) and \( Y = \text{Product} + \text{Assessor} + \text{Product} \times \text{Assessor} \). If the Session option is active, the two possible models are \( Y = \text{Product} + \text{Juge} + \text{Session} \) and \( Y = \text{Product} + \text{Assessor} + \text{Session} + \text{Product} \times \text{Assessor} + \text{Product} \times \text{Session} + \text{Session} \times \text{Assessor} \).

Sort the adjusted means table: activate this option if you want the adjusted means to be sorted so that similar products and descriptors are close to each other. A principal component analysis is applied to find the best positioning.

Significance level (%): enter the significance level for the confidence intervals.

Missing data tab:

Remove observations: Activate this option to remove the observations with missing data.

- Check each Y separately: Activate this option to remove observations for each descriptor separately (the sample size will vary from one model to another).

- For all Y: Activate this option to remove all observations with missing data.
Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

- **Nearest neighbour**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.

Charts tab:

Sensory profiles: Activate this option to display the chart of the sensory profiles.

- **Biplot**: Activate this option to display simultaneously products and Y variables (descriptors).

- **Filter out non discriminating descriptors**: Activate this option to ignore the descriptors that have been identified as non-discriminating in the previous analyses. You can enter the threshold above which a descriptor is considered as non discriminating and should be removed.

Results

Summary statistics: The tables of descriptive statistics show the simple statistics for all the variables selected. The number of observations, missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the descriptors. For qualitative explanatory variables the names of the various categories are displayed together with their respective frequencies.

Discriminating power by descriptor: This table shows the ordered descriptors from the most discriminating on the products to the least discriminating. Associated V-test and p-values are also displayed.

Model coefficients: This table displays the various coefficients of the chosen model for each combination product-descriptor. Adjusted mean, t test, p-value and confidence interval for each combination are also displayed. Graphics for each product with the coefficients are then displayed.

Adjusted means by product: This table shows the adjusted mean for each combination product-descriptor. The color corresponds to a significant positive effect for the blue color and a significant negative effect for the red color.

Chart with confidence ellipses for the sensory profiles obtained by PCA: this biplot, created following the method described by Husson et al (2005) allows to visualize on the same
graph the descriptors, as well as the products with a confidence ellipse whose orientation and surface depend on the ratings given by different assessors. These ellipses are calculated using a resampling method. The tables with the coordinates of the products and the corresponding cosines are displayed to avoid misleading interpretations.

Example

An example of product characterization is available at the Addinsoft website:

http://www.xlstat.com/demo-decat.htm

References


Penalty analysis

Use this tool to analyze the results of a survey run using a five level JAR (Just About Right) scale, on which the intermediary level 3 corresponds to the preferred value for the consumer.

Description

Penalty analysis is a method used in sensory data analysis to identify potential directions for the improvement of products, on the basis of surveys performed on consumers or experts.

Two types of data are used:

- Preference data (or liking scores) that correspond to a global satisfaction index for a product (for example, liking scores on a 10 point scale for a chocolate bar), or for a characteristic of a product (for example, the comfort of a car rated from 1 to 10).

- Data collected on a JAR (Just About Right) 5 point scale. These correspond to ratings ranging from 1 to 5 (or 1 to 7, or 1 to 9) for one or more characteristics of the product of interest. In the case of a 5 points JAR scale, 1 corresponds not « Not enough at all », 2 to « Not enough », 3 to « JAR » (Just About Right), an ideal for the consumer, 4 to « Too much » and 5 to « Far too much ». For example, for a chocolate bar, one can rate the bitterness, and for the comfort of the car, the sound volume of the engine.

The method, based on multiple comparisons such as those used in ANOVA, consists in identifying, for each characteristic studied on the JAR scale, if the rankings on the JAR scale are related to significantly different results in the liking scores. For example, if a chocolate is too bitter, does that significantly impact the liking scores?

The word penalty comes from the fact that we are looking for the characteristics which can penalize the consumer satisfaction for a given product. The penalty is the difference between the mean of the liking scores for the JAR category, and the mean of the scores for the other categories.

Penalty analysis is subdivided into three phases:

The data of the JAR scale are aggregated: for example, in the case of a 5 points JAR scale, on one hand, categories 1 and 2 are grouped, and on the other hand categories 4 and 5 are grouped, which leads to a three point scale. We now have three levels: “Not enough”, “JAR”, and “Too much”.

We then compute and compare the means of the liking scores for the three categories, to identify significant differences. The difference between the means of the 2 non-JAR categories and the JAR category is called mean drops.

We compute the penalty and test if it is significantly different from 0.
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

Reload: Click this button to reload the default options.

Delete: Click this button to delete the data selections.

**General tab:**

Liking scores: Select the preference data. Several columns can be selected. If a column header has been selected, check that the "Column labels" option has been activated.

Just about right data: Select the data measured on the JAR scale. Several columns can be selected. If a column header has been selected, check that the "Column labels" option has been activated.

- **Scale:** Select the scale that corresponds to the data (1 -> 5, 1 -> 7, 1 -> 9).

Labels of the 3 JAR levels: Activate this option if you want to use labels for the 3 point JAR scale. There must be three rows and as many columns as in the Just about right data selection. If a column header has been selected, check that the "Column labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.
**Workbook**: Activate this option to display the results in a new workbook.

**Column labels**: Activate this option if the first row of the data selections (Liking scores, Just about right data, labels of the 3 JAR levels) includes a header.

**Weights**: Activate this option if the observations are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Column labels" option is activated.

**Options** tab:

**Threshold for population size**: Enter the % of the total population that should represent a category to be taken into account for multiple comparisons.

**Missing data** tab:

**Do not accept missing data**: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

**Remove observations**: Activate this option to ignore the observations that contain missing data.

**Ignore missing data**: Activate this option to ignore missing data.

**Estimate missing data**: Activate this option to estimate the missing data by using the mean of the variables.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

- **Nearest neighbour**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.

**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the selected variables.

**Correlations**: Activate this option to display the matrix of correlations of the selected dimensions. If all data are ordinal, it is recommended to use the Spearman coefficient of correlation.
3 levels table: Activate this option to display the JAR data once they are collapsed from 5 to 3 categories.

Penalty table: Activate this option to display the table that displays the mean drops for the non-JAR categories, as well as the penalties.

Multiple comparisons: Activate this option to run the multiple comparisons tests on the difference between means. Several methods are available, grouped into two categories: multiple pairwise comparisons, and multiple comparisons with a control, the latter being here the JAR category.

- Significance level (%): Enter the significance level used to determine if the differences are significant or not.

Charts tab:

Stacked bars: Activate this option to display a stacked bars chart that allows visualizing the relative frequencies of the various categories of the JAR scale.

- 3D: Activate this option to display the stacked bars in three dimensions.

Summary: Activate this option to display the charts that summarize the multiple comparisons of the penalty analysis.

Mean drops vs %: Activate this option do display the chart that displays the mean drops as a function of the corresponding % of the population of testers.

Results

After the display of the basic statistics and the correlation matrix for the liking scores and the JAR data, XLSTAT displays a table that shows for each JAR dimension the frequencies for the 5 levels (or 7 or 9 depending on the selected scale). The corresponding stacked bar diagram is then displayed.

The table of the collapsed data on three levels is then displayed, followed by the corresponding relative frequencies table and the stacked bar diagram.

The penalty table allows to visualize the statistics for the 3 point scale JAR data, including the means, the mean drops, the penalties and the results of the multiple comparisons tests.

Last, the summary charts allow to quickly identify the JAR dimensions for which the differences between the JAR category and the 2 non-JAR categories (“Not enough”, “Too much”) are significantly different: when the difference is significant, the bars are displayed in red color,
whereas they are displayed in green color when the difference is not significant. The bars are displayed in grey when the size of a group is lower than the select threshold (see the Options tab of the dialog box).

The **mean drop vs %** chart displays the mean drops as a function of the corresponding % of the population of testers. The threshold % of the population over which the results are considered significant is displayed with a dotted line.

**Example**

A tutorial on penalty analysis is available on the Addinsoft website:

http://www.xlstat.com/demo-pen.htm

**References**

CATA data analysis

Use this function to analyse CATA (check-all-that-apply) data quickly and efficiently. If the CATA survey includes preference data, this tool can be used to identify drivers of liking or attributes that consumers consider as negative.

Description

CATA (check-all-that-apply) surveys have become more and more popular for sensory product characterization since 2007, when it was presented by Adams et al. CATA surveys allow to focus on consumers, more representative of the market, instead of trained assessors. They are easy to set up and easy for participants to answer. The principle is that each assessor receives a questionnaire with attributes or descriptors that the respondent may feel, or not, that they apply to one or more products. If it does, he simply needs to check the attribute, otherwise he does not need to do anything. Other questions on different scales may be added to relate the attributes to preferences and liking scores. If participants are asked to give an overall rating to each product of the study, then further analyses and preference modelling is possible. Ares et al. (2014) recommend to randomize the order of the CATA questions between assessors to improve the reproducibility.

The CATA data analysis tool of XLSTAT has been developed to automate the analysis of CATA data. Let us consider that N assessors were surveyed for P products (one of the products can be a virtual, often ideal, product) on K attributes. The CATA data for the K attributes are assumed to be recorded in a binary format (1 for checked, 0 for not checked). Three formats are currently accepted by XLSTAT:

1. Horizontal format (P x K x N): XLSTAT expects that you have in Excel, a table with P rows, and N groups of K columns all next to each other. You will then only need to specify the value of N, from which XLSTAT will guess K. If you asked each assessor to give his liking, you can add that column within each group of K columns at a position you can let XLSTAT know. In that case each group will have K+1 columns. If one of the products is an ideal product, you can specify its position.

2. Horizontal format (N x K x P): XLSTAT expects that you have in Excel, a table with N rows, and P groups of K columns all next to each other. You will then only need to specify the value of P, from which XLSTAT will guess K. If you asked each assessor to give his liking, you can add that column within each group of K columns at a position you can let XLSTAT know. In that case each group will have K+1 columns. If one of the products is an ideal product, you can specify its position.

3. Vertical format ((N x P) x K): XLSTAT expects that you have in Excel, a table with P x N rows, and K columns. You will then need to select that table. In two additional fields, you need
to select the product identifier and the assessor identifier. If you asked each assessor to rate
the products, you need to select the column that corresponds to the preference data. If one of
the products is an ideal product, you can specify its name so that XLSTAT identifies it.

The analyses performed by XLSTAT on CATA data are based on the article by Meyners et al.
(2013) who investigated in depth the possibilities offered by CATA data.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from
the selection of data to the display of results. You will find below the description of the various
elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- ![Reload](reload_icon.png): Click this button to reload the default options.
- ![Delete](delete_icon.png): Click this button to delete the data selections.

**General** tab:

**CATA data (0/1)**: Select the CATA data (0/1).

**Data format**: Choose the format of the data that corresponds to the layout of the CATA data. It
can be either **horizontal** or **vertical** (see the description section for further details). If column
headers have been selected, check that the "Labels included" option has been activated.

If the format is **horizontal**:

- *(P x K x N)*
  - **Number of assessors**: Enter the number of assessors (N). XLSTAT will guess the
    number of attributes (K).

- *(N x K x P)*
  - **Number of products**: Enter the number of products (P). XLSTAT will guess the
    number of attributes (K).
**Position of the ideal product**: Choose if the ideal product is at a given position in the CATA table, or if it is at the last position.

**Preference data**: Choose if the preference (liking) data are at a given position in the CATA table, or if it is at the last position. There must be one preference column for each assessor and one value for each product. It can be missing for the ideal product.

**Product labels**: Activate this option if Product labels are available. Then select the corresponding data. If the "Labels included" option is activated you need to include a header in the selection.

**Assessor labels**: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection.

If the format is **vertical** \( ((N \times P) \times K) \):

**Products**: Select the data corresponding to the tested products. Only one column has to be selected. If column headers have been selected, check that the "Labels included" option has been activated.

**Assessors**: Select the data corresponding to the assessors. Only one column has to be selected. If column headers have been selected, check that the "Labels included" option has been activated.

**Preference data**: If preference data are available, activate this option and select the data. Only one column has to be selected. If column headers have been selected, check that the "Labels included" option has been activated.

**Ideal product**: Activate this option if the assessors have qualified an ideal product, and specify how the ideal product is named in the Products field.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Labels included**: Activate this option if the data selections include a header.

**Options (1) tab:**

**Cochran's Q test**: Activate this option to run a Cochran’s Q test.
Multiple pairwise comparisons: Select the method to use for the multiple pairwise comparisons: McNemar (Bonferroni) for testing each pair with a McNemar test where the significance level is amended using the Bonferroni approach, or Marascuilo to use the procedure suggested by Marascuilo and McSweeney (1977).

Filter out non significant attributes: Activate this option to remove the attributes for which the Cochran’s Q tests is not significant for a threshold you can choose.

Correspondence analysis:

Distance: Select the distance to be used for the correspondence analysis (CA): Chi-Square for classical CA, or Hellinger if some terms have low frequencies.

Independence test: Activate this option to run an independence test on the contingency table.

Significance level (%): Enter the significance level for the test. This value is also used to determine when Cochran’s Q tests are significant.

Filter factors: You can activate one of the two following options in order to reduce the number of factors displayed:

- Minimum %: Activate this option and then enter the minimum percentage that should be reached to determine the number of factors to display.
- Maximum number: Activate this option to set the maximum number of factors to take into account when displaying the results.

Options (2) tab:

Filter out products: Activate this option to be able to choose on which products the CATA analysis is performed.

Filter out assessors: Activate this option to be able to choose on which assessors the CATA analysis is performed.

Threshold for population size: Enter the % of the total population that should represent a category to be taken into account for the mean impact analysis within penalty analysis.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Replace missing data by 0: Activate this option if you consider that missing data are equivalent to 0.
Remove the observations: Activate this option to remove the observations with missing data.

Results

Cochran’s Q test

Cochran’s Q tests are ran on the Assessors x Products table, independently for each attribute. The first column of the results table gives for each attribute (in rows) the p-values. Pairwise comparisons based on the McNemar-Bonferroni or Marascuilo approach are performed. The next columns give the proportion of assessors that checked the product for the given assessor. The letters in parentheses are only important to be considered if the p-value is significant. They can be used to identify the products responsible of a rejection of the null hypothesis that there is no difference between products. The Cochran’s Q test is equivalent to a McNemar test if there are only two products.

Correspondence Analysis

CATA data are summarized in a contingency table (sum of the N individual CATA tables (the maximum value for each cell is N)). A Correspondance Analysis (CA) is ran to visualize the contingency table. The CA can be based on the chi-square distance or the Hellinger distance (also known as the Bhattacharya distance, which is how it is referred to in the similarity/dissimilarity tool of XLSTAT). The Hellinger distance between two samples depends only on the profiles of these two samples. Hence, the analysis based on the Hellinger distance might be used when the dataset includes terms with low frequency (Meyners et al., 2013). Attributes with null marginal sum are removed from the correspondence analysis. The following results are displayed: contingency table, test of independence between the rows and the columns, eigenvalues and percentages of inertia, and symmetric or asymmetric plot (respectively for the chi-square and Hellinger option).

Principal Coordinate Analysis

The tetrachoric correlations (well suited for binary data) between attributes and, when liking scores are available, the biserial correlations (developed to measure the correlation between a binary and a quantitative variable) between liking and attributes are computed and visualized using a Principal Coordinate Analysis (PCOA). The eigenvalues and percentage of inertia and the principal coordinates together with a graphical representation are displayed. The proximities between attributes can be analysed.

Penalty Analysis

If liking scores are available, a penalty analysis is performed. When an ideal product has been evaluated, two analysis are ran, for the must have attributes (P(No)|(Yes) and P(Yes)|(Yes)) and the nice to have attributes (P(Yes)|(No) and P(No)|(No)). In the case where there is no ideal product, these analyses are substituted by a single analysis of presence and absence of the attributes.
A summary table shows the frequencies with which the two situations (P(No)|(Yes) and P(Yes)|(Yes) or P(Yes)|(No) and P(No)|(No) or presence and absence) occurs for each attribute.

The comparison table displays the mean drops in liking between the two situations for each attribute and their significances. This table is illustrated with the mean impact display plot and the mean drops vs % plot. In the case where there is an ideal product the must have and the nice to have analysis are summarized in one mean drops vs % plot.

**Attribute analysis**

A set of K (one for each attribute) 2x2 tables is displayed, with on the left, the values recorded for the ideal product and at the top, the values obtained for the surveyed products. The table contains the average liking (averaged over the assessors and the products) and the % of all records that correspond to this combination of 0s and/or 1s).

<table>
<thead>
<tr>
<th>Ideal\Products</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>6.2 (12%)</td>
<td>7.4 (8%)</td>
</tr>
<tr>
<td>1</td>
<td>5.1 (39%)</td>
<td>7.2 (41%)</td>
</tr>
</tbody>
</table>

For a given attribute,

- If the attribute is checked for the ideal product (second row), then if the preference for the products that are checked (cell [1,1]) is higher than when it is not checked (cell [1,0]), then the attribute is a "must have".
- Symmetrically, if the attribute is not checked for the ideal product (first row), then if the preference for the products that are not checked (cell [0,0]) is higher than when it is not checked (cell [0,1]), then the attribute is a "must not have".
- If the attribute is not checked for the ideal product (first row), and if the preference for the products that are checked (cell [0,1]) is about the same (in XLSTAT we have set this as an absolute difference less than one) as when it is not checked (cell [0,0]), then the attribute is a "does not harm".

Some tables could correspond to 3 cases. XLSTAT will only associate each table to one case, but you might want to control the results. XLSTAT will try to relate each 2x2 table to one of the rules defined above in the same order.

**Example**

A tutorial on CATA data analysis is available on the Addinsoft website:

[http://www.xlstat.com/demo-catadata.htm](http://www.xlstat.com/demo-catadata.htm)
References


Sensory shelf life analysis

This tool enables you to run sensory shelf life test using assessors’ judgments. It is used to find the optimal period for consuming a product using sensorial judgments. XLSTAT-MX uses the parametric survival models to model the shelf life of a product.

Description

Sensory shelf life analysis is used to evaluate the ideal period for consumption of a product using sensory evaluation of assessors at different times/dates.

It may happen that the physico-chemical properties of a product are not sufficient to assess the quality of a product with respect to the period in which it is consumed. Frequently, adding sensory evaluation of the product will highlight the best consumption period. In the example of a yogurt, you may have a product that is suitable for consumption but in a sensory evaluation will be too acid or after a certain period will look less attractive.

Methods conventionally used in the analysis of survival data are applicable in this case.

Generally, when conducting this type of sensory tests, the assessors taste the same product at different times/dates. This can be done in different sessions, but it is generally recommended to prepare a protocol that allows to obtain products with different seniority for the test day.

Each assessor will express its opinion on the tested product (like / do not like) and we thus obtain a table of preferences per assessor for each date.

Two formats of input can be used in XLSTAT-MX:

- An assessor x date table: each column represents a date, each row represents an assessor. There will be two different values depending on the preference of the assessor (like / not like).

- A date column and a column with the assessors’ names. For each assessor, one enters the date when the change in his preference has been observed. We assume that all judges like the product for the first tasting.

XLSTAT-MX then uses a parametric survival model to estimate a model for the shelf life of the product.

As the exact dates the assessor has change his preference are not known, we use the notion of censoring to set these dates. Thus, if preferences have been collected each week, if an assessor does not like a product after 3 weeks, this assessor is censored by interval between the 2nd and 3rd week. Assume that an assessor appreciates the product all long the study, this assessor is right censored at the last date of the study. Finally, if the assessor likes the
product then does not like it and likes it again later in the study, we consider this assessor is left censored at the last date he changed his preference.

For more details on parametric survival models and censoring, please see the chapter dedicated to these methods. XLSTAT-MX can use an exponential, a Weibull or a log-normal distribution.

As outputs, you will find graphics and parameters of the model.

XLSTAT-MX can also add external information to the model using qualitative or quantitative variables associated to each assessor.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- ![Reload Default Options](image)
  - Click this button to reload the default options.
- ![Delete Data Selections](image)
  - Click this button to delete the data selections.

**General** tab:

Two data format are available (please see the description chapter of this help).

**For the “one column per date” option:**

**Assessor x Date table**: Select the table corresponding to the assessors’ preference for each date. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

**Date data**: Select the data that correspond to the times/dates that have been recorded. If a column header has been selected on the first row, check that the "Column labels" option has been activated.
**Positive code:** Enter the code used to identify an assessor that appreciated the product. Default value is 1.

**Negative code:** Enter the code used to identify an assessor that did not appreciate the product. Default value is 0.

For the “one row per assessor” option:

**Date data:** Select the data that correspond to the times or the dates when it has been observed that an assessor has changed his preference. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

**Assessors:** Select the data that identify the assessor associated to the event. If a column header has been selected on the first row, check that the “Column labels” option has been activated.

**Explanatory variables:**

**Quantitative:** Activate this option if you want to include one or more quantitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The data selected may be of the numerical type. If the variable header has been selected, check that the "Column labels" option has been activated.

**Qualitative:** Activate this option if you want to include one or more qualitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Column labels" option has been activated (see description).

**Range:** Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Activate this option to display the results in a new worksheet of the active workbook.

**Workbook:** Activate this option to display the results in a new workbook.

**Column labels:** Activate this option if the first row of the data selections (time, status and explanatory variables labels) includes a header.

**Distribution:** Select the distribution to be used to fit your model. XLSTAT-MX offers different distributions including Weibull, exponential, extreme value…

**Assessors’ labels:** In the case of the “one column per date” format, activate this option if you want to select the assessors’ names. If a column header has been selected, check that the "Variable labels” option is activated.
Options tab:

Significance level (%): Enter the significance level for the comparison tests (default value 5%). This value is also used to determine the confidence intervals around the estimated statistics.

Initial parameters: Activate this option if you want to take initial parameters into account. If you do not activate this option, the initial parameters are automatically obtained. If a column header has been selected, check that the "Variable labels" option is activated.

Tolerance: Activate this option to prevent the initial regression calculation algorithm taking into account variables which might be either constant or too correlated with other variables already used in the model (0.0001 by default).

Constraints: When qualitative explanatory variables have been selected, you can choose the constraints used on these variables:

a1 = 0: Choose this option so that the parameter of the first category of each factor is set to 0.
an = 0: Choose this option so that the parameter of the last category of each factor is set to 0.

Stop conditions:

- **Iterations**: Enter the maximum number of iterations for the Newton-Raphson algorithm. The calculations are stopped when the maximum number if iterations has been exceeded. Default value: 100.

- **Convergence**: Enter the maximum value of the evolution of the log of the likelihood from one iteration to another which, when reached, means that the algorithm is considered to have converged. Default value: 0.000001.

Model selection: Activate this option if you want to use one of the two selection methods provided:

- **Forward**: The selection process starts by adding the variable with the largest contribution to the model. If a second variable is such that its entry probability is greater than the entry threshold value, then it is added to the model. This process is iterated until no new variable can be entered in the model.

- **Backward**: This method is similar to the previous one but starts from a complete model.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.
**Remove observations:** Activate this option to remove the observations with missing data.

**Outputs** tab:

**Descriptive statistics:** Activate this option to display descriptive statistics for the variables selected.

**Dates statistics:** Activate this option to display statistics for each time/date.

**Goodness of fit statistics:** Activate this option to display the table of goodness of fit statistics for the model.

**Model coefficients:** Activate this option to display the table of coefficients for the model. The last columns display the hazard ratios and their confidence intervals (the hazard ratio is calculated as the exponential of the estimated coefficient).

**Residuals and predictions:** Activate this option to display the residuals for all the observations (standardized residuals, Cox-Snell residuals). The value of the estimated cumulative distribution function, the hazard function and the cumulative survival function for each observation are displayed.

**Quantiles:** Activate this option to display the quantiles for different values of the percentiles (1, 5, 10, 25, 50, 75, 90, 95 and 99 %).

**Charts** tab:

**Preference plot:** Activate this option to display the chart corresponding to the number of assessors that likes the product at each date/time.

**Preference distribution function:** Activate this option to display the charts corresponding to the cumulative preference distribution function (equivalent to the cumulative survival function).

**Residuals:** Activate this option to display the residual charts.

**Results**

XLSTAT displays a large number of tables and charts to help in analysing and interpreting the results.

**Assessors removed from the analysis:** This table displays the assessors that have been removed from the analysis due to a bad coding.

**Summary statistics:** This table displays descriptive statistics for all the variables selected. For the quantitative variables, the number of missing values, the number of non-missing values,
the mean and the standard deviation (unbiased) are displayed. For qualitative variables, the categories with their respective frequencies and percentages are displayed.

**Dates statistics:** This table displays the number of judges that like the product at each date/time. The associated percentage is also displayed.

**Summary of the variables selection:** When a selection method has been chosen, XLSTAT displays the selection summary.

**Goodness of fit coefficients:** This table displays a series of statistics for the independent model (corresponding to the case where there is no impact of covariates, beta=0) and for the adjusted model.

- **Observations:** The total number of observations taken into;
- **DF:** Degrees of freedom;
- **-2 Log(Like.):** The logarithm of the likelihood function associated with the model;
- **AIC:** Akaike's Information Criterion;
- **SBC:** Schwarz's Bayesian Criterion;
- **Iterations:** Number of iterations until convergence.

**Model parameters:** The parameter estimate, corresponding standard deviation, Wald's Chi², the corresponding p-value and the confidence interval are displayed for each variable of the model. Confidence intervals are also displayed.

The residual and predictions table shows, for each observation, the time variable, the censoring variable, the value of the residuals, the cumulative distribution function, the cumulative survival function and the hazard function.

The quantiles associated to the preference curve are presented in a specific table.

**Charts:** Depending on the selected options, charts are displayed. The cumulative preference function or the residuals’ plot can be displayed.

**Example**

A tutorial on how to test sensory shelf life is available on the Addinsoft website:

[http://www.xlstat.com/demo-shelflife.htm](http://www.xlstat.com/demo-shelflife.htm)
References


Generalized Bradley-Terry model

Use this tool to fit a Bradley-Terry model to data obtained from pairwise comparisons.

Description

The generalized Bradley-Terry model is used to describe possible outcomes when elements of a set are repeatedly compared with one another in pairs. Consider a set of K elements.

The generalized Bradley-Terry model

For two elements $i$ and $j$ compared in pairs, Bradley and Terry (1952) suggested the following model to evaluate the probability that $i$ is better than $j$ (or $i$ beats $j$):

$$P(i > j) = \frac{\lambda_i}{\lambda_i + \lambda_j},$$

where $\lambda_i$ is the skill rating of element $i$, $\lambda_i \geq 0$.

Several extensions have been proposed for this model. For instance, Agresti (1990) proposed to handle with home-field advantage and Rao and Kupper (1967) developed a model where ties are allowed.

To account for home-field advantage, Agresti (1990) added a parameter $\delta$ which measures the strength of this advantage.

$$P(i > j) = \begin{cases} \frac{\delta \lambda_i}{\delta \lambda_i + \lambda_j} & \text{if } i \text{ at home} \\ \frac{\lambda_i}{\lambda_i + \delta \lambda_j} & \text{if } j \text{ at home} \end{cases}$$

In the case where ties are allowed between two elements $i$ and $j$, Rao et Kupper (1967) proposed to include a parameter $\theta$ in the model such that:

$$P(i > j) = \frac{\lambda_i}{\lambda_i + \theta \lambda_j}$$

$$P(i = j) = \frac{(\theta^2 - 1)\lambda_i \lambda_j}{(\lambda_i + \theta \lambda_j)(\theta \lambda_i + \lambda_j)}$$

with $\theta > 1$. 

763
Inference of model parameters

In the case of a usual Bradley-Terry model, a maximum likelihood estimator of the parameters can be obtained using a simple iterative MM algorithm (Maximization-Minimization, Hunter (2004)). The model (with or without home-field advantage) can be rewritten as a logistic regression model. In this case, a numerical algorithm is used to determine an estimate of the parameters.

In 2012, by considering parameters as random variables, Caron and Doucet proposed a Bayesian approach to overcome the difficulties related to data sparsity. Two methods can be considered:

- Maximizing the log-likelihood by an EM algorithm. For a specific prior distribution on the parameters, this algorithm corresponds to the classical MM algorithm.
- Estimating a posterior distribution of the parameters via a Gibbs sampler.

These two approaches rely on the introduction of latent variables such that the complete likelihood can be written simply. Assume that \( \omega_{ij} \) is the number of comparisons where i beats j, \( \omega_i = \sum_{j=1, j\neq i}^K \omega_{ij} \) is the total number of wins of element i and \( n_{ij} = \omega_{ij} + \omega_{ji} \) the total number of comparisons between i and j. From the Thurstone interpretation (Diaconis (1988)), the Bradley-Terry model can be written as:

\[
P(Y_{ki} < Y_{kj}) = \frac{\lambda_i}{\lambda_i + \lambda_j},
\]

where \( Y_{ki} \sim \mathcal{E}(\lambda_i) \) and \( k \in \{1, \ldots, n_{ij}\} \). To simplify the complete likelihood, a new latent variable \( Z_{ij} \) is defined such that:

\[
Z_{ij} = \sum_{k=1}^{n_{ij}} \min(Y_{kj}, Y_{ki}) \sim \Gamma(n_{ij}, \lambda_i + \lambda_j).
\]

In a Bayesian framework, a prior distribution is defined for each parameter. Hence, we assume that the parameters \( \lambda_i \) are distributed according to a Gamma distribution with parameters a and b:

\[
P(\lambda) = \prod_{i=1}^K \Gamma(\lambda_i; a, b).
\]
The prior distribution of the home-field parameter $\delta$ is a Gamma $\Gamma(\delta; a_\delta, b_\delta)$ and a flat improper distribution on $[1, +\infty]$ is adopt for the ties parameter $\theta$.

**Bayesian EM:** This iterative approach aims at maximizing the expected log-likelihood.

**Usual model:**

At the $t$-th iteration, the estimate of the parameter $\lambda_i$ is given by:

$$
\lambda_i^{(t)} = \frac{a - 1 + \omega_i}{b + \sum_{j \neq i} \frac{n_{ij}}{\lambda_i^{(t-1)} + \lambda_j^{(t-1)}}},
$$

If $a = 1$ and $b = 0$ this estimate corresponds to the MM one.

**Model with home-field advantage:**

At the $t$-th iteration, the estimates of the parameters $\lambda_i$ and $\delta$ are:

$$
\lambda_i^{(t)} = \frac{a - 1 + \omega_i}{b + \sum_{j \neq i} \frac{\delta^{(t-1)} n_{ij}}{\lambda_i^{(t-1)} + \lambda_j^{(t-1)}}} + \frac{n_{ii}}{\lambda_i^{(t-1)} + \lambda_i^{(t-1)}}
$$

$$
\delta^{(t)} = \frac{a_\delta - 1 + c}{b_\delta + \sum_{j \neq i} \frac{\delta^{(t-1)} n_{ij}}{\lambda_i^{(t-1)} + \lambda_j^{(t-1)}}},
$$

where $c = \sum_{i \neq j} a_{ij}$ and $a_{ij}$ is the number of comparisons where $i$ beats $j$ when $i$ is at home.

**Model with ties:**

Denote by $t_{ij}$ the number of ties between $i$ and $j$. At the $t$-th iteration, the estimates of the parameters $\lambda_i$ and $\theta$ are:
\[ \lambda_i^{(t)} = \frac{a - 1 + s_i}{b + \sum_{j \neq i} \lambda_j^{(t-1)} + \theta^{(t-1)} \lambda_j^{(t-1)} + \theta^{(t-1)} \lambda_j^{(t-1)} + \lambda_j^{(t-1)}} , \]

where \( s_j = \omega_j + t_j \) and \( s_i = \sum_{j \neq i} s_j \). We have:

\[ \theta^{(t)} = \frac{1}{2c^{(t)}} + \sqrt{1 + \frac{1}{4c^{(t)^2}}} , \]

\[ c^{(t)} = \frac{2}{T} \sum_{j \neq i} \frac{s_j \lambda_j^{(t)}}{\lambda_j^{(t-1)} + \theta^{(t-1)} \lambda_j^{(t-1)}} , \]

with \( T = \frac{1}{2} \sum_{j \neq i} t_j \) the total number of ties.

**Sampling:** this approach is based on the Gibbs sampler.

**Usual model:**

We used the following algorithm to estimate parameter \( \lambda_i \):

1. For \( 1 \leq i < j \leq K \) s.t. \( n_{ij} > 0 \),
   \[ Z_{ij}^{(t)} | X, \lambda_i^{(t-1)} \sim \Gamma(n_{ij}, \lambda_i^{(t-1)} + \lambda_j^{(t-1)}) \]

2. For \( 1 \leq i \leq K \),
   \[ \lambda_i^{(t)} | X, Z^{(t)} \sim \Gamma(a + \omega_i, b + \sum_{i < j, n_{ij} > 0} Z_{ij}^{(t)} + \sum_{j < i, n_{ij} > 0} Z_{ji}^{(t)}) \]

**Model with home-field advantage:**

We used the following algorithm to estimate the parameters \( \lambda_i \) and \( \delta \):

1. For \( 1 \leq i < j \leq K \) s.t. \( n_{ij} > 0 \),
   \[ Z_{ij}^{(t)} | X, \lambda_i^{(t-1)}, \delta^{(t-1)} \sim \Gamma(n_{ij}, \delta^{(t-1)} \lambda_i^{(t-1)} + \lambda_j^{(t-1)}) \]

2. For \( 1 \leq i \leq K \),
\[
\lambda^{(t)} \mid X, Z^{(t)}, \lambda^{(t-1)} \sim \Gamma(a + \omega_\lambda, b + \delta^{(t-1)} \sum_{i \neq j, i \neq j > 0} Z_{ij}^{(t)} + \sum_{j \neq j \neq j > 0} Z_{jj}^{(t)})
\]

3. Then,

\[
\delta^{(t)} \mid X, Z^{(t)}, \lambda^{(t-1)} \sim \Gamma(a_\delta + c + b_\delta + \sum_{i=1}^{K} \lambda_i^{(t)} \sum_{j \neq j \neq j > 0} Z_{ij}^{(t)})
\]

**Model with ties:**

We used the following algorithm to estimate the parameters \( \lambda_i \) and \( \theta \):

1. For \( 1 \leq i < j \leq K \) s.t. \( n_{ij} > 0 \),

\[
Z_{ij}^{(t)} \mid X, \lambda_i^{(t-1)}, \theta^{(t-1)} \sim \Gamma(s_\theta, \lambda_i^{(t-1)} + \theta^{(t-1)} \lambda_j^{(t-1)})
\]

2. For \( 1 \leq i \leq K \),

\[
\lambda^{(t)} \mid X, Z^{(t)}, \theta^{(t-1)} \sim \Gamma(a + s, b + \sum_{j \neq i, j \neq i > 0} Z_{ij}^{(t)} + \theta^{(t-1)} \sum_{j \neq i, j \neq i > 0} Z_{ji}^{(t)})
\]

3. Then,

\[
\theta^{(t)} \mid X, Z^{(t)}, \lambda^{(t)} \sim P(\theta \mid X, Z^{(t)}, \lambda^{(t)})
\]

With

\[
P(\theta \mid X, Z^{(t)}, \lambda^{(t)}) \propto (\theta^2 - 1)^T \exp(-\sum_{i \neq j, i \neq j > 0} Z_{ij} \theta)
\]

These two methods lead to posterior distributions on the model parameters. However, only the sampling approach allows to estimate the parameters of the complete model.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.

- **Cancel**: Click this button to close the dialog box without doing any computation.
Help: Click this button to display the help.

Reload: Click this button to reload the default options.

Delete: Click this button to delete the data selections.

Change: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

**Data format:** Select the format of the data

- **Two-way table**: Activate this option to select data in a contingency table (wins in rows and losses in column). Only the classical model can be used.

- **Pairs/Variables table**: Activate this option to select data presented in the form of two tables. The pairs table corresponds to the meetings between the elements. The variables table contains the results of each meeting. The first column is the number of wins of the first element and the second column its number of losses. A third optional column can contain the number of ties.

If headers have been selected with the data, make sure the "Labels" option is checked.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Labels**: Activate this option if headers have been selected with the input data.

**Options** tab:

**Inference method**: select the inference method

- **Numerical**: The model is rewritten as a logistic regression (see section [description](#)). Ties are not allowed.
• **Bayesian EM**: The parameters are supposed to be distributed as a Gamma distribution. The inference is done via an EM algorithm which aims at updating the prior distributions. The parameters of the complete model (with home-field advantage and ties) cannot be inferred with this algorithm.

• **Sampling**: The parameters are supposed to be distributed as a Gamma distribution. The posterior distribution is obtained by a Gibbs sampler.

**Options:**

• **Home**: Select this option to take home-field advantage into account. In this case, the order of the elements in the pairs table is of importance. The first element is supposed to be at home.

• **Ties**: Select this option if ties are allowed. If the option is enabled, the variables table must have 3 columns.

**Confidence interval (%)**: Enter the confidence level of the confidence interval of the parameters.

**Stop conditions:**

• **Iterations /Number of simulations**: Maximal number of iterations.

• **Temps maximum**: Maximal allocated time (in second).

• **Convergence**: Threshold of convergence.

**Prior parameter**: This option is active only if the inference method is Bayesian EM or Sampling.

• **Scale**: Scale parameter of the Gamma distribution

• **Shape**: Shape parameter of the Gamma distribution.

**Outputs** tab:

**Descriptive statistics**: Activate this option to compute and display the statistics that correspond to each element.

**Likelihood-based criterion**: Activate this option to calculate and to display the likelihood, the BIC (Bayesian Information Criterion) and AIC (Akaike Information Criterion).
Probabilities of winning: Activate this option to calculate and to display the probabilities of winning according to model options.

Charts tab:

Convergence graph: Activate this option to display the evolution of model parameters for the Sampling approach.

Results

Summary statistics: This table displays the descriptive statistics for each element

Estimated parameters: the estimates of the model parameters are given in this table. The standard error and the confidence interval are also provided for each parameter.

Likelihood-based criterion: In this table, several likelihood-based criteria are given (-2*log(Likelihood), BIC, AIC).

Probabilities of winning: This table provides the probability that element $i$ (in row) beats element $j$ (in column), given the model parameters.

Convergence graph: This chart displays for each parameter the evolution of the parameter and the corresponding confidence interval.

Example

An example of the use of Bradley-Terry model is available on Addinsoft website:

http://www.xlstat.com/demo-bradley.htm

References


**Generalized Procrustes Analysis (GPA)**

Use Generalized Procrustes Analysis (GPA) to transform several multidimensional configurations so that they become as much alike as possible. A comparison of transformed configurations can then be carried out.

**Description**

Procrustes (or Procustes), which in ancient Greek means "the one who lengthens while stretching", is a character of the Greek mythology. The name of the gangster Procrustes is associated to the bed that he used to torture the travelers to whom he proposed the lodging. Procrustes installed his future victim on a bed with variable dimensions: short for the tall ones and long for the small ones. According to case's, he cut off with a sword what exceeded out of the bed or stretched the body of the traveler until bringing the size of the traveler to that of the bed, by using a mechanism that Hephaistos had manufactured for him. In both cases the torment was appalling. Theseus, while traveling to Athens, met the robber, discovered the trap and laid down slantwise on the bed. When Procrustes adjusted the body of Theseus, he did not understand the situation immediately and remained perplexed giving Theseus the time to slit with his sword the brigand in two equal parts.

**Concept**

We define by configuration an n x p matrix that corresponds to the description of n objects (or individuals/cases/products) on p dimensions (or attributes/variables/criteria/descriptors).

We name consensus configuration the mean configuration computed from the m configurations. Procrustes Analysis is an iterative method that allows to reduce, by applying transformations to the configurations (rescaling, translations, rotations, reflections), the distance of the m configurations to the consensus configuration, the latter being updated after each transformation.

Let us take the example of 5 experts rating 4 cheeses according to 3 criteria. The ratings can go from 1 to 10. One can easily consider that an expert tends to be harder in his notation, leading to a shift to the bottom of the ratings, or that another expert tends to give ratings around the average, without daring to use extreme ratings. To work on an average configuration could lead to false interpretations. One can easily see that a translation of the ratings of the first expert is necessary, or that rescaling the ratings of the second expert would make his ratings possibly closer to those of the other experts.

Once the consensus configuration has been obtained, it is possible to run a PCA (Principal Components Analysis) on the consensus configuration in order to allow an optimal visualization in two or three dimensions.
Structure of the data

There exist two cases:

1. If the number and the designation of the p dimensions are identical for the m configurations, one speaks in sensory analysis about conventional profiles.

2. If the number p and the designation of the dimensions varies from one configuration to the other, one speaks in sensory analysis about free profiles, and the data can then only be represented by a series of m matrices of size n x p(k), k=1,2, ..., m.

For the entering of the data, XLSTAT expects an n x (p x m) table, corresponding to m contiguous configurations. If the number of dimensions varies from one configuration to the other, and if P is the maximum number of dimensions over the whole set of configurations, you need to previously add columns of 0 for the missing dimensions of each configuration, so that there are P x m columns in the table. These dummy dimensions are not displayed on the correlations circle chart.

If the labels of the dimensions vary from one configuration to the other, XLSTAT indicates by Var(i) the ith dimension of the configurations, but it keeps the original labels when displaying the correlations circle chart.

Data transposition

It sometimes occurs that the number (m x p) of columns exceeds the limits of Excel. To get around that drawback, XLSTAT allows you to use transposed tables. To use transposed tables (in that case all tables that you want to select need to be transposed), you only need to click the blue arrow at the bottom left of the dialog box, which then becomes red.

Algorithms

XLSTAT is the unique product offering the choice between the two main available algorithms: the one based on the works initiated by John Gower (1975), and the later one described in the thesis of Jacques Commandeur (1991). Which algorithm performs best (in terms of least squares) depends on the dataset, but the Commandeur algorithm is the only one that allows to take into account missing data; by missing data we mean here that for a given configuration and a given observation or row, the values were not recorded for all the dimensions of the configuration. The later case can happen in sensory data analysis if one of the judges has not evaluated a product.
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload**: Click this button to reload the default options.
- **Delete**: Click this button to delete the data selections.
- **Switch**: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab**:

- **Configurations**: Select the data that correspond to the configurations. If a column header has been selected, check that the "Dimension labels" option has been activated.

- **Number of configurations**: Enter the number of contiguous configurations in the configurations table.

- **Number of variables per table**:
  - **Equal**: Choose this option if the number of variables is identical for all the tables. In that case XLSTAT determines automatically the number of variables in each table.
  - **User defined**: Choose this option to select a column that contains the number of variables contained in each table. If the "Variable labels" option has been activated, the first row must correspond to a header.

- **Configuration labels**: Check this option if you want to use the available configuration labels. If you do not check this option, labels will be created automatically (C1, C2, etc.). If a column header has been selected, check that the "Dimension labels" option has been activated.
**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Dimension labels**: Activate this option if the first row (or column if in transposed mode) of the selected data (configurations, configuration labels, object labels) contains a header.

**Object labels**: Check this option if you want to use the available configuration labels. If you do not check this option, labels will be created automatically (Obs1, Obs 2, etc.). If a column header has been selected, check that the "Dimension labels" option has been activated.

**Method**: Select the algorithm you want to use:

- **Commandeur**: Activate this option to use the Commandeur algorithm (see the section description section for further details).

- **Gower**: Activate this option to use the Gower algorithm (see the section description section for further details).

**Options** tab:

**Scaling**: Activate this option to run rescale the matrices during the GPA.

**Rotation/Reflection**: Activate this option to perform the rotation/reflection steps of the GPA.

**PCA**: Activate this option to run a PCA at the end of the GPA steps.

**Filter factors**: You can activate one of the following two options in order to reduce the number of factors which are taken into account after the PCA.

- **Minimum %**: Activate this option then enter the minimum percentage of the total variability that the chosen factors must represent.

- **Maximum Number**: Activate this option to set the number of factors to take into account.

**Tests**:

775
- **Consensus test**: Activate this option pour to use a permutation test that allows to determine if a consensus is reached after the GPA transformations.

- **Dimensions test**: Activate this option pour to use a permutation test that allows to determine what the appropriate number of factors to keep is.

**Number of permutations**: Enter the number of permutations to perform for the tests (Default value: 300)

**Significance level (%)**: Enter the significance level for the tests.

**Stop conditions**:  
- **Iterations**: Enter the maximum number of iterations for the algorithm. The calculations are stopped when the maximum number of iterations has been exceeded. Default value: 100.  
- **Convergence**: Enter the maximum value of the evolution in the convergence criterion from one iteration to another which, when reached, means that the algorithm is considered to have converged. Default value: 0.00001.

**Missing data** tab:

- **Do not accept missing data**: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.  
- **Replace missing data by 0**: Activate this option to replace missing data by 0.  
- **Remove the observations**: Activate this option to remove observations with missing data.

**Outputs** tab:

- **PANOVA table**: Activate this option to display the PANOVA table.  
- **Residuals by object**: Activate this option to display the residuals for each object.  
- **Residuals by configuration**: Activate this option to display the residuals for each configuration.  
- **Scaling factors**: Activate this option to display the scaling factors applied to each configuration.  
- **Rotation matrices**: Activate this option to display the rotation matrices corresponding to each configuration.
The following options are available only if a PCA has been requested:

**Eigenvalues**: Activate this option to display the eigenvalues of the PCA.

**Consensus configuration**: Activate this option to display the coordinates of the dimensions for the consensus configuration.

**Configurations**: Activate this option to display the coordinates of the dimensions for each configuration.

**Objects coordinates**: Activate this option to display the coordinates of the objects after the transformations.

- **Presentation by configuration**: Activate this option to display one table of coordinates per configuration.
- **Presentation by object**: Activate this option to display one table of coordinates per object.

**Charts (PCA) tab:**

The following options are available only if a PCA has been requested:

**Eigenvalues**: Activate this option to display the scree plot.

**Correlations charts**: Activate this option to display the correlations charts for the consensus configuration and individual configurations.

- **Vectors**: Activate this option to display the dimensions in the form of vectors.

**Objects coordinates**: Activate this option to display the maps showing the objects.

- **Presentation by configuration**: Activate this option to display a chart where the color depends on the configuration.
- **Presentation by object**: Activate this option to display a chart where the color depends on the object.

**Full biplot**: Activate this option to display the biplot showing both the objects and the dimensions of all configurations.

**Colored labels**: Activate this option to show variable and observation labels in the same color as the corresponding points. If this option is not activated the labels are displayed in black color.
Type of biplot: Choose the type of biplot you want to display. See the description section of the PCA for more details.

- **Correlation biplot**: Activate this option to display correlation biplots.

- **Distance biplot**: Activate this option to display distance biplots.

- **Symmetric biplot**: Activate this option to display symmetric biplots.

- **Coefficient**: Choose the coefficient whose square root is to be multiplied by the coordinates of the variables. This coefficient lets you to adjust the position of the variable points in the biplot in order to make it more readable. If set to other than 1, the length of the variable vectors can no longer be interpreted as standard deviation (correlation biplot) or contribution (distance biplot).

Charts tab:

**Residuals by object**: Activate this option to display the bar chart of the residuals for each object.

**Residuals by configuration**: Activate this option to display the bar chart of the residuals for each configuration.

**Scaling factors**: Activate this option to display the bar chart of the scaling factors applied to each configuration.

**Test histograms**: Activate this option to display the histograms that correspond to the consensus and dimensions tests.

Results

**PANOVA table**: Inspired from the format of the analysis of variance table of the linear model, this table allows to evaluate the relative contribution of each transformation to the evolution of the variance. In this table are displayed the residual variance before and after the transformations, the contribution to the evolution of the variance of the rescaling, rotation and translation steps. The computing of the Fisher’s F statistic allows to compare the relative contributions of the transformations. The corresponding probabilities allow to determine whether the contributions are significant or not.

**Residuals by object**: This table and the corresponding bar chart allow to visualize the distribution of the residual variance by object. Thus, it is possible to identify for which objects the GPA has been the less efficient, in other words, which objects are the farther from the consensus configuration.
**Residuals by configuration:** This table and the corresponding bar chart allow to visualize the distribution of the residual variance by configuration. Thus, it is possible to identify for which configurations the GPA has been the less efficient, in other words, which configurations are the farther from the consensus configuration.

**Scaling factors for each configuration:** This table and the corresponding bar chart allow to compare the scaling factors applied to the configurations. It is used in sensory analysis to understand how the experts use the rating scales.

**Rotation matrices:** The rotation matrices that have been applied to each configuration are displayed if requested by the user.

**Results of the consensus test:** This table displays the number of permutations that have been performed, the value of $R_c$ which corresponds to the proportion of the original variance explained by the consensus configuration, and the quantile corresponding to $R_c$, calculated using the distribution of $R_c$ obtained from the permutations. To evaluate if the GPA is effective, one can set a confidence interval (typically 95%), and if the quantile is beyond the confidence interval, one concludes that the GPA significantly reduced the variance.

**Results of the dimensions test:** This table displays for each factor retained at the end of the PCA step, the number of permutations that have been performed, the $F$ calculated after the GPA ($F$ is here the ratio of the variance between the objects, on the variance between the configurations), and the quantile corresponding to $F$ calculated using the distribution of $F$ obtained from the permutations. To evaluate if a dimension contributes significantly to the quality of the GPA, one can set a confidence interval (typically 95%), and if the quantile is beyond the confidence interval, one concludes that factor contributes significantly. As an indication are also displayed, the critical values and the p-value that corresponds to the Fisher’s $F$ distribution for the selected alpha significance level. It may be that the conclusions resulting from the Fisher’s $F$ distribution is very different from what the permutations test indicates: using Fisher’s $F$ distribution requires assuming the normality of the data, which is not necessarily the case.

**Results for the consensus configuration:**

**Objects coordinates before the PCA:** This table corresponds to the mean over the configurations of the objects coordinates, after the GPA transformations and before the PCA.

**Eigenvalues:** If a PCA has been requested, the table of the eigenvalues and the corresponding scree-plot are displayed. The percentage of the total variability corresponding to each axis is computed from the eigenvalues.

**Correlations of the variables with the factors:** These results correspond to the correlations between the variables of the consensus configuration before and after the transformations (GPA and PCA if the latter has been requested). These results are not displayed on the circle of correlations as they are not always interpretable.
**Objects coordinates**: This table corresponds to the mean over the configurations of the objects coordinates, after the transformations (GPA and PCA if the latter has been requested). These results are displayed on the objects charts.

Results for the configurations after transformations:

**Variance by configuration and by dimension**: This table allows to visualize how the percentage of total variability corresponding to each axis is divided up for the configurations.

**Correlations of the variables with the factors**: These results, displayed for all the configurations, correspond to the correlations between the variables of the configurations before and after the transformations (GPA and PCA if the latter has been requested). These results are displayed on the circle of correlations.

**Objects coordinates (presentation by configuration)**: This series of tables corresponds to the objects coordinates for each configuration after the transformations (GPA and PCA if the latter has been requested). These results are displayed on the first series of objects charts.

**Objects coordinates (presentation by object)**: This series of tables corresponds to the objects coordinates for each configuration after the transformations (GPA and PCA if the latter has been requested). These results are displayed on the second series of objects charts.

**Example**

A tutorial on Generalized Procrustean Analysis is available on the Addinsoft website. To view this tutorial go to:

http://www.xlstat.com/demo-gpa.htm

**References**


**Semantic differential charts**

Use this method to easily visualize on a chart, ratings given to objects by a series of judges on a series of dimensions.

**Description**

Psychologist Charles E. Osgood has developed the visualization method *Semantic differential* in order to plot the differences between individuals' connotations for a given word. When applying the method, Osgood asked survey participants to describe a word on a series of scales ranging from one extreme to the other (for example favorable/unfavorable). When patterns were significantly different from one individual to the other or from one group of individuals to the other, Osgood could then interpret the Semantic Differential as a mapping of the psychological or even behavioral distance between the individuals or groups.

This method can also be used for a variety of applications:

- Analysis of the experts' perceptions for a product (for example a yogurt) described by a series of criteria (for example, acidity, saltiness, sweetness, softness) on similar scales (either from one extreme to the other, or on a similar likert scale for each criterion). A Semantic differential chart will allow to quickly see which experts agree, and if significantly different patterns are obtained.

- Survey analysis after a customer satisfaction survey.

- Profile analysis of candidates during a recruitment session.

This tool can also be used in sensory data analysis. Here are two examples:

A panel of experts rates (from 1 to 5) a chocolate bar (the object) on three criteria (the "attributes") namely the flavor, the texture, the odor. In this case, the input table contains in cell (i,j) the rating given by the i<sup>th</sup> judge to the product on the j<sup>th</sup> criterion. The semantic differential chart allows to quickly compare the judges.

A panel of experts rates (from 1 to 5) a series of chocolate bars (the objects) on three criteria (the "attributes") namely the flavor, the texture, the odor. In this case, the input table contains in cell (i,j) the average rating given by the judges to i<sup>th</sup> the product on the j<sup>th</sup> criterion. The semantic differential chart allows to quickly compare the objects.
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload options**: Click this button to reload the default options.
- **Delete data**: Click this button to delete the data selections.

**General tab:**

**Data**: Select the data on the Excel worksheet. If a column header has been selected, check that the "Descriptor labels" option has been activated.

Rows correspond to:

- **Objects**: Choose this option to create a chart where values correspond to the abscissa, descriptors to ordinates, and the objects to the lines on the chart.
- **Descriptors**: Choose this option to create a chart where objects correspond to the abscissa, descriptors to ordinates, and the descriptors to the lines on the chart.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Descriptor labels**: Activate this option if the first row of the selected data (data, observation labels) contains a header.

**Observation labels**: Activate this option if observations labels are available. Then select the corresponding data. If the "Descriptor labels" option is activated you need to include a header...
in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 ...).

**Charts** tab:

**Color:** Activate this option to use different colors when displaying the lines corresponding to the various objects/descriptors.

**Grid:** Activate this option to display the grid on the chart.

**Values:** Activate this option to display the values on the chart.

**Results**

The result that is displayed is the Semantic Differential chart. As it is an Excel chart, you can modify it as much as you want.

**Example**

An example of Semantic Differential Charts on the Addinsoft web site:


**References**


TURF Analysis

Use this tool to run a TURF (Total Unduplicated Reach and Frequency) analysis to highlight a group of products that will reach better market share.

Description

The TURF (Total Unduplicated Reach and Frequency) method is used in marketing to highlight a line of products from a complete range of products in order to have the highest market share. From all the products of a brand, we can obtain a subset, which should be the line of products with the maximum reach.

For example, let's consider an ice cream manufacturer producing 30 different flavors and who wants to put forward a line of six flavors that will reach as many consumers as possible. Thus, he submitted a questionnaire to a panel of 500 consumers who scored each flavor on a scale from 1 to 10. The manufacturer believes that the consumer will be satisfied and inclined to choose the flavor if he gives a score above 8. TURF analysis will look for the combination of 6 flavors with greatest reach and frequency.

This method is a simple statistical method. It is based on a questionnaire (with scores on a fixed scale). The analysis runs through every possible combination of products and records for each combination (1) the percentage of those that desire at least 1 product in the given combination (i.e. reach), and (2) the total number of times products are desired in the given combination (i.e. frequency).

XLSTAT offers a variety of techniques to find the best combination of products: The enumeration method will test all the combinations but may be time consuming; the greedy algorithm is very fast but can stop on a local optimum and the fast search algorithm is close from the enumeration method but it is faster and does not guarantee the optimal solution.

Methods

The data used are data from a questionnaire: one row per consumer and one column per product. They should be in the form of scores (Likert scales). XLSTAT allows you to define different scales. However, all notes must be on the same scale. The user chooses an interval in which he considers that the goal is reached (eg scores greater than 8 in 10).

XLSTAT allows you to use three different algorithms to find the right product line:
- **The enumeration method**: All combinations of k products on the p products (k<p) are tested. We retain the combinations that have the highest reach and frequency.

   The reach is defined by:

   \[
   \text{Reach} = \text{Number of consumers that desire at least 1 product in the given combination}
   \]

   This method is accurate but can be highly time consuming if p and k become large (eg for p = 40 and k = 12, there will be 5,586,853,480 combinations).

- **The greedy algorithm**: This algorithm is a simple heuristic that can find a good result very quickly by maximizing the reach.

   It works as follows:

   - Find the product hitting the target more often (highest frequency)
   - Select this product in the combination
   - Remove the observations for which this product has reached the target
   - Repeat until there is no observation to remove or until you are not able to withdraw any observation

   This algorithm is repeated many times with different initial conditions to minimize the risk of falling into local optima. Its advantage lies in its quickness but it does not warrant obtaining the global maximum.

- **The fast algorithm**: This algorithm is based on the same principle as the enumeration method but when no improvement is found after a number of steps, a jump in the combinations is done in order to avoid some "unnecessary" combinations. This algorithm does not guarantee to find the local optimum but it will explore more possibilities than the greedy algorithm and it is less costly than the enumeration method.

**Constraints in TURF**

XLSTAT allows to add constraints on the products in the framework of a TURF analysis. Two kinds of constraints are available:

- Constraints on the products: In that case, you can force a product to be included in a line of products.
Constraints by group: In that case you select a supplementary variable associating a group to each product. When the algorithm runs, at least one product of each group will be included in the line.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload default options**: Click this button to reload the default options.
- **Delete data selections**: Click this button to delete the data selections.
- **Change data handling**: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab**:

**Data**: Select the data in the form of scores on a common scale for all products. If headers have been selected, please check the option "Labels samples" is enabled.

**Subset size**: Select the size of the subset (that is to say the number of products that must be integrated into the product line).

**Other groups**: Select this option if you want to select another group of products in which a subset will also be obtained. Sometimes two categories of products have to be represented in the line. We want a certain number of products for each category. (In the case of the previous example, we may want three ice creams and three sherbets.)

**Scale**: Choose the scale used to rate products. If you select "other", then you must enter a minimum and the maximum of your scale.
**Range:** Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Check this option to display the results in a new worksheet in the active workbook.

**Workbook:** Check this option to display the results in a new workbook.

**Variable labels:** Check this option if the first line of the selections (data, other group) contains a label.

**Observation labels:** Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection.

**Observation weights:** Activate this option if observations weights are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection.

**Constraints on the products:** Activate this option if you want to include some products in all the line of products. If this option is activated, when you press Ok, a new dialog box will enable you to choose which product to include in your lines.

**Constraints by group:** Activate this option if you want that at least one of the products of each group is included in the generated line of products. Select a column in which the group associated to each product is given. This column should have as many elements as the number of products. If the "Variable labels" option is activated you need to include a header in the selection.

**Options** tab:

**Number of combinations displayed:** Enter the number of combinations you want to keep. It is sometimes interesting to look at several combinations that get good reaches and frequencies to select the best product line.

**The objective is attained for scores between __ and ___:** Enter the lower and upper bounds of the scores that will be considered as reaching the objective.

**Method:** Select the method you want to use for analysis TURF. For the list, you can enter a maximum time for the algorithm stops automatically. If the number of combinations is reduced, XLSTAT automatically opts for the enumeration method because it is accurate.

**Missing data** tab:
Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Remove observations: Activate this option to remove the observations that contain missing data.

Outputs tab:

Frequencies: Activate this option to display the frequency at which each product has been chosen by the individuals.

Product lines obtained with the TURF analysis: Activate this option to display the table that summarizes the reach, total frequency and products’ names of each line of product.

Product lines obtained with the TURF analysis (%): Activate this option to display the table with percentage of observations that are reached and the frequency of each product in the line of products.

Contribution to the reach: Activate this option to display a table with the contribution to the reach of each product in a product line.

Two-way table: Activate this option to display a two-way table that summarizes the TURF analysis.
Results

Frequencies by product: This table displays the frequency with which the objective has been reached for each product.

Product lines obtained with the TURF analysis: This table displays for each selected combination: the Reach, the frequency and the name of each product.

Product lines obtained with the TURF analysis (%): This table displays for each selected combination: the percentage of observations for which the objective has been reached, the frequency in percentage, and the frequency in percentage for each product in each combination.

Contribution to the REACH: This table displays the contribution to the reach of each product in each line of product. The contribution is computed as follows: the product that reaches the highest number of respondents is selected, and then the respondents selected with the preceding product are withdrawn from the sample. The product that reaches the highest number of respondents in the new sample is selected, and then the respondents selected with the preceding product are withdrawn from the sample. The method continues until all products of the line of products have been selected. 3 measures are obtained: the contribution to the reach which is the number of respondent reached by that product (removing the respondent already reached). The sum of these values is equal to the reach of the line of products. The percentage of respondent is then displayed. Finally, the relative contribution is displayed; it supposes that 100% is equivalent to the number of respondents that are reached by the line of products.

Two-way table (TURF): This table crosses all the products (rows) and all the line of products (columns). The first and second rows are the reach and reach in percentage for each line. The last column displays the frequency (in percentage) that each product has been included in the generated lines of products. When a product is included in a line, the value in the table is the global frequency divided by the reach (multiplied by 100).

Example

An example of TURF analysis is available at the Addinsoft website:

http://www.xlstat.com/demo-turf.htm

References

Design of experiments for sensory data analysis

Use this tool to create an experimental design in the context of sensory data analysis.

Description

Designing an experiment is a fundamental step for anyone who wants to ensure that data collected will be statistically usable in the best possible way. No use to evaluate products from a panel of judges if the products cannot be compared under statistically reliable conditions. It is also not necessary to have each judge evaluate all products to compare products between them.

This tool is designed to provide specialists in sensory analysis to provide a simple and powerful tool to prepare a sensory evaluation where judges (experts and/or consumers) evaluate a set of products.

When you want a panel of consumers to evaluate a set of products, say 9, the first issue that arises is what is the appropriate number of consumers that should be involved, knowing that there may be technical constraints (a limited number of trained consumers is available), or budgetary constraints. Once the number of consumer is defined, for example 82, arises the question of the maximum number of products that a consumer can evaluate during each session. Rarely for budgetary reasons, but mostly because of physiological constraints: a consumer, even trained, may not necessarily retain its sensory capabilities to rate too many products at once. Imagine that the experiment shows that three products is a maximum for a session and for organizational reasons, only two sessions can be arranged. Each consumer can then evaluate a maximum of 6 products.

It remains to determine which products will be evaluated by each of the 82 consumers in each session, and in what order. It is possible that the order has an influence (this is not the issue here, but a proper design of experiments would allow verifying or invalidating this assumption). To avoid penalizing certain products we should ensure that products are seen as often as possible in the three different positions during each session. Furthermore, it is possible that some sequences of products also have a bearing on sensory assessments. We restrict here to consider pairs of products (carry-over of order 2). As for the order, we will also ensure that different ordered pairs, 72 in our example, be present at a frequency as homogeneous as possible in the design.

When generating the plan we therefore try to reconcile the following three requirements:

- Products must be seen by as many judges as possible and with an overall frequency of the different products as homogeneous as possible,

- Each product must be seen in the different orders during each session, with an overall frequency for each pair (order, product) as homogeneous as possible
- The different ordered pairs of products must be present in the design of experiments with a frequency as homogeneous as possible.

**Measuring the performance of the design**

Let \( N \) be the matrix with rows for products and columns for judges, and containing the frequency in which each judge sees each product during each session. In the sensory designs we are dealing with sensory, \( N \) contains either 0 or 1's (we impose that a judge can evaluate only once a given product during a session) and the marginal sums for columns are constant and equal to \( k \) (the same number of products is being evaluated by each judge).

Matrix \( M = NN' \) contains on its diagonal the frequency in the design of each product, and on the triangular parts the number of times each unordered pair of products has been evaluated by a same judge. This matrix is called the **concurrence matrix**.

Matrix \( A^* = I - qNN'q/k \) (\( q \) is a diagonal matrix with the inverse of the square root of the frequency of each product in the design) is directly related to the products information matrix, and as as consequence, to the variances and covariances of the parameters of the products in the ANOVA model we can compute once the ratings have been recorded. If we want to ensure that the variances of the differences between the parameters associated with the products are as uniform as possible, we need to ensure that the eigenvalues of the \( A^* \) matrix are close to each other.

We define the **A-efficiency** as the harmonic mean of the at most \( p-1 \) nonzero eigenvalues of matrix \( A^* \), and the **D-efficiency** as the geometric mean of the same values. The two criteria are equal in the ideal case where all eigenvalues are equal.

**Balanced Incomplete Block Designs**

A block design is a design in which we study the influence of two factors on one or more phenomena. We know that one factor has an impact that we cannot control, but that is not of interest. So we want to ensure that this factor does not disturb the analysis that we perform once the collected data. For this we make sure that the various levels of other factors are well represented in each block.

In our case, the blocking factor corresponds to the judges, and the factor of interest we want to study corresponds to the products.

A complete block design is a design in which all levels of the factors of interest are present once within each block. For a sensory design, this corresponds to a design where all products are seen once by each judge.

In an incomplete block design, all levels of the factors of interest are not present for all levels of the blocking factor. It is balanced if each level of the factor of interest are present a same
number of times \( r \) in the design, and if each pair of levels of each factor is present the same number of times \( \lambda \).

If \( v \) is the number of products studied, \( b \) the number of judges, \( k \) the number of products seen by each judge, we show that the following conditions are necessary (but not sufficient) to have a balanced incomplete block design:

\[
bk = vr \\
r(k-1) = \lambda(v-1)
\]

In the case where a balanced incomplete block design exists, the optimal value of the two criteria (A and D-efficiency) is known. We have:

\[
E = \frac{v(k-1)}{k(v-1)}
\]

XLSTAT allows users to search an optimal design within the meaning of the A-efficiency or the D-efficiency, and whether in the case of complete plans or in the case of incomplete block designs, whether balanced or not.

**Algorithm to obtain a design**

XLSTAT relies on two different techniques for generating designs. If the “Fast” option is selected by the user and if \((b/v)\) is integer, then XLSTAT uses the cyclic plans initiated by Williams (1949), and studied in great detail by John and Williams (1995). If \((b/v)\) is not an integer or if the “Fast” option has not been required, XLSTAT uses a proprietary (unpublished) algorithm to very quickly generate an appropriate solution and, starting from this solution, it searches by simulated annealing a better solution for a time defined by the user. If the design is a complete or balanced incomplete block design and if the optimum is found, the search is interrupted before the available time runs out.

**Algorithms to improve the column frequency or the carry-over**

Once the design is found (the matrix \( N \) is known), we need to order products to optimize the in terms of column frequency and carry-over (Périnel and Pagès, 2004). We want that each product is present the same number of times at a given position, and that each ordered pair is also present the same number of times. In order to obtain that, XLSTAT uses two matrices: the matrix of column frequencies and the matrix of carry-over. The goal of algorithm is to make as homogeneous as possible these matrices. The matrix of the carry-over is a matrix that shows the position \( ij \) the number of times that product \( i \) precedes product \( j \) in the design. We define a parameter \( \lambda \) which will allow us to favor either obtaining a good carry-over (\( \lambda \) close to 0) or obtaining a column frequency matrix positions closest to the constant matrix (\( \lambda \) close to 1).
The optimization algorithm is iterative. It permutes the ranks of the products for each judge to maximize the following criterion:

\[ \lambda \sum_{i,j} r_{ij}^2 + (1 - \lambda) \sum_{i,j} s_{ij}^2 \]

where \( r_{ij} \) are the elements of the column frequency matrix and \( s_{ij} \) are the elements of the carry-over matrix. As soon as an optimal value is reached or as the maximum number of iterations is reached the algorithm stops.

XLSTAT uses two criteria to verify the quality of the two matrices:

- **MDR (mean deviation of R):** \( MDR = \sum_{i,j} \left( r_{ij} - \bar{r} \right) \), which is the deviation for the elements of the column frequency matrix.

- **MDS (mean deviation of S):** \( MDS = \sum_{i,j} \left( s_{ij} - \bar{s} \right) \), which is the deviation for the elements of the carry-over matrix.

For incomplete block designs, one can calculate the optimal value that should be reached.

**Resolvable designs and improved presentation**

A design is said to be resolvable if it can be divided into \( g \) groups of judges that are such that, within each group, we have a unique instance of each product. Some balanced incomplete block designs have this property. Presenting a design using such a subdivision into groups has the advantage that if some judges do not appear, it is not necessary to rebuild a design of experiment, but simply to ensure that the last experiments are canceled. This approach is also particularly interesting when one wants to put in place several evaluation sessions (see below). A condition for a balanced incomplete block design to be resolvable is that \( v/k \) must be an integer.

Even when they are balanced or not resolvable, XLSTAT tries to present the incomplete block designs so that the products are present at most twice and, if possible once in a group of size \( <v/k> \) (where \(<i> \) equals \( i \) if \( i \) is an integer, and the rounding to the next integer value otherwise). Thus, if judges were finally absent, it does not penalize too much the quality of the design.

**Sessions**

It is sometimes necessary to split sensory evaluations into sessions. To generate a design that takes into account the need for sessions, XLSTAT uses the same initial design for each session and then applies permutations to both rows and columns, while trying to keep as even as possible column frequencies and carry-over. When the designs are resolvable or near
resolvable, the same judge will not be testing twice the same product during two different sessions.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- ![Reload][1]: Click this button to reload the default options.
- ![Delete][2]: Click this button to delete the data selections.

**General** tab:

**Product**: Enter the number of products involved in the experiment.

**Products/Judge**: Enter the number of products that should evaluate each judge. If the session option is activated, you need to enter the number or products evaluated by each judge during each session.

**Judges**: Enter the number of judges evaluating the products.

**Sessions**: Activate this option if the design should comprise more than one tasting session.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.
Judge labels: Activate this option if you want to select on an Excel sheet the labels that should be used for the judges when displaying the results.

Options tab:

Method: Choose the method to use to generate the design.

- **Fast**: Activate this option to use a method that reduces as much as possible the time spent to find a fine design.
- **Search**: Activate this option to define the time allocated to the search for an optimal design. The maximum time must be entered in seconds.

Criterion: Choose the criterion to maximise when searching for the optimal design.

- **A-efficiency**: Activate this option to search for a design that maximizes the A-efficiency.
- **D-efficiency**: Activate this option to search for a design that maximizes the D-efficiency.

** Carry-over vs frequency**: Define here your preference regarding what should be the priority in the second phase when generating the design: choose among homogenizing the frequency of the products rankings (the order in which they are evaluated), or homogenizing the number of times two products are evaluated one after the other (**carry-over**).

- **Lambda**: Let this parameter vary between 0 (priority given to carry-over) and 1 (priority given to column frequency).
- **Iterations**: Enter the maximum number of iterations that can be used for the algorithm that searches for the best solutions.

Product codes: Select how the product codes should be generated.

- **Product ID**: Activate this option to use a simple product identifier (P1,P2, ...).
- **Random code**: Activate this option to use a random three letters code generated by XLSTAT.
- **User defined**: Activate this option to select on an Excel sheet the product codes you want to use. The number of codes you select must correspond to the number of products.
**Outputs** tab:

**Judges x Products table**: Activate this option to display the binary table that shows if a judge rated (value 1) or not (value 0) a product.

**Concurrence table**: Activate this option to display the concurrence that shows how many times two products have been rated by the same judge.

**Judges x Ranks table**: Activate this option to display the table that shows, for each judge, which product is being rated at each step of the experiment.

**Column frequency table**: Activate this option to display the tableau that shows how many times each product has been rated at a given step of the experiment.

**Carry-over table**: Activate this option to display the table that shows how many times each product has been rated just after another one.

**Design table**: Activate this option to display the table that can later be used for an ANOVA, once the ratings given by the judges have been recorded.

**Results**

Once the calculations are completed, XLSTAT indicates the time spent looking for the optimal plan. The two criteria A and D-efficiency are displayed. XLSTAT indicates if the optimal plan has been found (case of a balanced incomplete block design). Similarly, if the plan is resolvable, it is indicated and the group size is specified.

If sessions have been requested, a first set of results is displayed with results taking into account all the sessions. The results for each session are then displayed.

The **Judges x Products table** is displayed to show whether a judge has assessed (value 1) or not (value 0) a product.

The **concurrence table**: shows how many times two products have been rated by the same judge.

The **MDS/MDR table** displays the criteria that allow assessing the quality of the column frequencies and carry-over that have been obtained, compared to the optimal values.

The **Judges x Ranks table** shows, for each judge, which product is being rated at each step of the experiment.

The **column frequency table** shows how many times each product has been rated at a given step of the experiment.

The **carry-over table** shows how many times each product has been rated just after another one.
The **design table** can later be used for an ANOVA, once the ratings given by the judges have been recorded.

**Example**

An example showing how to generate a DOE for sensory data analysis is available at the Addinsoft website:

http://www.xlstat.com/demo-doesenso.htm

**References**


Design of experiments for sensory discrimination tests

Use this tool to create an experimental design in the context of sensory discrimination tests. This tool allows you to generate the setting for a variety of discrimination tests among which the triangle test, the duo-trio test or the tetrad test.

Description

Designing an experiment is a fundamental step for anyone who wants to ensure that data collected will be statistically usable in the best possible way. No use to evaluate products from a panel of assessors if the products cannot be compared under statistically reliable conditions. It is also not necessary to have each assessor evaluate all products to compare products between them.

This tool is designed to provide specialists in sensory analysis a simple and powerful tool to prepare a sensory discrimination test where assessors (experts and/or consumers) evaluate a set of samples.

Before introducing a new product on the market, discrimination testing is an important step. XLSTAT allows you to prepare the tests. XLSTAT allows you to generate combination of products to be presented to your assessors so that they are in the correct settings for that kind of test.

Sensory discrimination test are based on comparing two products that are presented in a specific setting.

When creating your design, you have to know which test you want to apply, the number of assessors and, if possible, the products’ names.

XLSTAT allows you to run:

- Triangle test: 3 products are presented to each assessor in different orders. Within these products, two are similar and the third one is different. Assessors have to identify the product that is different from the others.

- Duo-trio test: Assessors taste a reference product. Then they taste two different products. Assessors must identify the product that is similar to the reference product.

- Two out of five test: five products are presented to the assessors. These products are separated into two groups, the first one with 3 identical products and the second one with 3 identical products. The assessors have to identify the group with 2 identical products.

- 2-AFC test: 2 products are presented to each assessor. The assessors have to tell which product has the highest intensity for a particular characteristic.
- 3-AFC test: 3 samples are presented to each assessor. Two are similar and the third one is different. The assessors have to tell which product has the highest intensity on a particular characteristic.

- Tetrad test: Four products grouped into two groups, with identical products within each group are presented to each assessor. The assessors are asked to distinguish the two groups.

For each test, you can generate a design of experiments obtained using randomization of the available combinations.

You can specify more than one session and add labels to the assessors and products.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- ![ Reload default options ]: Click this button to reload the default options.
- ![ Delete data selections ]: Click this button to delete the data selections.

**General tab**:

**Type of test**: Select the name of the discrimination test you want to use.

**Judges**: Enter the number of judges evaluating the products.

**Sessions**: Activate this option if the design should comprise more than one tasting session.

**Judge labels**: Activate this option if you want to select on an Excel sheet the labels that should be used for the judges when displaying the results.
**Range:** Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Activate this option to display the results in a new worksheet of the active workbook.

**Workbook:** Activate this option to display the results in a new workbook.

**Product codes:** Select how the product codes should be generated.

- **Product ID:** Activate this option to use a simple product identifier (P1,P2, …).
- **Random code:** Activate this option to use a random three letters code generated by XLSTAT.
- **User defined:** Activate this option to select on an Excel sheet the product codes you want to use. The number of codes you select must correspond to the number of products.

**Results**

Once the calculations are completed, XLSTAT displays the question to be asked to the assessors specific to the chosen test.

The next table displays the product that should be tasted by each assessor (one row = one assessor, one column = one sample). The last column is left empty to allow you to enter the result of the tasting.

**Example**

An example showing how to generate a DOE for discrimination test together with the analysis of the results is available at the Addinsoft website:

http://www.xlstat.com/demo-sensotest.htm

**References**


Sensory discrimination tests

Use this tool to perform discrimination test, among which the triangle test, the duo-trio test or the tetrad test.

Description

Before introducing a new product on the market, discrimination testing is an important step. XLSTAT allows you to prepare the tests (see design of experiments for discriminantion tests) and to analyze the results of these tests.

Two models can be used to estimate the parameters of these tests:

- The guessing model;
- The Thurstonian model.

XLSTAT allows you to run:

- Triangle test: 3 products are presented to each assessor in different orders. Within these products, two are similar and the third one is different. Assessors have to identify the product that is different from the others.

- Duo-trio test: Assessors taste a reference product. Then they taste two different products. Assessors must identify the product that is similar to the reference product.

- Two out of five test: five products are presented to the assessors. These products are separated into two groups, the first one with 3 identical products and the second one with 3 identical products. The assessors have to identify the group with 2 identical products.

- 2-AFC test: 2 products are presented to each assessor. The assessors have to tell which product has the highest intensity for a particular characteristic.

- 3-AFC test: 3 samples are presented to each assessor. Two are similar and the third one is different. The assessors have to tell which product has the highest intensity on a particular characteristic.

- Tetrad test: Four products grouped into two groups, with identical products within each group are presented to each assessor. The assessors are asked to distinguish the two groups.

Each of these tests has its own advantages and drawbacks. A complete review on the subject if available in the book by Bi (2008).
Some concepts should be introduced: pC is the probability of a correct answer, pD is a probability of discrimination, pG is the guessing probability, d’ is the d-prime also called Thurstonian delta. These concepts are detailed below.

**Models**

Two models are commonly used in discrimination testing:

The guessing model assumes that consumers are either discriminators or non-discriminators. Discriminators always find the correct answer. Non-discriminators are guessing the answer with a known guessing probability (which depend on the test used). Someone who does not taste a difference will still have 1 chance out of 3 for the triangle test. The proportion of discriminators is the proportion of people who are able to actually detect a difference between the products.

This concept can be expressed as $p_D = \frac{(p_C - p_G)}{(1 - p_G)}$ where pC is the probability of a correct answer and pG is the guessing probability.

In the Thurstonian model, the required parameter is not a probability of discrimination pD but a d’ (d-prime). It is the sensory distance between the two products, where one unit represents a standard deviation.

The assumptions are that the sensory representations of the products are following two normal distributions and that the consumers are not categorized as discriminators/non-discriminators. Consumers are always correct, translating what they perceive. Thus an incorrect answer is translated into closeness between products that leads to an incorrect perception. If d’ is close to 0, then products cannot be discriminated.

For each test, you will have the guessing probability (as in the guessing model) and a psychometric function that link d’ and the proportion of correct answers. These parameters are specific to each test.

We have $p_C = f_{test}(d')$

**Guessing probability**

For each test the guessing probability which is the probability to obtain the correct answer by guessing is equal to:

Triangle test: pG=1/3

Duo-trio test: pG=1/2

Two out of five test: pG=1/10
2-AFC: \( pG=1/2 \)

3-AFC: \( pG=1/3 \)

Tetrad test: \( pG=1/4 \)

**Psychometric functions**

For each test the psychometric function which is the link between \( d' \) and \( pC \) (the probability of a correct answer) is defined by:

**Triangle test:**

\[
pC = f_{\text{triangle}}(d') = 2 \int_{0}^{\infty} \left\{ \Phi \left[ -x\sqrt{3} + d'\sqrt{2/3} \right] + \Phi \left[ -x\sqrt{3} - d'\sqrt{2/3} \right] \right\} \phi(x) dx
\]

**Duo-trio test:**

\[
pC = f_{\text{duo-trio}}(d') = -\Phi \left( d'/\sqrt{2} \right) - \Phi \left( d'/\sqrt{6} \right) + 2\Phi \left( d'/\sqrt{2} \right) \Phi \left( d'/\sqrt{6} \right)
\]

**2-AFC:**

\[
pC = f_{\text{2-AFC}}(d') = \Phi \left( d'/\sqrt{2} \right)
\]

**3-AFC:**

\[
pC = f_{\text{3-AFC}}(d') = \int_{-\infty}^{\infty} \phi(x-d') \Phi[x]^2 dx
\]

**Tetrad test:**

\[
pC = f_{\text{tetrad}}(d') = \int_{-\infty}^{\infty} \phi(x) \Phi[x][1 - \Phi[x-d']]^2 dx
\]

These functions are estimated using the Gauss-Legendre or Gauss-Hermite algorithm for numerical integration.

**Calculating p-value and power**

P-value and power for these tests are obtained using the binomial or normal distribution based on the estimated \( pC \).

**Standard error and confidence intervals for the Thurstonian model parameters**

When using the Thurstonian model, you can obtain standard error and confidence interval for the parameters of interest.

For the probability of a correct answer \( pC \), we have:

\[
SE(pC) = \sqrt{pC(1-pC)/N}
\]

For the probability of discrimination \( pD \), we have:
\[
SE(pD) = \frac{1}{1-Pg} SE(pC)
\]

For the \(d'\), we have:

\[
SE(d') = \frac{1}{f'_{test}(d')} SE(pC)
\]

Where \(f'\) is the derivative of the psychometric function with respect to \(d'\) (Brockhoff and Christensen, 2010).

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload**: Click this button to reload the default options.
- **Delete**: Click this button to delete the data selections.

**General tab:**

**Type of test**: Select the test you want to analyze.

**Method**: Select the model to be used between the Thurstonian model and the guessing model.

**Input data**: Select the type of data you want to select as input. Three options are available depending on the chosen option, others options will be displayed.

**Data selection case:**
**Test results:** Select a column with as many rows as assessors and in which each cell gives the result of the test for each assessor.

**Code for correct:** Enter the code used to identify a correct answer in the selected data.

**Sample size case:**

**Number of assessors:** enter the total number of assessors in the study.

**Number of correct answers:** enter the number of assessors that gave a correct answer to the test.

**Proportion case:**

**Number of assessors:** enter the total number of assessors in the study.

**Proportion of correct answers:** enter the proportion of assessors that gave a correct answer to the test.

The following options will appear only if the Thurstone model is selected.

**Options for the Thurstone model:**

**D-prime:** activate this option if you want to enter a fixed value for $d'$. You can then enter the value in the available textbox.

**pD:** activate this option if you want to enter a fixed value for the proportion of distinguishers. You can then enter the value in the available textbox.

**Estimate:** activate this option if you want XLSTAT to estimate these values using the model.

**Distribution:** select the distribution to be used for the tests.

**Range:** Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Check this option to display the results in a new worksheet in the active workbook.

**Workbook:** Check this option to display the results in a new workbook.

**Column labels:** Activate this option if the first row of the selected data contains a label.

**Significance level (%):** Enter the significance level for the test (default value: 5%).
Results

Summary of selected options: This table displays the parameters selected in the dialog box. The confidence interval for the proportion of discriminating assessors is then displayed. The results that are displayed correspond to the test statistic, the p-value and the power for the test. A quick interpretation is also given.

If the Thurstone model was selected, estimated probabilities and d' are displayed together with their standard error and confidence intervals.

Example

An example of discrimination test in sensory analysis is available on the Addinsoft website at http://www.xlstat.com/demo-sensotest.htm

References


Design of experiments for conjoint analysis

Use this tool to generate a design for a classical conjoint analysis based on full profiles.

Description

The principle of conjoint analysis is to present a set of products (also known as profiles) to the individuals who will rank, rate, or choose some of them.

In an "ideal" analysis, individuals should test all possible products. But it is soon impossible; the capacity of each being limited and the number of combinations increases very rapidly with the number of attributes (if one wants to study five attributes with three categories each, that means already 243 possible products). We therefore use the methods of experimental design to obtain a acceptable number of profiles to be judged while maintaining good statistical properties.

XLSTAT-Conjoint includes two different methods of conjoint analysis: the full profile analysis and the choice based conjoint (CBC) analysis.

Full profiles conjoint analysis

The first step in a conjoint analysis requires the selection of a number of factors describing a product. These factors should be qualitative. For example, if one seeks to introduce a new product in a market, we can choose as differentiating factors: the price, the quality, the durability ... and for each factor, we must define a number of categories (different prices, different lifetimes ...). This first step is crucial and should be done together with experts of the studied market.

Once this first step done, the goal of a conjoint analysis is to understand the mechanism of choice. Why people choose one product over another?

To try to answer this question, we will propose a number of products (combining different categories of the studied factors). We can not offer all possible products, so we will select products by using design of experiments before presenting them to people who will rate them or rank them.

The full profile method is the oldest methods of conjoint analysis; we seek to build an experimental design that includes a limited number of full profiles that each individual interviewed will then rank or rate.

XLSTAT-Conjoint uses fractional factorial designs in order to generate profiles that will then be presented to respondents. When no design is available, XLSTAT-Conjoint uses algorithms to search for D-optimal designs (see description of the module XLSTAT-DOE).
As part of the traditional conjoint analysis, the questionnaires used are based on the rating or ranking of a number of complete profiles.

You have to select the attributes of interest for your product and the categories associated with these attributes. XLSTAT-Conjoint then generates profiles to be ranked / rated by each respondent.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- ![Reload Default Options](image): Click this button to reload the default options.
- ![Delete Data Selections](image): Click this button to delete the data selections.

**General tab:**

**Analysis name**: Enter the name of the analysis you want to perform.

**Number of attributes**: Select the number of attributes that will be tested during this analysis (number of variables).

**Maximum number of profiles**: Enter the maximum number of profiles to be presented to the individuals.

**Number of responses**: Enter the number of expected individuals who will respond to the conjoint analysis.
Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Options tab:

Design of experiments: The generation of the design of experiments is automatic using fractional factorial designs or D-optimal designs (see the chapter on screening design of the XLSTAT-DOE module).

Initial partition: XLSTAT-Conjoint uses a random initial partition. You can decide how many repetitions are needed to obtain your design. XLSTAT-Conjoint will choose the best obtained design.

Stop conditions: The number of iterations and the convergence criterion to obtain the design can be modified.

Factors tab:

Manuel selection: Select this option to enter details on the factors manually. This option is only available if the number of factors is less than 6.

- Short name: Enter the short name of each factor.
- Long name: Enter the long name of each factor.
- Number of categories: Enter the number of categories for each factor.
- Labels: Activate this option if you want to select the names associated with each category. The names will be distributed in a column for each factor.

Selection in a sheet: Select this option to select details on the factors in a sheet.

- Short name: Select a data column in which the short names of the factors are listed.
- Long name: Select a data column in which the long names of the factors are listed.
- Number of categories: Select a data column in which the number of categories for each factor is listed.
- Labels: Activate this option if you want to select the names associated with each modality. The names should be divided by columns in a table.
**Outputs** tab:

**Optimization summary**: Activate this option to display the optimization summary for generating the design.

**Print individual sheets**: Activate this option to print individual sheets for each respondent. Each sheet will include all generated profiles. The respondent has to fill the last column of the table with the rates or ranks associated to each generated profile. Two assignment options are available; the fixed option displays the profiles in the same order for all individuals; the random option displays the profiles in random orders (different from one respondent to another).

**Include references**: Activate this option to include references between the main sheet and the individual sheets. When an individual enter his chosen rating / ranking in the individual sheet, the value is automatically displayed in the main sheet of the analysis.

**Design for conjoint analysis** dialog box:

**Selection of experimental design**: This dialog box lets you select the design of experiments you want to use. A list of fractional factorial designs is presented with their respective distance to the design that was to be generated. If you select a design and you click “Select”, then the selected design will appear in your conjoint analysis. If no design fits your needs, click on the “optimize” button, and an algorithm will give you a design corresponding exactly to the selected factors.

**Results**

**Variable information**: This table displays all the information relative to the used factors.

**Conjoint analysis design**: This table displays the generated profiles. Empty cells associated to each individual respondent are also displayed. If the options “print individual sheets” and “include references” have been activated, then formulas with reference to the individual sheets are included in the empty cells.

**Optimization details**: This table displays the details of the optimization process when a search for a D-optimal design has been selected.

**Individual _Res sheets**: When the “Print individual sheets” option is activated, these sheets include the name of the analysis, the individual number and a table associated to the profiles to be rated / ranked. Individual respondents should fill the last column of this table.
Example

An example of full profile based conjoint analysis is available at the Addinsoft website:

http://www.xlstat.com/demo-conjoint.htm

References


Design for choice based conjoint analysis

Use this tool to generate a design of experiments for a Choice-Based Conjoint analysis (CBC).

Description

The principle of conjoint analysis is to present a set of products (also known as profiles) to the individuals who will note, class, or choose some of them.

In an "ideal" analysis, individuals should test all possible products. But it is soon impossible; the capacity of each being limited and the number of combinations increases very rapidly with the number of attributes (if one wants to study five attributes with three categories each, that means already 243 possible products). We therefore use the methods of experimental design to obtain a acceptable number of profiles to be judged while maintaining good statistical properties.

XLSTAT-Conjoint includes two different methods of conjoint analysis: the full profiles analysis and the choice based conjoint (CBC) analysis.

Choice Based Conjoint analysis (CBC)

The principle of choice based conjoint (CBC) analysis is based on choices in a group of profiles. The individual respondent chooses between different products offered instead of rating or ranking products.

The process of CBC is based on comparisons of profiles. These profiles are generated using the same methods as for full profile conjoint analysis. Then, these profiles are put together in many comparison groups (with a fixed size). The individual respondent then chooses the profile that he would select compared to the other profiles included in the comparison.

The statistical process is separated into 2 steps:

- Fractional factorial designs or D-optimal designs are used to generate the profiles.

- Once the profiles have been generated they are allocated in the comparison groups using incomplete block designs.

The first step in a conjoint analysis requires the selection of a number of factors describing a product. These factors should be qualitative. For example, if one seeks to introduce a new product in a market, we can choose as differentiating factors: the price, the quality, the durability ... and for each factor, we must define a number of categories (different prices, different lifetimes ...). This first step is crucial and should be done together with experts of the studied market.
Once past this first step, the goal of a conjoint analysis is to understand the mechanism for choosing one product over another. Instead of proposing all profiles to the individual respondents and asking to rate or rank them, CBC is based on a choice after a comparison of some of the profiles. Groups of profiles are presented to the individual respondents and they have to indicate which profile they would choose (a no choice option is also available in XLSTAT-Conjoint).

This method combines two designs of experiments, the fractional factorial design to select the profiles to be compared and the incomplete block design to generate the comparisons to be presented. For more details on these methods, please see the screening design chapter of the DOE module help and the DOE for sensory analysis chapter of the MX module help.

XLSTAT-Conjoint enables you to add the no choice option if the individual respondent would not choose any of the proposed profiles.

XLSTAT-Conjoint enables to obtain a global table for CBC analysis but also individual tables for each respondent and each comparison in separated Excel sheets. References are also included so that when a respondent select a profile in an individual sheet, it is directly reported in the main table.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- ![Reload](image)
  : Click this button to reload the default options.
- ![Delete](image)
  : Click this button to delete the data selections.

**General tab:**
**Analysis name:** Enter the name of the analysis you want to perform.

**Number of attributes:** Select the number of attributes that will be tested during this analysis (number of variables).

**Maximum number of profiles:** Enter the maximum number of profiles to be presented to the individuals.

**Number of responses:** Enter the number of expected individuals who will respond to the conjoint analysis.

**Maximum number of comparisons:** Enter the maximum number of comparison to be presented to the individual respondents. This number has to be greater than the number of profiles.

**Number of profiles per comparison:** Enter the number of profiles per comparison.

**Range:** Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Check this option to display the results in a new worksheet in the active workbook.

**Workbook:** Check this option to display the results in a new workbook.

**Options** tab:

**Design of experiments:** The generation of the design of experiments is automatic using fractional factorial designs or D-optimal design (see the chapter on screening design of the XLSTAT-DOE module). For the comparison design, incomplete block designs are used.

**Initial partition:** XLSTAT-Conjoint uses a random initial partition. You can decide how many repetitions are needed to obtain your design. XLSTAT-Conjoint will choose the best obtained design.

**Stop conditions:** The number of iterations and the convergence criterion to obtain the design can be modified.

**Factors** tab:

**Manuel selection:** Select this option to enter details on the factors manually. This option is only available if the number of factors is less than 6.

- **Short name:** Enter the short name of each factor.
- **Long name:** Enter the long name of each factor.
• **Number of categories**: Enter the number of categories for each factor.

• **Labels**: Activate this option if you want to select the names associated with each category. The names will be distributed in a column for each factor.

**Selection in a sheet**: Select this option to select details on the factors in a sheet.

• **Short name**: Select a data column in which the short names of the factors are listed.

• **Long name**: Select a data column in which the long names of the factors are listed.

• **Number of categories**: Select a data column in which the number of categories for each factor is listed.

• **Labels**: Activate this option if you want to select the names associated with each modality. The names should be divided by columns in a table.

**Outputs** tab:

**Optimization summary**: Activate this option to display the optimization summary for generating the design.

**Print individual sheets**: Activate this option to print individual sheets for each respondent. Each sheet will include a table for each comparison. The respondent has to enter the code associated to the profile he would choose in the box at the bottom of each table. Two assignment options are available; the fixed option displays the comparisons in the same order for all individuals; the random option displays the comparisons in random orders (different from one respondent to another).

**Include references**: Activate this option to include references between the main sheet and the individual sheets. When an individual enter his chosen code in the individual sheet, the result is automatically displayed in the main sheet of the analysis.

**Include the no choice option**: Activate this option to include a no choice option for each comparison in the individual sheets.

**Design for conjoint analysis** dialog box:

**Selection of experimental design**: This dialog box lets you select the design of experiment you want to use. Thus, a list of fractional factorial designs is presented with their respective distance to the design that was to be generated. If you select a design and you click Select, then the selected design will appear in your conjoint analysis. If no design fits your needs, click on the “optimize” button, and an algorithm will give you a design corresponding exactly to the selected factors.
Results

Variable information: This table displays all the information relative to the used factors.

Profiles: This table displays the generated profiles using the design of experiments tool.

Conjoint analysis design: This table displays the comparisons presented to the respondent. Each row is associated to a comparison of profiles. The numbers in the rows are associated to the profiles numbers in the profiles tables. Empty cells associated to each individual respondent are also displayed. Respondent have to enter the code associated to the choice made (1 to number of profiles per comparisons; or 0 if the no choice option is selected).

Optimization details: This table displays the details of the optimization process when a search for a D-optimal design has been selected.

Individual _Res sheets: When the “Print individual sheets” option is activated, these sheets include the name of the analysis, the individual number and tables associated to the comparisons with the profiles to be compared. Individual respondents should enter the code associated to their choice in the bottom right of each table.

Example

An example of choice based conjoint (CBC) analysis is available at the Addinsoft website:

http://www.xlstat.com/demo-cbc.htm

References


Conjoint analysis

Use this tool to run a Full Profile Conjoint analysis. This tool is included in the XLSTAT-Conjoint module; it must be applied on design of experiments for conjoint analysis generated with XLSTAT-Conjoint.

Description

Conjoint analysis is a comprehensive method for the analysis of new products in a competitive environment.

This tool allows you to carry out the step of analyzing the results obtained after the collection of responses from a sample of people. It is the fourth step of the analysis, once the attributes have been defined, the design has been generated and the individual responses have been collected.

Full profile conjoint analysis is based on ratings or rankings of profiles representing products with different characteristics. These products have been generated using a design of experiments and can be real or virtual.

The analysis is done using two statistical methods:

- Analysis of variance based on ordinary least squares (OLS).
- Monotone analysis of variance (Kruskal, 1964) that uses monotonic transformations of the responses to better adjust the analysis of variance (MONANOVA).

Both approaches are described in detail in the chapters "Analysis of variance" and "Monotone regression (MONANOVA)" of the help of XLSTAT.

Conjoint analysis therefore provides for each individual what is called partial utilities associated with each category of the variables. These utilities provide a rough idea of the impact of each modality on the process of choosing a product.

In addition to utilities, conjoint analysis provides an importance associated with each variable. It shows how each variable in the selection process associated with each individual is important.

The full profile conjoint analysis details the results for each individual separately, which preserves the heterogeneity of the results.
XLSTAT-Conjoint also proposes to make classifications on the individuals. Using the utilities, XLSTAT-Conjoint will obtain classes of individuals that can be analyzed and be useful for further research. Classification methods used in XLSTAT-Conjoint are the agglomerative hierarchical classification (see the chapter on this subject in the help of XLSTAT) and the k-means method (see the chapter on this subject in the help of XLSTAT).

**Type of data**

XLSTAT-Conjoint offers two types of input data for the conjoint analysis: rankings and ratings. The type of data must be indicated because the treatment used is slightly different.

Indeed, with rankings, the best profile will have the lowest value, whereas with a rating, it will have the highest value.

If the ranking option is selected, XLSTAT-Conjoint transforms the answers in order to reverse this arrangement and so that utilities can be interpreted easily.

**Interactions**

By interaction is meant an artificial factor (not measured) which reflects the interaction between at least two measured factors. For example, if we carry out treatment on a plant, and tests are carried out under two different light intensities, we will be able to include in the model an interaction factor treatment*light which will be used to identify a possible interaction between the two factors. If there is an interaction between the two factors, we will observe a significantly larger effect on the plants when the light is strong and the treatment is of type 2 while the effect is average for weak light, treatment 2 and strong light, treatment 1 combinations.

To make a parallel with linear regression, the interactions are equivalent to the products between the continuous explanatory values although here obtaining interactions requires nothing more than simple multiplication between two variables. However, the notation used to represent the interaction between factor A and factor B is A*B.

The interactions to be used in the model can be easily defined in XLSTAT.

**Constraints**

During the calculations, each factor is broken down into a sub-matrix containing as many columns as there are categories in the factor. Typically, this is a full disjunctive table. Nevertheless, the breakdown poses a problem: if there are g categories, the rank of this sub-matrix is not g but g-1. This leads to the requirement to delete one of the columns of the sub-matrix and possibly to transform the other columns. Several strategies are available depending on the interpretation we want to make afterwards:
1) \(a_1=0\): the parameter for the first category is null. This choice allows us to force the effect of the first category as a standard. In this case, the constant of the model is equal to the mean of the dependent variable for group 1.

2) \(a_n=0\): the parameter for the last category is null. This choice allows us to force the effect of the last category as a standard. In this case, the constant of the model is equal to the mean of the dependent variable for group \(n\).

3) \(\text{Sum}(a_i) = 0\): the sum of the parameters is null. This choice forces the constant of the model to be equal to the mean of the dependent variable when the design is balanced.

Note: even if the choice of constraint influences the values of the parameters, it has no effect on the predicted values and on the different fitting statistics.

**Generating a market**

XLSTAT-Conjoint includes a small tool to automatically generate a market that is to be then simulated using the XLSTAT-Conjoint simulation tool. This tool allows to build the market table using the attributes’ names and the categories’ names. The obtained table can then be used for simulation purposes in a conjoint simulation.

You only need to select the names of the attributes, the names of the categories in a table and the number of products to include in the market (it is also possible to enter the products ID).

Once this information is entered into the dialog box, just click OK, and for each attribute of each product, you will be asked to choose the category to add. When an entire product has been defined, you can either continue with the next product or stop building the table and obtain a partial market table.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
Click this button to reload the default options.

Click this button to delete the data selections.

Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

**Responses:** Select the responses that have been given by the respondents. If headers have been selected, please check the option "Variable labels" is enabled. This selection corresponds to the right part of the conjoint analysis design table generated with the "design of conjoint analysis" tool of XLSTAT-Conjoint.

**Response type:** select the type of response given by the respondents (ratings or rankings).

**Profiles:** Select the profiles that have been generated. If headers have been selected, please check the option "Variable labels" is enabled. This selection corresponds to the right part of the conjoint analysis design table generated with the "design of conjoint analysis" tool of XLSTAT-Conjoint. Do not select the first column of the table.

**Range:** Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Check this option to display the results in a new worksheet in the active workbook.

**Workbook:** Check this option to display the results in a new workbook.

**Variable labels:** Check this option if the first line of the selections (data, other group) contains a label.

**Profiles weights:** Activate this option if profiles weights are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection.

**Response weights:** Activate this option if response weights are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection.
Options tab:

Method: Select the method to be used for estimation.

Interactions / Level: Activate this option to include interactions in the model then enter the maximum interaction level (value between 1 and 4).

Tolerance: Activate this option to prevent the OLS regression calculation algorithm taking into account variables which might be either constant or too correlated with other variables already used in the model (0.0001 by default).

Confidence interval (%): Enter the percentage range of the confidence interval to use for the various tests and for calculating the confidence intervals around the parameters and predictions. Default value: 95.

Constraints: Details on the various options are available in the description section.

- a1 = 0: Choose this option so that the parameter of the first category of each factor is set to 0.
- an = 0: Choose this option so that the parameter of the last category of each factor is set to 0.
- Sum (ai) = 0: for each factor, the sum of the parameters associated with the various categories is set to 0.

Segmentation: Activate this option if you want XLSTAT-Conjoint to apply an individuals based clustering method on the on the partial utilities. Two methods are available: agglomerative hierarchical classification and k-means classification.

- Number of classes: Enter the number of classes to be created by the algorithm for the k-means.
- Truncation: Activate this option if you want XLSTAT to automatically define the truncation level, and therefore the number of classes to retain, or if you want to define the number of classes to create, or the level at which the dendrogram is to be truncated.

Stop conditions: the number of iterations and the convergence criterion for the MONANOVA algorithm can be modified.

Missing data tab:
**Remove observations**: Activate this option to remove the observations with missing data.

**Estimate missing data**: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbour**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.

**Outputs** tab:

**Goodness of fit statistics**: Activate this option to display the table of goodness of fit statistics for the model.

**Type III analysis**: Activate this option to display the type III analysis of variance table.

**Standardized coefficients**: Activate this option if you want the standardized coefficients (beta coefficients) for the model to be displayed.

**Predictions and residuals**: Activate this option to display the predictions and residuals for all the observations.

**Charts** tab:

**Regression charts**: Activate this option to display regression chart:

- **Standardized coefficients**: Activate this option to display the standardized parameters for the model with their confidence interval on a chart.

**Transformation plot**: Activate this option to display the monotone transformation of the responses plot.

**Results**

**Variable information**: This table displays all the information relative to the used factors.

**Utilities (individual data)**: This table displays utilities associated to each category of the factors for each respondent.

**Importance (individual data)**: This table displays importance for each factor of the analysis for each respondent.
**Goodness of fit statistics**: The statistics relating to the fitting of the regression model are shown in this table:

- **Observations**: The number of observations used in the calculations. In the formulas shown below, \( n \) is the number of observations.

- **Sum of weights**: The sum of the weights of the observations used in the calculations. In the formulas shown below, \( W \) is the sum of the weights.

- **DF**: The number of degrees of freedom for the chosen model (corresponding to the error part).

- **\( R^2 \)**: The determination coefficient for the model. This coefficient, whose value is between 0 and 1, is only displayed if the constant of the model has not been fixed by the user. Its value is defined by:

\[
R^2 = 1 - \frac{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} w_i (y_i - \bar{y})^2}, \quad \text{where} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} w_i y_i,
\]

The \( R^2 \) is interpreted as the proportion of the variability of the dependent variable explained by the model. The nearer \( R^2 \) is to 1, the better is the model. The problem with the \( R^2 \) is that it does not take into account the number of variables used to fit the model.

- **Adjusted \( R^2 \)**: The adjusted determination coefficient for the model. The adjusted \( R^2 \) can be negative if the \( R^2 \) is near to zero. This coefficient is only calculated if the constant of the model has not been fixed by the user. Its value is defined by:

\[
\hat{R}^2 = 1 - (1 - R^2) \frac{W - 1}{W - p - 1}
\]

The adjusted \( R^2 \) is a correction to the \( R^2 \) which takes into account the number of variables used in the model.

- **MSE**: The mean squared error (MSE) is defined by:

\[
MSE = \frac{1}{W - p} \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2
\]

- **RMSE**: The root mean square of the errors (RMSE) is the square root of the MSE.

- **MAPE**: The *Mean Absolute Percentage Error* is calculated as follows:

\[
MAPE = \frac{100}{W} \sum_{i=1}^{n} w_i \left| \frac{y_i - \hat{y}_i}{y_i} \right|
\]
• **DW**: The Durbin-Watson statistic is defined by:

\[
DW = \frac{\sum_{i=2}^{n} \left( (y_i - \hat{y}_i) - (y_{i-1} - \hat{y}_{i-1}) \right)^2}{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}
\]

This coefficient is the order 1 autocorrelation coefficient and is used to check that the residuals of the model are not autocorrelated, given that the independence of the residuals is one of the basic hypotheses of linear regression. The user can refer to a table of Durbin-Watson statistics to check if the independence hypothesis for the residuals is acceptable.

• **Cp**: Mallows Cp coefficient is defined by:

\[
Cp = \frac{SSE}{\hat{\sigma}} + 2p^* - W
\]

where SSE is the sum of the squares of the errors for the model with p explanatory variables and \( \hat{\sigma} \) is the estimator of the variance of the residuals for the model comprising all the explanatory variables. The nearer the Cp coefficient is to \( p^* \), the less the model is biased.

• **AIC**: Akaike's Information Criterion is defined by:

\[
AIC = W \ln \left( \frac{SSE}{W} \right) + 2p^*
\]

This criterion, proposed by Akaike (1973) is derived from the information theory and uses Kullback and Leibler's measurement (1951). It is a model selection criterion which penalizes models for which adding new explanatory variables does not supply sufficient information to the model, the information being measured through the MSE. The aim is to minimize the AIC criterion.

• **SBC**: Schwarz's Bayesian Criterion is defined by:

\[
SBC = W \ln \left( \frac{SSE}{W} \right) + \ln (W) p^*
\]

This criterion, proposed by Schwarz (1978) is similar to the AIC and, like this, the aim is to minimize it.

• **PC**: Amemiya's Prediction Criterion is defined by:

\[
PC = \frac{(1 - R^2)(W + p^*)}{W - p^*}
\]
This criterion, proposed by Amemiya (1980) is used, like the adjusted $R^2$ to take account of the parsimony of the model.

- **Press RMSE**: Press' statistic is only displayed if the corresponding option has been activated in the dialog box. It is defined by:

$$
\text{Press} = \sum_{i=1}^{n} w_i \left( y_i - \hat{y}_{i(-i)} \right)^2
$$

where $\hat{y}_{i(-i)}$ is the prediction for observation $i$ when the latter is not used for estimating parameters. We then get:

$$
\text{Press RMSE} = \sqrt{\frac{\text{Press}}{W - p^*}}
$$

Press's RMSE can then be compared to the RMSE. A large difference between the two shows that the model is sensitive to the presence or absence of certain observations in the model.

- **Iteration**: Number of iteration until convergence of the ALS algorithm.

**Utilities (descriptive statistics)**: This table displays minimum, maximum, mean and standard error of the partial utilities associated to each category of the factors.

**Importance (descriptive statistics)**: This table displays minimum, maximum, mean and standard error of the importance for each factor of the analysis.

**Standard deviations table**: This table displays the standard deviation for each utility and each respondent together with the model error. It is useful to apply the RFC-Bolse Approach for market simulation (see the conjoint analysis simulation chapter).

**Goodness of fit coefficients (MONANOVA)**: In this table are shown the statistics for the fit of the regression model specific to the case of MONANOVA. These statistics are the Wilks' lambda, the Pillai's trace, the trace of Hotelling-Lawlet and the largest root of Roy. For more details on these statistics, see the help on the conditional logit model.

If the Type I/II/III SS (SS: Sum of Squares) option is activated, the corresponding tables are displayed.

The table of **Type I SS** values is used to visualize the influence that progressively adding explanatory variables has on the fitting of the model, as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's $F$, or the probability associated with Fisher's $F$. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. The sums of squares in the Type I table always
add up to the model SS. Note: the order in which the variables are selected in the model influences the values obtained.

The table of **Type II SS** values is used to visualize the influence that removing an explanatory variable has on the fitting of the model, all other variables being retained, as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. Note: unlike Type I SS, the order in which the variables are selected in the model has no influence on the values obtained. Type II SS are not recommended in unbalanced designs but we display them as some users might need them. It is identical to Type III for balanced designs.

The table of **Type III SS** values is used to visualize the influence that removing an explanatory variable has on the fitting of the model, all other variables being retained, as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. Note: unlike Type I SS, the order in which the variables are selected in the model has no influence on the values obtained. While Type II SS depends on the number of observations per cell (cell means combination of categories of the factors), Type III does not and is therefore preferred.

The **analysis of variance table** is used to evaluate the explanatory power of the explanatory variables. Where the constant of the model is not set to a given value, the explanatory power is evaluated by comparing the fit (as regards least squares) of the final model with the fit of the rudimentary model including only a constant equal to the mean of the dependent variable. Where the constant of the model is set, the comparison is made with respect to the model for which the dependent variable is equal to the constant which has been set.

The **parameters of the model** table displays the estimate of the parameters, the corresponding standard error, the Student's t, the corresponding probability, as well as the confidence interval.

The table of **standardized coefficients** (also called beta coefficients) are used to compare the relative weights of the variables. The higher the absolute value of a coefficient, the more important the weight of the corresponding variable. When the confidence interval around standardized coefficients has value 0 (this can be easily seen on the chart of standardized coefficients), the weight of a variable in the model is not significant.

The **predictions and residuals** table shows, for each observation, its weight, the observed value of the dependent variable, the transformed value of the dependant variable, the model's prediction, the residuals, and the confidence intervals. Two types of confidence interval are displayed: a confidence interval around the mean (corresponding to the case where the prediction would be made for an infinite number of observations with a set of given values for the explanatory variables) and an interval around the isolated prediction (corresponding to the case of an isolated prediction for the values given for the explanatory variables). The second interval is always greater than the first, the random values being larger.
The chart which follows shows the transformation of the dependant variable.

Example

An example conjoint analysis is available at the Addinsoft website:

http://www.xlstat.com/demo-conjoint.htm

References


Choice based conjoint analysis

Use this tool to run a Choice-Based Conjoint analysis (CBC). This tool is included in the XLSTAT-Conjoint module; it must be applied on design of experiments for choice based conjoint analysis generated with XLSTAT-Conjoint.

Description

Conjoint analysis is a comprehensive method for the analysis of new products in a competitive environment.

This tool allows you to carry out the step of analyzing the results obtained after the collection of responses from a sample of people. It is the fourth step of the analysis, once the attributes have been defined, the design has been generated and the individual responses have been collected.

In the case of CBC models, individuals have to choose between selections of profiles. Thus, a number of choices are given to all individuals (we will select a product from a number of products generated).

Analysis of these choices is made using:

- A multinomial logit model based on a specific conditional logit model. For more details see the help on the conditional logit model. In this case, we obtain aggregate utilities, that is to say, one utility for each category of each variable associated with all the individuals. It is impossible to make classifications based on the individuals.

- A hierarchical Bayes algorithm which gives individual results. Parameters are estimated at the individual level using an iterative method (Gibbs sampling) taking into account each individual’s choice but also the global distribution of the choices. The obtained individual utilities will give better market simulation as the classical CBC algorithm.

XLSTAT-Conjoint proposes to include a segmentation variable when using the classical CBC algorithm that will build separate models for each group defined by the variable. When CBC/HB is used, since individual utilities are obtained, you can apply a clustering method on the individuals.

In addition to utilities, conjoint analysis provides the importance associated with each variable.

Interactions
By interaction is meant an artificial factor (not measured) which reflects the interaction between at least two measured factors. For example, if we carry out treatment on a plant, and tests are carried out under two different light intensities, we will be able to include in the model an interaction factor treatment*light which will be used to identify a possible interaction between the two factors. If there is an interaction between the two factors, we will observe a significantly larger effect on the plants when the light is strong and the treatment is of type 2 while the effect is average for weak light, treatment 2 and strong light, treatment 1 combinations.

To make a parallel with linear regression, the interactions are equivalent to the products between the continuous explanatory values although here obtaining interactions requires nothing more than simple multiplication between two variables. However, the notation used to represent the interaction between factor A and factor B is A*B.

The interactions to be used in the model can be easily defined in XLSTAT.

**Constraints**

During the calculations, each factor is broken down into a sub-matrix containing as many columns as there are categories in the factor. Typically, this is a full disjunctive table. Nevertheless, the breakdown poses a problem: if there are g categories, the rank of this sub-matrix is not g but g-1. This leads to the requirement to delete one of the columns of the sub-matrix and possibly to transform the other columns. Several strategies are available depending on the interpretation we want to make afterwards:

1) \( a_1 = 0 \): the parameter for the first category is null. This choice allows us force the effect of the first category as a standard.

2) \( a_n = 0 \): the parameter for the last category is null. This choice allows us force the effect of the last category as a standard.

3) \( \text{Sum (ai)} = 0 \): the sum of the parameters is null.

Note: even if the choice of constraint influences the values of the parameters, it has no effect on the predicted values and on the different fitting statistics.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

![OK Button](Click this button to start the computations.)
### General tab:

**Responses**: Select the responses that have been given by respondents. If headers have been selected, please check the option "Variable labels" is enabled. This selection corresponds to the right part of the conjoint analysis design table generated with the "design of choice based conjoint analysis" tool of XLSTAT-Conjoint.

**Choice table**: Select the choices that have been presented to the respondents. If headers have been selected, please check the option "Variable labels" is enabled. This selection corresponds to the left part of the conjoint analysis design table generated with the "design of choice based conjoint analysis" tool of XLSTAT-Conjoint. Do not select the first column of the table.

**Profiles**: Select the profiles that have been generated. If headers have been selected, please check the option "Variable labels" is enabled. This selection corresponds to profiles table generated with the "design of choice based conjoint analysis" tool of XLSTAT-Conjoint. Do not select the first column of the table.

**Range**: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Check this option to display the results in a new worksheet in the active workbook.

**Workbook**: Check this option to display the results in a new workbook.

**Variable labels**: Check this option if the first line of the selections (data, other group) contains a label.
**Group variable:** Activate this option then select a column containing the group identifiers. If a header has been selected, check that the "Variable labels" option has been activated.

**Response weights:** Activate this option if response weights are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection.

**Options tab:**

**Method:** Select the method to be used for estimation.

**Interactions / Level:** Activate this option to include interactions in the model then enter the maximum interaction level (value between 1 and 4).

**Tolerance:** Activate this option to prevent the OLS regression calculation algorithm taking into account variables which might be either constant or too correlated with other variables already used in the model (0.0001 by default).

**Confidence interval (%):** Enter the percentage range of the confidence interval to use for the various tests and for calculating the confidence intervals around the parameters and predictions. Default value: 95.

**Constraints:** Details on the various options are available in the description section.

*a1 = 0*: Choose this option so that the parameter of the first category of each factor is set to 0.

*an = 0*: Choose this option so that the parameter of the last category of each factor is set to 0.

*Sum (ai) = 0*: for each factor, the sum of the parameters associated with the various categories is set to 0.

**Bayesian options (only using CBC/HB algorithm):** the number of iterations for the burn-in period and the maximal time for the hierarchical Bayes algorithm can be modified.

**Segmentation (only using CBC/HB algorithm):** Activate this option if you want to apply an individual based clustering method on the partial utilities. Two methods are available: agglomerative hierarchical classification and k-means classification.

- **Number of classes:** Enter the number of classes to be created by the algorithm for the k-means.
- **Truncation**: Activate this option if you want XLSTAT to automatically define the truncation level, and therefore the number of classes to retain, or if you want to define the number of classes to create, or the level at which the dendrogram is to be truncated.

**Stop conditions**: the number of iterations and the convergence criterion until convergence of the Newton-Raphson algorithm can be modified.

**Missing data** tab:

**Remove observations**: Activate this option to remove the observations with missing data.

**Outputs** tab:

**Goodness of fit statistics**: Activate this option to display the table of goodness of fit statistics for the model.

**Type III analysis**: Activate this option to display the type III analysis of variance table.

**Model coefficients**: Activate this option to display the model’s coefficients also called aggregated utilities.

**Standardized coefficients**: Activate this option if you want the standardized coefficients (beta coefficients) for the model to be displayed.

**Predictions and residuals**: Activate this option to display the predictions and residuals obtained with the aggregated utilities.

**Observation details**: activate this option to display the characteristics of the posterior distribution for each individual when using CBC/HB algorithm.

**Charts** tab:

**Regression charts**: Activate this option to display regression chart:

- **Standardized coefficients**: Activate this option to display the standardized parameters for the model with their confidence interval on a chart.

**Convergence graph**: Activate this option to display the evolution of model parameters for each individual when using CBC/HB algorithm.
Results

Variable information: This table displays all the information relative to the used factors. XLSTAT displays a large number tables and charts to help in analyzing and interpreting the results.

Utilities: This table displays utilities associated to each category of the factors with their respective standard error.

Importance: This table displays importance for each factor of the analysis.

Goodness of fit coefficients: This table displays a series of statistics for the independent model (corresponding to the case where the linear combination of explanatory variables reduces to a constant) and for the adjusted model.

- Observations: The total number of observations taken into account (sum of the weights of the observations);
- Sum of weights: The total number of observations taken into account (sum of the weights of the observations multiplied by the weights in the regression);
- DF: Degrees of freedom;
- -2 Log(Like.): The logarithm of the likelihood function associated with the model;
- $R^2$ (McFadden): Coefficient, like the $R^2$, between 0 and 1 which measures how well the model is adjusted. This coefficient is equal to 1 minus the ratio of the likelihood of the adjusted model to the likelihood of the independent model;
- $R^2$(Cox and Snell): Coefficient, like the $R^2$, between 0 and 1 which measures how well the model is adjusted. This coefficient is equal to 1 minus the ratio of the likelihood of the adjusted model to the likelihood of the independent model raised to the power 2/Sw, where Sw is the sum of weights.
- $R^2$(Nagelkerke): Coefficient, like the $R^2$, between 0 and 1 which measures how well the model is adjusted. This coefficient is equal to ratio of the $R^2$ of Cox and Snell, divided by 1 minus the likelihood of the independent model raised to the power 2/Sw;
- AIC: Akaike's Information Criterion;
- SBC: Schwarz’s Bayesian Criterion.
- Iteration: Number of iteration to reach convergence.
- rlh: root likelihood. This value varies between 0 and 1, the value of 1 being a perfect fit. It is only available for the CBC/HB algorithm.
Goodness of fit indexes (conditional logit): In this table are shown the goodness of fit statistics specific to the case of the conditional logit model. For more details on these statistics, see the description part of this help.

Test of the null hypothesis H0: Y=p0: The H0 hypothesis corresponds to the independent model which gives probability p0 whatever the values of the explanatory variables. We seek to check if the adjusted model is significantly more powerful than this model. Three tests are available: the likelihood ratio test (-2 Log(Like.)), the Score test and the Wald test. The three statistics follow a Chi^2 distribution whose degrees of freedom are shown.

The table of standardized coefficients (also called beta coefficients) are used to compare the relative weights of the variables. The higher the absolute value of a coefficient, the more important the weight of the corresponding variable. When the confidence interval around standardized coefficients has value 0 (this can easily be seen on the chart of standardized coefficients), the weight of a variable in the model is not significant.

Example

An example of choice based conjoint (CBC) analysis is available at the Addinsoft website:

http://www.xlstat.com/demo-cbc.htm

References


Conjoint analysis simulation tool

Use this tool to run market simulations based on the results of a conjoint analysis (full profile or choice-based) obtained with XLSTAT-Conjoint.

Description

Conjoint analysis is a comprehensive method for the analysis of new products in a competitive environment.

Once the analysis has been performed, the major advantage of conjoint analysis is its ability to perform market simulations using the obtained utilities. The products included in the market do not have to be part of the tested products.

Outputs from conjoint analysis include utilities which can be partial (associated to each individual in full profile conjoint analysis) or aggregate (associated to all the individuals in CBC). These utilities allow computing a global utility associated to any product that you want to include in your simulated market.

Four estimation methods are proposed in XLSTAT-Conjoint: first choice, logit, Bradley-Terry-Luce and randomized first choice. These methods are described below.

The obtained market shares can then be analyzed to assess the possible introduction of a new product on the market. The results of these simulations are nevertheless dependent on the knowledge of the real market and the fact that all important factors associated with each product in the conjoint analysis have been taken into account.

XLSTAT-Conjoint can also add weights to the categories of the factors or to the individuals.

XLSTAT-Conjoint can also take into account groups of individuals when a group variable (segmentation) is available. It can be obtained, for example, with the segmentation tool associated with the conjoint analysis.

Data type

XLSTAT-Conjoint proposes two models for conjoint analysis. In a full profile analysis, a constant is associated to the utilities and there are as many utilities as individuals in the study. You have to select all the utilities and their constant (without the column with the names of the categories). In the case of CBC, there is no constant and you have to select one column of utilities without the labels associated to the name of the categories.
In XLSTAT-Conjoint, you have to entirely select the variable information table provided by the conjoint analysis tool. On the other hand, the market to be simulated must be generated "by hand" using the categories of the factors in the model.

**Simulation methods**

XLSTAT-Conjoint offers four methods for simulation of market share.

The first step consists of calculating the global utility associated with each new product. Thus, for a CBC analysis for analyzing men's shoes with three factors: the price (50 dollars, 100 dollars, 150 dollars), their finishing (canvas, leather, suede) and the color (brown, black). We have a table with 8 partial utilities rows and one column.

We want to simulate a market with a black leather shoe with price equal USD 100. The utility of this product is: \( U_{P1} = U_{\text{price-100}} + U_{\text{F-Leather}} + U_{\text{C-Black}} \)

We calculate the utility for each product in the market and we seek the probability of choosing this product using different estimation methods:

- **First choice**: it is the most basic; you select the product with maximum utility with a probability of 1.

- **Logit**: this method is based on the exponential function to find the probability, it is more accurate than the method first choice and it is generally preferred. It has the disadvantage of the IIA assumption (assumption of independence of irrelevant alternatives). It is calculated for the product \( P_1 \): \( P_{p1} = \frac{e^{U_{p1}\beta}}{\sum e^{U_{pi}\beta}} \) with beta = 1 or 2.

- **Bradley-Terry-Luce** is a method close to the logit method without using the exponential function. It always involves the assumption of IIA and demands positive utilities (if beta = 1). It is calculated for the product \( P_1 \): \( P_{p1} = \frac{U_{p1}^\beta}{\sum U_{pi}^\beta} \) with beta = 1 or 2.

- **Randomized first choice**: it is a method midway between logit and First Choice. It has the advantage of not assuming the IIA assumption and is based on a simple principle: it generates a large number of numbers from a Gumbel distribution and creates a new set of utilities using the initial utilities adding the numbers generated. For each set of utilities created, the first choice method is used to select one of the products. So we will accept slight variations around the calculated values of the utilities. This method is the most advanced but also more suited to the case of conjoint analysis.

- **RFC-Bolse**: In the case of profile-based conjoint analysis, the Randomized First Choice BOLSE (RFC-BOLSE) was introduced to overcome the problems of the RFC method. Indeed, RFC is based on a Gumbel law that do not fit the full profile method. This approach is based on the same principle as Randomized First Choice but it uses a
different distribution function to generate the simulated numbers. The RFC model adds unique random error (variation) to the part-worths and computes market shares using the First Choice rule. The centered normal distribution is used with standard error equal to the standard error of the parameters of the regression model and a global error term associated to the entire model. For each set of utilities created, the first choice method is used to select one of the products. So we will accept slight variations around the calculated values of the utilities. This method is the most advanced but also more suited to the case of profile based conjoint analysis.

When more than one column of utilities (with a conjoint analysis with full profiles) are selected, XLSTAT-Conjoint uses the mean of the probabilities.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload**: Click this button to reload the default options.
- **Delete**: Click this button to delete the data selections.
- **Switch**: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab:**
Utilities table: Select the utilities obtained with XLSTAT-Conjoint. If headers have been selected, please check the option "Variable labels" is enabled. Do not select the name of the categories.

Variables information: Select the variable information table generated with XLSTAT-Conjoint. If headers have been selected, please check the option "Variable labels" is enabled.

Model: Choose the type of conjoint analysis that you used (full profile or CBC).

Simulated market: Select the market to be simulated. The products will be distributed in a table with a product per line and a variable per column. If headers have been selected, please check the option "Variable labels" is enabled.

Method: Choose the method to use to compute market shares.

Product ID: Activate this option if products ID are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Variable labels: Check this option if the first line of the selections (data, other group) contains a label.

Categories weights: Activate this option if categories weights are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection.

Group variable: Activate this option then select a column containing the group identifiers. If a header has been selected, check that the "Variable labels" option has been activated.

Response weights: Activate this option if response weights are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection.

Options tab:

Interactions / Level: Activate this option if interactions were selected in the conjoint analysis. Then, enter the maximum level of interaction (value between 1 and 3).
**Number of simulations:** Enter the number of simulations to be generated with the “randomized first choice” option.

**Charts tab:**

**Market share plot:** Activate this option to display market share plots:
- **Pie charts:** Activate this option to display market share pie charts.
- **Compare to the total sample:** If groups have been selected, activate this option to compare the market shares of sub-samples with those of the complete sample.

**Results**

**Variable information:** This table displays the summary of the information on the selected factors.

**Simulated market:** This table displays the products used to perform the simulation.

Market shares: This table displays the obtained market shares. If groups have been selected, the first column is associated with the global market and the following columns are associated with each group.

**Market share plots:** The first pie chart is associated to the global market. If groups have been selected, the following diagrams are associated with the different groups. If the option “compare to the total sample” is selected, the plots are superimposed; in the background the global market shares are displayed and in front, market shares associated to the group of individuals studied are shown.

**Utilities / Market shares:** This table, which appears only if no groups are selected, displays products utilities, market shares as well as standard deviations (when possible) associated with each product from the simulated market.

**Market shares (individual):** This table, which appears only if no groups are selected and when full profile conjoint analysis is selected, displays market shares obtained for each individual.

**Example**

An example of conjoint analysis is available at the Addinsoft website:

An example of choice based conjoint (CBC) analysis is available at the Addinsoft website:

http://www.xlstat.com/demo-cbc.htm

References


Design for MaxDiff

Use this tool to generate a design of experiments for MaxDiff analysis (best-worst model).

Description

MaxDiff or Maximum Difference Scaling is a method introduced by Jordan Louvière (1991) that allows obtaining importance of attributes. Attributes are presented to a respondent who must choose to best and worst attributes (most important / least important).

Two steps are needed to apply that method. First, a design must be generated so that each attribute is presented with other attributes an equal number of times. Then, once the respondent has selected for each choice the best and worst attribute, a model is applied in order to obtain the importance of each attribute. A Hierarchical Bayes model is applied to obtain individual values of the importance.

To obtain the design, design of experiments is used. An incomplete block design is used to generate the choices to be presented. For more details on these methods, please see the DOE for sensory analysis chapter of the MX module help.

The number of comparisons and the number of attributes per comparison should be chosen depending on the number of attributes. Keep in mind that too many attributes can lead to problems and that too many choices can be problematic for the respondent.

XLSTAT-Conjoint allows obtaining a global table for MaxDiff analysis but also individual tables for each respondent and each comparison in separated Excel sheets. References are also included so that when a respondent select a profile in an individual sheet, it is directly reported in the main table.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

OK: Click this button to start the computations.
Cancel: Click this button to close the dialog box without doing any computation.
General tab:

Analysis name: Enter the name of the analysis you want to perform.

Attributes: Select the attributes that will be tested during this analysis.

Number of responses: Enter the number of expected individuals who will respond to the MaxDiff analysis.

Maximum number of comparisons: Enter the maximum number of comparison to be presented to the individual respondents. This number has to be greater than the number of attributes.

Number of profiles per comparison: Enter the number of attributes per comparison.

Terminology: Choose among the alternatives offered, the terms that best correspond to your case.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Variable labels: Check this option if the first line of the selections contains a label.

Outputs tab:

Print individual sheets: Activate this option to print individual sheets for each respondent. Each sheet will include a table for each comparison. The respondent has to enter any value close to the best (on the right) and worst (on the left) attributes. Two assignment options are available; the fixed option displays the comparisons in the same order for all individuals; the random option displays the comparisons in random orders (different from one respondent to another).
Include references: Activate this option to include references between the main sheet and the individual sheets. When an individual enter his chosen code in the individual sheet, the result is automatically displayed in the main sheet of the analysis.

Results

Variable information: This table displays all the information relative to the attributes.

MaxDiff analysis design: This table displays the comparisons presented to the respondent. Each row is associated to a comparison of attributes. Empty cells associated to each individual respondent are also displayed. Respondent have to enter the code associated to the choice made (1 to number of attributes per comparisons). Two columns per respondent have to be filled (best and worst).

Individual _Res sheets: When the “Print individual sheets” option is activated, these sheets include the name of the analysis, the individual number and tables associated to the comparisons with the profiles to be compared. Individual respondents should enter the code associated to their choice in the bottom right of each table.

Example

An example of MaxDiff analysis is available at the Addinsoft website:

http://www.xlstat.com/demo-maxdiff.htm

References


MaxDiff analysis

Use this tool to run a MaxDiff analysis. This tool is included in the XLSTAT-Conjoint module; it must be applied on design of experiments for MaxDiff analysis generated with XLSTAT-Conjoint.

Description

MaxDiff or Maximum Difference Scaling is a method introduced by Jordan Louvière (1991) that allows obtaining importance of attributes. Attributes are presented to a respondent who must choose to best and worst attributes (most important / least important).

This tool allows you to carry out the step of analyzing the results obtained after the collection of responses from a sample of people. This analysis can only be done once the attributes have been defined, the design has been generated, and the individual responses have been collected.

In the case of MaxDiff models, individuals must choose between selections of attributes. Thus, a number of choices is given to all individuals (we select an attribute from a number of attributes).

Analysis of these choices can be done using a conditional logit model or a hierarchical Bayes algorithm which gives individual results.

Hierarchical Bayes model

Parameters are estimated at the individual level using an iterative method (Gibbs sampling) taking into account each individual's choice but also the global distribution of the choices. The obtained individual importance will be more precise.

The MaxDiff analysis allows obtaining individual MaxDiff score for each respondent and each attribute.

The model coefficients are obtained using the HB model with X as input for best choices and –X for worst choices. Then, these coefficients are transformed to obtain MaxDiff scores. They are centered then transformed using that formula: exp(beta)/(exp(beta)+nb_alter-1 with nb_alter being the number of alternatives proposed in each choice task. Then the scores are rescaled in order to sum to 100.
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload default options**: Click this button to reload the default options.
- **Delete data selections**: Click this button to delete the data selections.
- **Change data handling**: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab:**

**Responses**: Select the responses that have been given by respondents. If headers have been selected, please check the option "Variable labels" is enabled. This selection corresponds to the right part of the MaxDiff analysis design table generated with the "design of MaxDiff analysis" tool of XLSTAT-Conjoint.

**Choice table**: Select the choices that have been presented to the respondents. If headers have been selected, please check the option "Variable labels" is enabled. This selection corresponds to the left part of the Max-Diff analysis design table generated with the "design of Max-Diff analysis" tool of XLSTAT-Conjoint. Do not select the first column of the table.

**Terminology**: Choose among the alternatives offered, the terms that best correspond to your case.

**Range**: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Check this option to display the results in a new worksheet in the active workbook.
Workbook: Check this option to display the results in a new workbook.

Variable labels: Check this option if the first line of the selections contains a label.

Response weights: Activate this option if response weights are available. Then select the corresponding data. If the “Variable labels” option is activated you need to include a header in the selection.

Options tab:

Method: Select the method to be used for estimation. Hierarchical Bayes in that case.

Confidence interval (%): Enter the percentage range of the confidence interval to use for the various tests and for calculating the confidence intervals around the parameters and predictions. Default value: 95.

Bayesian options: the number of iterations for the burn-in period and the maximal time for the hierarchical Bayes algorithm can be modified.

Stop conditions: the number of iterations and the convergence criterion until convergence of the algorithm can be modified.

Missing data tab:

Remove observations: Activate this option to remove the observations with missing data.

Outputs tab:

Observation details: activate this option to display the characteristics of the posterior distribution for each individual.

Results

Counts analysis: These tables summarize the results of the MaxDiff survey by showing globally and then for each respondent how many times each attribute has been chosen as best and worst. The third column of these tables correspond to the difference.

The following results are only displayed in the case of a hierarchical Bayes model.

Variable information: This table displays all the information relative to the used attributes.
MaxDiff scores: This table displays MaxDiff scores for each attribute of the analysis for each respondent. Individual values and descriptive statistics are available.

Model coefficients: This table displays the HB model coefficients.

Goodness of fit coefficients: This table displays a series of statistics for the independent model (corresponding to the case where the linear combination of explanatory variables reduces to a constant) and for the adjusted model.

- **Observations**: The total number of observations taken into account (sum of the weights of the observations);
- **Sum of weights**: The total number of observations taken into account (sum of the weights of the observations multiplied by the weights in the regression);
- **-2 Log(Like.)**: The logarithm of the likelihood function associated with the model;
- **rlh**: root likelihood. This value varies between 0 and 1, the value of 1 being a perfect fit.

Individual results are then displayed.

Example

An example of MaxDiff analysis is available at the Addinsoft website:

http://www.xlstat.com/demo-maxdiff.htm

References


Monotone regression (MONANOVA)

Use this tool to apply a monotone regression or MONANOVA model. Advanced options let you choose the constraints on the model and take into account interactions between factors. This tool is included in the module XLSTAT-Conjoint.

Description

The MONANOVA model is part of the XLSTAT-Conjoint module.

Monotone regression and the MONANOVA model differ only in the fact that the explanatory variables are either quantitative or qualitative. These methods are based on iterative algorithms based on the ALS (alternating least squares) algorithm. Their principle is simple, it consists of alternating between a conventional estimation using linear regression or ANOVA and a monotonic transformation of the dependent variables (after searching for optimal scaling transformations).

The MONANOVA algorithm was introduced by Kruskal (1965) and the monotone regression and the works on the ALS algorithm are due to Young et al. (1976).

These methods are commonly used as part of the full profile conjoint analysis. XLSTAT-Conjoint allows applying them within a conjoint analysis (see chapter on conjoint analysis based on full profiles) as well as independently.

The monotone regression tool (MONANOVA) combines a monotonic transformation of the responses to a linear regression as a way to improve the linear regression results. It is well suited to ordinal dependent variables.

XLSTAT-Conjoint allows you to add interactions and to vary the constraints on the variables.

Method

Monotone regression combines two stages: an ordinary linear regression between the explanatory variables and the response variable and a transformation step of the response variables to maximize the quality of prediction.

The algorithm is:

1- Run an OLS regression between the response variable Y and the explanatory variables X. We obtain the beta coefficients.
2- Calculation of the predicted values of $Y$: $\text{Pred} (Y) = \beta \cdot X$

3- Transformation of $Y$ using a monotonic transformation (Kruskal, 1965) so that $\text{Pred} (Y)$ and $Y$ are close (using optimal scaling methods).

4- Run an OLS Regression between $Y_{\text{trans}}$ and the explanatory variables $X$. This gives new values for the $\beta$.

5- Steps 2 through 4 are repeated until the change in $R^2$ from one stage to another is smaller than the convergence criterion.

**Goodness of fit (MONANOVA)**

In the context of MONANOVA, additional results are available. These results are generally associated with a multivariate analysis but as we are in the case of a transformation of the responses, their presence is necessary. Instead of using the squared canonical correlations between measures, we use the $R^2$. XLSTAT-Conjoint calculates the Wilks' lambda, Pillai's trace, the trace of Hotelling-Lawlet and Roy largest root using a matrix with largest eigenvalue equal to the $R^2$ and 0 for other eigenvalues. The largest root of Roy gives a lower bound for the $p$-value of the model. Other statistics are upper bounds on the $p$-value of the model.

**Interactions**

By interaction is meant an artificial factor (not measured) which reflects the interaction between at least two measured factors. For example, if we carry out treatment on a plant, and tests are carried out under two different light intensities, we will be able to include in the model an interaction factor treatment*$light$ which will be used to identify a possible interaction between the two factors. If there is an interaction between the two factors, we will observe a significantly larger effect on the plants when the light is strong and the treatment is of type 2 while the effect is average for weak light, treatment 2 and strong light, treatment 1 combinations.

To make a parallel with linear regression, the interactions are equivalent to the products between the continuous explanatory values although here obtaining interactions requires nothing more than simple multiplication between two variables. However, the notation used to represent the interaction between factor A and factor B is $A^*B$.

The interactions to be used in the model can be easily defined in XLSTAT-Conjoint.

**Constraints for qualitative predictors**

During the calculations, each factor is broken down into a sub-matrix containing as many columns as there are categories in the factor. Typically, this is a full disjunctive table. Nevertheless, the breakdown poses a problem: if there are $g$ categories, the rank of this sub-matrix is not $g$ but $g-1$. This leads to the requirement to delete one of the columns of the sub-
matrix and possibly to transform the other columns. Several strategies are available depending on the interpretation we want to make afterwards:

1) \(a_1=0\): the parameter for the first category is null. This choice allows us to force the effect of the first category as a standard.

2) \(a_n=0\): the parameter for the last category is null. This choice allows us to force the effect of the last category as a standard.

3) \(\text{Sum (ai) = 0}\): the sum of the parameters is null. This choice forces the constant of the model to be equal to the mean of the dependent variable when the design is balanced.

Note: even if the choice of constraint influences the values of the parameters, it has no effect on the predicted values and on the different fitting statistics.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload**: Click this button to reload the default options.
- **Delete**: Click this button to delete the data selections.
- **Change**: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab:**

Y / Dependent variables:
Quantitative: Select the response variable(s) you want to model. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

X / Explanatory variables:

Quantitative: Select the quantitative explanatory variables in the Excel worksheet. The data selected must be of type numeric. If the variable header has been selected, check that the "Variable labels" option has been activated.

Qualitative: Select the qualitative explanatory variables (the factors) in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Variable labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 ...).

Observation weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will all be taken as 1. Weights must be greater than or equal to 0. A weight of 2 is equivalent to repeating the same observation twice. If a column header has been selected, check that the "Variable labels" option has been activated.

Options tab:

Fixed constant: Activate this option to fix the constant of the regression model to a value you then enter (0 by default).
**Tolerance**: Activate this option to prevent the OLS regression calculation algorithm taking into account variables which might be either constant or too correlated with other variables already used in the model (0.0001 by default).

**Interactions / Level**: Activate this option to include interactions in the model then enter the maximum interaction level (value between 1 and 4).

**Confidence interval (%)**: Enter the percentage range of the confidence interval to use for the various tests and for calculating the confidence intervals around the parameters and predictions. Default value: 95.

**Constraints**: Details on the various options are available in the description section.

**a1 = 0**: Choose this option so that the parameter of the first category of each factor is set to 0.

**an = 0**: Choose this option so that the parameter of the last category of each factor is set to 0.

**Sum (ai) = 0**: for each factor, the sum of the parameters associated with the various categories is set to 0.

**Stop conditions**:

- **Iterations**: Enter the maximum number of iterations for the ALS algorithm. The calculations are stopped when the maximum number if iterations has been exceeded. Default value: 100.

- **Convergence**: Enter the maximum value of the evolution of $R^2$ from one iteration to another which, when reached, means that the algorithm is considered to have converged. Default value: 0.000001.

**Missing data** tab:

**Remove observations**: Activate this option to remove the observations with missing data.

- **Check for each Y separately**: Choose this option to remove the observations with missing data in the select Y (dependent) variables, only if the Y of interest has a missing data.

- **Across all Ys**: Choose this option to remove the observations with missing data in the Y (dependent) variables, even if the Y of interest has no missing data.

**Estimate missing data**: Activate this option to estimate missing data before starting the computations.
• **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

• **Nearest neighbour**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.

**Outputs** tab:

**Observation details**: Activate this option to display detailed outputs for each respondent.

**Correlations**: Activate this option to display the correlation matrix for quantitative variables (dependent or explanatory).

**Analysis of variance**: Activate this option to display the analysis of variance table.

**Type I/II/III SS**: Activate this option to display the Type I, Type II, and Type III sum of squares tables. Type II table is only displayed if it is different from Type III.

**Standardized coefficients**: Activate this option if you want the standardized coefficients (beta coefficients) for the model to be displayed.

**Predictions and residuals**: Activate this option to display the predictions and residuals for all the observations.

**Charts** tab:

**Regression charts**: Activate this option to display regression chart:

  • **Standardized coefficients**: Activate this option to display the standardized parameters for the model with their confidence interval on a chart.*

**Transformation plot**: Activate this option to display the monotone transformation of the response plot.

**Results**

XLSTAT displays a large number tables and charts to help in analyzing and interpreting the results.

**Summary statistics**: This table displays descriptive statistics for all the variables selected. For the quantitative variables, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed. For qualitative variables,
including the dependent variable, the categories with their respective frequencies and percentages are displayed.

**Correlation matrix**: This table displays the correlations between the explanatory variables.

**Goodness of fit statistics**: The statistics relating to the fitting of the regression model are shown in this table:

- **Observations**: The number of observations used in the calculations. In the formulas shown below, \( n \) is the number of observations.

- **Sum of weights**: The sum of the weights of the observations used in the calculations. In the formulas shown below, \( W \) is the sum of the weights.

- **DF**: The number of degrees of freedom for the chosen model (corresponding to the error part).

- **\( R^2 \)**: The determination coefficient for the model. This coefficient, whose value is between 0 and 1, is only displayed if the constant of the model has not been fixed by the user. Its value is defined by:

\[
R^2 = 1 - \frac{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} w_i (y_i - \bar{y})^2}, \quad \text{where} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} w_i y_i ,
\]

The \( R^2 \) is interpreted as the proportion of the variability of the dependent variable explained by the model. The nearer \( R^2 \) is to 1, the better is the model. The problem with the \( R^2 \) is that it does not take into account the number of variables used to fit the model.

- **Adjusted \( R^2 \)**: The adjusted determination coefficient for the model. The adjusted \( R^2 \) can be negative if the \( R^2 \) is near to zero. This coefficient is only calculated if the constant of the model has not been fixed by the user. Its value is defined by:

\[
\hat{R}^2 = 1 - \left(1 - R^2\right) \frac{W - 1}{W - p - 1}
\]

The adjusted \( R^2 \) is a correction to the \( R^2 \) which takes into account the number of variables used in the model.

- **MSE**: The mean squared error (MSE) is defined by:

\[
MSE = \frac{1}{W - p} \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2
\]

- **RMSE**: The root mean square of the errors (RMSE) is the square root of the MSE.
MAPE: The *Mean Absolute Percentage Error* is calculated as follows::

\[
MAPE = \frac{100}{W} \sum_{i=1}^{n} w_i \left| \frac{y_i - \hat{y}_i}{y_i} \right|
\]

DW: The Durbin-Watson statistic is defined by:

\[
DW = \frac{\sum_{i=2}^{n} \left[ (y_i - \hat{y}_i) - (y_{i-1} - \hat{y}_{i-1}) \right]^2}{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}
\]

This coefficient is the order 1 autocorrelation coefficient and is used to check that the residuals of the model are not autocorrelated, given that the independence of the residuals is one of the basic hypotheses of linear regression. The user can refer to a table of Durbin-Watson statistics to check if the independence hypothesis for the residuals is acceptable.

Cp: Mallows Cp coefficient is defined by:

\[
Cp = \frac{SSE}{\hat{\sigma}} + 2 p^* - W
\]

AIC: Akaike's Information Criterion is defined by:

\[
AIC = W \ln \left( \frac{SSE}{W} \right) + 2 p^*
\]

This criterion, proposed by Akaike (1973) is derived from the information theory and uses Kullback and Leibler's measurement (1951). It is a model selection criterion which penalizes models for which adding new explanatory variables does not supply sufficient information to the model, the information being measured through the MSE. The aim is to minimize the AIC criterion.

SBC: Schwarz’s Bayesian Criterion is defined by:

\[
SBC = W \ln \left( \frac{SSE}{W} \right) + \ln (W) p^*
\]

This criterion, proposed by Schwarz (1978) is similar to the AIC and, like this, the aim is to minimize it.
- **PC**: Amemiya's Prediction Criterion is defined by:

\[
PC = \frac{(1 - R^2)(W + p^*)}{W - p^*}
\]

This criterion, proposed by Amemiya (1980) is used, like the adjusted $R^2$ to take account of the parsimony of the model.

- **Press RMSE**: Press' statistic is only displayed if the corresponding option has been activated in the dialog box. It is defined by:

\[
Press = \sum_{i=1}^{n} w_i (y_i - \hat{y}_{i(-i)})^2
\]

where $\hat{y}_{i(-i)}$ is the prediction for observation $i$ when the latter is not used for estimating parameters. We then get:

\[
Press \text{ RMSE} = \sqrt{\frac{Press}{W - p^*}}
\]

Press's RMSE can then be compared to the RMSE. A large difference between the two shows that the model is sensitive to the presence or absence of certain observations in the model.

- **Iteration**: Number of iteration until convergence of the ALS algorithm.

**Goodness of fit coefficients (MONANOVA)**: In this table are shown the statistics for the fit of the regression model specific to the case of MONANOVA. These statistics are the Wilks' lambda, the Pillai's trace, the trace of Hotelling-Lawlet and the largest root of Roy. For more details on these statistics, see the description part of this help.

If the Type I/II/III SS (SS: Sum of Squares) option is activated, the corresponding tables are displayed.

The table of **Type I SS** values is used to visualize the influence that progressively adding explanatory variables has on the fitting of the model, as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. The sums of squares in the Type I table always add up to the model SS. Note: the order in which the variables are selected in the model influences the values obtained.

The table of **Type II SS** values is used to visualize the influence that removing an explanatory variable has on the fitting of the model, all other variables being retained, as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable...
to the model, all the other variables already being in the model. Note: unlike Type I SS, the order in which the variables are selected in the model has no influence on the values obtained. Type II SS are not recommended in unbalanced designs but we display them as some users might need them. It is identical to Type III for balanced designs.

The table of **Type III SS** values is used to visualize the influence that removing an explanatory variable has on the fitting of the model, all other variables being retained, as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. Note: unlike Type I SS, the order in which the variables are selected in the model has no influence on the values obtained. While Type II SS depends on the number of observations per cell (cell means combination of categories of the factors), Type III does not and is therefore preferred.

The **analysis of variance table** is used to evaluate the explanatory power of the explanatory variables. Where the constant of the model is not set to a given value, the explanatory power is evaluated by comparing the fit (as regards least squares) of the final model with the fit of the rudimentary model including only a constant equal to the mean of the dependent variable. Where the constant of the model is set, the comparison is made with respect to the model for which the dependent variable is equal to the constant which has been set.

The **parameters of the model** table displays the estimate of the parameters, the corresponding standard error, the Student’s t, the corresponding probability, as well as the confidence interval.

The table of **standardized coefficients** (also called beta coefficients) are used to compare the relative weights of the variables. The higher the absolute value of a coefficient, the more important the weight of the corresponding variable. When the confidence interval around standardized coefficients has value 0 (this can be easily seen on the chart of standardized coefficients), the weight of a variable in the model is not significant.

The **predictions and residuals** table shows, for each observation, its weight, the observed value of the dependent variable, the transformed value of the dependant variable, the model's prediction, the residuals, and the confidence intervals. Two types of confidence interval are displayed: a confidence interval around the mean (corresponding to the case where the prediction would be made for an infinite number of observations with a set of given values for the explanatory variables) and an interval around the isolated prediction (corresponding to the case of an isolated prediction for the values given for the explanatory variables). The second interval is always greater than the first, the random values being larger.

The **charts** which follow show the transformation of the dependant variable.
Example

An example of MONANOVA is available at the Addinsoft website:

http://www.xlstat.com/demo-monanova.htm

References


**Conditional logit model**

Use conditional logit model to model a binary variable using quantitative and/or qualitative explanatory.

**Description**

The conditional logit model is part of the XLSTAT-Conjoint module.

The conditional logit model is based on a model similar to that of the logistic regression. The difference is that all individuals are subjected to different situations before expressing their choice (modeled using a binary variable which is the dependent variable). The fact that the same individuals are used is taken into account by the conditional logit model (NB: the observations are not independent within a block corresponding to the same individual).

The conditional logit model is a method mostly used in conjoint analysis, it is nevertheless useful when analyzing a certain type of data. It is McFadden (1973) who introduced this model. Instead of having one line per individual like in the classical logit model, there will be one row for each category of the variable of interest. If one seeks to study transportations, for example, there will be four types of transports (car / train / plane / bike), each type of transport have characteristics (their price, their environmental costs...) but an individual can choose only one of four transportations. As part of a conditional logit model, all four options are presented to each individual and the individual chooses his preferred option. We have for N individuals, N * 4 rows with 4 rows for each individual associated with each transportation. The binary response variable will indicate the choice of the individual (1) and 0 if the individual did not choose this option.

In XLSTAT-Conjoint, you will also have to select a column associated with the name of the individuals (with 4 lines per individual in our example). The explanatory variables will also have N * 4 lines.

**Method**

The conditional logit model is based on a model similar to that of the logistic regression except that instead of having individual characteristics, there will be characteristics of the different alternatives proposed to the individuals.

The probability that individual i chooses product j is given by:

\[ P_{ij} = \frac{e^{\beta_j z_{ij}}}{\sum_k e^{\beta_k z_{ik}}} \]
From this probability, we calculate a likelihood function:

\[ l(\beta) = \sum_{i=1}^{n} \sum_{j=1}^{J} y_{ij} \log(P_j) \]

With \( y \) being a binary variable indicating the choice of individual \( i \) for product \( j \) and \( J \) being the number of choices available to each individual.

To estimate the model parameters \( \beta \) (the coefficients of the linear function), it seeks to maximize the likelihood function. Unlike linear regression, an exact analytical solution does not exist. It is therefore necessary to use an iterative algorithm. XLSTAT-Conjoint uses a Newton-Raphson algorithm.

**Goodness of fit (conditional logit)**

Some specific goodness of fit indexes are displayed for the conditional logit model.

- **Likelihood ratio R**: \( R = -2 \left( \log(L) - \log(L_0) \right) \)

- **Upper bound of the likelihood ratio U**: \( U = -2 \log(L_0) \)

- **Aldrich-Nelson**: \( AN = \frac{R}{R + N} \)

- **Cragg-Uhler 1**: \( CU_1 = 1 - e^{\frac{-R}{N}} \)

- **Cragg-Uhler 2**: \( CU_2 = \frac{1 - e^{\frac{-R}{N}}}{1 - e^{\frac{-U}{N}}} \)

- **Estrella**: \( Estrella = 1 - \left( 1 - \frac{R}{U} \right)^{\frac{U}{N}} \)

- **Adjusted Estrella**: \( Adj.Estrella = 1 - \left( \frac{\log(L) - k}{\log(L_0)} \right)^{\frac{2}{n} \log(L_0)} \)

- **Veall-Zimmermann**: \( VZ = \frac{R(U + N)}{U(R + N)} \)
With N being the sample size and K being the number of predictors.

**Constraints for qualitative predictors**

During the calculations, when qualitative predictors are selected, each factor is broken down into a sub-matrix containing as many columns as there are categories in the factor. Typically, this is a full disjunctive table. Nevertheless, the breakdown poses a problem: if there are g categories, the rank of this sub-matrix is not g but g-1. This leads to the requirement to delete one of the columns of the sub-matrix and possibly to transform the other columns. Several strategies are available depending on the interpretation we want to make afterwards:

1) $a_1 = 0$: the parameter for the first category is null. This choice allows us force the effect of the first category as a standard.

2) $\text{Sum } (a_i) = 0$: the sum of the parameters is null.

Note: even if the choice of constraint influences the values of the parameters, it has no effect on the predicted values and on the different fitting statistics.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- ![Reload Default Options](image)
  - Click this button to reload the default options.
- ![Delete Data Selections](image)
  - Click this button to delete the data selections.
- ![Change Data Handling](image)
  - Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.
**General tab:**

**Response variable:** Select the response variable you want to model. If headers have been selected, please check the option "Variable labels" is enabled. This variable has to be a binary variable.

**Subject variable:** Select the subject variable corresponding to the name of the individuals. If headers have been selected, please check the option "Variable labels" is enabled.

**Explanatory variables:**

**Quantitative:** Activate this option if you want to include one or more quantitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The data selected may be of the numerical type. If the variable header has been selected, check that the "Variable labels" option has been activated.

**Qualitative:** Activate this option if you want to include one or more qualitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Variable labels" option has been activated.

**Range:** Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Check this option to display the results in a new worksheet in the active workbook.

**Workbook:** Check this option to display the results in a new workbook.

**Variable labels:** Check this option if the first line of the selections (data, other group) contains a label.

**Observation weights:** Activate this option if observations weights are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection.

**Options tab:**

**Tolerance:** Enter the value of the tolerance threshold below which a variable will automatically be ignored.
**Interactions / Level**: Activate this option to include interactions in the model then enter the maximum interaction level (value between 1 and 4).

**Confidence interval (%)**: Enter the percentage range of the confidence interval to use for the various tests and for calculating the confidence intervals around the parameters and predictions. Default value: 95.

**Stop conditions**:
- **Iterations**: Enter the maximum number of iterations for the Newton-Raphson algorithm. The calculations are stopped when the maximum number if iterations has been exceeded. Default value: 100.
- **Convergence**: Enter the maximum value of the evolution of the log of the likelihood from one iteration to another which, when reached, means that the algorithm is considered to have converged. Default value: 0.000001.

**Missing data** tab:
- **Remove observations**: Activate this option to remove the observations with missing data.
- **Estimate missing data**: Activate this option to estimate missing data before starting the computations.
  - **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
  - **Nearest neighbour**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.

**Outputs** tab:
- **Descriptive statistics**: Activate this option to display descriptive statistics for the variables selected.
- **Correlations**: Activate this option to display the explanatory variables correlation matrix.
- **Goodness of fit statistics**: Activate this option to display the table of goodness of fit statistics for the model.
- **Type III analysis**: Activate this option to display the type III analysis of variance table.
- **Standardized coefficients**: Activate this option if you want the standardized coefficients (beta coefficients) for the model to be displayed.
**Predictions and residuals**: Activate this option to display the predictions and residuals for all the observations.

**Charts** tab:

**Regression charts**: Activate this option to display regression chart:

- **Standardized coefficients**: Activate this option to display the standardized parameters for the model with their confidence interval on a chart.

**Results**

XLSTAT displays a large number of tables and charts to help in analyzing and interpreting the results.

**Summary statistics**: This table displays descriptive statistics for all the variables selected. For the quantitative variables, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed. For qualitative variables, including the dependent variable, the categories with their respective frequencies and percentages are displayed.

**Correlation matrix**: This table displays the correlations between the explanatory variables.

**Goodness of fit coefficients**: This table displays a series of statistics for the independent model (corresponding to the case where the linear combination of explanatory variables reduces to a constant) and for the adjusted model.

- **Observations**: The total number of observations taken into account (sum of the weights of the observations);

- **Sum of weights**: The total number of observations taken into account (sum of the weights of the observations multiplied by the weights in the regression);

- **DF**: Degrees of freedom;

- **-2 Log(Like.)**: The logarithm of the likelihood function associated with the model;

- **R² (McFadden)**: Coefficient, like the $R^2$, between 0 and 1 which measures how well the model is adjusted. This coefficient is equal to 1 minus the ratio of the likelihood of the adjusted model to the likelihood of the independent model;

- **R²(Cox and Snell)**: Coefficient, like the $R^2$, between 0 and 1 which measures how well the model is adjusted. This coefficient is equal to 1 minus the ratio of the likelihood of the
adjusted model to the likelihood of the independent model raised to the power 2/Sw, where Sw is the sum of weights.

- **R²(Nagelkerke):** Coefficient, like the R², between 0 and 1 which measures how well the model is adjusted. This coefficient is equal to ratio of the R² of Cox and Snell, divided by 1 minus the likelihood of the independent model raised to the power 2/Sw;

- **AIC:** Akaike’s Information Criterion;

- **SBC:** Schwarz’s Bayesian Criterion.

- **Iteration:** Number of iteration to reach convergence.

**Goodness of fit indexes (conditional logit):** In this table are shown the goodness of fit statistics specific to the case of the conditional logit model. For more details on these statistics, see the description part of this help.

**Test of the null hypothesis H0: Y=p0:** The H0 hypothesis corresponds to the independent model which gives probability p0 whatever the values of the explanatory variables. We seek to check if the adjusted model is significantly more powerful than this model. Three tests are available: the likelihood ratio test (-2 Log(Like.)), the Score test and the Wald test. The three statistics follow a Chi² distribution whose degrees of freedom are shown.

**Type III analysis:** This table is only useful if there is more than one explanatory variable. Here, the adjusted model is tested against a test model where the variable in the row of the table in question has been removed. If the probability Pr > LR is less than a significance threshold which has been set (typically 0.05), then the contribution of the variable to the adjustment of the model is significant. Otherwise, it can be removed from the model.

The table of **standardized coefficients** (also called beta coefficients) are used to compare the relative weights of the variables. The higher the absolute value of a coefficient, the more important the weight of the corresponding variable. When the confidence interval around standardized coefficients has value 0 (this can easily be seen on the chart of standardized coefficients), the weight of a variable in the model is not significant.

The **predictions and residuals** table shows, for each observation, its weight, the observed value of the dependent variable, the model's prediction, the same values divided by the weights, the standardized residuals and a confidence interval.
Example

An example of conditional logit model is available at the Addinsoft website:

http://www.xlstat.com/demo-clogit.htm

References


Time series visualization

Use this tool to create in three clicks as many charts as you have time series.

Description

This tool allows to create in three clicks as many charts as you have time series. It also allows you to group the series on a single graph. Finally, an option allows you to link charts to the input data: If you choose that option, charts are automatically updated when there is a change in the input data.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- OK: Click this button to start the computations.
- Cancel: Click this button to close the dialog box without doing any computation.
- Help: Click this button to display the help.
- : Click this button to reload the default options.
- : Click this button to delete the data selections.
- : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Times series: Select the data that correspond to the time series. If a header is available on the first row, make sure you activate the "Series labels" option.
**Date data**: Activate this option if you want to select date or time data. These data must be available either in the Excel date/time formats or in a numerical format. If this option is not activated, XLSTAT creates its own time variable ranging from 1 to the number of data.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Series labels**: Activate this option if the first row of the selected series includes a header.

---

**Charts** tab:

**Link the chart to the input data**: Activate this option so that a change in the input data directly results in an update of the chart.

**Display all series on a single chart**: Activate this option to display the data on a single chart.

---

**Results**

Charts are displayed for all the selected series.

---

**Exemple**

An example of time series visualization is available at the Addinsoft website:


---

**References**

Descriptive analysis (Times Series)

Use this tool to compute the descriptive statistics that are specially suited for time series analysis.

Description

One of the key issues in time series analysis is to determine whether the value we observe at time t depends on what has been observed in the past or not. If the answer is yes, then the next question is how.

The sample autocovariance function (ACVF) and the autocorrelation function (ACF) give an idea of the degree of dependence between the values of a time series. The visualization of the ACF or of the partial autocorrelation function (PACF) helps to identify the suitable models to explain the passed observations and to do predictions. The theory shows that the PACF function of an AR(p) – an autoregressive process of order p - is zero for lags greater than p.

The cross-correlations function (CCF) allows to relate two time series, and to determine if they co-vary and to which extent.

The ACVF, the ACF, the PACF and CCF are computed by this tool.

One important step in time series analysis is the transformation of time series (see Transforming time series) which goal is to obtain a white noise. Obtaining a white noise means that all deterministic and autocorrelations components have been removed. Several white noise tests, based on the ACF, are available to test whether a time series can be assumed to be a white noise or not.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
General tab:

Times series: Select the data that correspond to the time series. If a header is available on the first row, make sure you activate the "Series labels" option.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Series labels: Activate this option if the first row of the selected series includes a header.

Options tab:

Time steps: the number of time steps for which the statistics are computed can be automatically determined by XLSTAT, or set by the user.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Replace by the average of the previous and next values: Activate this option to estimate the missing data by the mean of the first preceding non missing value and of the first next non missing value.

Ignore missing data: Activate this option to ignore missing data.

Outputs tab:
**Descriptive statistics**: Activate this option to display the descriptive statistics of the selected series.

**Autocorrelations**: Activate this option to estimate the autocorrelation function of the selected series (ACF).

**Autocovariances**: Activate this option to estimate the autocovariance function of the selected series.

**Partial autocorrelations**: Activate this option to compute the partial autocorrelations of the selected series (PACF).

**Cross-correlations**: Activate this option to compute the estimate of the cross-correlation function (CCF).

**Confidence interval (%)**: Activate this option to display the confidence intervals. The value you enter (between 1 and 99) is used to determine the confidence intervals for the estimated values. Confidence intervals are automatically displayed on the charts.

- **White noise assumption**: Activate this option if you want that the confidence intervals are computed using the assumption that the time series is a white noise.

**White noise tests**: Activate this option if you want XLSTAT to display the results of the normality test and the white noise tests.

- **h1**: Enter the minimum number of lags to compute the white noise tests.
- **h2**: Enter the maximum number of lags to compute the white noise tests.
- **s**: Enter the number of lags between two series of white noise tests. s must be a multiple of (h2-h1).

**Charts tab**:

**Autocorrelogram**: Activate this option to display the autocorrelogram of the selected series.

**Partial autocorrelogram**: Activate this option to display the partial autocorrelogram of the selected series.

**Cross-correlations**: Activate this option to display the cross-correlations diagram in the case where several series have been selected.
Results

For each series, the following results are displayed:

**Summary statistics:** This table displays for the selected variables, the number of observations, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased).

**Normality and white noise tests:** Table displaying the results of the various tests. The Jarque-Bera normality test is computed once on the time series, while the other tests (Box-Pierce, Ljung-Box and McLeod-Li) are computed at each selected lag. The degrees of freedom (DF), the value of the statistics and the p-value computed using a Chi-Square(DF) distribution are displayed. For the Jarque-Bera test, the lower the p-value, the more likely the normality of the sample. For the three other tests, the lower the p-value, the less likely the randomness of the data.

**Descriptive functions for the series:** Table displaying for each time lag the values of the various selected descriptive functions, and the corresponding confidence intervals.

**Charts:** For each selected function, a chart is displayed if the "Charts" option has been activated in the dialog box.

If several time series have been selected and if the "cross-correlations" option has been selected the following results are displayed:

**Normality and white noise tests:** Table displaying the results of the various tests, Box-Pierce, Ljung-Box and McLeod-Li, which are computed at each selected lag. The degrees of freedom (DF), the value of the statistics and the p-value computed using a Chi-Square(DF) distribution are displayed. The lower the p-value, the less likely the randomness of the data.

**Cross-correlations:** Table displaying for each time lag the value of the cross-correlation function.

Example

A tutorial explaining how to use descriptive analysis with a time series is available on the Addinsoft web site. To consult the tutorial, please go to:

http://www.xlstat.com/demo-desc.htm
References


Mann-Kendall Tests

Use this tool to determine with a nonparametric test if a trend can be identified in a series, even if there is a seasonal component in the series.

Description

A nonparametric trend test has first been proposed by Mann (1945) then further studied by Kendall (1975) and improved by Hirsch et al (1982, 1984) who allowed to take into account a seasonality.

The null hypothesis $H_0$ for these tests is that there is no trend in the series. The three alternative hypotheses that there is a negative, non-null, or positive trend can be chosen.

The Mann-Kendall tests are based on the calculation of Kendall's tau measure of association between two samples, which is itself based on the ranks with the samples.

Mann-Kendall trend test

In the particular case of the trend test, the first series is an increasing time indicator generated automatically for which ranks are obvious, which simplifies the calculations. The $S$ statistic used for the test and its variance are given by:

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \text{Sgn}(x_j - x_i)$$

$$\text{Var}(S) = \frac{n(n-1)(2n+5)}{18}$$

where $n$ is the number of observations and $x_i (i=1\ldots n)$ are the independent observations.

To calculate the p-value of this test, XLSTAT can calculate, as in the case of the Kendall tau test, an exact p-value if there are no ties in the series and if the sample size is less than 50. If an exact calculation is not possible, a normal approximation is used, for which a correction for continuity is optional but recommended.

Taking into account the autocorrelations

The Mann-Kendall trend test requires that the observations are independent (meaning the correlation between the series with itself with a given lag should not be significant). In the case
where there is some autocorrelation in the series, the variance of the S statistic has been shown to be underestimated. Therefore, several improvements have been suggested. XLSTAT offers two alternative methods, the first one published by Hamed and Rao (1998) and the second by Yue and Wang (2004). The first method performs well in the case of no trend in the series (it avoids identifying a trend when it is in fact due to the autocorrelation) and the second has the advantage of performing better when there are both a trend and an autocorrelation. Before running a Mann-Kendall trend test, it is of course recommended to first check the autocorrelations of a series using the corresponding feature of XLSTAT-Time.

**Seasonal Mann-Kendall test**

In the case of seasonal Mann-Kendall test, we take into account the seasonality of the series. This means that for monthly data with seasonality of 12 months, one will not try to find out if there is a trend in the overall series, but if from one month of January to another, and from one month February and another, and so on, there is a trend.

For this test, we first calculate all Kendall's tau for each season, then calculate an average Kendall's tau. The variance of the statistic can be calculated assuming that the series are independent (eg values of January and February are independent) or dependent, which requires the calculation of a covariance. XLSTAT allows both (serial dependence or not).

To calculate the p-value of this test, XLSTAT uses a normal approximation to estimate the distribution of the average Kendall tau. A continuity correction can be used.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- Click this button to start the computations.
- Click this button to close the dialog box without doing any computation.
- Click this button to display the help.
- Click this button to reload the default options.
- Click this button to delete the data selections.
General tab:

Times series: Select the data that correspond to the time series. If a header is available on the first row, make sure you activate the "Series labels" option.

Date data: Activate this option if you want to select date or time data. These data must be available either in the Excel date/time formats or in a numerical format. If this option is not activated, XLSTAT creates its own time variable ranging from 1 to the number of data.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Series labels: Activate this option if the first row of the selected series includes a header.

Mann-Kendall trend test: Activate this option to run this test.

Seasonal Mann-Kendall test: Activate this option to run this test. Then enter the value of the period (number of lags between two seasons). Specify if you consider that there is serial dependence or not.

Options tab:

Alternative hypothesis: Choose the alternative hypothesis to be used for the test (see description).

Significance level (%): Enter the significance level for the test (default value: 5%).

Exact p-values: Activate this option if you want XLSTAT to calculate the exact p-value as far as possible (see description).

Continuity correction: Activate this option if you want XLSTAT to use the continuity correction if the exact p-values calculation has not been requested or is not possible (see description).

Autocorrelations: Activate one of the two options Hamed and Rao or Yue and Wang to into account for autocorrelations in the series. For the Hamed and Rao option you can filter out the
autocorrelations for which the p-value is not below a given level that you can set (default value: 10%).

**Missing data** tab:

**Do not accept missing data:** Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

**Remove observations:** Activate this option to remove the observations with missing data.

**Replace by the average of the previous and next values:** Activate this option to estimate the missing data by the mean of the first preceding non missing value and of the first next non missing value.

**Ignore missing data:** Activate this option to ignore missing data.

**Outputs** tab:

**Descriptive statistics:** Activate this option to display the descriptive statistics of the selected series.

**Results**

For each series, the following results are displayed:

**Summary statistics:** This table displays for the selected variables, the number of observations, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased).

**Mann-Kendall trend test:** Results of the Mann-Kendall trend test are displayed if the corresponding option has been activated. It is followed by an interpretation of the results.

**Mann-Kendall trend test:** Results of the seasonal Mann-Kendall test are displayed if the corresponding option has been activated. It is followed by an interpretation of the results.

**Example**

A tutorial explaining how to use the Mann-Kendall trend tests with a time series is available on the Addinsoft web site. To consult the tutorial, please go to:
References


Homogeneity tests

Use this tool to determine using one of four proposed tests (Pettitt, Buishand, SNHT, or von Neumann), if we may consider a series is homogeneous over time, or if there is a time at which a change occurs.

Description

Homogeneity tests involve a large number of tests for which the null hypothesis is that a time series is homogenous between two given times.

The variety of the tests comes from the fact that there are many possible alternative hypotheses: change in distribution, changes in average (one or more times) or presence of trend.

The tests presented in this tool correspond to the alternative hypothesis of a single shift. For all tests, XLSTAT provides p-values using Monte Carlo resamplings. Exact calculations are either impossible or too costly in computing time.

When presenting the various tests, by Xi (i=1, 2, ...,T) we refer to a series of T variables for which we observe xi (i=1,2,3, ..., T) at T successive times. Let \( \hat{\mu} \) be the mean of the T observed values and let \( \hat{\sigma} \) be the biased estimator of their standard deviation (we divide by T).

Note 1: If you have a clear idea of the time when the shift occurs, one can use the tests available in the parametric or nonparametric tests sections. For example, assuming that the variables follow normal distributions, one can use the test z (known variance) or the Student t test (estimated variance) to test the presence of a change at time \( \tau \). If one believes that the variance changes, you can use a comparison test of variances (F-test in the normal case, for example, or Kolmogorov-Smirnov in a more general case).

Note 2: The tests presented below are sensitive to a trend (for example a linear trend). Before applying these tests, you need to be sure you want to identify a time at which there is a shift between two homogeneous series.

Pettitt’s test

The Pettitt's test is a nonparametric test that requires no assumption about the distribution of data. The Pettitt's test is an adaptation of the tank-based Mann-Whitney test that allows identifying the time at which the shift occurs.

In his article of 1979 Pettitt describes the null hypothesis as being that the T variables follow the same distribution \( F \), and the alternative hypothesis as being that at a time \( \tau \) there is a change of distribution. Nevertheless, the Pettitt test does not detect a change in distribution if
there is no change of location. For example, if before the time $\tau$, the variables follow a normal $N(0,1)$ distribution and from time $\tau$ a $N(0,3)$ distribution, the Pettitt test will not detect a change in the same way a Mann-Whitney would not detect a change of position in such a case. In this case, one should use a Kolmogorov Smirnov based test or another method able to detect a change in another characteristic than the location. We thus reformulate the null and alternative hypotheses:

- H0: The T variables follow one or more distributions that have the same location parameter.
- Two-tailed test: Ha: There exists a time $\tau$ from which the variables change of location parameter.
- Left-tailed test: Ha: There exists a time $\tau$ from which the variables location is reduced by $\Delta$.
- Left-tailed test: Ha: There exists a time $\tau$ from which the variables location is augmented by $\Delta$.

The statistic used for the Pettitt’s test is computed as follows:

Let $D_{ij} = -1$ if $(x_i-x_j)>0$, $D_{ij} = 0$ if $(x_i-x_j)=0$, $D_{ij}=1$ if $(x_i-x_j)>0$

We then define $U_{i,T} = \sum_{i=1}^{T} \sum_{j=i+1}^{T} D_{ij}$

The Pettitt’s statistic for the various alternative hypotheses is given by:

$K_T^- = \max_{\frac{t}{15.t<T}} |U_{i,T}|$, for the two-tailed case

$K_T^- = -\min_{\frac{t}{15.t<T}} U_{i,T}$, for the left-tailed case

$K_T^+ = \max_{\frac{t}{15.t<T}} U_{i,T}$, for the right-tailed case

XLSTAT evaluates the p-value and an interval around the p-value by using a Monte Carlo method.

**Alexandersson’s SNHT test**

The SNHT test (Standard Normal Homogeneity Test) was developed by Alexandersson (1986) to detect a change in a series of rainfall data. The test is applied to a series of ratios that compare the observations of a measuring station with the average of several stations. The ratios are then standardized. The series of $X_i$ corresponds here to the standardized ratios. The null and alternative hypotheses are determined by:
- H0: The T variables $X_i$ follow a N(0,1) distribution.

- Ha: Between times 1 and $\tau$ the variables follow an N($\mu_1$, 1) distribution, and between $\tau+1$ and T they follow an N($\mu_2$,1) distribution.

The Petitt statistic is defined by:

$$T_0 = \max_{1 \leq i < T} \left[ n \bar{z}_i^2 + (n - \nu) \bar{z}_2^2 \right]$$

with

$$\bar{z}_1 = \frac{1}{\nu} \sum_{i=1}^{\nu} x_i$$

$$\bar{z}_2 = \frac{1}{n - \nu} \sum_{j=n+1}^{T} x_j$$

The $T_0$ statistic derives from a calculation comparing the likelihood of the two alternative models. The model corresponding to Ha implies that $\mu_1$ and $\mu_2$ are estimated while determining the $\nu$ parameter maximizing the likelihood.

XLSTAT evaluates the p-value and an interval around the p-value by using a Monte Carlo method.

Note: if $\nu$ is known, it is enough to run a z test on the two series of ratios. The SNHT test allows identifying the most likely $\nu$.

**Buishand’s test**

The Buishand’s test (1982) can be used on variables following any type of distribution. But its properties have been particularly studied for the normal case. In his article, Buishand focuses on the case of the two-tailed test, but for the $Q$ statistic presented below the one-sided cases are also possible. Buishand has developed a second statistic $R$, for which only a bilateral hypothesis is possible.

In the case of the $Q$ statistic, the null and alternative hypotheses are given by:

- H0: The T variables follow one or more distributions that have the same mean.

- Two-tailed test: Ha: There exists a time $\tau$ from which the variables change of mean.

- Left-tailed test: Ha: There exists a time $\tau$ from which the variables mean is reduced by $\Delta$.

- Left-tailed test: Ha: There exists a time $\tau$ from which the variables mean is augmented by $\Delta$. 
We define $S^*_0 = 0, S^*_k = \sum_{i=1}^{k} (x_i - \hat{\mu}), k = 1, 2, ..., T$

and $S^{**}_k = S^*_k / \hat{\sigma}$

The Buishand’s $Q$ statistics are computed as follows:

$$Q = \max_{1 \leq k < T} |S^{**}_k|, \text{ for the two-tailed case}$$

$$Q^- = \max_{1 \leq k < T} (S^{**}_k), \text{ for the left-tailed case}$$

$$Q^+ = -\min_{1 \leq k < T} (S^{**}_k), \text{ for the right-tailed case}$$

XLSTAT evaluates the p-value and an interval around the p-value by using a Monte Carlo method.

In the case of the $R$ statistic ($R$ stands for Range), the null and alternative hypotheses are given by:

- H0: The T variables follow one or more distributions that have the same mean.
- Two-sided test: Ha: The T variables are not homogeneous for what concerns their mean.

The Buishand’s $R$ statistic is computed as:

$$R = \max_{1 \leq k < T} (S^{**}_k) - \min_{1 \leq k < T} (S^{**}_k)$$

XLSTAT evaluates the p-value and an interval around the p-value by using a Monte Carlo method.

Note: The $R$ test does not allow detecting the time at which the change occurs.

**von Neumann’s ratio test**

The von Neumann ratio is defined by:

$$N = \frac{1}{T \hat{\sigma}} \sum_{i=1}^{T-1} (x_i - x_{i+1})^2$$

We show that the expectation of $N$ is 2 when the $X_i$ have the same mean.

XLSTAT evaluates the p-value and an interval around the p-value by using a Monte Carlo method.

Note: This test does not allow detecting the time at which the change.
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload**: Click this button to reload the default options.
- **Delete**: Click this button to delete the data selections.

**General tab:**

- **Times series**: Select the data that correspond to the time series. If a header is available on the first row, make sure you activate the "Series labels" option.
- **Date data**: Activate this option if you want to select date or time data. These data must be available either in the Excel date/time formats or in a numerical format. If this option is not activated, XLSTAT creates its own time variable ranging from 1 to the number of data.
  - **Check intervals**: Activate this option so that XLSTAT checks that the spacing between the date data is regular.
- **Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.
- **Sheet**: Activate this option to display the results in a new worksheet of the active workbook.
- **Workbook**: Activate this option to display the results in a new workbook.
- **Series labels**: Activate this option if the first row of the selected series includes a header.
**Pettitt's test**: Activate this option to run this test (see the description section for more details).

**SNHT test**: Activate this option to run this test (see the description section for more details).

**Buishand's test**: Activate this option to run this test (see the description section for more details).

**von Neumann's test**: Activate this option to run this test (see the description section for more details).

**Options** tab:

**Alternative hypothesis**: Choose the alternative hypothesis to be used for the test (see the description section for more details).

**Significance level (%)**: Enter the significance level for the test (default value: 5%).

**Monte Carlo method**: Activate this option to compute the p-value using Monte Carlo simulations. Enter the maximum number of simulations to perform and the maximum computing time (in seconds) not to exceed.

**Missing data** tab:

**Do not accept missing data**: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

**Remove observations**: Activate this option to remove the observations with missing data.

**Replace by the average of the previous and next values**: Activate this option to estimate the missing data by the mean of the first preceding non missing value and of the first next non missing value.

**Outputs** tab:

**Descriptive statistics**: Activate this option to display the descriptive statistics of the selected series.

**Charts** tab:
Display charts: Activate this option to display the charts of the series before and after transformation.

Results

For each series, the following results are displayed:

Summary statistics: This table displays for the selected variables, the number of observations, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased).

The results of various tests are then displayed. For the Pettitt’s test, the SNHT the Buishand’s Q test, charts are displayed with means $\mu_1$ and $\mu_2$ if a change-point is detected and $\mu$ if no change-point is detected.

Example

A tutorial explaining how to use the homogeneity tests is available at the Addinsoft web site. To consult the tutorial, please go to:

http://www.xlstat.com/demo-homogeneity.htm

References


Durbin-Watson test

Use this tool to check if the residuals of a linear regression are autocorrelated.

Description

Developed by J. Durbin and G. Watson (1950, 1951), the Durbin-Watson test is used to detect the autocorrelation in the residuals from a linear regression.

Denote by $Y$ the dependent variable, $X$ the matrix of explanatory variables, and the coefficients and $\varepsilon$ the error term. Consider the following model:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

In practice, the errors are often autocorrelated, it leads to undesirable consequences such as sub-optimal least-squares estimates. The Durbin-Watson test is used to detect autocorrelations in the error terms.

Assume that the $\{\varepsilon_t\}$ are stationary and normally distributed with mean 0. The null and alternative hypotheses of the Durbin-Watson test are:

$\text{H}_0$: The errors are uncorrelated.

$\text{H}_a$: The errors are AR($p$), where $p$ is the order of autocorrelation.

The Durbin-Watson D statistic writes:

$$D = \frac{\sum_{t=1}^{n} (\varepsilon_i - \varepsilon_{i-r})^2}{\sum_{t=1}^{n} \varepsilon_i^2}$$

In the context of the Durbin-Watson test, the main problem is the evaluation of the p-values which cannot be computed directly. XLSTAT-Time uses the Pan (1968) algorithm for time series with less than 70 observations and the Imhof (1961) procedure when there are more than 70 observations.
**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload default options**: Click this button to reload the default options.
- **Delete data selections**: Click this button to delete the data selections.
- **Change data handling**: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

- **Residuals**: Select the residuals from the linear regression. If the variable header has been selected, check that the "Variable labels" option has been activated.

- **X / Explanatory variables**: Select the quantitative explanatory variables in the Excel worksheet. The data selected must be of numeric type. If the variable header has been selected, check that the "Variable labels" option has been activated.

- **Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

- **Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

- **Workbook**: Activate this option to display the results in a new workbook.

- **Variable labels**: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.
**Options** tab:

**Significance level (%)**: Enter the significance level for the test (default value: 5%)

**Order**: Enter the order, i.e. the number of lags for the residuals (default value: 1)

**Missing** tab:

**Do not accept missing data**: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

**Remove observations**: Activate this option to remove the observations which include missing data.

**Replace by the average of the previous and next values**: Activate this option to estimate the missing data by the mean of the first preceding non missing value and of the first next non missing value.

**Ignore missing data**: Activate this option to ignore missing data.

**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the residuals.

**Results**

**Summary statistics**: The tables of descriptive statistics show the simple statistics for the residuals. The number of observations, missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed.

The results of the Durbin-Watson test are then displayed.

**Example**

A tutorial on the Durbin-Watson test is available on the Addinsoft website:

References


Cochrane-Orcutt estimation

Use this tool to account for serial correlation in the error term of a linear model.

Description

Developed by D. Cochrane and G. Orcutt in 1949, the Cochrane-Orcutt estimation is a well-known econometric approach to account for serial correlation in the error term of a linear model. In case of serial correlation, usual linear regression is invalid because the standard errors are not unbiased.

Denote by $Y$ the dependent variable, $X$ the matrix of explanatory variables, $\alpha$ and $\beta$ the coefficients and $\varepsilon$ the error term. Consider the following model:

$$y_t = \alpha + \beta x_t + \varepsilon_t$$

And suppose that the error term $\varepsilon$ is generated by a stationary first-order autoregressive process such that:

$$\varepsilon_t = \rho \varepsilon_{t-1} + e_t, \text{ with } |\rho| < 1$$

with $\{e_t\}$ as a white noise.

To estimate the coefficients, the Cochrane-Orcutt procedure is based on the following transformed model:

$$\forall t \geq 2, y_t - \rho y_{t-1} = \alpha (1 - \rho) + \beta (X_t - \rho X_{t-1}) + e_t$$

By introducing 3 new variables such as

$$Y^* = y_t - \rho y_{t-1}, \quad X^* = X_t - \rho X_{t-1}, \quad \lambda^* = 1 - \rho,$$

we have:

$$\forall t \geq 2, \quad y_t^* = \alpha \lambda^* + \beta X_t^* + e_t$$

Since $\{e_t\}$ is a white noise, usual statistical inference can now be used.
**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload options**: Click this button to reload the default options.
- **Delete data selections**: Click this button to delete the data selections.
- **Change data handling**: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab:**

**Y / Dependent variables:**

- **Quantitative**: Select the response variable(s) you want to model. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

**X / Explanatory variables:**

- **Quantitative**: Select the quantitative explanatory variables in the Excel worksheet. The data selected must be of numeric type. If the variable header has been selected, check that the "Variable labels" option has been activated.
- **Date data**: Activate this option if you want to select date or time data. These data must be available either in the Excel date/time formats or in a numerical format. If this option is not activated, XLSTAT creates its own time variable ranging from 1 to the number of data.
Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

Observation weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will all be taken as 1. Weights must be greater than or equal to 0. A weight of 2 is equivalent to repeating the same observation twice. If a column header has been selected, check that the “Variable labels” option has been activated.

Regression weights: Activate this option if you want to carry out a weighted least squares regression. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the “Variable labels” option is activated.

Options tab:

Tolerance: Activate this option to prevent the OLS regression calculation algorithm taking into account variables which might be either constant or too correlated with other variables already used in the model (0.0001 by default).

Confidence interval (%): Enter the percentage range of the confidence interval to use for the various tests and for calculating the confidence intervals around the parameters and predictions. Default value: 95.

Validation tab:

Validation: Activate this option if you want to use a sub-sample of the data to validate the model.

Validation set: Choose one of the following options to define how to obtain the observations used for the validation:

- Random: The observations are randomly selected. The “Number of observations” N must then be specified.
- N last rows: The N last observations are selected for the validation. The “Number of observations” N must then be specified.
• **N first rows**: The N first observations are selected for the validation. The “Number of observations” N must then be specified.

• **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to use for the validation.

**Prediction** tab:

**Prediction**: Activate this option if you want to select data to use them in prediction mode. If you activate this option, you need to make sure that the prediction dataset is structured as the estimation dataset: same variables with the same order in the selections. On the other hand, variable labels must not be selected: the first row of the selections listed below must correspond to data.

**X / Explanatory variables**: Select the quantitative explanatory variables. The first row must not include variable labels.

**Observations labels**: activate this option if observations labels are available. Then select the corresponding data. If this option is not activated, the observations labels are automatically generated by XLSTAT (PredObs1, PredObs2 …).

**Missing data** tab:

**Remove observations**: Activate this option to remove the observations with missing data.

• **Check for each Y separately**: Choose this option to remove the observations with missing data in the select Y (dependent) variables, only if the Y of interest has a missing data.

• **Across all Ys**: Choose this option to remove the observations with missing data in the Y (dependent) variables, even if the Y of interest has no missing data.

**Estimate missing data**: Activate this option to estimate missing data before starting the computations.

• **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

• **Nearest neighbour**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.

**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the variables selected.
Correlations: Activate this option to display the correlation matrix for quantitative variables (dependent or explanatory).

Analysis of variance: Activate this option to display the analysis of variance table.

Standardized coefficients: Activate this option if you want the standardized coefficients (beta coefficients) for the model to be displayed.

Predictions and residuals: Activate this option to display the predictions and residuals for all the observations.

Charts tab:

Regression charts: Activate this option to display regression chart.

- Standardized coefficients: Activate this option to display the standardized parameters for the model with their confidence interval on a chart.

- Predictions and residuals: Activate this option to display the following charts.
  
  (1) Line of regression: This chart is only displayed if there is only one explanatory variable and this variable is quantitative.
  
  (2) Explanatory variable versus standardized residuals: This chart is only displayed if there is only one explanatory variable and this variable is quantitative.
  
  (3) Dependent variable versus standardized residuals.
  
  (4) Predictions for the dependent variable versus the dependent variable.
  
  (5) Bar chart of standardized residuals.

  - Confidence intervals: Activate this option to have confidence intervals displayed on charts (1) and (4).

Results

Summary statistics: The tables of descriptive statistics show the simple statistics for all the variables selected. The number of observations, missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the dependent variables (in blue) and the quantitative explanatory variables. For qualitative explanatory variables the names of the various categories are displayed together with their respective frequencies.
Correlation matrix: This table is displayed to give you a view of the correlations between the various variables selected.

Summary of the variables selection: Where a selection method has been chosen, XLSTAT displays the selection summary. For a stepwise selection, the statistics corresponding to the different steps are displayed. Where the best model for a number of variables varying from p to q has been selected, the best model for each number or variables is displayed with the corresponding statistics and the best model for the criterion chosen is displayed in bold.

Goodness of fit statistics: The statistics related to the fitting of the regression model are shown in this table:

- **Observations**: The number of observations used in the calculations. In the formulas shown below, \( n \) is the number of observations.
- **Sum of weights**: The sum of the weights of the observations used in the calculations. In the formulas shown below, \( W \) is the sum of the weights.
- **DF**: The number of degrees of freedom for the chosen model (corresponding to the error part).
- **R²**: The determination coefficient for the model. This coefficient, whose value is between 0 and 1, is only displayed if the constant of the model has not been fixed by the user. Its value is defined by:
  \[
  R^2 = 1 - \frac{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} w_i (y_i - \bar{y})^2}, \quad \text{where} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} w_i y_i,
  \]
  The \( R^2 \) is interpreted as the proportion of the variability of the dependent variable explained by the model. The nearer \( R^2 \) is to 1, the better is the model. The problem with the \( R^2 \) is that it does not take into account the number of variables used to fit the model.
- **Adjusted R²**: The adjusted determination coefficient for the model. The adjusted \( R^2 \) can be negative if the \( R^2 \) is near to zero. This coefficient is only calculated if the constant of the model has not been fixed by the user. Its value is defined by:
  \[
  \hat{R}^2 = 1 - (1 - R^2) \frac{W - 1}{W - p - 1}
  \]
  The adjusted \( R^2 \) is a correction to the \( R^2 \) which takes into account the number of variables used in the model.
- **MSE**: The mean squared error (MSE) is defined by:
  \[
  MSE = \frac{1}{W - p} \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2
  \]
• **RMSE**: The root mean square of the errors (RMSE) is the square root of the MSE.

• **MAPE**: The *Mean Absolute Percentage Error* is calculated as follows:

\[
MAPE = \frac{100}{W} \sum_{i=1}^{n} w_i \left| \frac{y_i - \hat{y}_i}{y_i} \right|
\]

• **DW**: The Durbin-Watson statistic is defined by:

\[
DW = \frac{\sum_{i=2}^{n} \left[ (y_i - \hat{y}_i) - (y_{i-1} - \hat{y}_{i-1}) \right]^2}{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}
\]

This coefficient is the order 1 autocorrelation coefficient and is used to check that the residuals of the model are not autocorrelated, given that the independence of the residuals is one of the basic hypotheses of linear regression. The user can refer to a table of Durbin-Watson statistics to check if the independence hypothesis for the residuals is acceptable.

• **Cp**: Mallows Cp coefficient is defined by:

\[
Cp = \frac{SSE}{\hat{\sigma}} + 2p^* - W
\]

where SSE is the sum of the squares of the errors for the model with p explanatory variables and \( \hat{\sigma} \) is the estimator of the variance of the residuals for the model comprising all the explanatory variables. The nearer the Cp coefficient is to \( p^* \), the less the model is biased.

• **AIC**: Akaike's Information Criterion is defined by:

\[
AIC = W \ln \left( \frac{SSE}{W} \right) + 2p^*
\]

This criterion, proposed by Akaike (1973) is derived from the information theory and uses Kullback and Leibler's measurement (1951). It is a model selection criterion which penalizes models for which adding new explanatory variables does not supply sufficient information to the model, the information being measured through the MSE. The aim is to minimize the AIC criterion.

• **SBC**: Schwarz’s Bayesian Criterion is defined by:

\[
SBC = W \ln \left( \frac{SSE}{W} \right) + \ln(W) p^*
\]
This criterion, proposed by Schwarz (1978) is similar to the AIC, and the aim is to minimize it.

- **PC**: Amemiya’s Prediction Criterion is defined by:
  \[
  PC = \frac{(1 - R^2)(W + p^*)}{W - p^*}
  \]

  This criterion, proposed by Amemiya (1980) is used, like the adjusted R² to take account of the parsimony of the model.

- **Press RMSE**: Press' statistic is only displayed if the corresponding option has been activated in the dialog box. It is defined by:
  \[
  \text{Press} = \sum_{i=1}^{n} w_i \left( y_i - \hat{y}_{i(-i)} \right)^2
  \]
  where \( \hat{y}_{i(-i)} \) is the prediction for observation i when the latter is not used for estimating parameters. We then get:
  \[
  \text{Press RMSE} = \sqrt{\frac{\text{Press}}{W - p^*}}
  \]
  Press's RMSE can then be compared to the RMSE. A large difference between the two shows that the model is sensitive to the presence or absence of certain observations in the model.

The **analysis of variance table** is used to evaluate the explanatory power of the explanatory variables. Where the constant of the model is not set to a given value, the explanatory power is evaluated by comparing the fit (as regards least squares) of the final model with the fit of the rudimentary model including only a constant equal to the mean of the dependent variable. Where the constant of the model is set, the comparison is made with respect to the model for which the dependent variable is equal to the constant which has been set.

The **parameters of the model** table display the estimate of the parameters, the corresponding standard error, the Student’s t, the corresponding probability, as well as the confidence interval. The autocorrelation coefficient \( \rho \) is also displayed.

The **equation of the model** is then displayed to make it easier to read or re-use the model.

Auto-correlation coefficient: The estimated value of the autocorrelation coefficient \( \rho \).
The table of standardized coefficients (also called beta coefficients) is used to compare the relative weights of the variables. The higher the absolute value of a coefficient, the more important the weight of the corresponding variable. When the confidence interval around standardized coefficients has value 0 (this can be easily seen on the chart of normalized coefficients), the weight of a variable in the model is not significant.

The predictions and residuals table shows, for each observation, its weight, the value of the qualitative explanatory variable, if there is only one, the observed value of the dependent variable, the model's prediction, the residuals and the confidence intervals with the fitted prediction. Two types of confidence intervals are displayed: a confidence interval around the mean (corresponding to the case where the prediction would be made for an infinite number of observations with a set of given values for the explanatory variables) and an interval around the isolated prediction (corresponding to the case of an isolated prediction for the values given for the explanatory variables). The second interval is always greater than the first, the random values being larger. If the validation data have been selected, they are displayed at the end of the table.

The charts which follow show the results mentioned above. If there is only one explanatory variable in the model, the first chart displayed shows the observed values, the regression line and both types of confidence interval around the predictions. The second chart shows the normalized residuals as a function of the explanatory variable. In principle, the residuals should be distributed randomly around the X-axis. If there is a trend or a shape, this shows a problem with the model.

The three charts displayed next show respectively the evolution of the standardized residuals as a function of the dependent variable, the distance between the predictions and the observations (for an ideal model, the points would all be on the bisector), and the standardized residuals on a bar chart. The last chart quickly shows if an abnormal number of values are outside the interval ]-2, 2[ given that the latter, assuming that the sample is normally distributed, should contain about 95% of the data.

If you have selected the data to be used for calculating predictions on new observations, the corresponding table is displayed next.

Example

A tutorial on the Cochrane-Orcutt estimation is available on the Addinsoft website:
http://www.xlstat.com/demo-cochorcutt.htm
References

**Heteroscedasticity tests**

Use this tool to determine whether the residuals from a linear regression can be considered as having a variance that is independent of the observations.

**Description**

The concept of heteroscedasticity - the opposite being homoscedasticity - is used in statistics, especially in the context of linear regression or for time series analysis, to describe the case where the variance of errors or the model is not the same for all observations, while often one of the basic assumption in modeling is that the variances are homogeneous and that the errors of the model are identically distributed.

In linear regression analysis, the fact that the errors of the model (also named residuals) are not homoskedastic has the consequence that the model coefficients estimated using ordinary least squares (OLS) are neither unbiased nor those with minimum variance. The estimation of their variance is not reliable.

If it is suspected that the variances are not homogeneous (a representation of the residuals against the explanatory variables may reveal heteroscedasticity), it is therefore necessary to perform a test for heteroscedasticity. Several tests have been developed, with the following null and alternative hypotheses:

- **H0**: The residuals are homoscedastic
- **Ha**: The residuals are heteroscedastic

**Breusch-Pagan test**

This test has been developed by Breusch and Pagan (1979), and later improved by Koenker (1981) - which is why this test is sometimes named the Breusch-Pagan and Koenker test - to allow identifying cases of heteroscedasticity, which make the classical estimators of the parameters of the linear regression unreliable. If e is the vector of the errors of the model, the null hypothesis **H0** can write:

\[
\text{Var}(u/x) = \sigma^2
\]

\[
\text{H0}: \text{Var}(u/x) = \text{E}(e^2/x) = \text{E}(e^2/x_1, x_2, ..., x_k) = \text{E}(e^2) = \sigma^2
\]

To verify that the quadratic errors are independent of the explanatory variables, which can translate into many functional forms, the simplest is to regress the squared errors by the explanatory variables. If the data are homoskedastic, the coefficient of determination R² should then not be equal to 0. If H0 is not rejected we can conclude that heteroscedasticity, if it exists,
does not take the functional form used. Practice shows that heteroscedasticity is not a problem if H0 is not accepted. If H0 is rejected, it is likely that there is heteroscedasticity and that it takes the functional form described above.

The statistic used for the test, proposed by Koenker (1981) is:

\[ LM = nR^2 \]

where LM stands for Lagrange multiplier. This statistic has the advantage of asymptotically following a Chi-square distribution with p degrees of freedom, where p is the number of explanatory variables.

If the null hypothesis is rejected, it will be necessary to transform the data before doing the regression, or using modeling methods to take into account the variability of the variance.

**White test and modified White test (Wooldridge)**

This test was developed by White (1980) to identify cases of heteroscedasticity making classical estimators of the parameters of linear regression unreliable. The idea is similar to that of Breusch and Pagan, but it relies on weaker assumptions as for the form that heteroscedasticity takes. This results in a regression of the quadratic errors by the explanatory variables and by the squares and cross-products of the latter (for example for two regressors, we take x1, x2, x1², x2², x1x2 to model squared errors). The statistic used is the same as the test-Breusch Pagan, but due to the presence of many more regressors, there are here 2p + p * (p-1) / 2 degrees of freedom for the Chi-square distribution.

In order to avoid losing too many degrees of freedom, Wooldridge (2009) proposed to regress the squared errors by the model predictions and by their square. This reduces to 2 the number of degrees of freedom for the Chi-square.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
General tab:

Residuals: Select the residuals from the linear regression. If the variable header has been selected, check that the "Variable labels" option has been activated.

X / Explanatory variables: Select the quantitative explanatory variables in the Excel worksheet. The data selected must be of numeric type. If the variable header has been selected, check that the "Labels included" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Labels included: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

Breusch-Pagan test: Activate this option to run a Breusch-Pagan test.

White test: Activate this option to run a White test. Activate the "Wooldridge" option if you want to use the modified version of the test (see the description chapter for further details).

Options tab:

Significance level (%): Enter the significance level for the test (default value: 5%).

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.
Outputs tab:

Descriptive statistics: Activate this option to display the descriptive statistics of the selected series.

Charts tab:

Display charts: Activate this option to display the scatter plot of the residuals versus the explanatory variable.

Results

Summary statistics: This table displays for the selected variables, the number of observations, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased).

The results of the selected tests are then displayed.

Example

A tutorial explaining how to use the heteroscedasticity tests is available on the Addinsoft website. To consult the tutorial, please go to:

http://www.xlstat.com/demo-whitetest.htm

References


**Unit root and stationarity tests**

Use this tool to determine whether a series is stationary or not.

**Description**

A time series $Y_t$ ($t=1,2...$) is said to be stationary (in the week sense) if its statistical properties do not vary with time (expectation, variance, autocorrelation). The white noise is an example of a stationary time series, with for example the case where $Y_t$ follows a normal distribution $N(\mu, \sigma^2)$ independent of $t$. An example of a non-stationary series is the random walk defined by:

$$Y_t = Y_{t-1} + \varepsilon_t$$

where $\varepsilon_t$ is a white noise.

Identifying that a series is not stationary allows to afterwards study where the non-stationarity comes from. A non-stationary series can, for example, be stationary in difference: $Y_t$ is not stationary, but the $Y_t - Y_{t-1}$ difference is stationary. It is the case of the random walk. A series can also be stationary in trend. It is the case with the series defined by:

$$Y_t = 0.5X_{t-1} + 1.4t + \varepsilon_t$$

where $\varepsilon_t$ is a white noise, that is not stationary.

On the other hand, the series $Y_t - 1.4t = 0.5Y_{t-1} + \varepsilon_t$ is stationary. $Y_t$ is also stationary in difference.

Stationarity tests allow verifying whether a series is stationary or not. There are two different approaches: some tests consider as null hypothesis $H_0$ that the series is stationary (KPSS test, Leybourne and McCabe test), and for other tests, on the opposite, the null hypothesis is on the contrary that the series is not stationary (Dickey-Fuller test, augmented Dickey-Fuller test, Phillips-Perron test, DF-GLS test). XLSTAT includes the KPSS test, the Dickey-Fuller test and its augmented version and the Phillips-Perron test.

**Dickey-Fuller test**

This test has been developed by Dickey and Fuller (1979) to allow identifying a unit root in a time series for which one thinks there is an order 1 autoregressive component, and may be as well a trend component linearly related to the time. As a reminder, an order 1 autoregressive model (noted AR(1)), can be written as follows:

$$X_t = \rho X_{t-1} + \varepsilon_t, t=1,2..., $$

where the $\varepsilon_t$ are independent identically distributed variables that follow an $N(0, \sigma^2)$ normal distribution.

The series is stationary if $|\rho|<1$. It is not stationary and corresponds to a random walk if $\rho=1$.

If one adds a constant and a trend to the model, the model writes:
\[ X_t = \rho X_{t-1} + \alpha + \beta t + \varepsilon_t, \quad t=1,2,..., \] where the \( \varepsilon_t \) are independent identically distributed variables that follow an \( N(0, \sigma^2) \) normal distribution.

Dickey and Fuller decided to take as null hypothesis \( \rho=1 \) because it has an immediate operational impact: if the null hypothesis is not rejected, then, in order to be able to analyze the time series and if necessary to make predictions, it is necessary to transform the series, using differencing (see the Time series transformation and ARIMA tools).

The two possible alternative hypotheses are:

- Ha(1): \(|\rho|<1\), the series is stationary
- Ha(2): \(|\rho|>1\), the series is explosive

The statistics used in the Dickey-Fuller test are computed using a linear regression model, and correspond to the \( t \) statistic computed by dividing the coefficient of the model by its standard error. Dickey and Fuller define:

- \( \text{AR}(1) \) model:
  \[ \hat{\tau} = \left( \hat{\rho} - 1 \right) / \sqrt{S^2_1 c_1} \]

- \( \text{AR}(1) \) model with constant \( \mu \):
  \[ \hat{\tau}_\mu = \left( \hat{\rho}_\mu - 1 \right) / \sqrt{S^2_2 c_2} \]

- \( \text{AR}(1) \) model with constant \( \mu \) and a linear trend function of \( t \):
  \[ \hat{\tau}_\gamma = \left( \hat{\rho}_\gamma - 1 \right) / \sqrt{S^2_3 c_3} \]

The \( S^2 \) correspond to the mean squared error and the \( c_k \) to variances.

While these statistics are straightforward to compute, their exact and asymptotic distributions are complex. The critical values have been estimated through Monte Carlo simulations by the authors, with several improvements over time, as the machines allowed more simulations. MacKinnon (1996) has proposed an approach based on numerous Monte Carlo simulations that allows to compute p-values and critical values for various sample sizes. XLSTAT estimates critical values and p-values either by running a predefined set of Monte Carlo simulations for the considered sample size or the surface regression approach proposed by MacKinnon (1996).

Dickey et Fuller have shown that these distributions do not depend on the distribution of the \( \varepsilon_t \) and on the initial value of the series, \( Y_0 \).

Fuller (1976) had already shown that this approach can be generalized to AR(p) models to determine whether there exists a unit root while not being able to say from which term in the model the non-stationarity comes from.
Augmented Dickey-Fuller test

This test has been developed by Said et Dickey (1984) and complements the Dickey-Fuller test by generalizing the approach valid for AR(p) models to ARMA(p, q) models, for which we assume that it is in fact an ARIMA(p, d, q), with d\geq 1 under the null hypothesis H0. Said and Dickey show that it is not necessary to know p, d and q to apply the Dickey-Fuller test presented above. However, a k parameter, corresponding to the horizon to consider for the mobile mean of the model must be provided by the user so that the test can be run. By default, XLSTAT recommends the following value:

\[ k = \text{INT}((n - 1)^{1/3}) \]

where INT() is the integer part

Said and Dickey show that the statistic \( \tau \) of the Dickey-Fuller test can be used. Its asymptotic distribution is the same as the one of the Dickey-Fuller test.

Phillips-Perron test

An alternative generalization of the Dickey-Fuller test to more complex data generation processes was introduced by Phillips (1987a) and further developed in Perron (1988) and Phillips and Perron (1988).

As for the DF test, three possible regressions are considered in the Phillips-Perron (PP) test, namely, without an intercept, with an intercept and with an intercept and a time trend. Those are given in the following equations, respectively.

\[ X_t = \rho X_{t-1} + \varepsilon_t \]
\[ X_t = \rho X_{t-1} + \alpha + \varepsilon_t \]
\[ X_t = \rho X_{t-1} + \alpha + \beta \cdot (t - T / 2) + \varepsilon_t \]

It should be noted that within the PP test, the error term \( \varepsilon_t \) is expected to have a null average but it can be serially correlated and/or heteroscedastic.

Unlike the augmented Dickey-Fuller (ADF) test, the PP test does not deal with serial correlation at the regression level. Instead, a non parametric correction is applied to the statistic itself to account for potential effects of heteroscedasticity and serial correlations on the adjustment residuals. The statistic noted \( Z_\tau \) is given by:
\[ Z_t = \left( \frac{\sigma}{\hat{\lambda}} \right) T^{\rho/2} - \frac{1}{2} \left( \frac{\hat{\lambda} - \sigma}{\hat{\lambda}} \right) \left( \frac{T \times SE(\rho)}{\sigma} \right)^{1/2} \]

Where \( \hat{\lambda} \) and \( \sigma \) are consistent estimates of the variance parameters:

\[ \lambda^2 = \lim_{n \to +\infty} T^{-1} \sum_{t=1}^{T} E \left[ T^{-1} \left( \sum_{i=1}^{T} \epsilon_i \right)^2 \right], \quad \sigma^2 = \lim_{n \to +\infty} T^{-1} \sum_{t=1}^{T} E[\epsilon_i^2] \]

And \( t^\rho = \frac{\rho-1}{SE(\rho)} \)

The estimator \( \hat{\lambda} \) is the one proposed by Newey and West (1987). It guarantees the robustness of the statistic against heteroscedasticity and serial correlations.

- Short (default option): the number of steps considered for the computation of the Newey-West estimator is given by \( k = ENT\left( 4 \left( \frac{T}{100} \right)^{2/9} \right) \)

- Long: for series resulting from a higher-order MA process, the number of steps is given by \( k = ENT\left( 12 \left( \frac{T}{100} \right)^{2/9} \right) \)

Where \( ENT(\cdot) \) is the integer part.

The PP test uses the same distribution as the DF or ADF t-statistic. Critical value and p-value estimates are made following the surface regression approach proposed by MacKinnon (1996) or using Monte Carlo simulations.

One of the advantages of the PP test over the ADF test is to allow for heteroscedasticity in the data generation process of \( \epsilon_i \). Furthermore, there is no need for a sensitive parametrization of the Newey-West estimator as for the ADF test.

**KPSS test of stationarity**

This test takes its name from its authors, Kwiatkowski, Phillips, Schmidt and Shin (1991). Contrary to the Dickey-Fuller tests, this test allows testing the null hypothesis that the series is stationary. Consider the model where
\[ Y_t = \xi + r_t + \varepsilon_t, \quad t=1,2,\ldots, \]

where \( \varepsilon_t \) is a stationary error, and \( r_t \) is a random walk defined by

\[ r_t = r_{t-1} + u_t, \]

where \( r_0 \) is a constant, and the \( u_t \) are independent identically distributed variables with mean 0 and variance \( \sigma^2 \).

The \( Y_t \) series is stationary in the case where the \( \sigma^2 \) variance is null. It is stationary in trend if \( \xi \) is not null, and stationary in level (around \( r_0 \)) if \( \xi = 0 \).

Let \( n \) be the number of time steps available for the series. Let \( e_t \) be the residuals, when regressing the \( y_t \) by the time and a constant, when one wants to test stationarity in trend, or when comparing the series with its mean when testing for stationarity in level.

We define:

\[ s^2(l) = \frac{1}{n} \sum_{t=1}^{n} e_t^2 + \frac{2}{n} \sum_{s=1}^{l} w(s,l) \sum_{t=s+1}^{n} e_t e_{t-s} \]

with

\[ w(s,l) = 1 - s(l + 1) \]

Let \( S^2 \) be the mean of squared errors between times 1 and \( t \). The statistic used for the "Level" stationarity test is given by:

\[ \eta_\mu = \frac{1}{n} \sum_{t=1}^{n} S_t^2 / s^2(l) \]

For the "Trend" stationarity test we use:

\[ \eta_\tau = \frac{1}{n} \sum_{t=1}^{n} S_t^2 / s^2(l) \]

the difference between both comes from the different residuals.

As with the Dickey-Fuller test, these statistics are easy to compute, but their exact and asymptotic distributions are complex. Kwiatkowski et al. computed the asymptotic critical values using Monte Carlo simulations. XLSTAT allows to compute critical values and p-values adapted to the size of the sample, using Monte Carlo simulations for each new run.

**Weighting with the Newey-West method**

The Newey-West (1987) estimator is used to reduce the effect of dependence (correlation, autocorrelation) and heteroscedasticity (non homogeneous variances) of the error terms of a model. The idea is to balance the model errors in the calculation of statistics involving them. If \( L \) is the number of steps taken into account, the weight of each error is given by:
The KPSS test uses linear regressions that assume the homoscedasticity of the errors. The use of the Newey-West weighting is recommended by the authors and is available in XLSTAT. XLSTAT recommends for the value of $L$:

- Short: $L = \lfloor 3 \times \sqrt{n} / 13 \rfloor$
- Long: $L = \lfloor 10 \times \sqrt{n} / 14 \rfloor$

where $\lfloor \cdot \rfloor$ is the integer part.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload**: Click this button to reload the default options.
- **Delete**: Click this button to delete the data selections.

**General** tab:

**Times series**: Select the data that correspond to the time series. If a header is available on the first row, make sure you activate the "Series labels" option.

**Date data**: Activate this option if you want to select date or time data. These data must be available either in the Excel date/time formats or in a numerical format. If this option is not activated, XLSTAT creates its own time variable ranging from 1 to the number of data.
Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Series labels: Activate this option if the first row of the selected series includes a header.

Dickey-Fuller test: Activate this option to run a Dickey-Fuller test. Choose the type of test you want to use (see the description section for further details).

Phillips-Perron test: Activate this option to run a Phillips-Perron test. Choose the type of test you want to use (see the description section for further details).

KPSS test: Activate this option to run a KPSS test. Choose the type of test you want to use (see the description section for further details).

Options tab:

Significance level (%): Enter the significance level for the test (default value: 5%).

Method: choose the method to use for the p-value and critical value estimates
- Monte Carlo: selects Monte Carlo simulations based estimates.

Dickey-Fuller test: In the case of a Dickey-Fuller test, you can use the default value of k (see the "Description" section for more details) or enter your own value.

Phillips-Perron test: for a Phillips-Perron test, you should select either the short (default value) or the long number of steps (see the "Description" section for more details).

KPSS test: Choose whether you want to use the Newey-West weighting to remove the impact of possible autocorrelations in the residuals of the model. For the lag to apply, you can choose between short, long, or you can enter your own value for L (see the "Description" section for more details).

Missing tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.
Remove observations: Activate this option to remove the observations with missing data.

Replace by the average of the previous and next values: Activate this option to estimate the missing data by the mean of the first preceding non missing value and of the first next non missing value.

Ignore missing data: Activate this option to ignore missing data.

Outputs tab:

Descriptive statistics: Activate this option to display the descriptive statistics of the selected series.

Charts tab:

Display charts: Activate this option to display the charts of the series.

Results

Summary statistics: This table displays for the selected variables, the number of observations, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased).

The results of the selected tests are then displayed.

Example

A tutorial explaining how to perform unit root tests or stationarity test is available on the Addinsoft web site. To consult the tutorial, please go to:

http://www.xlstat.com/demo-unitroot.htm

References


Cointegration tests

Use this module to perform VAR-based cointegration tests on a group of two or more I(1) time series using the approach proposed by Johansen (1991, 1995).

Description

Economic theory often suggests long term relationship between two or more economic variables. Although those variables can derive from each other on a short term basis, the economic forces at work should restore the original equilibrium between them in the long run. Examples of such relationships in economics include money with income, prices and interest rates or exchange rate with foreign and domestic prices. In finance, such relationships are expected for instance between the prices of the same asset on different market places.

The term of cointegration was first introduced by Engle and Granger (1987) after the work of Granger and Newbold (1974) on spurious regression. It identifies a situation where two or more non stationary time series are bound together in such a way that they cannot deviate from some equilibrium in the long term. In other words, there exists one or more linear combination of those I(1) time series (or integrated of order 1, see unit root test) that is stationary (or I(0)). Those stationary combinations are called cointegrating equations.

One of the most interesting approaches for testing for cointegration within a group of time series is the maximum likelihood methodology proposed by Johansen (1988, 1991). This approach, implemented in XLSTAT, is based on Vector Autoregressive (VAR) models and can be described as follows.

First consider the levels VAR($P$) model for $Y_t$, a $K$-vector of I(1) time series:

$$Y_t = \Phi D_t + \Pi_1 Y_{t-1} + ... + \Pi_p Y_{t-p} + \epsilon_t \quad \text{for } t = 1, ..., T$$

Where $D_t$ contains deterministic terms such as constant or trend and $\epsilon_t$ is the vector of innovations.

The parameter $P$ is the VAR order and is one of the input parameter to Johansen’s methodology for testing cointegration. If you don’t know which value this parameter should take for you data set, you should select the option automatic in the General tab. You will then have to specify the model that best describes your data in the option tab (no trend nor intercept, intercept, trend or trend and intercept), set a maximum number of lags to evaluate and choose the discriminating criterion among the 4 proposed (AIC, FPE, HQ, BIC). XLSTAT will then estimate the parameter $P$ following the approach detailed in Lüktépohl (2005) and perform subsequent analysis. Detailed results are provided at the end of the analysis for further control.

According to the Granger representation theorem, a VAR($P$) model with I(1) variables can equivalently be represented as a Vector Error Correction Model (VECM):
\[ \Delta Y_t = \Phi D_t + \Pi Y_{t-1} + \Gamma_1 \Delta Y_{t-1} + \ldots + \Gamma_{p-1} \Delta Y_{t-p+1} + \varepsilon_t \]

Where \( \Delta \) denotes the difference operator, \( \Pi = \Pi_1 + \ldots + \Pi_{p-1} - I_K \) and \( \Gamma_i = -\sum_{j=l+1}^{p} \Pi_j \) for \( l = 1, \ldots, P - 1 \).

In this representation, \( \Delta Y_t \) and its lags are all I(0). The term \( Y_{t-1} \) is the only potentially non-stationary component. Therefore for the above equation to hold (a linear combination of I(0) terms is also I(0)), the term \( \Pi Y_{t-1} \) must contain the cointegration relationship if it exists.

Three cases can be considered:

- the matrix \( \Pi \) equals 0 (\( \text{rank}(\Pi) = 0 \)), then no cointegration exists,
- the matrix \( \Pi \) has full rank (\( \text{rank}(\Pi) = K \)), then each independent component of \( Y_{t-1} \) is I(0) (which violates or first assumption of I(1) series),
- the matrix \( \Pi \) is neither null nor of full rank \( 0 < \text{rank}(\Pi) < K \), then \( Y_{t-1} \) is I(1) with \( r \) linearly independent cointegrating vectors and \( K - r \) common stochastic trends.

In the latter case, the matrix \( \Pi \) can be written as the product:

\[
\Pi_{(K \times K)} = \alpha (K \times r) \beta'^{(r \times K)}
\]

Where \( \text{rank}(\alpha) = \text{rank}(\beta) = r \).

The matrix \( \beta \) is the cointegrating matrix and its columns form a basis for the cointegrating coefficients. The matrix \( \alpha \) also known as the adjustment matrix (or loading matrix) controls the speed at which the effect of \( Y_{t-1} \) propagates to \( \Delta Y_t \). It is important to note that the factorization \( \Pi = \alpha \beta' \) is not uniquely defined and may require some arbitrary normalization to obtain unique values of \( \alpha \) and \( \beta \). Values reported in XLSTAT use the normalization \( \beta' S_{11} \beta = I_r \) proposed by Johansen (1995).

The test methodology estimates the matrix \( \Pi \) and constructs successive likelihood ratio (LR) tests for its reduced rank on the estimated eigenvalues of \( \Pi : \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_K \). The reduced rank of \( \Pi \) is equal to the number of non-zero eigenvalues. It is also the rank of cointegration of the system (or equivalently the number of cointegrating equations).

Two sequential procedures proposed by Johansen are implemented to evaluate the cointegration rank \( r_0 \):

- the \( \lambda_{\text{max}} \)-test (or lambda max) uses the statistic \( LR_{\text{max}}(r_0) = -T \ln(1 - \hat{\lambda}_{r_0+1}) \),

\[ 919 \]
- the trace test for which the statistic is \( LR_{\text{trace}}(r_0) = -T \sum_{i=r_0+1}^{n} \ln(1 - \hat{\lambda}_i) \).

Starting from the Null hypothesis that no cointegration relationship exists \( (r_0 = 0) \), the \( \hat{\lambda}_{r_0+1} \)-test will test that the \( (r_0 + 1)^{th} \) eigenvalue can be accepted to be zero. If the hypothesis of \( \hat{\lambda}_{r_0+1} \approx 0 \) is rejected, then the next level of cointegration can be tested. Similarly, \( LR_{\text{trace}} \) of the trace test should be close to zero if the rank of \( \Pi \) equals \( r_0 \) and large if it is greater than \( r_0 \).

The asymptotic distributions of those LR tests are non standard and depend on the assumption made on the deterministic trends of \( \Delta Y_t \) which can be rewritten as:

\[
\Delta Y_t = c_1 + d_1 t + \alpha (\beta^\prime Y_{t-1} + c_0 + d_0 t) + \Gamma_1 \Delta Y_{t-1} + \ldots + \Gamma_{p-1} \Delta Y_{t-p+1} + \varepsilon_t
\]

5 types of restriction are considered depending on the trending nature of both \( Y_t \) and \( \beta^\prime Y_t \) (the cointegrating relationships):

- \( H_2 \) \( (c_0 = c_1 = d_0 = d_1 = 0) \): the series in \( Y_t \) are \( I(1) \) with no deterministic trends in levels and \( \beta^\prime Y_t \) have means zero. In practice, this case is rarely used.

- \( H_{1^*} \) \( (c_1 = d_0 = d_1 = 0) \): the series in \( Y_t \) are \( I(1) \) with no deterministic trends in levels and \( \beta^\prime Y_t \) have non-zero means.

- \( H_1 \) \( (d_0 = d_1 = 0) \): the series in \( Y_t \) are \( I(1) \) with linear trends in levels and \( \beta^\prime Y_t \) have non-zero means.

- \( H^* \) \( (d_1 = 0) \): the series in \( Y_t \) and \( \beta^\prime Y_t \) have linear trends.

- \( H \) (unconstrained): the series in \( Y_t \) are \( I(1) \) with quadratic trends in levels and \( \beta^\prime Y_t \) have linear trends. Again, this case is hardly used in practice.

To perform a cointegration test in XLSTAT, you have to choose one of the above assumptions. The choice should be motivated by the specific nature of your data and the considered economics model. However, if it is unclear which restriction applies best, a good strategy might be to evaluate the robustness of the result by successively selecting a different assumption among \( H_{1^*} \), \( H_1 \) and \( H^* \) (the remaining 2 options being very specific and easily identifiable). Critical values and p-values for both the \( \hat{\lambda}_{r_0+1} \)-test and the trace test are computed in XLSTAT as proposed by MacKinnon-Haug-Mechelis (1998).
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

{bmct ok.bmp}: Click this button to start the computations.

{bmct cancel.bmp}: Click this button to close the dialog box without doing any computation.

{bmct help.bmp}: Click this button to display the help.

{bmct reset56.bmp}: Click this button to reload the default options.

{bmct erase.bmp}: Click this button to delete the data selections.

General tab:

Times series: Select the data that correspond to the time series. If a header is available on the first row, make sure you activate the "Series labels" option.

Date data: Activate this option if you want to select date or time data. These data must be available either in the Excel date/time formats or in a numerical format. If this option is not activated, XLSTAT creates its own time variable ranging from 1 to the number of data.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Series labels: Activate this option if the first row of the selected series includes a header.

Model: Select between H2, H1*, H1, H* and H the type of restriction that best describes your data set (see the description for further details).

VAR order: Select the automatic option for an automatic estimation of the $P$ parameter (see the description for further details) or select the user defined option and enter your own value.

Options tab:
Significance level (%): Enter the significance level for the test (default value: 5%).

VAR order estimation: If the automatic option is selected for the VAR order on the General tab, you must set three parameters: the model, the selection criterion and the maximum number of lag.

Model: Select between None, Intercept, Trend and Intercept + trend the model that best describes your time series.

Selection criterion: Select between the four criteria computed (AIC, FPE, HQ and BIC), the one XLSTAT will use to select the VAR order.

Maximum number of lag: Select the maximum number of lag that will be computed by XLSTAT to select the VAR order.

Missing tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Replace by the average of the previous and next values: Activate this option to estimate the missing data by the mean of the first preceding non missing value and of the first next non missing value.

Outputs tab:

Descriptive statistics: Activate this option to display the descriptive statistics of the selected series.

Results

Summary statistics: This table displays for the selected variables, the number of observations, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased).

VAR order estimation: If the automatic option is selected for the VAR order, this table displays the four criteria values for the VAR order estimation. Each line corresponds to the evaluation of one number of lags from 1 up to the maximum number of lag. The discriminating criterion is in bold.
**Lambda max test**: This table displays for each rank of cointegration tested the corresponding eigenvalue, the lambda max test statistic and the associated critical value and p-values.

**Trace test**: This table displays for each rank of cointegration tested the corresponding eigenvalue, the trace test statistic and the associated critical value and p-values.

**Adjustment coefficients (alpha)**: This table displays the resulting loading matrix $\alpha$ (see description for further details).

**Cointegration coefficients (beta)**: This table displays the cointegrating matrix $\beta$ (see description for further details).

**Example**

A tutorial explaining how to perform cointegration analysis on time series is available on the Addinsoft web site. To consult the tutorial, please go to:


**References**


Time series transformation

Use this tool to transform a time series A into a time series B that has better properties: removed trend, reduced seasonality, and better normality.

Description

XLSTAT offers four different possibilities for transforming a time series \{X_t\} into \{Y_t\}, (t=1,…,n):

**Box-Cox transformation** to improve the normality of the time series; the Box-Cox transformation is defined by the following equation:

\[
Y_t = \begin{cases} 
\frac{X_t^{\lambda} - 1}{\lambda}, & (X_t > 0, \lambda \neq 0) \\
\ln(X_t), & X_t > 0, \lambda = 0
\end{cases}
\]

XLSTAT accepts a fixed value of \(\lambda\), or it can find the value that maximizes the likelihood, the model being a simple linear model with the time as sole explanatory variable.

**Differencing**, to remove trend and seasonalities and to obtain stationarity of the time series. The difference equation writes:

\[
Y_t = (1 - B)^d (1 - B^s) X_t
\]

where \(d\) is the order of the first differencing component, \(s\) is the period of the seasonal component, \(D\) is the order of the seasonal component, and \(B\) is the lag operator defined by:

\[
BX_i = X_{i-1}
\]

The values of \((d, D, s)\) can be chosen in a trial and error process, or guessed by looking at the descriptive functions (ACF, PACF). Typical values are \((1,1,s), (2,1,s)\). \(s\) is 12 for monthly data with a yearly seasonality, 0 when there is no seasonality.

**Detrending** and **deseasonalizing**, using the classical decomposition model which writes:

\[
X_t = m_t + s_t + e_t
\]

where \(m_t\) is the trend component and \(s_t\) the seasonal component, and \(e_t\) is a N(0,1) white noise component. XLSTAT allows to fit this model in two separate and/or successive steps:

1 – Detrending model:
\[ X_i = m_i + \epsilon_i = \sum_{j=0}^{k} a_j t^j + \epsilon_i \]

where \( k \) is the polynomial degree. The \( a_i \) parameters are obtained by fitting a linear model to the data. The transformed time series writes:

\[ Y_i = \epsilon_i = X_i - \sum_{j=0}^{p} a_j t^j \]

2 – Deseasonalization model:

\[ X_i = s_i + \epsilon_i = \mu + b_i + \epsilon_i , \quad i = t \mod p \]

where \( p \) is the period. The \( b_i \) parameters are obtained by fitting a linear model to the data. The transformed time series writes:

\[ Y_i = \epsilon_i = X_i - b_i - \mu \]

Note: there exist many other possible transformations. Some of them are available in the transformations tool of XLSTAT-Pro (see the "Preparing data" section). Linear filters may also be applied. Moving average smoothing methods which are linear filters are available in the "Smoothing" tool of XLSTAT.

**Seasonal decomposition**, from a user defined period \( P \), the seasonal decomposition estimates and decomposes the time series into 3 components (trend, seasonal and random).

If the chosen model type is additive, the model can be expressed as follows:

\[ X_i = m_i + s_{t \mod p} + \epsilon_i \]

with \( X_i \) the initial time series, \( m_i \) the trend component, \( s_{t \mod p} \) the seasonal component and \( \epsilon_i \) the random component.

First, the trend component is estimated by applying a centered moving average filter to \( X_i \):

\[ m_i = \sum_{j=-P/2}^{P/2} w_j X_{i+j} \]

where \( P/2 \) is the integer division of \( P \) by 2 and the coefficients \( w_j \) are defined as follows:
Each seasonal index $s_i$ is computed from the difference $s_i = X_i - \hat{m}$, as the average of the elements of $st$ for which $t \mod P = i$. Their values are then centered as shown below:

$$s_i = s_i - \frac{1}{P} \sum_{j=1}^{P} s_j$$

Finally, the random component is estimated as follows:

$$e_i = X_i - m_i - s_{t \mod P}$$

If the multiplicative type of decomposition is chosen, the model is given by:

$$X_i = m_i \times s_{t \mod P} \times e_i$$

The trend component is estimated as given for the additive decomposition.

The seasonal indices $s_i$ are computed as the average of the elements of $s_i = X_i / \hat{m}$, for which $t \mod P = i$.

They are then normalized as follows:

$$s_i = s_i \times \left( \prod_{j=1}^{P} s_j \right)^{-1/P}$$

Finally, the estimated random component is given by:

$$e_i = \frac{X_i}{s_{t \mod P} \times m_i}$$

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.
OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

Reload: Click this button to reload the default options.

Delete: Click this button to delete the data selections.

**General** tab:

**Times series:** Select the data that correspond to the time series. If a header is available on the first row, make sure you activate the "Series labels" option.

**Date data:** Activate this option if you want to select date or time data. These data must be available either in the Excel date/time formats or in a numerical format. If this option is not activated, XLSTAT creates its own time variable ranging from 1 to the number of data.

- **Check intervals:** Activate this option so that XLSTAT checks that the spacing between the date data is regular.

**Range:** Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Activate this option to display the results in a new worksheet of the active workbook.

**Workbook:** Activate this option to display the results in a new workbook.

**Series labels:** Activate this option if the first row of the selected series includes a header.

**Options** tab:

**Box-Cox:** Activate this option to compute the Box-Cox transformation. You can either fix the value of the Lambda parameter, or decide to let XLSTAT optimize it (see the description for further details).

**Differencing:** Activate this option to compute differenced series. You need to enter the differencing orders (d, D, s). See the description for further details.
**Polynomial regression**: Activate this option to detrend the time series. You need to enter polynomial degree. See the description for further details.

**Deseasonalization**: Activate this option to remove the seasonal components using a linear model. You need to enter the period of the series. See the description for further details.

**Seasonal decomposition**: Activate this option to compute the seasonal indices and decompose the time series. You need to select a model type, additive or multiplicative and enter the period of the series. See the description for further details.

Missing data tab:

- **Do not accept missing data**: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.
- **Remove observations**: Activate this option to remove the observations with missing data.
- **Replace by the average of the previous and next values**: Activate this option to estimate the missing data by the mean of the first preceding non missing value and of the first next non missing value.
- **Ignore missing data**: Activate this option to ignore missing data.

Outputs tab:

- **Descriptive statistics**: Activate this option to display the descriptive statistics of the selected series.

Charts tab:

- **Display charts**: Activate this option to display the charts of the series before and after transformation.

Results

**Summary statistics**: This table displays for the selected variables, the number of observations, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased).
**Box-Cox transformation:**

*Estimates of the parameters of the model:* This table is available only if the Lambda parameter has been optimized. It displays the three parameters of the model, which are Lambda, the Intercept of the model and slope coefficient.

*Series before and after transformation:* This table displays the series before and after transformation. If Lambda has been optimized, the transformed series corresponds to the residuals of the model. If it hasn’t then the transformed series is the direct application of the Box-Cox transformation.

**Differencing**

*Series before and after transformation:* This table displays the series before and after transformation and the differenced series. The first d+D+s data are not available in the transformed series because of the lag due to the differencing itself.

**Detrending (Polynomial regression)**

*Goodness of fit coefficients:* This table displays the goodness of fit coefficients.

*Estimates of the parameters of the model:* This table displays the parameters of the model.

*Series before and after transformation:* This table displays the series before and after transformation. The transformed series corresponds to the residuals of the model.

**Deseasonalization**

*Goodness of fit coefficients:* This table displays the goodness of fit coefficients.

*Estimates of the parameters of the model:* This table displays the parameters of the model. The intercept is equal to the mean of the series before transformation.

*Series before and after transformation:* This table displays the series before and after transformation. The transformed series corresponds to the residuals of the model.

**Example**

A tutorial explaining how to transform time series is available on the Addinsoft web site. To consult the tutorial, please go to:
References


Smoothing

Use this tool to smooth a time series and make predictions, using moving averages, exponential smoothing, Fourier smoothing, Holt or Holt-Winter’s methods.

Description

Several smoothing methods are available. We define by \( \{Y_t\}, \ t=1,\ldots,n \), the time series of interest, by \( P_tY_{t+h} \) the predictor of \( Y_{t+h} \) with minimum mean square error, and \( \varepsilon_t \) a N(0,1) white noise. The smoothing methods are described by the following equations:

**Simple exponential smoothing**

This model is sometimes referred to as Brown's Simple Exponential Smoothing, or the exponentially weighted moving average model. The equations of the model write:

\[
\begin{align*}
Y_t &= \mu_t + \varepsilon_t \\
P_tY_{t+h} &= \mu_t, \quad h = 1,2,\ldots \\
S_t &= \alpha Y_t + (1 - \alpha)S_{t-1}, \quad 0 < \alpha < 2 \\
\hat{Y}_{t+h} &= P_tY_{t+h} = S_t, \quad h = 1,2,\ldots
\end{align*}
\]

The region for \( \alpha \) corresponds to additivity and invertibility.

Exponential smoothing is useful when one needs to model a value by simply taking into account past observations. It is called "exponential” because the weight of past observations decreases exponentially. This method it is not very satisfactory in terms of prediction, as the predictions are constant after \( n+1 \).

**Double exponential smoothing**

This model is sometimes referred to as Brown's Linear Exponential Smoothing or Brown's Double Exponential Smoothing. It allows to take into account a trend that varies with time. The predictions take into account the trend as it is for the last observed data. The equations of the model write:
Holt’s linear exponential smoothing

This model is sometimes referred to as the Holt-Winters non seasonal algorithm. It allows to take into account a permanent component and a trend that varies with time. This model adapts itself quicker to the data compared with the double exponential smoothing. It involves a second parameter. The predictions for t>n take into account the permanent component and the trend component. The equations of the model write:

\[
\begin{align*}
Y_i &= \mu_t + \beta_i t + \epsilon_i \\
P_{iY_{t+h}} &= \mu_t + \beta_t t \\
S_i &= \alpha Y_i + (1 - \alpha) S_{i-1}, \quad 0 < \alpha < 2 \\
T_i &= \alpha S_i + (1 - \alpha) T_{i-1} \\
\hat{Y}_{t+h} &= P_{iY_{t+h}} = \left(2 + \frac{\alpha h}{1 - \alpha}\right) S_i - \left(1 + \frac{\alpha h}{1 - \alpha}\right) T_i, \quad \alpha \neq 1 \quad h = 1,2,\ldots \\
\hat{Y}_{t+h} &= P_{iY_{t+h}} = Y_i, \quad \alpha = 0 \quad h = 1,2,\ldots
\end{align*}
\]

The region for \(\alpha\) corresponds to additivity and invertibility.

Holt-Winters seasonal additive model

This method allows to take into account a trend that varies with time and a seasonal component with a period \(p\). The predictions take into account the trend and the seasonality. The model is called additive because the seasonality effect is stable and does not grow with time. The equations of the model write:

\[
\begin{align*}
Y_i &= \mu_t + \beta_t t + \epsilon_i \\
P_{iY_{t+h}} &= \mu_t + \beta_t t \\
S_i &= \alpha Y_i + (1 - \alpha) (S_{i-1} + T_{i-1}), \quad 0 < \alpha < 2 \\
T_i &= \beta (S_i - S_{i-1}) + (1 - \beta) T_{i-1}, \quad 0 < \beta < \frac{4}{\alpha} - 2 \\
\hat{Y}_{t+h} &= P_{iY_{t+h}} = S_i + hT_i, \quad h = 1,2,\ldots
\end{align*}
\]

The region for \(\alpha\) and \(\beta\) corresponds to additivity and invertibility.
\[
Y_t = \mu_t + \beta_t t + s_{p}(t) + \varepsilon_t \\
PY_{t+h} = \mu_t + \beta_t t + s_{p}(t) \quad h = 1,2,... \\
S_t = \alpha(Y_t - S_{t-1}) + (1-\alpha)(S_{t-1} + T_{t-1}) \\
T_t = \beta(S_t - S_{t-1}) + (1-\beta)T_{t-1} \\
D_t = \gamma(Y_t - S_t) + (1-\gamma)D_{t-p} \\
\hat{Y}_{t+h} = PY_{t+h} = (S_t + hT_t)D_{t-p+h}, \quad h = 1,2,...
\]

For the definition of the additive-invertible region please refer to Archibald (1990).

**Holt-Winters seasonal multiplicative model**

This method allows to take into account a trend that varies with time and a seasonal component with a period \( p \). The predictions take into account the trend and the seasonality. The model is called multiplicative because the seasonality effect varies with time. The more the discrepancies between the observations are high, the more the seasonal component increases. The equations of the model write:

\[
Y_t = (\mu_t + \beta_t t)s_p(t) + \varepsilon_t \\
P_Y_{t+h} = (\mu_t + \beta_t t)s_p(t) \quad h = 1,2,... \\
S_t = \alpha(Y_t / S_{t-p}) + (1-\alpha)(S_{t-1} + T_{t-1}) \\
T_t = \beta(S_t - S_{t-1}) + (1-\beta)T_{t-1} \\
D_t = \gamma(Y_t / S_t) + (1-\gamma)D_{t-p} \\
\hat{Y}_{t+h} = PY_{t+h} = (S_t + hT_t)D_{t-p+h}, \quad h = 1,2,...
\]

For the definition of the additive-invertible region please refer to Archibald (1990).

**Note 1:** for all the above models, XLSTAT estimates the values of the parameters that minimize the mean square error (MSE). However, it is also possible to maximize the likelihood, as, apart from the Holt-Winters multiplicative model, it is possible to write these models as ARIMA models. For example, the simple exponential smoothing is equivalent to an ARIMA(0,1,1) model, and the Holt-Winters additive model is equivalent to an ARIMA (0,1,p+1)(0,1,0)_p. If you prefer to maximize the likelihood, we advise you to use the ARIMA procedure of XLSTAT.

**Note 2:** for all the above models, initial values for \( S, T \) and \( D \), are required. XLSTAT offers several options, including backcasting to set these values. When backcasting is selected, the algorithm reverses the series, starts with simple initial values corresponding to the \( Y(x) \) option, then computes estimates and uses these estimates as initial values. The values corresponding to the various options for each method are described hereunder:
Simple exponential smoothing:
\[ Y(1) : \quad S_1 = Y_1 \]
Mean(6): \[ S_i = \frac{\sum_{j=1}^{6} Y_j}{6} \]
Backcasting
Optimized

Double exponential smoothing:
\[ Y(1) : \quad S_1 = Y_1, \quad T_1 = Y_1 \]
Mean(6): \[ S_i = \frac{\sum_{j=1}^{6} Y_j}{6}, \quad T_i = S_i \]
Backcasting

Holt’s linear exponential smoothing:
0: \[ S_1 = 0 \]
Backcasting

Holt-Winters seasonal additive model:
\[ Y(1+p) : \quad S_{1+p} = \sum_{i=1}^{p} Y_i / p, \quad T_{1+p} = 0, \quad D_i = Y_i - \left( Y_i + T_{1+p}(i-1) \right), i = 1, \ldots, p \]
Backcasting

Holt Winters seasonal multiplicative model:
\[ Y(1+p) : \quad S_{1+p} = \sum_{i=1}^{p} Y_i / p, \quad T_{1+p} = 0, \quad D_i = Y_i / \left( Y_i + T_{1+p}(i-1) \right), i = 1, \ldots, p \]
Backcasting

Moving average
This model is a simple way to take into account past and optionally future observations to predict values. It works as a filter that is able to remove noise. While with the smoothing methods defined below, an observation influences all future predictions (even if the decay is exponential), in the case of the moving average the memory is limited to \( q \). If the constant \( i \) is set to zero, the prediction depends on the past \( q \) values and on the current value, and if \( i \) is set to one, it also depends on the next \( q \) values. Moving averages are often used as filters, and
not as way to do accurate predictions. However XLSTAT enables you to do predictions based on the moving average model that writes:

\[
\begin{align*}
  \hat{Y}_t &= \hat{\mu}_l + \varepsilon_t, \\
  \hat{\mu}_l &= \frac{\sum_{i=-q}^{q} w_i Y_{t+i}}{\\sum_{i=-q}^{q} w_i} \\
  \end{align*}
\]

where \( l \) is a constant, which, when set to zero, allows the prediction to depend on the \( q \) previous values and on the current value. If \( l \) is set to one, the prediction also depends on the \( q \) next values. The \( w_i \) (\( i=1…q \)) are the weights. Weights can be either constant, fixed by the user, or based on existing optimal weights for a given application. XLSTAT allows to use the Spencer 15-points model that passes polynomials of degree 3 without distortion.

**Fourier smoothing**

The concept of the Fourier smoothing is to transform a time series into its Fourier coordinates, then remove part of the higher frequencies, and then transform the coordinates back to a signal. This new signal is a smoothed series.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload default options**: Click this button to reload the default options.
- **Delete data selections**: Click this button to delete the data selections.
Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

**Times series**: Select the data that correspond to the time series. If a header is available on the first row, make sure you activate the "Series labels" option.

**Date data**: Activate this option if you want to select date or time data. These data must be available either in the Excel date/time formats or in a numerical format. If this option is not activated, XLSTAT creates its own time variable ranging from 1 to the number of data.

- **Check intervals**: Activate this option so that XLSTAT checks that the spacing between the date data is regular.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Series labels**: Activate this option if the first row of the selected series includes a header.

**Model**: Select the smoothing model you want to use (see description for more information on the various models).

**Options** tab:

**Method**: Select the method for the selected model (see description for more information on the various models).

Stop conditions:

- **Iterations**: Enter the maximum number of iterations for the algorithm. The calculations are stopped when the maximum number if iterations has been exceeded. Default value: 500.
- **Convergence**: Enter the maximum value of the evolution in the convergence criterion from one iteration to another which, when reached, means that the algorithm is considered to have converged. Default value: 0.00001.

**Confidence interval (%)**: The value you enter (between 1 and 99) is used to determine the confidence intervals for the predicted values. Confidence intervals are automatically displayed on the charts.

**S1**: Choose an estimation method for the initial values. See the description for more information on that topic.

Depending on the model type, and on the method you have chosen, different options are available in the dialog box. In the description section, you can find information on the various models and on the corresponding parameters.

In the case of exponential or Holt-Winters models, you can decide to set the parameters to a given value, or to optimize them. In the case of the Holt-Winters seasonal models, you need to enter the value of the **period**.

In the case of the Fourier smoothing, you need to enter to the proportion $p$ of the spectrum that needs to be kept after the high frequencies are removed.

For the moving average model, you need to specify the number $q$ of time steps that must be taken into account to compute the predicted value. You can decided to only consider the previous $q$ steps (the left part) of the series.

**Validation** tab:

**Validation**: Activate this option to use some data for the validation of the model.

**Time steps**: Enter the number the number of data at the end of the series that need to be used for the validation.

**Prediction** tab:

**Prediction**: Activate this option to use the model to do some forecasting.

**Time steps**: Enter the number the number of time steps for which you want XLSTAT to compute a forecast.

**Missing data** tab:
Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Replace by the average of the previous and next values: Activate this option to estimate the missing data by the mean of the first preceding non missing value and of the first next non missing value.

Ignore missing data: Activate this option to ignore missing data.

Outputs tab:

Descriptive statistics: Activate this option to display the descriptive statistics of the selected series.

Goodness of fit coefficients: Activate this option to display the goodness of fit statistics.

Model parameters: Activate this option to display the table of the model parameters.

Predictions and residuals: Activate this option to display the table of the predictions and the residuals.

Charts tab:

Display charts: Activate this option to display the charts of the series before and after smoothing, as well as the bar chart of the residuals.

Results

Goodness of fit coefficients: This table displays the goodness of fit coefficients which include the number of degrees of freedom (DF), the DDL, the sum of squares of errors (SSE) the mean square of errors (MSE), the root of the MSE (RMSE), the mean absolute percentage error (MAPE), the mean percentage error (MPE) the mean absolute error (MAE) and the coefficient of determination (R²). Note: all these statistics are computed for the observations involved in the estimation of the model only; the validation data are not taken into account.

Model parameters: This table displays the estimates of the parameters, and, if available, the standard error of the estimates. Note: to S1 corresponds the first computed value of the S series, and to T1 corresponds the first computed value of the series T. See the description for more information.
Series before and after smoothing: This table displays the series before and after smoothing. If some predictions have been computed (t>n), and if the confidence intervals option has been activated, the confidence intervals are computed for the predictions.

Charts: The first chart displays the data, the model, and the predictions (validation + prediction values) as well as the confidence intervals. The second chart corresponds to the bar chart of the residuals.

Example

A tutorial explaining how to do forecasting with the Holt-Winters method is available on the Addinsoft web site. To consult the tutorial, please go to:

http://www.xlstat.com/demo-hw.htm

References


ARIMA

Use this tool to fit an ARMA (Autoregressive Moving Average), an ARIMA (Autoregressive Integrated Moving Average) or a SARIMA (Seasonal Autoregressive Integrated Moving Average) model, and to compute forecasts using the model which parameters are either known or to be estimated.

Description

The models of the ARIMA family allow to represent in a synthetic way phenomena that vary with time, and to predict future values with a confidence interval around the predictions.

The mathematical writing of the ARIMA models differs from one author to the other. The differences concern most of the time the sign of the coefficients. XLSTAT is using the most commonly found writing, used by most software.

If we define by \{X_t\} a series with mean \(\mu\), then if the series is supposed to follow an ARIMA\((p,d,q)(P,D,Q)_s\) model, we can write:

\[
Y_t = (1 - B)^d (1 - B^s)^P X_t - \mu
\]

\[
\phi(B) \Phi(B^s) Y_t = \theta(B) \Theta(B^s) Z_t, \quad Z_t \sim N(0, \sigma^2)
\]

with

\[
\phi(z) = 1 - \sum_{i=1}^{p} \phi_i z^i, \quad \Phi(z) = 1 - \sum_{i=1}^{P} \Phi_i z^i
\]

\[
\theta(z) = 1 + \sum_{i=1}^{q} \theta_i z^i, \quad \Theta(z) = 1 + \sum_{i=1}^{Q} \Theta_i z^i
\]

\(p\) is the order of the autoregressive part of the model.

\(q\) is the order of the moving average part of the model.

\(d\) is the differencing order of the model.

\(D\) is the differencing order of the seasonal part of the model.

\(s\) is the period of the model (for example 12 if the data are monthly data, and if one noticed a yearly periodicity in the data).

\(P\) is the order of the autoregressive seasonal part of the model.

\(Q\) is the order of the moving average seasonal part of the model.
Remark 1: the \( \{Y_t\} \) process is causal if and only if for any \( z \) such that \( |z| \leq 1 \), \( \phi(z) \neq 0 \) and \( \theta(z) \neq 0 \).

Remark 2: if \( D=0 \), the model is an ARIMA\((p,d,q)\) model. In that case, \( P \), \( Q \) and \( s \) are considered as null.

Remark 3: if \( d=0 \) and \( D=0 \), the model simplifies to an ARMA\((p,q)\) model.

Remark 4: if \( d=0 \), \( D=0 \) and \( q=0 \), the model simplifies to an AR\((p)\) model.

Remark 5: if \( d=0 \), \( D=0 \) and \( p=0 \), the model simplifies to an MA\((q)\) model.

**Explanatory variables**

XLSTAT allows you to take into account explanatory variables through a linear model. Three different approaches are possible:

1. **OLS**: A linear regression model is fitted using the classical linear regression approach, then the residuals are modeled using an \((S)\)ARIMA model.

2. **CO-LS**: If \( d \) or \( D \) and \( s \) are not zero, the data (including the explanatory variables) are differenced, then the corresponding ARMA model is fitted at the same time as the linear model coefficients using the Cochrane and Orcutt (1949) approach.

3. **GLS**: A linear regression model is fitted, then the residuals are modeled using an \((S)\)ARIMA model, then we loop back to the regression step, in order to improve the likelihood of the model by changing the regression coefficients using a Newton-Raphson approach.

   Note: if no differencing is requested \((d=0 \text{ and } D=0)\), and if there are no explanatory variables in the model, the constant of the model is estimated using CO-LS.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
Help: Click this button to display the help.

Refresh: Click this button to reload the default options.

Clear: Click this button to delete the data selections.

Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

**Times series**: Select the data that correspond to the time series. If a header is available on the first row, make sure you activate the "Series labels" option.

**Center**: Activate this option to center the data after the differencing.

**Variance**: Activate this option to set the value of the variance of the errors.

**Date data**: Activate this option if you want to select date or time data. These data must be available either in the Excel date/time formats or in a numerical format. If this option is not activated, XLSTAT creates its own time variable ranging from 1 to the number of data.

- **Check intervals**: Activate this option so that XLSTAT checks that the spacing between the date data is regular.

**X / Explanatory variables**: Activate this option if you want to include one or more quantitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The data selected must be of the numerical type. If a variable header has been selected, check that the "Variable labels" option has been activated.

- **Mode**: Choose the way you want to take into account the explanatory variables (the three modes OLS, CO-LS, GLS, are described in the description section).

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.
Workbook: Activate this option to display the results in a new workbook.

Series labels: Activate this option if the first row of the selected series includes a header.

Model parameters: Enter orders of the model:

- **p**: Enter the order of the autoregressive part of the model. For example, enter 1 for an AR(1) model or for an ARMA(1,2) model.
- **d**: Enter the differencing order of the model. For example, enter 1 for an ARIMA(0,1,2) model.
- **q**: Enter the order of the moving average part of the model. For example, enter 2 for a MA(2) model or for an ARIMA(1,1,2) model.
- **P**: Enter the order of the autoregressive seasonal part of the model. For example, enter 1 for an ARIMA(1,1,0)(1,1,0)¹² model. You can modify this value only if D*0. If D=0, XLSTAT considers that P=0.
- **D**: Enter the differencing order for the seasonal part of the model. For example, enter 1 for an ARIMA(0,1,1)(0,1,1)¹² model.
- **Q**: Enter the order of the moving average seasonal part of the model. For example, enter 1 for an ARIMA(0,1,1)(0,1,1)¹² model. You can modify this value only if D*0. If D=0, XLSTAT considers that P=0.
- **s**: Enter the period of the model. You can modify this value only if D*0. If D=0, XLSTAT considers that s=0.

Options tab:

Preliminary estimation: Activate this option if you want to use a preliminary estimation method. This option is available only if D=0.

- **Yule-Walker**: Activate this option to estimate the coefficients of the autoregressive AR(p) model using the Yule-Walker algorithm.
- **Burg**: Activate this option to estimate the coefficients of the autoregressive AR(p) model using the Burg’s algorithm.
- **Innovations**: Activate this option to estimate the coefficients of the moving average MA(q) model using the Innovations algorithm.
- **Hannan-Rissanen**: Activate this option to estimate the coefficients of the ARMA(p,q) model using the Hannan-Rissanen algorithm.
- **m/Automatic**: If you choose to use the Innovations or the Hannan-Rissanen algorithm, you need to either enter the m value corresponding to the algorithm or to let XLSTAT determine automatically (select Automatic) what is an appropriate value for m.

**Initial coefficients**: Activate this option to select the initial values of the coefficients of the model.

- **Phi**: Select here the value of the coefficients corresponding to the autoregressive part of the model (including the seasonal part). The number of values to select is equal to p+P.

- **Theta**: Select here the value of the coefficients corresponding to the moving average part of the model (including the seasonal part). The number of values to select is equal to q+Q.

**Optimize**: Activate this option to estimate the coefficients using one of the two available methods:

- **Likelihood**: Activate this option pour maximize the likelihood of the parameters knowing the data.

- **Least squares**: Activate this option to minimize the sum of squares of the residuals.

**Stop conditions**:

- **Iterations**: Enter the maximum number of iterations for the algorithm. The calculations are stopped when the maximum number of iterations has been exceeded. Default value: 500.

- **Convergence**: Enter the maximum value of the evolution in the convergence criterion from one iteration to another which, when reached, means that the algorithm is considered to have converged. Default value: 0.00001.

**Find the best model**: Activate this option to explore several combines of orders. If you activate this option, the minimum order is the one given in the “General” tab, and the maximum orders need to be defined using the following options:

- **Max(p)**: Enter the maximum value of p to explore.

- **Max(q)**: Enter the maximum value of q to explore.

- **Max(P)**: Enter the maximum value of P to explore.
- **Max(Q)**: Enter the maximum value of Q to explore.

- **AICC**: Activate this option to use the AICC (Akaike Information Criterion Corrected) to identify the best model.

- **SBC**: Activate this option to use the SBC (Schwarz’s Bayesian Criterion) to identify the best model.

**Validation** tab:

**Validation**: Activate this option to use some data for the validation of the model.

**Time steps**: Enter the number of data at the end of the series that need to be used for the validation.

**Prediction** tab:

**Prediction**: Activate this option to use the model to do some forecasting.

**Time steps**: Enter the number of time steps for which you want XLSTAT to compute a forecast.

**Missing data** tab:

**Do not accept missing data**: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

**Remove observations**: Activate this option to remove the observations with missing data.

**Replace by the average of the previous and next values**: Activate this option to estimate the missing data by the mean of the first preceding non missing value and of the first next non missing value.

**Outputs** tab:

**Descriptive statistics**: Activate this option to display the descriptive statistics of the selected series.

**Goodness of fit coefficients**: Activate this option to display the goodness of fit statistics.

**Model parameters**: Activate this option to display the table of the model parameters.

**Predictions and residuals**: Activate this option to display the table of the predictions and the residuals.
Confidence interval (%): The value you enter (between 1 and 99) is used to determine the confidence intervals for the predicted values. Confidence intervals are automatically displayed on the charts.

Charts tab:

Display charts: Activate this option to display the chart that display the input data together with the model predictions, as well the bar chart of the residuals.

Results

Summary statistics: This table displays for the selected variables, the number of observations, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased).

If a preliminary estimation and an optimization have been requested the results for the preliminary estimation are first displayed followed by the results after the optimization. If initial coefficients have been entered the results corresponding to these coefficients are displayed first.

Goodness of fit coefficients:

- **Observations**: The number of data used for the fitting of the model.
- **SSE**: Sum of Squares of Errors. This statistic is minimized if the "Least Squares" option has been selected for the optimization.
- **WN variance**: The white noise variance is equal to the SSE divided by N. In some software, this statistic is named sigma2 (sigma-square).
- **WN variance estimate**: This statistic is usually equal to the previous. In the case of a preliminary estimation using the Yule-Walker or Burg's algorithms, a slightly different estimate is displayed.
- **-2Log(Like.)**: This statistic is minimized if the "Likelihood" option has been selected for the optimization. It is equal to 2 times the natural logarithm of the likelihood.
- **FPE**: Akaike's Final Prediction Error. This criterion is adapted to autoregressive models.
- **AIC**: The Akaike Information Criterion.
- **AICC**: This criterion has been suggested by Brockwell (Akaike Information Criterion Corrected).
• **SBC**: Schwarz’s Bayesian Criterion.

Model parameters:

The first table of parameters shows the coefficients of the linear model fitted to the data (a constant if no explanatory variable was selected).

The next table gives the estimator for each coefficient of each polynomial, as well as the standard deviation obtained either directly from the estimation method (preliminary estimation), or from the Fisher’s information matrix (Hessian). The asymptotical standard deviations are also computed. For each coefficient and each standard deviation, a confidence interval is displayed. The coefficients are identified as follows:

AR(i): that corresponds to the order i coefficient of the $\phi(z)$ polynomial.

SAR(i): coefficient that corresponds to the order i coefficient of the $\Phi(z)$ polynomial.

MA(i): coefficient that corresponds to the order i coefficient of the $\theta(z)$ polynomial.

SMA(i): coefficient that corresponds to the order i coefficient of the $\Theta(z)$ polynomial.

**Data, Predictions and Residuals**: This table displays the data, the corresponding predictions computed with the model, and the residuals. If the user requested it, predictions are computed for the validation data and forecasts for future values. Standard deviations and confidence intervals are computed for validation predictions and forecasts.

**Charts**: Two charts are displayed. The first chart displays the data, the corresponding values predicted by the model, and the predictions corresponding to the values for the validation and/or prediction time steps. The second chart corresponds to the bar chart of residuals.

**Example**

A tutorial explaining how to do fit an ARIMA model and to use the model to do forecasting is available on the Addinsoft web site. To consult the tutorial, please go to:

http://www.xlstat.com/demo-arima.htm
References


Spectral analysis

Use this tool to transform a time series into its coordinates in the space of frequencies, and then to analyze its characteristics in this space.

Description

This tool allows to transform a time series into its coordinates in the space of frequencies, and then to analyze its characteristics in this space. From the coordinates we can extract the magnitude, the phase, build representations such as the periodogram, the spectral density, and test if the series is stationary. By looking at the spectral density, we can identify seasonal components, and decide to which extent we should filter noise. Spectral analysis is a very general method used in a variety of domains.

The spectral representation of a time series \( \{X_t\} \), \( (t=1,\ldots,n) \), decomposes \( \{X_t\} \) into a sum of sinusoidal components with uncorrelated random coefficients. From there we can obtain decomposition the autocovariance and autocorrelation functions into sinusoids.

The spectral density corresponds to the transform of a continuous time series. However, we usually have only access to a limited number of equally spaced data, and therefore, we need to obtain first the discrete Fourier coordinates (cosine and sine transforms), and then the periodogram. From the periodogram, using a smoothing function, we can obtain a spectral density estimate which is a better estimator of the spectrum.

Using fast and powerful methods, XLSTAT automatically computes the Fourier cosine and sine transforms of \( \{X_t\} \), for each Fourier frequency, and then the various functions that derive from these transforms.

With \( n \) being the sample size, and \( [i] \) being the largest integer less than or equal to \( i \), the Fourier frequencies write:

\[
\omega_k = \frac{2\pi k}{n}, \quad k = -\left[\frac{n-1}{2}\right], \ldots, \left\lfloor \frac{n}{2} \right\rfloor
\]

The Fourier cosine and sine coefficients write:

\[
a_k = \frac{2}{n} \sum_{t=1}^{n} X_t \cos(\omega_k (t-1))
\]

\[
b_k = \frac{2}{n} \sum_{t=1}^{n} X_t \sin(\omega_k (t-1))
\]

The periodogram writes:
\[ I_k = \frac{n}{2} \sum_{t=1}^{n} (a_k^2 + b_k^2) \]

The spectral density estimate (or discrete spectral average estimator) of the time series \( \{X_t\} \) writes:

\[ \hat{f}_k = \sum_{i=-p}^{p} w_i J_{k+i} \]

with \( J_{k+i} = I_{k+i}, \quad 0 \leq k + i \leq n \)

\[ J_{k+i} = I_{-(k+i)}, \quad k + i < 0 \]

\[ J_{k+i} = I_{n-(k+i)}, \quad k + i > n \]

where \( p \), the bandwidth, and \( w_i \), the weights, are either fixed by the user, or determined by the choice of a kernel. XLSTAT suggests the use of the following kernels:

If we define \( p = c.q^e \), \( q = \lfloor n/2 \rfloor + 1 \), and \( \lambda_i = i / p \)

Bartlett:

\[
\begin{cases}
  c = 1/2, & e = 1/3 \\
  w_i = 1 - |\lambda_i| & \text{if } |\lambda_i| \leq 1 \\
  w_i = 0 & \text{otherwise}
\end{cases}
\]

Parzen:

\[
\begin{cases}
  c = 1, & e = 1/5 \\
  w_i = 1 - 6|\lambda_i|^2 + 6|\lambda_i|^3 & \text{if } |\lambda_i| \leq 0.5 \\
  w_i = 2(1-|\lambda_i|)^3 & \text{if } 0.5 \leq |\lambda_i| \leq 1 \\
  w_i = 0 & \text{otherwise}
\end{cases}
\]

Quadratic spectral:

\[
\begin{cases}
  c = 1/2, & e = 1/5 \\
  w_i = \frac{25}{12\pi^2 \lambda_i^2} \left( \frac{\sin(6\pi\lambda_i / 5)}{6\pi \lambda_i / 5} - \cos(6\pi\lambda_i / 5) \right)
\end{cases}
\]

Tukey-Hanning:

\[
\begin{cases}
  c = 2/3, & e = 1/5 \\
  w_i = (1 + \cos(\pi \lambda_i)) / 2 & \text{if } |\lambda_i| \leq 1 \\
  w_i = 0 & \text{otherwise}
\end{cases}
\]

Truncated:
\[
\begin{cases}
c = 1/4, & e = 1/5 \\
w_i = 1 & \text{if } |\lambda_i| \leq 1 \\
w_i = 0 & \text{otherwise}
\end{cases}
\]

Note: the bandwidth \( p \) is a function of \( n \), the size of the sample. The weights \( w_i \) must be positive and must sum to one. If they don’t, XLSTAT automatically rescales them.

If a second time series \( \{Y_t\} \) is available, several additional functions can be computed to estimate the cross-spectrum:

The real part of the cross-periodogram of \( \{X_t\} \) and \( \{Y_t\} \) writes:

\[
\text{Real}_k = \frac{n}{2} \sum_{i=1}^{n} \left( a_{X,k} a_{Y,k} + b_{X,k} b_{Y,k} \right)
\]

The imaginary part of the cross-periodogram of \( \{X_t\} \) and \( \{Y_t\} \) writes:

\[
\text{Imag}_k = \frac{n}{2} \sum_{i=1}^{n} \left( a_{X,k} b_{Y,k} - b_{X,k} a_{Y,k} \right)
\]

The cospectrum estimate (real part of the cross-spectrum) of the time series \( \{X_t\} \) and \( \{Y_t\} \) writes:

\[
C_k = \sum_{i=-p}^{p} w_i R_{k+i}
\]

with \( R_{k+i} = \text{Real}_{k+i} \), \( 0 \leq k + i \leq n \)

\( R_{k+i} = \text{Real}_{-(k+i)} \), \( k + i < 0 \)

\( R_{k+i} = \text{Real}_{n-(k+i)} \), \( k + i > n \)

The quadrature spectrum (imaginary part of the cross-periodogram) estimate of the time series \( \{X_t\} \) and \( \{Y_t\} \) writes:

\[
Q_k = \sum_{i=-p}^{p} w_i H_{k+i}
\]

with \( H_{k+i} = \text{Imag}_{k+i} \), \( 0 \leq k + i \leq n \)

\( H_{k+i} = \text{Imag}_{-(k+i)} \), \( k + i < 0 \)

\( H_{k+i} = \text{Imag}_{n-(k+i)} \), \( k + i > n \)

The phase of the cross-spectrum of \( \{X_t\} \) and \( \{Y_t\} \) writes:

\[
\Phi_k = \arctan(Q_k / C_k)
\]
The amplitude of the cross-spectrum of \{X_t\} and \{Y_t\} writes:

\[ A_k = \sqrt{C_k^2 + Q_k^2} \]

The squared coherency estimate between the \{X_t\} and \{Y_t\} series writes:

\[ K_k = \frac{A_k^2}{\hat{f}_{X,k}\hat{f}_{Y,k}} \]

**White noise tests**: XLSTAT optionally displays two test statistics and the corresponding p-values for white noise: Fisher’s Kappa and Bartlett’s Kolmogorov-Smirnov statistic.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload**: Click this button to reload the default options.
- **Delete**: Click this button to delete the data selections.

**General** tab:

**Times series**: Select the data that correspond to the time series for which you want to compute the various spectral functions.

**Date data**: Activate this option if you want to select date or time data. These data must be available either in the Excel date/time formats or in a numerical format. If this option is not activated, XLSTAT creates its own time variable ranging from 1 to the number of data.
- **Check intervals**: Activate this option so that XLSTAT checks that the spacing between the date data is regular.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Series labels**: Activate this option if the first row of the selected series includes a header.

**Missing data** tab:

**Do not accept missing data**: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

**Remove observations**: Activate this option to remove the observations with missing data.

**Replace by the average of the previous and next values**: Activate this option to estimate the missing data by the mean of the first preceding non missing value and of the first next non missing value.

**Outputs (1)** tab:

**White noise tests**: Activate this option if you want to display the results of the white noise tests.

**Cosine part**: Activate this option if you want to display the Fourier cosine coefficients.

**Sine part**: Activate this option if you want to display the Fourier sine coefficients.

**Amplitude**: Activate this option if you want to display the amplitude of the spectrum.

**Phase**: Activate this option if you want to display the phase of the spectrum.

**Spectral density**: Activate this option if you want to display the estimate of spectral density.

- **Kernel weighting**: Select the type of kernel. The kernel functions are described in the description section.
  
  - **c**: Enter the value of the c parameter. This parameter is described in the description section.
- **e**: Enter the value of the e parameter. This parameter is described in the description section.

- **Fixed weighting**: Select on an Excel sheet the values of the fixed weights. The number of weights must be odd. Symmetric weights are recommended (Example: 1,2,3,2,1).

**Outputs (2) tab:**

- **Cross-spectrum**: Activate this option to analyze the cross-spectra. The computations are only done if at least two series have been selected.
  - **Real part**: Activate this option to display the real part of the cross-spectrum.
  - **Imaginary part**: Activate this option to display the imaginary part of the cross-spectrum.
  - **Cospectrum**: Activate this option to display the cospectrum estimate (real part of the cross-spectrum).
  - **Quadrature spectrum**: Activate this option to display the quadrature estimate (real part of the cross-spectrum).
  - **Squared coherency**: Activate this option to display the squared coherency.

**Charts tab:**

- **Periodogram**: Activate this option to display the periodogram of the series.
- **Spectral density**: Activate this option to display the chart of the spectral density.

**Results**

White noise tests: This table displays both the Fisher’s Kappa Bartlett’s Kolmogorov-Smirnov statistics and the corresponding p-values. If the p-values are lower than the significance level (typically 0.05), then you need to reject the assumption that the time series is just white noise.

A table is displayed for each selected time series. It displays various columns:

- **Frequency**: frequencies from 0 to $\pi$.
- **Period**: in time units.
- **Cosine part**: the cosine coefficients of the Fourier transform
**Sine part**: the sine coefficients of the Fourier transform

**Phase**: Phase of the spectrum.

**Periodogram**: value of the periodogram.

**Spectral density**: estimate of the spectral density.

**Charts**: XLSTAT displays the periodogram and the spectral density charts on both the frequency and period scales.

If two series or more have been selected, and if the cross-spectrum options have been selected, XLSTAT displays additional information:

**Cross-spectrum analysis**: This displays various the cross-spectrum information:

- **Frequency**: frequencies from 0 to $\pi$.
- **Period**: in time units.
- **Real part**: the real part of the cross-spectrum.
- **Imaginary part**: the imaginary part of the cross-periodogram.
- **Cospectrum**: The cospectrum estimate (real part of the cross-spectrum).
- **Quadrature spectrum**: The quadrature estimate (imaginary part of the cross-spectrum).
- **Amplitude**: amplitude of the cross-spectrum.
- **Squared coherency**: estimates of the squared coherency.

**Charts**: XLSTAT displays the amplitude of the estimate of the cross-spectrum on both the frequency and period scales.

**Example**

An example of Spectral analysis is available on the Addinsoft web site. To consult the tutorial, please go to:

References


Fourier transformation

Use this tool to transform a time series or a signal to its Fourier coordinates, or to do the inverse transformation.

Description

Use this tool to transform a time series or a signal to its Fourier coordinates, or to do the inverse transformation. While the Excel function is limited to powers of two for the length of the time series, XLSTAT is not restricted. Outputs optionally include the amplitude and the phase.

Dialog box

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- ![Reload default options](image)
  - Click this button to reload the default options.
- ![Delete data selections](image)
  - Click this button to delete the data selections.

**Real part**: Activate this option and then select the signal to transform, or the real part of the Fourier coordinates for an inverse transformation.

**Imaginary part**: Activate this option and then select the imaginary part of the Fourier coordinates for an inverse transformation.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.
Column labels: Activate this option if the first row of the data selections (real part, imaginary part) includes a header.

Inverse transformation: Activate this option if you want to compute the inverse Fourier transform.
Amplitude: Activate this option if you want to compute and display the amplitude of the spectrum.
Phase: Activate this option if you want to compute and display the phase of the spectrum.

Results
Real part: This column contains the real part after the transform or the inverse transform.
Imaginary part: This column contains the real part after the transform or the inverse transform.
Amplitude: Amplitude of the spectrum.
Phase: Phase of the spectrum.

References
**XLSTAT-Sim**

XLSTAT-Sim is an easy to use and powerful solution to create and run simulation models.

**Introduction**

XLSTAT-Sim is a module that allows to build and compute simulation models, an innovative method for estimating variables, whose exact value is not known, but that can be estimated by means of repeated simulation of random variables that follow certain theoretical laws. Before running the model, you need to create the model, defining a series of input and output (or result) variables.

Simulation models allow to obtain information, such as mean or median, on variables that do not have an exact value, but for which we can know, assume or compute a distribution. If some “result” variables depend of these “distributed” variables by the way of known or assumed formulae, then the “result” variables will also have a distribution. XLSTAT-Sim allows you to define the distributions, and then obtain through simulations an empirical distribution of the input and output variables as well as the corresponding statistics.

Simulation models are used in many areas such as finance and insurance, medicine, oil and gas prospecting, accounting, or sales prediction.

Four elements are involved in the construction of a simulation model:

- **Distributions** are associated to random variables. XLSTAT gives a choice of more than 20 distributions to describe the uncertainty on the values that a variable can take (see the chapter “Define a distribution” for more details). For example, you can choose a triangular distribution if you have a quantity for which you know it can vary between two bounds, but with a value that is more likely (a mode). At each iteration of the computation of the simulation model, a random draw is performed in each distribution that has been defined.

- **Scenario variables** allow to include in the simulation model a quantity that is fixed in the model, except during the tornado analysis where it can vary between two bounds (see the
chapter “Define a scenario variable” for more details, and the section on tornado analysis below).

- **Result variables** correspond to outputs of the model. They depend either directly or indirectly, through one or more Excel formulae, on the random variables to which distributions have been associated and if available on the scenario variables. The goal of computing the simulation model is to obtain the distribution of the result variables (see the chapter “Define a result variable” for more details).

- **Statistics** allow to track a given statistic a result variable. For example, we might want to monitor the standard deviation of a result variable (see the chapter “Define a statistic” for more details).

A correct model should comprise at least one distribution and one result. Models can contain any number of these four elements.

A model can be limited to a single Excel sheet or can use a whole Excel folder.

Simulation models can take into account the dependencies between the input variables described by distributions. If you know that two variables are usually related such that the correlation coefficient between them is 0.4, then you want that, when you do simulations, the sampled values for both variables have the same property. This is possible in XLSTAT-Sim by entering in the Run dialog box the correlation or covariance matrix between some or all the input random variables used in the model.

**Outputs**

When you run the model, a series of results is displayed. While giving the critical statistics such are information on the distribution of the input and result variables, it also allows interpreting relationships between variables. Sensitivity analysis is also available if scenario variables have been included.

Descriptive statistics:

The report that is generated after the simulation contains information on the distributions of the model. The user may choose from a range of descriptive statistics the most important indicators that should be integrated into the report in order to easily interpret the results. A selection of charts is also available to graphically display the relationships.

Details and formulae relative to the descriptive statistics are available in the description section of the “Descriptive statistics” tool of XLSTAT.
Charts:

The following charts are available to display information on the variables:

- **Box plots**: These univariate representations of quantitative data samples are sometimes called "box and whisker diagrams". It is a simple and quite complete representation since in the version provided by XLSTAT the minimum, 1st quartile, median, mean and 3rd quartile are displayed together with both limits (the ends of the "whiskers") beyond which values are considered anomalous. The mean is displayed with a red +, and a black line corresponds to the median. Limits are calculated as follows:

  **Lower limit**: $L_{\text{inf}} = X(i)$ such that $(X(i) - [Q1 - 1.5 (Q3 - Q1)])$ is minimum and $X(i) \geq Q1 - 1.5 (Q3 - Q1)$.

  **Upper limit**: $L_{\text{sup}} = X(i)$ such that $(X(i) - [Q3 + 1.5 (Q3 - Q1)])$ is minimum and $X(i) \leq Q3 + 1.5 (Q3 - Q1)$

Values that are outside the $[Q1 - 3 (Q3 - Q1); Q3 + 3 (Q3 - Q1)]$ interval are displayed with the * symbol. Values that are in the $[Q1 - 3 (Q3 - Q1); Q1 - 1.5 (Q3 - Q1)]$ or the $[Q3 + 1.5 (Q3 - Q1); Q3 + 3 (Q3 - Q1)]$ intervals are displayed with the "o" symbol.

- **Scattergrams**: These univariate representations give an idea of the distribution and possible plurality of the modes of a sample. All points are represented together with the mean and the median.

- **P-P Charts (normal distribution)**: P-P charts (for Probability-Probability) are used to compare the empirical distribution function of a sample with that of a normal variable for the same mean and deviation. If the sample follows a normal distribution, the data will lie along the first bisector of the plan.

- **Q-Q Charts (normal distribution)**: Q-Q charts (for Quantile-Quantile) are used to compare the quantities of the sample with that of a normal variable for the same mean and deviation. If the sample follows a normal distribution, the data will lie along the first bisector of the plan.

Correlations:

Once the computations are over, the simulation report may contain information on the correlations between the different variables included in the simulation model. Three different correlation coefficients are available:

- **Pearson correlation coefficient**: This coefficient corresponds to the classical linear correlation coefficient. This coefficient is well suited for continuous data. Its value ranges from -1 to 1, and it measures the degree of linear correlation between two variables. Note: the squared Pearson correlation coefficient gives an idea of how much of the variability of a variable is explained by the other variable. The p-values that are computed for each
coefficient allow testing the null hypothesis that the coefficients are not significantly different from 0. However, one needs to be cautious when interpreting these results, as if two variables are independent, their correlation coefficient is zero, but the reciprocal is not true.

- **Spearman correlation coefficient (rho):** This coefficient is based on the ranks of the observations and not on their value. This coefficient is adapted to ordinal data. As for the Pearson correlation, one can interpret this coefficient in terms of variability explained, but here we mean the variability of the ranks.

- **Kendall correlation coefficient (tau):** As for the Spearman coefficient, it is well suited for ordinal variables as it is also based on ranks. However, this coefficient is conceptually very different. It can be interpreted in terms of probability: it is the difference between the probabilities that the variables vary in the same direction and the probabilities that the variables vary in the opposite direction. When the number of observations is lower than 50 and when there are no ties, XLSTAT gives the exact p-value. If not, an approximation is used. The latter is known as being reliable when there are more than 8 observations.

**Sensitivity analysis:**

The sensitivity analysis displays information about the impact of the different input variables for one output variable. Based on the simulation results and on the correlation coefficient that has been chosen (see above), the correlations between the input random variables and the result variables are calculated and displayed in a declining order of impact on the result variable.

**Tornado and spider analyses:**

Tornado and spider analyses are not based on the iterations of the simulation but on a point by point analysis of all the input variables (random variables with distributions and scenario variables).

During the tornado analysis, for each result variable, each input random variable and each scenario variable are studied one by one. We make their value vary between two bounds and record the value of the result variable in order to know how each random and scenario variable impacts the result variables. For a random variable, the values explored can either be around the median or around the default cell value, with bounds defined by percentiles or deviation. For a scenario variable, the analysis is performed between two bounds specified when defining the variables. The number of points is an option that can be modified by the user before running the simulation model.

The spider analysis does not only display the maximum and minimum change of the result variable, but also the value of the result variable for each data point of the random and scenario variables. This is useful to check if the dependence between distribution variables and result variables is monotonous or not.
Toolbar

XLSTAT-Sim has a dedicated toolbar “XLSTAT-Sim”. The “XLSTAT-Sim” toolbar can be displayed by clicking the XLSTAT-Sim icon in the XLSTAT toolbar.

Click this icon to define a new distribution (see Define a distribution for more details).

Click this icon to define a new scenario variable (see Define a scenario variable for more details).

Click this icon to define a new result (see Define a result variable for more details).

Click this icon to define a new statistic (see Define a statistic for more details).

Click this icon to reinitialize the simulation model and do a first simulation iteration.

Click this icon to do one simulation step.

Click this icon to start the simulation and display a report.

Click this icon to export the simulation model. All XLSTAT-Sim functions are transformed to comments. The formulae in the cells are stored as cell comments and the formulae are either replaced by the default value or by the formula linking to other cells in the case of XLSTAT_SimRes.

Click this icon to import the simulation model. All XLSTAT-Sim functions are extracted from cell comments and exported as formulae in the corresponding cells.

Click this icon to display the XLSTAT-Sim options dialog box.
Options

To display the options dialog box, click the button of the “XLSTAT-SIM” toolbar. Use this dialog box to define the general options of the XLSTAT-SIM module.

**General** tab:

**Model limited to:** This option allows defining the size of the active simulation model. Limit if possible your model to a single Excel sheet. The following options are available:

- **Sheet:** Only the simulation functions in the active Excel sheet will be used in the simulation model. The other sheets are ignored.
- **Workbook:** All the simulation functions of the active workbook are included in the simulation model. This option allows using several Excel sheets for one model.

**Sampling method:** This option allows choosing the method of sample generation. Two possibilities are available:

- **Classic:** The samples are generated using Monte Carlo simulations.
- **Latin hypercubes:** The samples are generated using the Latin Hypercubes method. This method divides the distribution function of the variable into sections that have the same size and then generates equally sized samples within each section. This leads to a faster convergence of the simulation. You can enter the number of sections. Default value is 500.

**Single step memory:** Enter the maximum number of simulation steps that will be stored in the single step mode in order to calculate the statistics fields. When the limit is reached, the window moves forward (the first iteration is forgotten and the new one is stored). The default value is 500. This value can be larger, if necessary.

**Number of iterations by step:** Enter the value of the number of simulation iterations that are performed during one step. The default value is 1.

**Format** tab:

Use these options to set the format of the various model elements that are displayed on the Excel sheets:

- **Distributions:** You can define the color of the font and the color of the background of the cells where the definition of the input random variables and their corresponding distributions are stored.

- **Scenario variables:** You can define the color of the font and the color of the background of the cells where the scenario variables are stored.
- **Result variables**: You can define the color of the font and the color of the background of the cells where the result variables are stored.

- **Statistics**: You can define the color of the font and the color of the background of the cells where the statistics are stored.

**Convergence** tab:

**Stop conditions**: Activate this option to stop the simulation if the convergence criteria are reached.

- **Criterion**: Select the criterion that should be used for testing the convergence. There are three options available:
  
  o **Mean**: The means of the monitored “result variables” (see below) of the simulation model will be used to check if the convergence conditions are met.

  o **Standard deviation**: The standard deviation of the monitored “result variables” (see below) of the simulation model will be used to check if the convergence conditions are met.

  o **Percentile**: The percentiles of the monitored “result variables” (see below) of the simulation model will be used to check if the convergence conditions are met. Choose the **Percentile** to be used. Default value is 90%.

- **Test frequency**: Enter the number of iterations to perform before the convergence criteria are checked again. Default value: 100.

- **Convergence**: Enter the value in % of the evolution of the convergence criteria from one check to the next, which, when reached, means that the algorithm has converged. Default value: 3%.

- **Confidence interval (%)**: Enter the size in % of the confidence interval that is computed around the selected criterion. The upper bound of the interval is compared to the convergence value defined above, in order to determine if the convergence is reached or not. Default value: 95%.

- **Monitored results**: Select which result variables of the simulation model should be monitored for the convergence. There are two options available:

  o **All result variables**: All result variables of the simulation model will be monitored during the convergence test.

  o **Activated result variables**: Only result variables that have their ConvActive parameter equal to 1 are monitored.
**References tab:**

**Reference to Excel cells:** Select the way references to names of variables to the simulation models are generated:

- **Absolute reference:** XLSTAT creates absolute references (for example $A$4) to the cell.
- **Relative reference:** XLSTAT creates absolute references (for example A4) to the cell.

Note: The absolute reference will not be changed if you copy and paste the XLSTAT_Sim formula, contrary to the relative reference.

**Results tab:**

**Filter level for results:** Select the level of details that will be displayed in the report. This controls for the descriptive statistics tables and the histograms of the different model elements:

- **All:** Details are displayed for all elements of the model.
- **Activated:** Details are only displayed for the elements that have a value of the Visible parameter set to 1.
- **None:** No detail will be displayed for the elements of the model.

**Example**

Examples showing how to build a simulation model are available on the Addinsoft website at:

http://www.xlstat.com/demo-sim1.htm

http://www.xlstat.com/demo-sim2.htm


http://www.xlstat.com/demo-sim4.htm

**References**

Define a distribution

Use this tool in a simulation model when there is uncertainty on the value of a variable (or quantity) that can be described with a distribution. The distribution will be associated with the currently selected cell.

Description

This function is one of the essential elements of a simulation model. For a more detailed description on how a simulation model is constructed and calculated, please read the introduction on XLSTAT-Sim.

This tool allows to define the theoretical distribution function with known parameters that will be used to generate a sample of a given random variable. A wide choice of distribution functions is available.

To define the distribution that a given variable (physically, a cell on the Excel sheet) follows, you need to create a call to one of the XLSTAT_SimX functions or to use the dialog box that will generate for you the formula calling XLSTAT_SimX. X stands for the distribution (see the table below for additional details).

XLSTAT_SimX syntax:

XLSTAT_SimX(VarName, Param1, Param2, Param3, Param4, Param5, TruncMode, LowerBound, UpperBound, DefaultType, DefaultValue, Visible)

XLSTAT_SimX stands for one of the available distribution functions that are listed in the table below. A variable based on the corresponding distribution is defined. See the table below to see the available distributions.

VarName is a string giving the name of the variable for which the distribution is being defined. The name of the variable is used in the report to identify the variable.

Param1 is an optional input (default is 0) that gives the value of the first parameter of the distribution if relevant.

Param2 is an optional input (default is 0) that gives the value of the second parameter of the distribution if relevant.

Param3 is an optional input (default is 0) that gives the value of the third parameter of the distribution if relevant.

Param4 is an optional input (default is 0) that gives the value of the fourth parameter of the distribution if relevant.
**Param5** is an optional input (default is 0) that gives the value of the fifth parameter of the distribution if relevant.

**TruncMode** is an optional integer that indicates if and how the distribution is truncated. A 0 (default value) corresponds to no truncation. 1 corresponds to truncating the distribution between two bounds that must then be specified. 2 corresponds to truncating between two percentiles that must then be specified.

**TruncLower** is an optional value that gives the lower bound of the truncation.

**TruncUpper** is an optional value that gives the upper bound of the truncation.

**DefaultType** is an optional integer that chooses the default value of the variable: 0 (default value) corresponds to the theoretical expected mean; 1 to the value given by the DefaultValue argument.

**DefaultValue** is an optional value giving the default value displayed in the cell before any simulation is performed. When no simulation process is ongoing, the default value will be displayed in the Excel cell as the result of the function.

**Visible** is an optional input that indicates if the details of this variable should be displayed in the simulation report. This option is only taken into account when the “Filter level for results” in the Options dialog box of XLSTAT-Sim is set to “Activated” (see the Format tab). 0 deactivates the display and 1 activates the display. Default value is 1.
The function will associate to the cell where they are entered a normal distribution with mean 50000 and standard deviation 5000. The cell will show 50000 (the default value). If a report is generated afterwards the results corresponding to that cell will be identified by “Revenue Q1”. The Param3, Param4 and Param5 are not entered because the Normal distribution has only two parameters. As the other parameters are not entered, they are set to their default value.
Determination of the parameters

In general, the choice of law and parameters of the law is guided by an empirical knowledge of the phenomenon, the results already available or working hypothesis.

To select the best suited law and the corresponding parameters you can use the “Distribution fitting” tool of XLSTAT. If you have a sample of data, by the help of this tool you can find the best parameters for a given distribution.

Random distributions available in XLSTAT-Sim

XLSTAT provides the following distributions:

- **Arcsine (α):** the density function of this distribution (which is a simplified version of the Beta type I distribution) is given by:
  \[
  f(x) = \frac{\sin(\pi x)}{\pi x} \left(\frac{x}{1-x}\right)^{\alpha}, \quad \text{with } 0<\alpha<1, \quad x \in [0,1]
  \]
  We have \(E(X) = \alpha\) and \(V(X) = \alpha(1-\alpha)/2\)

- **Bernoulli (p):** the density function of this distribution is given by:
  \[
  P(X = 1) = p, \quad P(X = 0) = 1 - p \quad \text{with } p \in [0,1]
  \]
  We have \(E(X) = p\) and \(V(X) = p(1-p)\)
  The Bernoulli, named after the Swiss mathematician Jacob Bernoulli (1654-1705), allows to describe binary phenomena where only events can occur with respective probabilities of \(p\) and \(1-p\).

- **Beta (α, β):** the density function of this distribution (also called Beta type I) is given by:
  \[
  f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}, \quad \text{with } \alpha, \beta > 0, \quad x \in [0,1] \quad \text{and} \quad B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}
  \]
  We have \(E(X) = \alpha/(\alpha + \beta)\) and \(V(X) = \alpha\beta/[(\alpha + \beta + 1)(\alpha + \beta)^2]\)

- **Beta4 (α, β, c, d):** the density function of this distribution is given by:
  \[
  f(x) = \frac{1}{B(\alpha, \beta)} \frac{(x-c)^{\alpha-1}}{(d-c)^{\alpha+\beta-1}} \frac{(d-x)^{\beta-1}}{(d-c)^{\alpha+\beta-1}}, \quad \text{with } \alpha, \beta > 0, \quad x \in [c,d]
  \]
  \[
  c, d \in \mathbb{R}, \quad \text{and} \quad B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}
  \]
We have $E(X) = c+(c-d)/\alpha$ and $V(X) = (c-d)^2\alpha\beta/[(\alpha+\beta+1)(\alpha+\beta)^2]$.

Pour the type I beta distribution, $X$ takes values in the $[0,1]$ range. The beta distribution is obtained by a variable transformation such that the distribution is on a $[c, d]$ interval where $c$ and $d$ can take any value.

- **Beta $(a, b)$**: the density function of this distribution (also called Beta type I) is given by:
  \[ f(x) = \frac{1}{B(a,b)} x^{a-1} (1-x)^{b-1}, \quad \text{with } a,b>0, x \in [0,1] \] and $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$

  $E(X) = a/(a+b)$ and $V(X) = ab/[(a+b+1)(a+b)^2]$.

- **Binomial $(n, p)$**: the density function of this distribution is given by:
  \[ P(X = x) = C_n^x p^x (1-p)^{n-x}, \quad \text{with } x \in \mathbb{N}, n \in \mathbb{N}^+, p \in [0,1] \]

  $E(X)= np$ and $V(X) = np(1-p)$

  $n$ is the number of trials, and $p$ the probability of success. The binomial distribution is the distribution of the number of successes for $n$ trials, given that the probability of success is $p$.

- **Negative binomial type I $(n, p)$**: the density function of this distribution is given by:
  \[ P(X = x) = C_{n+x-1}^x p^x (1-p)^{n-x}, \quad \text{with } x \in \mathbb{N}, n \in \mathbb{N}^+, p \in [0,1] \]

  $E(X) = n(1-p)/p$ and $V(X) = n(1-p)/p^2$

  $n$ is the number of successes, and $p$ the probability of success. The negative binomial type I distribution is the distribution of the number of unsuccessful trials necessary before obtaining $n$ successes.

- **Negative binomial type II $(k, p)$**: the density function of this distribution is given by:
  \[ P(X = x) = \frac{\Gamma(k+x)}{x!\Gamma(k)(1+p)^{k+x}} p^x, \quad \text{with } x \in \mathbb{N}, k, p>0 \]

  $E(X) = kp$ and $V(X) = kp(p+1)$

  The negative binomial type II distribution is used to represent discrete and highly heterogeneous phenomena. As $k$ tends to infinity, the negative binomial type II distribution tends towards a Poisson distribution with $\lambda = kp$.

- **Chi-square $(df)$**: the density function of this distribution is given by:
\[
f(x) = \left(\frac{1/2}{\Gamma(df/2)}\right)^{df/2} x^{df/2 - 1} e^{-x/2}, \quad \text{with } x > 0, \ df \in \mathbb{N}^*
\]

E(X) = df and V(X) = 2df

The Chi-square distribution corresponds to the distribution of the sum of df squared standard normal distributions. It is often used for testing hypotheses.

- Erlang \((k, \lambda)\): the density function of this distribution is given by:
  \[
f(x) = \lambda^k x^{k-1} e^{-\lambda x} \frac{1}{(k-1)!}, \quad \text{with } x \geq 0 \text{ and } k, \lambda > 0 \text{ and } k \in \mathbb{N}
\]

E(X) = \(k/\lambda\) and V(X) = \(k/\lambda^2\)

k is the shape parameter and \(\lambda\) is the rate parameter.

This distribution, developed by the Danish scientist A. K. Erlang (1878-1929) when studying the telephone traffic, is more generally used in the study of queuing problems.

Note: When \(k=1\), this distribution is equivalent to the exponential distribution. The Gamma distribution with two parameters is a generalization of the Erlang distribution to the case where \(k\) is a real and not an integer (for the Gamma distribution the scale parameter \(\beta = 1/\lambda\) is used).

- Exponential(\(\lambda\)): the density function of this distribution is given by:
  \[
f(x) = \lambda \exp(-\lambda x), \quad \text{with } x > 0 \text{ and } \lambda > 0
\]

E(X) = \(1/\lambda\) and V(X) = \(1/\lambda^2\)

The exponential distribution is often used for studying lifetime in quality control.

- Fisher \((df_1, df_2)\): the density function of this distribution is given by:
  \[
f(x) = \frac{1}{xB}(df_1/2, df_2/2) \left(\frac{df_1 x}{df_1 x + df_2}\right)^{df_1/2} \left(1 - \frac{df_1 x}{df_1 x + df_2}\right)^{df_2/2},
  \quad \text{with } x \geq 0 \text{ and } df_1, df_2 \in \mathbb{N}^+
\]

E(X) = \(df_2/(df_2 - 2)\) if \(df_2>0\), and V(X) = \(2df_2^3(df_1+df_2 - 2)/[df_1(df_1-2)(df_2-4)]\)

Fisher's distribution, from the name of the biologist, geneticist and statistician Ronald Aylmer Fisher (1890-1962), corresponds to the ratio of two Chi-square distributions. It is often used for testing hypotheses.

- Fisher-Tippett \((\beta, \mu)\): the density function of this distribution is given by:
\[ f(x) = \frac{1}{\beta} \exp \left( -\frac{x - \mu}{\beta} \right) - \exp \left( -\frac{x - \mu}{\beta} \right), \quad \text{with } \beta > 0 \]

\[ \text{E}(X) = \mu + \beta \gamma \text{ and } \text{V}(X) = \frac{\pi \beta^2}{6} \] where \( \gamma \) is the Euler-Mascheroni constant.

The Fisher-Tippett distribution, also called the Log-Weibull or extreme value distribution, is used in the study of extreme phenomena. The Gumbel distribution is a special case of the Fisher-Tippett distribution where \( \beta = 1 \) and \( \mu = 0 \).

- **Gamma** \((k, \beta, \mu)\): the density of this distribution is given by:

\[ f(x) = (x - \mu)^{k-1} e^{-\frac{(x-\mu)}{\beta}} \frac{\exp \left( -\frac{x - \mu}{\beta} \right)}{\beta^k \Gamma(k)}, \quad \text{with } x > \mu \text{ and } k, \beta > 0 \]

\[ \text{E}(X) = \mu + k \beta \text{ and } \text{V}(X) = k \beta^2 \]

\( k \) is the shape parameter of the distribution and \( \beta \) the scale parameter.

- **GEV** \((\beta, k, \mu)\): the density function of this distribution is given by:

\[ f(x) = \frac{1}{\beta} \left( 1 - k \frac{x - \mu}{\beta} \right)^{1/k-1} \exp \left( -\left( 1 - k \frac{x - \mu}{\beta} \right)^{1/k} \right), \quad \text{with } \beta > 0 \]

We have \( \text{E}(X) = \mu + \frac{\beta}{k} \Gamma(1 + k) \) and \( \text{V}(X) = \left( \frac{\beta}{k} \right)^2 \left( \Gamma(1 + 2k) - \Gamma^2(1 + k) \right) \)

The GEV (Generalized Extreme Values) distribution is much used in hydrology for modeling flood phenomena. \( k \) lies typically between -0.6 and 0.6.

- **Gumbel**: the density function of this distribution is given by:

\[ f(x) = \exp \left( -x - \exp \left( -x \right) \right) \]

\[ \text{E}(X) = \gamma \text{ and } \text{V}(X) = \frac{\pi^2}{6} \] where \( \gamma \) is the Euler-Mascheroni constant (0.5772156649…).

The Gumbel distribution, named after Emil Julius Gumbel (1891-1966), is a special case of the Fisher-Tippett distribution with \( \beta = 1 \) and \( \mu = 0 \). It is used in the study of extreme phenomena such as precipitations, flooding and earthquakes.

- **Logistic** \((\mu, s)\): the density function of this distribution is given by:

\[ f(x) = \frac{e^{-\frac{(x-\mu)}{s}}}{s \left( 1 + e^{-\frac{(x-\mu)}{s}} \right)}, \quad \text{with } \mu \in \mathbb{R}, \text{ and } s > 0 \]
We have $E(X) = \mu$ and $V(X) = (\pi s)^3/3$

- **Lognormal ($\mu, \sigma$):** the density function of this distribution is given by:
  \[
  f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x) - \mu)^2}{2\sigma^2}}, \quad \text{with} \; x, \sigma > 0
  \]
  
  $E(X) = \exp(\mu + \sigma^2/2)$ and $V(X) = [\exp(\sigma^2)-1]\exp(2\mu + \sigma^2)$

- **Lognormal2 ($m, s$):** the density function of this distribution is given by:
  \[
  f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x) - \mu)^2}{2\sigma^2}}, \quad \text{with} \; x, \sigma > 0
  \]
  
  $\mu = \ln(m) - \ln(1 + s^2/m^2)/2$ and $\sigma^2 = \ln(1 + s^2/m^2)$
  
  $E(X) = m$ and $V(X) = s^2$
  
  This distribution is just a reparametrization of the Lognormal distribution.

- **Normal ($\mu, \sigma$):** the density function of this distribution is given by:
  \[
  f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x - \mu)^2}{2\sigma^2}}, \quad \text{with} \; \sigma > 0
  \]
  
  $E(X) = \mu$ and $V(X) = \sigma^2$

- **Standard normal:** the density function of this distribution is given by:
  \[
  f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}
  \]
  
  $E(X) = 0$ and $V(X) = 1$
  
  This distribution is a special case of the normal distribution with $\mu=0$ and $\sigma=1$.

- **Pareto (a, b):** the density function of this distribution is given by:
  \[
  f(x) = \frac{ab^a}{x^{a+1}}, \quad \text{with} \; a, b > 0 \; \text{and} \; x \geq b
  \]
  
  $E(X) = ab/(a-1)$ and $V(X) = ab^2/[((a-1)^2(a-2)]$

  The Pareto distribution, named after the Italian economist Vilfredo Pareto (1848-1923), is also known as the Bradford distribution. This distribution was initially used to represent the distribution of wealth in society, with Pareto's principle that 80% of the wealth was owned by 20% of the population.
• PERT (a, m, b): the density function of this distribution is given by:

\[
\begin{align*}
    f(x) &= \frac{1}{B(\alpha, \beta)} \left( \frac{x-a}{b-a} \right)^{\alpha-1} \left( \frac{b-x}{b-a} \right)^{\beta-1}, \quad \text{with } \alpha, \beta > 0, \ x \in [a, b] \\
    \alpha &= \frac{4m + b - 5a}{b - a} \\
    \beta &= \frac{5b - a - 4m}{b - a}
\end{align*}
\]

We have \(E(X) = (b-a)\alpha/(\alpha+\beta)\) and \(V(X) = (b-a)^2\alpha\beta/[(\alpha+\beta+1)(\alpha+\beta)]\)

The PERT distribution is a special case of the beta4 distribution. It is defined by its definition interval \([a, b]\) and \(m\) the most likely value (the mode). PERT is an acronym for Program Evaluation and Review Technique, a project management and planning methodology. The PERT methodology and distribution were developed during the project held by the US Navy and Lockheed between 1956 and 1960 to develop the Polaris missiles launched from submarines. The PERT distribution is useful to model the time that is likely to be spent by a team to finish a project. The simpler triangular distribution is similar to the PERT distribution in that it is also defined by an interval and a most likely value.

• Poisson (\(\lambda\)): the density function of this distribution is given by:

\[
P(X = x) = \frac{\exp(-\lambda) \lambda^x}{x!}, \quad \text{with } x \in \mathbb{N} \text{ and } \lambda > 0
\]

\(E(X) = \lambda\) and \(V(X) = \lambda\)

Poisson's distribution, discovered by the mathematician and astronomer Siméon-Denis Poisson (1781-1840), pupil of Laplace, Lagrange and Legendre, is often used to study queuing phenomena.

• Student (\(df\)): the density function of this distribution is given by:

\[
f(x) = \frac{\Gamma\left(\frac{df + 1}{2}\right)}{\sqrt{\pi df} \Gamma\left(\frac{df}{2}\right)} \left(1 + \frac{x^2}{df}\right)^{-\frac{df+1}{2}}, \quad \text{with } df > 0
\]

\(E(X) = 0\) if \(df>1\) and \(V(X) = df/(df - 2)\) if \(df>2\)

The English chemist and statistician William Sealy Gosset (1876-1937), used the nickname Student to publish his work, in order to preserve his anonymity (the Guinness brewery forbade its employees to publish following the publication of
confidential information by another researcher). The Student’s t distribution is the
distribution of the mean of \( df \) variables standard normal variables. When \( df=1 \),
Student’s distribution is a Cauchy distribution with the particularity of having neither
expectation nor variance.

- Trapezoidal \((a, b, c, d)\): the density function of this distribution is given by:
  \[
  f(x) = \begin{cases} 
  \frac{2(x-a)}{(d+c-b-a)(b-a)} & x \in [a,b] \\
  \frac{2}{d+c-b-a} & x \in [b,c] \\
  \frac{2(d-x)}{(d+c-b-a)(d-c)} & x \in [a,b] \\
  0 & x < a, x > d
  \end{cases}
  \]
  with \( a < m < b \)

  We have \( E(X) = \frac{(d^2+c^2-b^2-a^2+cd-ab)}{3(d+c-b-a)} \)
  and \( V(X) = \frac{(c+d)(c^2+d^2)-(a+b)(a^2+b^2)}{6(d+c-b-a)}-E^2(X) \)

  This distribution is useful to represent a phenomenon for which we know that it can
take values between two extreme values \((a\text{ and } d)\), but that it is more likely to take
values between two values \((b\text{ and } c)\) within that interval.

- Triangular \((a, m, b)\): the density function of this distribution is given by:
  \[
  f(x) = \begin{cases} 
  \frac{2(x-a)}{(b-a)(m-a)} & x \in [a,m] \\
  \frac{2(b-x)}{(b-a)(b-m)} & x \in [m,b] \\
  0 & x < a, x > b
  \end{cases}
  \]
  with \( a < m < b \)

  We have \( E(X) = \frac{(a+m+b)}{3} \) and \( V(X) = \frac{(a^2+m^2+b^2 -ab-am-bm)}{18} \)

- TriangularQ \((q1, m, q2, p1, p2)\): the density function of this distribution is a
reparametrization of the Triangular distribution. A first step requires estimating the \( a \)
and \( b \) parameters of the triangular distribution, from the \( q1 \) and \( q2 \) quantiles to which
percentages \( p1 \) and \( p2 \) correspond. Once this is done, the distribution functions can be
computed using the triangular distribution functions.

- Uniform \((a, b)\): the density function of this distribution is given by:
\[ f(x) = \frac{1}{b-a}, \text{ with } b > a \text{ and } x \in [a,b] \]

E(X) = (a+b)/2 and V(X) = (b-a)^2/12

The uniform (0,1) distribution is much used for simulations. As the cumulative distribution function of all the distributions is between 0 and 1, a sample taken in a Uniform (0,1) distribution is used to obtain random samples in all the distributions for which the inverse can be calculated.

- **Uniform discrete (a, b):** the density function of this distribution is given by:
  \[
  f(x) = \frac{1}{b-a+1}, \text{ with } b > a, (a,b) \in N, x \in N, x \in [a,b] 
  \]
  We have E(X) = (a+b)/2 and V(X) = [(b-a+1)^2 -1]/12

  The uniform discrete distribution corresponds to the case where the uniform distribution is restricted to integers.

- **Weibull (\(\beta\)):** the density function of this distribution is given by:
  \[
  f(x) = \beta x^{\beta-1} \exp(-x^\beta), \text{ with } x > 0 \text{ and } \beta > 0
  \]
  We have E(X) = \(\Gamma\left(\frac{1}{\beta} + 1\right)\) and V(X) = \(\Gamma\left(\frac{2}{\beta} + 1\right) - \Gamma^2\left(\frac{1}{\beta} + 1\right)\)

  \(\beta\) is the shape parameter for the Weibull distribution.

- **Weibull (\(\beta, \gamma\)):** the density function of this distribution is given by:
  \[
  f(x) = \frac{\beta}{\gamma} \left(\frac{x}{\gamma}\right)^{\beta-1} e^{-\left(\frac{x}{\gamma}\right)^\beta}, \text{ with } x > 0, \text{ and } \beta, \gamma > 0
  \]
  We have E(X) = \(\gamma \Gamma\left(\frac{1}{\beta} + 1\right)\) and V(X) = \(\gamma^2 \left[ \Gamma\left(\frac{2}{\beta} + 1\right) - \Gamma^2\left(\frac{1}{\beta} + 1\right) \right]\)

  \(\beta\) is the shape parameter of the distribution and \(\gamma\) the scale parameter. When \(\beta=1\), the Weibull distribution is an exponential distribution with parameter \(1/\gamma\).

- **Weibull (\(\beta, \gamma, \mu\)):** the density function of this distribution is given by:
  \[
  f(x) = \frac{\beta}{\gamma} \left(\frac{x-\mu}{\gamma}\right)^{\beta-1} e^{-\left(\frac{x-\mu}{\gamma}\right)^\beta}, \text{ with } x > \mu, \text{ and } \beta, \gamma > 0
  \]
We have $\mathbb{E}(X) = \mu + \gamma \Gamma \left( \frac{1}{\beta} + 1 \right)$ and $\text{V}(X) = \gamma^2 \left[ \Gamma \left( \frac{2}{\beta} + 1 \right) - \Gamma^2 \left( \frac{1}{\beta} + 1 \right) \right]$.

The Weibull distribution, named after the Swede Ernst Hjalmar Waloddi Weibull (1887-1979), is much used in quality control and survival analysis. $\beta$ is the shape parameter of the distribution and $\gamma$ the scale parameter. When $\beta=1$ and $\mu=0$, the Weibull distribution is an exponential distribution with parameter $1/\gamma$.

### Dialog box

- **OK**: click this button to create the variable.
- **Cancel**: click this button to close the dialog box without doing any modification.
- **Help**: click this button to display help.
- **reload** button: click this button to reload the default options.
- **delete** button: click this button to delete the data selections.

### General tab:

**Variable name**: Enter the name of the random variable or select a cell where the name is available. If you select a cell, an absolute reference (for example $A4$) or a relative reference (for example $A4$) to the cell is created, depending on your choice in the XLSTAT options. (See the Options section for more details)

**Distributions**: Select the distribution that you want to use for the simulation. See the description section for more information on the available distributions.

**Parameters**: Enter the value of the parameters of the distribution you selected.

**Truncation**: Activate this option to truncate the distribution.

- **Absolute**: Select this option, if you want to enter the lower and upper bound of the truncation as absolute values.
• **Percentile**: Select this option, if you want to enter the lower and upper bound of the truncation as percentile values.

• **Lower bound**: Enter the value of the lower bound of the truncation.

• **Upper bound**: Enter the value of the upper bound of the truncation.

**Options** tab:

**Default cell value**: Choose the default value of the random variable. This value will be returned when no simulation model is running. The value may be defined by one of the following three methods:

• **Expected value**: This option selects the expected value of the distribution as the default cell value.

• **Fixed value**: Enter the default value.

• **Reference**: Choose a cell in the active Excel sheet that contains the default value.

**Display results**: Activate this option to display the detailed results for the random variable in the simulation report. This option is only active if you selected the “Activated” filter level in the simulation preferences. (See the Options section for more details).

**Results**

The result is function call to XLSTAT_SimX with the selected parameters. The following formula is generated in the active Excel cell:

```excel
= XLSTAT_SimX(VarName, Param1, Param2, Param3, Param4, Param5, TruncMode, LowerBound, UpperBound, DefaultType, DefaultValue, Visible)
```

The background color and the font color in the Excel cell are applied according to your choices in the XLSTAT-Sim options.
Define a scenario variable

Use this tool to define a variable which value varies between two known bounds during the tornado analysis.

Description

This function allows to build a scenario variable that is used during the tornado analysis. For a more detailed description on how a simulation model is constructed, please read the introduction on XLSTAT-Sim.

A scenario variable is used for tornado analysis. This function gives you the possibility to define a scenario variable by letting XLSTAT know the bounds between which it varies. To define the scenario variable (physically, a cell on the Excel sheet), you need to create a call to the XLSTAT_SimSVar function or to use the dialog box that will generate for you the formula calling XLSTAT_SimSVar.

**XLSTAT_SimSVar syntax**

XLSTAT_SimSVar(VarName, LowerBound, UpperBound, Type, Step, DefaultType, DefaultValue, Visible)

*VarName* is a string that contains the name of the scenario variable. This can be a reference to a cell in the same Excel sheet. The name is used during the report to identify the cell.

*LowerBound* corresponds to the lower bound of the interval of possible values for the scenario variable.

*UpperBound* corresponds to the upper bound of the interval of possible values for the scenario variable.

*Type* is an integer that indicates the data type of the scenario variable. 1 stands for a continuous variable and 2 for a discrete variable. This input is optional with default value 1.

*Step* is a number that indicates in the case of a discrete variable the step size between two values to be examined during the tornado analysis. This input is optional with default value 1.

*DefaultType* is an optional integer that chooses the default value of the variable: 0 (default value) corresponds to the theoretical expected mean; 1 to the value given by the DefaultValue argument.

*DefaultValue* is a value that that corresponds to the default value of the scenario variable. The default value is returned as the result of this function.
Visible is an optional input that indicates if the details of this variable should be displayed in the simulation report. This option is only taken into account when the “Filter level for results” in the options dialog box of XLSTAT-Sim is set to “Activated” (see the Format tab). 0 deactivates the display and 1 activates the display. Default value is 1.

Dialog box

- **OK**: click this button to create the variable.
- **Cancel**: click this button to close the dialog box without doing any modification.
- **Help**: click this button to display help.
- **Reload**: click this button to reload the default options.
- **Delete**: click this button to delete the data selections.

**General** tab:

**Variable name**: Enter the name of the scenario variable or select a cell where the name is available. If you select a cell, an absolute reference (for example $A$4) or a relative reference (for example A4) to the cell is created, depending on your choice in the XLSTAT options. (See the Options section for more details)

**Lower bound**: Enter the value of the lower bound or select a cell in the active Excel sheet that contains the value of the lower bound of the interval in which the scenario variable varies.

**Upper bound**: Enter the value of the upper bound or select a cell in the active Excel sheet that contains the value of the upper bound of the interval in which the scenario variable varies.

**Data type**:

- **Continuous**: Choose this option to define a continuous scenario variable that can take any value between the lower and upper bounds.

- **Discrete**: Choose this option to define a discrete scenario variable.
  - **Step**: Enter the value of the step or select a cell in the active Excel sheet that contains the value of the step.
**Options** tab:

**Default cell value:** Choose the default value of the random variable. This value will be returned when no simulation model is running. The value may be defined by one of the following three methods:

- **Expected value:** This option returns the center of the interval as the default cell value.
- **Fixed value:** Enter the default value.
- **Reference:** Choose a cell in the active Excel sheet that contains the default value.

**Display results:** Activate this option to display the detailed results for the random variable in the simulation report. This option is only active if you selected the “Activated” filter level in the simulation preferences. (See the Options section for more details).

**Results**

The result is function call to XLSTAT_SimSVar with the selected parameters. The following formula is generated in the active Excel cell:

\[
\text{=XLSTAT\_SimSVar(SVarName, LowerBound, UpperBound, Type, Step, DefaultType, DefaultValue, Visible)}
\]

The background color and the font color in the Excel cell are applied according to your choices in the XLSTAT-Sim options.
Define a result variable

Use this tool in a simulation model to define a result variable which calculation is the real aim of the simulation model.

Description

This result variable is one of the two essential elements of a simulation model. For a more detailed description on how a simulation model is constructed and calculated, please read the introduction on XLSTAT-Sim.

Result variables can be used to define when a simulation process should stop during a run. If, in the XLSTAT-Sim Options dialog box, you asked that the "Activated result variables" are used the stop the simulations when, for example the mean has converged, then, if the ConvActiv parameter of the result variable is set to 1, the mean of the variable will used to determine if the simulation process has converged or not.

To define the result variable (physically, a cell on the Excel sheet), you need to create a call to the XLSTAT_SimRes function or to use the dialog box that will generate for you the formula calling XLSTAT_SimRes.

**XLSTAT_SimRes syntax:**

`XLSTAT_SimRes (ResName, Formula, DefaultValue, ConvActiv, Visible)`

- **ResName** is a string that contains the name of the result variable or a reference to a cell where the name is located. The name is used during the report to identify the result variable.

- **Formula** is a string that contains the formula that is used to calculate the results. The formula links directly or indirectly the random input variables and, if available the scenario variables, to the result variable. This corresponds to an Excel formula without the leading "=".

- **DefaultValue** of type number is optional and contains the default value of the result variable. This value is not used in the computations.

- **ConvActiv** is an integer that indicates if this result is checked during the convergence tests. This option is only active, if the "Activated result variables" convergence option is activated in the XLSTAT-Sim options dialog box.

- **Visible** is an optional input that indicates if the details of this variable should be displayed in the simulation report. This option is only taken into account when the "Filter level for results" in the options dialog box of XLSTAT-Sim is set to "Activated" (see the Format tab). 0 deactivates the display and 1 activates the display. Default value is 1.
Example:

=XLSTAT_SimRes( "Forecast N+1", B3+B4-B5)

This function defines in the active cell a result variable called “Forecast N +1” calculated as the sum of cells B3 and B4 minus B5. The Visible parameter is not entered because it is only necessary when the “Filter level for the results” is set to “Activated” (see the Options dialog box) and because we want the result to be anyway visible.

Dialog box

- OK: click this button to create the variable.
- Cancel: click this button to close the dialog box without doing any modification.
- Help: click this button to display help.
- : click this button to reload the default options.
- : click this button to delete the data selections.

General tab:

Variable name: Enter the name of the random variable or select a cell where the name is available. If you select a cell, it depends on the selection in the options, whether an absolute (for example $A$4) or a relative reference (for example A4) to the cell is created. (See the Options section for more details)

Use to monitor convergence: Activate this option to include this result variable in the result variables that are used to test for convergence. This option is only active, if you selected the “Activated results variables” option in the XLSTAT-Sim convergence options. ConvActiv should be 1 if you want the variable to be used to monitor the results. Default value is 1.

Display Results: Activate this option to display the detailed results for the result variable in the simulation report. This option is only active, if you selected the restricted filter level in the simulation preferences. (See the XLSTAT-Sim options for more details).
Results

A function call to XLSTAT_SimRes with the selected parameters and the following syntax will be generated in the active Excel cell:

=XLSTAT_SimRes (ResName, Formula, DefaultValue, ConvActiv, Visible)

The background color and the font color in the Excel cell are applied according to your choices in the XLSTAT-Sim options.
Define a statistic

Use this tool in a simulation model to define a statistic based on a variable of the simulation model. The statistic is updated after each iteration of the simulation process. Results relative to the defined statistics are available in the simulation report. A wide choice of statistics is available.

Description

This function is one of the four elements of a simulation model. For a more detailed description on how a simulation model is constructed and calculated, please read the introduction on XLSTAT-Sim.

This tool allows to create a function that calculates a statistic after each iteration of the simulation process. The statistic is computed and stored. During the step by step simulations, you can track how the statistic evolves. In the simulation report you can optionally see details on the statistic. A wide choice of statistics is available.

To define the statistic function (physically, a cell on the Excel sheet), you need to create a call to a XLSTAT_SimStatX/TheoX/SPCX function or to use the dialog box that will generate for you the formula calling the function. X stands for the statistic as defined in the tables below. A variable based on the corresponding statistic is created.

**XLSTAT_SimStat/Theo/SPC Syntax**

```
XLSTAT_SimStatX(StatName, Reference, Visible)
XLSTAT_SimTheoX(StatName, Reference, Visible)
XLSTAT_SimSPCX(StatName, Reference, Visible)
```

X stands for one of the selected statistic. The available statistics are listed in the tables below.

**StatName** is a string that contains the name of the statistic or a reference to a cell where the name is located. The name is used during the report to identify the statistic.

**Reference** indicates the model variable to be tracked. This is a reference to a cell in the same Excel sheet.

**Visible** is an optional input that indicates if the details of this statistic should be displayed when the “Filter level for results” in the Options dialog box of XLSTAT-Sim is set to “Activated” (see the Format tab). 0 deactivates the display and 1 activates the display. Default value is 1.
Descriptive statistics

The following descriptive statistics are available:

<table>
<thead>
<tr>
<th>Statistic name</th>
<th>XLSTAT Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of observations</td>
<td>XLSTAT_SimStatNbrObs</td>
</tr>
<tr>
<td>Number of missing values</td>
<td>XLSTAT_SimStatNbrMiss</td>
</tr>
<tr>
<td>Sum of weights</td>
<td>XLSTAT_SimStatSumOfWeights</td>
</tr>
<tr>
<td>Minimum</td>
<td>XLSTAT_SimStatMinimum</td>
</tr>
<tr>
<td>Maximum</td>
<td>XLSTAT_SimStatMaximum</td>
</tr>
<tr>
<td>Frequency of minimum</td>
<td>XLSTAT_SimStatFreqMin</td>
</tr>
<tr>
<td>Frequency of maximum</td>
<td>XLSTAT_SimStatFreqMax</td>
</tr>
<tr>
<td>Range</td>
<td>XLSTAT_SimStatAmplitude</td>
</tr>
<tr>
<td>1st quartile</td>
<td>XLSTAT_SimStat1stQuartile</td>
</tr>
<tr>
<td>Median</td>
<td>XLSTAT_SimStatMedian</td>
</tr>
<tr>
<td>3rd quartile</td>
<td>XLSTAT_SimStat3rdQuartile</td>
</tr>
<tr>
<td>Sum</td>
<td>XLSTAT_SimStatSum</td>
</tr>
<tr>
<td>Mean</td>
<td>XLSTAT_SimStatMean</td>
</tr>
<tr>
<td>Variance n</td>
<td>XLSTAT_SimStatVarianceN</td>
</tr>
<tr>
<td>Variance n-1</td>
<td>XLSTAT_SimStatVarianceN1</td>
</tr>
<tr>
<td>Standard deviation n</td>
<td>XLSTAT_SimStatStdevN</td>
</tr>
<tr>
<td>Standard deviation n-1</td>
<td>XLSTAT_SimStatStdevN1</td>
</tr>
<tr>
<td>Variation coefficient</td>
<td>XLSTAT_SimStatVariation</td>
</tr>
<tr>
<td>Skewness (Pearson)</td>
<td>XLSTAT_SimStatSkewnessPearson</td>
</tr>
<tr>
<td>Skewness (Fisher)</td>
<td>XLSTAT_SimStatSkewnessFisher</td>
</tr>
<tr>
<td>Skewness (Bowley)</td>
<td>XLSTAT_SimStatSkewnessBowley</td>
</tr>
<tr>
<td>Kurtosis (Pearson)</td>
<td>XLSTAT_SimStatKurtosisPearson</td>
</tr>
<tr>
<td>Kurtosis (Fisher)</td>
<td>XLSTAT_SimStatKurtosisFisher</td>
</tr>
<tr>
<td>Standard error of the mean</td>
<td>XLSTAT_SimStatStdErrorOfMean</td>
</tr>
<tr>
<td>Lower bound on mean</td>
<td>XLSTAT_SimStatLowerBounce95Perc</td>
</tr>
<tr>
<td>Upper bound on mean</td>
<td>XLSTAT_SimStatUpperBounce95Perc</td>
</tr>
<tr>
<td>Standard deviation of skewness</td>
<td>XLSTAT_SimStatStdErrorOfSkewness</td>
</tr>
<tr>
<td>Standard deviation of kurtosis</td>
<td>XLSTAT_SimStatStdErrorOfKurtosis</td>
</tr>
<tr>
<td>Mean absolute deviation</td>
<td>XLSTAT_SimStatMeanAbsDeviation</td>
</tr>
<tr>
<td>Median absolute deviation</td>
<td>XLSTAT_SimStatMedianAbsDeviation</td>
</tr>
<tr>
<td>Geometric mean</td>
<td>XLSTAT_SimStatGeometricMean</td>
</tr>
<tr>
<td>Geometric standard deviation</td>
<td>XLSTAT_SimStatGSD</td>
</tr>
<tr>
<td>Harmonic mean</td>
<td>XLSTAT_SimStatHarmonicMean</td>
</tr>
</tbody>
</table>

Details and formulae relative to the above statistics are available in the description section of the “Descriptive statistics” tool of XLSTAT.

Theoretical statistics

These statistics are based on the theoretical computation of the mean, variance and standard deviation of the distribution, as opposed to the empirical computation based on the simulated samples.
SPC

Statistics from the domain of SPC (Statistical Process Control) are listed hereunder. These statistics are only available and calculated, if you have a valid license for the XLSTAT-SPC module.

<table>
<thead>
<tr>
<th>Statistic name</th>
<th>XLSTAT Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>XLSTAT_SimTheoMean</td>
</tr>
<tr>
<td>Variance</td>
<td>XLSTAT_SimTheoVariance</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>XLSTAT_SimTheoStddev</td>
</tr>
</tbody>
</table>
General tab:

Name: Enter the name of the statistic or select a cell where the name is available. If you select a cell, it depends on the selection in the options, whether an absolute (for example $A$4) or a relative reference (for example A4) to the cell is created. (See the Options section for more details).

Reference: Choose a cell in the active Excel sheet that contains the simulation model variable that you want to track with the selected statistic.

Statistic: Activate one of the following options and choose the statistic to compute:

- Descriptive: Select one of the available statistics (See description section for more details).
- Theoretical: Select one of the available statistics (See description section for more details).
- SPC: Select one of the available statistics (See description section for more details).

Display Results: Activate this option to display the detailed results for statistic in the simulation report. This option is only active, if you selected the restricted filter level in the simulation preferences (See the XLSTAT-Sim options section for more details).

Results

A function call to XLSTAT_SimStat/Theo/SPC with the selected parameters and the following syntax will be generated in the active Excel cell:

=XLSTAT_SimStat/Theo/SPC(DistName, Reference, Visible)

The background color and the font color in the Excel cell are applied according to your choices in the XLSTAT-Sim options.
Run

Once you have designed the simulation model using the four tools “define a distribution”, “define a scenario variable”, “define a result”, and “define a statistic”, you can click the icon of “XLSTAT-SIM” toolbar to display the “Run” dialog box that lets you define additional options before running the simulation model and displaying the report. A description of the results is available below.

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- ![Reload](image): Click this button to reload the default options.
- ![Delete](image): Click this button to delete the data selections.
- ![Switch Data Handling](image): Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab:**

**Number of simulations**: Enter the number of simulations to perform for the model (Default value: 300).

**Correlation/Covariance matrix**: Activate this option to include a correlation or covariance matrix in the simulation model. Column and row headers must be selected as they are used by XLSTAT to know which variables are involved. As a matter of fact, column and row labels must be identical to the names of the corresponding distribution fields of the simulation model.
Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Labels included: Activate this option if the row and column labels are selected.

Options tab:

Tornado/Spider: Choose the options for the calculation of the tornado and spider analysis.

- **Number of points**: Choose the number of points between the two bounds of the intervals that are used for the tornado analysis.

- **Standard value**: Choose how the standard value around which the intervals to check during the tornado and spider analysis needs to be computed for each variable.
  - **Median**: The default value of the distribution fields is the median of the simulated values.
  - **Default cell value**: The default value defined for the variables is used.

- **Interval definition**: Choose an option for the definition of the limits of the intervals of the variables that are checked during the tornado/spider analyses.
  - **Percentile of variable**: Choose which two percentiles need to be used to determine the bounds of the intervals for the tornado/spider analyses. You can choose between [25%, 75%], [10%, 90%], and [5%, 95%]. This option is only available if the median is the central value.
  - **% of deviation of value**: Choose which bounds; computed as % of the central value should be used as the bounds for the intervals. You can choose between [-25%, 25%], [-10%, 10%], and [-5%, 5%].

SPC tab:

Calculate Process capabilities: Activate this option to calculate process capabilities for input random variables, result variables and statistics.

- **Variable names**: Select the data that correspond to the names of the variables for which you want to calculate process capabilities.
• **LSL**: Select the data that correspond to the lower specification limit (LSL) of the process for the variables for which the names have been selected.

• **USL**: Select the data that correspond to the upper specification limit (USL) of the process for the variables for which the names have been selected.

• **Target**: Select the data that correspond to the target of the process for the variables for which the names have been selected.

• **Confidence interval (%)**: If the calculation of the process capabilities is activated, please enter the percentage range of the confidence interval to use for calculating the confidence interval around the parameters. Default value: 95.

**Outputs** tab:

**Correlations**: Activate this option to display the correlation matrix between the variables. If the "significant correlations in bold" option is activated, the correlations that are significant at the selected significance level are displayed in bold.

• **Type of correlation**: Choose the type of correlation to use for the computations (see the description section for more details).

• **Significance level (%)**: Enter the significance level for the test of on the correlations (default value: 5%).

• **p-values**: Activate this option to display the p-values corresponding to the correlations.

• **Sensitivity**: Activate this option to display the results of the sensitivity analysis.

**Tornado**: Activate this option to display the results of the tornado analysis.

**Spider**: Activate this option to display the results of the spider analysis.

**Simulation details**: Activate this option to display the details on the iterations of the simulation.

**Descriptive statistics**: Activate this option to compute and display descriptive statistics for the variables of the model.

• **All**: Click this button to select all.

• **None**: Click this button to deselect all.

• **Display vertically**: Check this option so that the table of descriptive statistics is displayed vertically (one line per descriptive statistic).
Charts tab:

This tab is divided into three sub-tabs.

Histograms tab:

Histograms: Activate this option to display the histograms of the samples. For a theoretical distribution, the density function is displayed.

- Bars: Choose this option to display the histograms with a bar for each interval.
- Continuous lines: Choose this option to display the histograms with a continuous line.

Cumulative histograms: Activate this option to display the cumulated histograms of the samples.

- Based on the histogram: Choose this option to display cumulative histograms based on the same interval definition as the histograms.
- Empirical cumulative distribution: Choose this option to display cumulative histograms which actually correspond to the empirical cumulative distribution of the sample.

Intervals: Select one of the following options to define the intervals of the histogram:

- Number: Choose this option to enter the number of intervals to create.
- Width: Choose this option to define a fixed width for the intervals.
- User defined: Select a column containing in increasing order the lower bound of the first interval, and the upper bound of all the intervals.
- Minimum: Activate this option to enter the minimum value of the histogram. If the Automatic option is chosen, the minimum is that of the sample. Otherwise, it is the value defined by the user.

Box plots tab:

Box plots: Check this option to display box plots (or box-and-whisker plots). See the description section for more details.

- Horizontal: Check this option to display box plots and scattergrams horizontally.
- Vertical: Check this option to display box plots and scattergrams vertically.
• **Group plots**: Check this option to group together the various box plots and scattergrams on the same chart to compare them.

• **Minimum/Maximum**: Check this option to systematically display the points corresponding to the minimum and maximum (box plots).

• **Outliers**: Check this option to display the points corresponding to outliers (box plots) with a hollowed-out circle.

**Scattergrams**: Check this option to display scattergrams. The mean (red +) and the median (red line) are always displayed.

**Normal P-P plots**: Check this option to display P-P plots.

**Normal Q-Q Charts**: Check this option to display Q-Q plots.

**Correlations** tab:

**Correlation maps**: Several visualizations of a correlation matrix are proposed.

• The "blue-red" option allows to represent low correlations with cold colors (blue is used for the correlations that are close to -1) and the high correlations are with hot colors (correlations close to 1 are displayed in red color).

• The "Black and white" option allows to either display in black the positive correlations and in white the negative correlations (the diagonal of 1s is display in grey color), or to display in black the significant correlations, and in white the correlations that are not significantly different from 0.

• The "Patterns" option allows to represent positive correlations by lines that rise from left to right, and the negative correlations by lines that rise from right to left. The higher the absolute value of the correlation, the large the space between the lines.

**Scatter plots**: Activate this option to display the scatter plots for all two by two combinations of variables.

• **Matrix of plots**: Check this option to display all possible combinations of variables in pairs in the form of a two-entry table with the various variables displayed in rows and in columns.
  
  o **Histograms**: Activate this option so that XLSTAT displays a histogram when the X and Y variables are identical.

  o **Q-Q plots**: Activate this option so that XLSTAT displays a Q-Q plot when the X and Y variables are identical.
Confidence ellipses: Activate this option to display confidence ellipses. The confidence ellipses correspond to a x% confidence interval (where x is determined using the significance level entered in the General tab) for a bivariate normal distribution with the same means and the same covariance matrix as the variables represented in abscissa and ordinates.

Results

The first results are general results that display information about the model:

Distributions: This table shows for each input random variable in the model, its name, the Excel cell where it is located, the selected distribution, the static value, the data type, the truncation mode and bounds and the parameters of the distribution.

Scenario variables: This table shows for each input random variable in the model, its name, the Excel cell where it is located, the default value, the type, the lower und upper limit and the step size.

Result variables: This table shows for each result variable in the model, its name, the Excel cell where it is located, and the formula for its calculation.

Statistics: This table shows for each statistic in the model, its name, the Excel cell that contains it and the selected statistic.

Correlation/covariance matrix: If the option correlation/covariance matrix in the simulation model has been activated, then this table displays the input correlation/covariance matrix.

Convergence: If the option convergence in the simulation options has been activated, then this table displays for each result variable that has been selected for convergence checking, the value and the variation of the lower and upper bound of the confidence interval for the selected convergence criterion. Under the matrix information about the selected convergence criterion, the corresponding threshold of variation, and the number of executed iterations of simulation are displayed.

In the following section, details for the different model elements, distributions, scenario variables, result variables and statistics, are displayed.

Descriptive statistics: For each type of variable, the statistics selected in the dialog box are displayed in a table.

Descriptive statistics for the intervals: This table displays for each interval of the histogram its lower bound, upper bound, the frequency (number of values of the sample within the
interval), the relative frequency (the number of values divided by the total number of values in the sample), and the density (the ratio of the frequency to the size of the interval).

**Sensitivity**: A table with the correlations, the contributions and the absolute value of the contributions between the input random variables is displayed for each result variable. The contributions are then plotted on a chart.

**Tornado**: This table displays the minimum, the maximum and the range of the result variable when the input random variables and the scenario variables vary in the defined ranges. Then the minimum and the maximum are shown on a chart.

**Spider**: This table displays for all the points that are evaluated during the tornado analysis the value of each result variable when the input random variables and scenario variables vary. These values are then displayed in a chart.

The **correlation matrix** and the table of the p-values are displayed so that you can see the relationships between the input variables and the output variables. The correlation maps allow identifying potential structures in the matrix, of to quickly identify interesting correlations.

**Simulation details**: A table showing the values of each variable at each iteration is displayed.
Compare means (XLSTAT-Power)

Use this tool to compute power and sample size in a statistical test comparing means. T test, z test and non parametric tests are available.

Description

XLSTAT-Pro includes several tests to compare means, namely the t test, the z test and other non parametric tests like Mann-Whitney test. XLSTAT-Power allows estimating the power of these tests and calculates the number of observations required to obtain sufficient power.

When testing a hypothesis using a statistical test, there are several decisions to take:

- The null hypothesis H0 and the alternative hypothesis Ha.
- The statistical test to use.
- The type I error also known as alpha. It occurs when one rejects the null hypothesis when it is true. It is set a priori for each test and is 5%.

The type II error or beta is less studied but is of great importance. In fact, it represents the probability that one does not reject the null hypothesis when it is false. We can not fix it upfront, but based on other parameters of the model we can try to minimize it. The power of a test is calculated as 1-beta and represents the probability that we reject the null hypothesis when it is false.

We therefore wish to maximize the power of the test. The XLSTAT-Power module calculates the power (and beta) when other parameters are known. For a given power, it also allows to calculate the sample size that is necessary to reach that power.

The statistical power calculations are usually done before the experiment is conducted. The main application of power calculations is to estimate the number of observations necessary to properly conduct an experiment.

XLSTAT allows to compare:

- A mean to a constant (with z-test, t-test and Wilcoxon signed rank test)
- Two means associated with paired samples (with z-test, t-test and Wilcoxon signed rank test)
- Two means associated with independent samples (with z-test, t-test and Mann-Whitney test)

We use the t-test when the variance of the population is estimated and the z-test when it is known. In each case, the parameters will be different and will be shown in the dialog box. The non parametric tests are used when the distribution assumption is not met.
Methods

The sections of this document dedicated to the t-test, the z-test and the non parametric tests describe in detail the methods themselves.

The power of a test is usually obtained by using the associated non-central distribution. Thus, for the t-test, the non-central Student distribution is used.

T-test for one sample

The power of this test is obtained using the non-central Student distribution with non-centrality parameter:

$$NCP = \left| \frac{\bar{X} - X_0}{SD} \right| \cdot \sqrt{N}$$

With $X_0$ theoretical mean and $SD$ standard deviation.

The part $\frac{\bar{X} - X_0}{SD}$ is called the effect size.

T-test for two paired samples

The same formula as for the one sample case applies, but the standard deviation is calculated differently, we have:

$$NCP = \left| \frac{X_1 - X_2}{SD_{Diff}} \right| \cdot \sqrt{N}$$

With $SD_{Diff} = \sqrt{(SD_1^2 + SD_2^2) - 2 \cdot Corr \cdot SD_1 \cdot SD_2}$ and $Corr$ is the correlation between the two samples.

The part $\frac{X_1 - X_2}{SD_{Diff}}$ is the effect size.

T-test for two independent samples

In the case of two independent samples, the standard deviation is calculated differently and we use the harmonic mean of the number of observations.
\[ NCP = \frac{\bar{X}_1 - \bar{X}_2}{SD_{\text{pooled}}} \cdot \sqrt{\frac{N_{\text{harmonised}}}{2}} \]

With \( SD_{\text{pooled}} = \sqrt{\frac{(N_1 - 1) \cdot SD_1^2 - (N_2 - 1) \cdot SD_2^2}{N_1 + N_2 - 2}} \).

The part \( \frac{\bar{X}_1 - \bar{X}_2}{SD_{\text{pooled}}} \) is called effect size.

**Z-test for one sample**

In the case of the z-test, using the classical normal distribution with a parameter added to shift the distribution.

\[ NCP = \frac{\bar{X} - X_0}{SD} \cdot \sqrt{N} \]

With \( X_0 \) being the theoretical mean and \( SD \) being the standard deviation.

The part \( \frac{\bar{X} - X_0}{SD} \) is called effect size.

**Z-test for two paired samples**

The same formula applies as for the one sample case, but the standard deviation is calculated differently, we have:

\[ NCP = \frac{X_1 - X_2}{SD_{\text{Diff}}} \cdot \sqrt{N} \]

With \( SD_{\text{Diff}} = \sqrt{(SD_1^2 + SD_2^2) - 2 \cdot Corr \cdot SD_1 \cdot SD_2} \) and \( Corr \) is the correlation between the two samples.

The part \( \frac{X_1 - X_2}{SD_{\text{Diff}}} \) is called effect size.

**Z-test for two independent samples**
In the case of two independent samples, the standard deviation is calculated differently and we use the harmonic mean of the number of observations.

\[
NCP = \frac{\bar{X}_1 - \bar{X}_2}{SD_{\text{pooled}}} \cdot \sqrt{\frac{N_{\text{harmo}}}{2}}
\]

With \(SD_{\text{pooled}} = \sqrt{\frac{(N_1 - 1) \cdot SD_1^2 + (N_2 - 1) \cdot SD_2^2}{N_1 + N_2 - 2}}\).

The part \(\frac{\bar{X}_1 - \bar{X}_2}{SD_{\text{pooled}}}\) is called effect size.

Non parametric tests

In the case of the non parametric cases, a method called ARE (asymptotic relative efficiency) is used. This method helps to relate formulas used for the power of a t-test to those of the non parametric approaches. It has been introduced by Lehmann (1975). A factor called ARE is used. It has been shown that for mean comparisons this minimum value of the ARE is 0.864. This value is equal to 0.955 if the data are normally distributed. XLSTAT-Power uses the minimum ARE for the computations.

To compute power of the test, the used H0 distribution is the central Student distribution \(t(N,k - 2)\). The used H1 distribution is the noncentral Student distribution \(t(N,k - 2,\delta)\), where the noncentrality parameter is given by: \(\delta = d^*\sqrt{(N_1 + N_2)(N_1 + N_2)}\). Parameter k represents the asymptotic relative efficiency and depends on the parent distribution. Parameter d is the effect size defined like in the t-test case depending on the type of sample studied (independent, paired or one-sample).

Calculating sample size

To calculate the number of observations required, XLSTAT uses an algorithm that searches for the root of a function. It is called the Van Wijngaarden-Dekker-Brent algorithm (Brent, 1973). This algorithm is adapted to the case where the derivatives of the function are not known. It tries to find the root of:

\[
\text{power (N)} - \text{expected_power}
\]

We then obtain the size N such that the test has a power as close as possible to the desired power.
**Effect size**

This concept is very important in power calculations. Indeed, Cohen (1988) developed this concept. The effect size is a quantity that will allow calculating the power of a test without entering any parameters but will tell if the effect to be tested is weak or strong.

In the context of comparisons of means, the conventions of magnitude of the effect size are:

- \(d=0.2\), the effect is small.
- \(d=0.5\), the effect is moderate.
- \(d=0.8\), the effect is strong.

XLSTAT-Power allows entering directly the effect size.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Recompute**: Click this button to reload the default options.
- **Delete**: Click this button to delete the data selections.

**General** tab:

- **Goal**: Choose between computing power and sample size estimation.
- **Statistical test**: Select the test you want to apply.
- **Alternative hypothesis**: Select the alternative hypothesis to be tested.
- **Theoretical mean** (when only one sample is used): Enter the value of the theoretical mean to be tested.
**Alpha:** Enter the value of the type I error (alpha, between 0.001 and 0.999).

**Power** (when sample size estimation has been selected): Enter the value of the power to be reached.

**Sample size (group 1)** (when power computation has been selected): Enter the size of the first sample.

**Sample size (group 2)** (when power computation has been selected): Enter the size of the second sample.

**N1/N2 ratio** (when sample size has been selected and when there are two samples): Enter the ratio between the sizes of the first and the second samples.

**Parameters:** Select this option to enter the test parameters directly.

**Effect size:** Select this option to directly enter the effect size D (see the description part of this help).

**Mean (group 1):** Enter the mean for group 1.

**Mean (group 2):** Enter the mean for group 2.

**Std error (group 1):** Enter the standard error for group 1.

**Std error (group 2):** Enter the standard error for group 2.

**Correlation (when using paired samples):** Enter the correlation between the groups.

**Range:** Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Check this option to display the results in a new worksheet in the active workbook.

**Workbook:** Check this option to display the results in a new workbook.

**Graphics** tab:

**Simulation plot:** Activate this option if you want to plot different parameters of the test. Two parameters can vary. All remaining parameters are used as they were defined in the General tab.
**Y axis:** Select the parameter to be used on the Y axis of the simulation plot. You can either choose the power or the sample size.

**X axis:** Select the parameter to be used on the X axis of the simulation plot. You can either choose the power or the sample size, the type I error (alpha) or the effect size.

**Interval size:** Enter the minimum, maximum and interval size for the X axis of the simulation plot.

**Results**

**Results:** This table displays the parameters of the test and the power or the required number of observations. The parameters obtained by the calculation are in bold format. An explanation is displayed below this table.

**Intervals for the simulation plot:** This table is composed of two columns: power and sample size or alpha depending on the parameters selected in the dialog box. It helps building the simulation plot.

**Simulation plot:** This plot shows the evolution of the parameters as defined in the graphics tab of the dialog box.

**Example**

An example of power calculation based on a test is available on the Addinsoft website at


An example of calculating the required sample size is available on the Addinsoft website at


**References**

Compare variances (XLSTAT-Power)

Use this tool to compute power and sample size in a statistical test comparing variances.

Description

XLSTAT-Pro includes several tests to compare variances. XLSTAT-Power can calculate the power or the number of observations required for a test based on Fisher's F distribution to compare variances.

When testing a hypothesis using a statistical test, there are several decisions to take:

- The null hypothesis $H_0$ and the alternative hypothesis $H_a$.
- The statistical test to use.
- The type I error also known as alpha. It occurs when one rejects the null hypothesis when it is true. It is set a priori for each test and is 5%.

The type II error or beta is less studied but is of great importance. In fact, it represents the probability that one does not reject the null hypothesis when it is false. We cannot fix it upfront, but based on other parameters of the model, we can try to minimize it. The power of a test is calculated as 1-beta and represents the probability that we reject the null hypothesis when it is false.

We therefore wish to maximize the power of the test. The XLSTAT-Power module calculates the power (and beta) when other parameters are known. For a given power, it also allows to calculate the sample size that is necessary to reach that power.

The statistical power calculations are usually done before the experiment is conducted. The main application of power calculations is to estimate the number of observations necessary to properly conduct an experiment.

XLSTAT allows to compare two variances. The parameters are shown in the dialog box.

Methods

The sections of this document dedicated to the tests used to compare variances test describe in detail the methods themselves.

The power of a test is usually obtained by using the associated non-central distribution. In that case, we use the F distribution.

Several hypotheses can be tested, but the most common are the following (two-tailed):
H0: The difference between the variances is equal to 0.
Ha: The difference between the variances is different from 0.

The power computation will give the proportion of experiments that reject the null hypothesis. The calculation is done using the F distribution with the ratio of the variances as parameter and the sample sizes – 1 as degrees of freedom.

Calculating sample size

To calculate the number of observations required, XLSTAT uses an algorithm that searches for the root of a function. It is called the Van Wijngaarden-Dekker-Brent algorithm (Brent, 1973). This algorithm is adapted to the case where the derivatives of the function are not known. It tries to find the root of:

\[ \text{power (N)} - \text{expected_power} \]

We then obtain the size N such that the test has a power as close as possible to the desired power.

Effect size

This concept is very important in power calculations. Indeed, Cohen (1988) developed this concept. The effect size is a quantity that will allow to calculate the power of a test without entering any parameters but will tell if the effect to be tested is weak or strong.

Within the comparison of variances, it is the ratio between two variances to compare.

XLSTAT-Power allows to enter directly the effect size.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

**OK**: Click this button to start the computations.

**Cancel**: Click this button to close the dialog box without doing any computation.
**General** tab:

**Goal**: Choose between computing power and sample size estimation.

**Statistical test**: Select the test you want to apply.

**Alternative hypothesis**: Select the alternative hypothesis to be tested.

**Theoretical mean** (when only one sample is used): Enter the value of the theoretical mean to be tested.

**Alpha**: Enter the value of the type I error (alpha, between 0.001 and 0.999).

**Power** (when sample size estimation has been selected): Enter the value of the power to be reached.

**Sample size (group 1)** (when power computation has been selected): Enter the size of the first sample.

**Sample size (group 2)** (when power computation has been selected): Enter the size of the second sample.

**N1/N2 ratio** (when sample size has been selected and when there are two samples): Enter the ratio between the sizes of the first and the second samples.

**Parameters**: Select this option to enter the test parameters directly.

**Effect size**: Select this option to directly enter the effect size D (see the description part of this help).

**Variance (group 1)**: Enter the variance for group 1.

**Variance (group 2)**: Enter the variance for group 2.

**Range**: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.
Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Graphics tab:

Simulation plot: Activate this option if you want to plot different parameters of the test. Two parameters can vary. All remaining parameters are used as they were defined in the General tab.

Y axis: Select the parameter to be used on the Y axis of the simulation plot. You can either choose the power or the sample size.

X axis: Select the parameter to be used on the X axis of the simulation plot. You can either choose the power or the sample size, the type I error (alpha) or the effect size.

Interval size: Enter the minimum, maximum and interval size for the X axis of the simulation plot.

Results

Results: This table displays the parameters of the test and the power or the required number of observations. The parameters obtained by the calculation are in bold format. An explanation is displayed below this table.

Intervals for the simulation plot: This table is composed of two columns: power and sample size or alpha depending on the parameters selected in the dialog box. It helps building the simulation plot.

Simulation plot: This plot shows the evolution of the parameters as defined in the graphics tab of the dialog box.

Example

An example of power calculation based on a test is available on the Addinsoft website at http://www.xlstat.com/demo-pwr.htm

An example of calculating the required sample size is available on the Addinsoft website at
References


Compare proportions (XLSTAT-Power)

Use this tool to compute power and sample size in a statistical test comparing proportions.

Description

XLSTAT-Pro includes parametric tests and nonparametric tests to compare proportions. Thus we can use the z-test, chi-square test, the sign test or the McNemar test. XLSTAT-Power can calculate the power or the number of observations necessary for these tests using either exact methods or approximations.

When testing a hypothesis using a statistical test, there are several decisions to take:

- The null hypothesis $H_0$ and the alternative hypothesis $H_a$.
- The statistical test to use.
- The type I error also known as alpha. It occurs when one rejects the null hypothesis when it is true. It is set a priori for each test and is 5%.

The type II error or beta is less studied but is of great importance. In fact, it represents the probability that one does not reject the null hypothesis when it is false. We can not fix it upfront, but based on other parameters of the model we can try to minimize it. The power of a test is calculated as $1 - \beta$ and represents the probability that we reject the null hypothesis when it is false.

We therefore wish to maximize the power of the test. The XLSTAT-Power module calculates the power (and beta) when other parameters are known. For a given power, it also allows to calculate the sample size that is necessary to reach that power.

The statistical power calculations are usually done before the experiment is conducted. The main application of power calculations is to estimate the number of observations necessary to properly conduct an experiment.

XLSTAT allows to compare:

- A proportion to a test proportion (z-test with different approximations).
- Two proportions (z-test with different approximations).
- Proportions in a contingency table (chi-square test).
- Proportions in a nonparametric way (the sign test and the McNemar test)

For each case, different input parameters are used and shown in the dialog box.
Methods

The sections of this document dedicated to the tests on proportions describe in detail the methods themselves.

The power of a test is usually obtained by using the associated non-central distribution. For this specific case we will use an approximation in order to compute the power.

Comparing a proportion to a test proportion

The alternative hypothesis in this case is: \( H_a: p_1 - p_0 \neq 0 \)

Various approximations are possible:

- Approximation using the normal distribution: In this case, we will use the normal distribution with means \( p_0 \) and \( p_1 \) and standard deviations \( \sqrt{\frac{p_0(1-p_0)}{N}} \) and \( \sqrt{\frac{p_1(1-p_1)}{N}} \).

- Exact calculation using the binomial distribution with parameters \( \sqrt{\frac{p_0(1-p_0)}{N}} \) and \( \sqrt{\frac{p_1(1-p_1)}{N}} \).

- Approximation using the beta distribution with parameters \( \left( (N-1)p_0, (N-1)(1-p_0) \right) \) and \( \left( (N-1)p_1, (N-1)(1-p_1) \right) \).

- Approximation using the method of the arcsin: This approximation is based on the arcsin transformation of proportions: \( H(p_0) \) and \( H(p_1) \). The power is obtained using the normal distribution: \( Z_p = \sqrt{N} \left( H(p_0) - H(p_1) \right) - Z_{req} \) with \( Z_{req} \) being the alpha-quantile of the normal distribution.

Comparing two proportions

The alternative hypothesis in this case is: \( H_a: p_1 - p_2 \neq 0 \)

Various approximations are possible:

- Approximation using the method of the arcsin: This approximation is based on the arcsin transformation of proportions: \( H(p_2) \) and \( H(p_1) \). The power is obtained using the normal distribution: \( Z_p = \sqrt{N} \left( H(p_2) - H(p_1) \right) - Z_{req} \) with \( Z_{req} \) being the alpha-quantile of the normal distribution.
- Approximation using the normal distribution: In this case, we will use the normal distribution with means \( p_1 \) and \( p_2 \) and standard deviations: \( \sqrt{\frac{p_1 (1 - p_1)}{N}} \) and \( \sqrt{\frac{p_2 (1 - p_2)}{N}} \).

Chi-square test

To calculate the power of the chi-square test in the case of a contingency table 2 * 2, we use the non-central chi-square distribution with the value of the chi-square as non-centrality parameter.

It therefore seeks to see whether two groups of observations have the same behavior based on a binary variable.

We have:

<table>
<thead>
<tr>
<th></th>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>( p_1 )</td>
<td>( p_2 )</td>
</tr>
<tr>
<td>Negative</td>
<td>( 1-p_1 )</td>
<td>( 1-p_2 )</td>
</tr>
</tbody>
</table>

\( p_1, N_1 \) and \( N_2 \) have to be entered in the dialog box (\( p_2 \) can be found from other parameters because the test has only one degree of freedom).

Sign test

The sign test is used to see if the proportion of cases in each group is equal to 50%. It has the same principle as the one proportion test against a constant. The constant being 0.5. Power is computed using an approximation by the normal distribution or an exact method with the binomial distribution.

We must therefore enter the sample size and the proportion in one group \( p_1 \) (the other proportion is such that \( p_2=1-p_1 \)).

McNemar test
The McNemar test on paired proportions is a specific case of testing a proportion against a constant. Indeed, one can represent the problem with the following table:

<table>
<thead>
<tr>
<th></th>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>PP</td>
<td>PN</td>
</tr>
<tr>
<td>Negative</td>
<td>P</td>
<td>NN</td>
</tr>
</tbody>
</table>

We have $PP + NN + PN + NP = 1$. We want to try to see the effect of a treatment; we are therefore interested in $NP$ and $PN$. The other values are not significant.

The test inputs are: $Proportion1 = NP$ and $Proportion2 = PN$. With necessarily $P1+P2<1$.

The effect is calculated only on a percentage of $NP + PN$ of the sample. The proportion of individuals that change from negative to positive is calculated as $NP / (NP + NP)$. So we will try to compare this figure to a value of 50% to see if we have more individuals who go from positive to negative than individuals who go from negative to positive.

This test is well suited for medical applications.

**Calculating sample size**

To calculate the number of observations required, XLSTAT uses an algorithm that searches for the root of a function. It is called the Van Wijngaarden-Dekker-Brent algorithm (Brent, 1973). This algorithm is adapted to the case where the derivatives of the function are not known. It tries to find the root of:

$$\text{power (N)} - \text{expected\_power}$$

We then obtain the size $N$ such that the test has a power as close as possible to the desired power.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.
Click this button to start the computations.

Click this button to close the dialog box without doing any computation.

Click this button to display the help.

Click this button to reload the default options.

Click this button to delete the data selections.

**General** tab:

**Goal**: Choose between computing power and sample size estimation.

**Statistical test**: Select the test you want to apply.

**Alternative hypothesis**: Select the alternative hypothesis to be tested.

**Theoretical mean** (when only one sample is used): Enter the value of the theoretical mean to be tested.

**Alpha**: Enter the value of the type I error (alpha, between 0.001 and 0.999).

**Power** (when sample size estimation has been selected): Enter the value of the power to be reached.

**Sample size (group 1)** (when power computation has been selected): Enter the size of the first sample.

**Sample size (group 2)** (when power computation has been selected): Enter the size of the second sample.

**N1/N2 ratio** (when sample size has been selected and when there are two samples): Enter the ratio between the sizes of the first and the second samples.

**Proportion 1**: Enter the proportion for group 1.

**Proportion 2**: Enter the proportion for group 2.

**Range**: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.
Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Graphics tab:

Simulation plot: Activate this option if you want to plot different parameters of the test. Two parameters can vary. All remaining parameters are used as they were defined in the General tab.

Y axis: Select the parameter to be used on the Y axis of the simulation plot. You can either choose the power or the sample size.

X axis: Select the parameter to be used on the X axis of the simulation plot. You can either choose the power or the sample size, the type I error (alpha) or the effect size.

Interval size: Enter the minimum, maximum and interval size for the X axis of the simulation plot.

Results

Results: This table displays the parameters of the test and the power or the required number of observations. The parameters obtained by the calculation are in bold format. An explanation is displayed below this table.

Intervals for the simulation plot: This table is composed of two columns: power and sample size or alpha depending on the parameters selected in the dialog box. It helps building the simulation plot.

Simulation plot: This plot shows the evolution of the parameters as defined in the graphics tab of the dialog box.

Example

An example of power calculation based on a test is available on the Addinsoft website at

http://www.xlstat.com/demo-pwr.htm

An example of calculating the required sample size is available on the Addinsoft website at
References


Compare correlations (XLSTAT-Power)

Use this tool to compute power and sample size in a statistical test comparing Pearson correlations.

Description

XLSTAT-Pro offers a test to compare correlations. XLSTAT-Power can calculate the power or the number of observations necessary for this test.

When testing a hypothesis using a statistical test, there are several decisions to take:

- The null hypothesis H0 and the alternative hypothesis Ha.
- The statistical test to use.
- The type I error also known as alpha. It occurs when one rejects the null hypothesis when it is true. It is set a priori for each test and is 5%.

The type II error or beta is less studied but is of great importance. In fact, it represents the probability that one does not reject the null hypothesis when it is false. We can not fix it upfront, but based on other parameters of the model we can try to minimize it. The power of a test is calculated as 1-beta and represents the probability that we reject the null hypothesis when it is false.

We therefore wish to maximize the power of the test. The XLSTAT-Power module calculates the power (and beta) when other parameters are known. For a given power, it also allows to calculate the sample size that is necessary to reach that power.

The statistical power calculations are usually done before the experiment is conducted. The main application of power calculations is to estimate the number of observations necessary to properly conduct an experiment.

XLSTAT allows to compare:

- One correlation to 0.
- One correlation to a constant.
- Two correlations.

Methods
The section of this document dedicated to the correlation tests describes in detail the methods themselves.

The power of a test is usually obtained by using the associated non-central distribution. For this specific case we will use an approximation in order to compute the power.

Comparing on correlation to 0

The alternative hypothesis in this case is: $H_a: r \neq 0$

The method used is an exact method based on the non-central Student distribution.

The non-centrality parameter used is the following: $NCP = \sqrt{\frac{r^2}{1-r^2}} \cdot \sqrt{N}$

The part $\frac{r^2}{1-r^2}$ is called effect size.

Comparing one correlation to a constant

The alternative hypothesis in this case is: $H_a: r \neq r_0$

The power calculation is done using an approximation by the normal distribution. We use the Fisher Z-transformation:

$$Z_r = \frac{1}{2} \log \left( \frac{1+r}{1-r} \right)$$

The effect size is: $Q = |Z_r - Z_{r_0}|$

The power is then found using the area under the curve of the normal distribution to the left of $Z_p$:

$$Z_p = Q \cdot \sqrt{N-3} - Z_{req}$$

where $z_{req}$ is the quantile of the normal distribution for alpha.

Comparing two correlations

The alternative hypothesis in this case is: $H_a: r_1 - r_2 \neq 0$

The power calculation is done using an approximation by the normal distribution. We use the Fisher Z-transformation:

$$Z_r = \frac{1}{2} \log \left( \frac{1+r}{1-r} \right)$$

The effect size is: $Q = |Z_{r_1} - Z_{r_2}|$
The power is then found using the area under the curve of the normal distribution to the left of $Z_p$: $Z_p = Q \cdot \sqrt{\frac{N' - 3}{2}} - Z_{req}$ where $Z_{req}$ is the quantile of the normal distribution for alpha and

$$N' = \frac{2(N_1 - 3)(N_2 - 3)}{N_1 + N_2 - 6} + 3.$$

**Calculating sample size**

To calculate the number of observations required, XLSTAT uses an algorithm that searches for the root of a function. It is called the Van Wijngaarden-Dekker-Brent algorithm (Brent, 1973). This algorithm is adapted to the case where the derivatives of the function are not known. It tries to find the root of:

$$\text{power (N)} - \text{expected_power}$$

We then obtain the size $N$ such that the test has a power as close as possible to the desired power.

**Effect size**

This concept is very important in power calculations. Indeed, Cohen (1988) developed this concept. The effect size is a quantity that will allow to calculate the power of a test without entering any parameters but will tell if the effect to be tested is weak or strong.

In the context of comparisons of correlations conventions of magnitude of the effect size are:

- $Q=0.1$, the effect is small.
- $Q=0.3$, the effect is moderate.
- $Q=0.5$, the effect is strong.

XLSTAT-Power allows to enter directly the effect size

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.
**General** tab:

**Goal**: Choose between computing power and sample size estimation.

**Statistical test**: Select the test you want to apply.

**Alternative hypothesis**: Select the alternative hypothesis to be tested.

**Theoretical mean** (when only one sample is used): Enter the value of the theoretical mean to be tested.

**Alpha**: Enter the value of the type I error (alpha, between 0.001 and 0.999).

**Power** (when sample size estimation has been selected): Enter the value of the power to be reached.

**Sample size (group 1)** (when power computation has been selected): Enter the size of the first sample.

**Sample size (group 2)** (when power computation has been selected): Enter the size of the second sample.

**N1/N2 ratio** (when sample size has been selected and when there are two samples): Enter the ratio between the sizes of the first and the second samples.

**Parameters**: Select this option to enter the test parameters directly.

**Effect size**: Select this option to directly enter the effect size D (see the description part of this help).

**Correlation (group 1)**: Enter the correlation for group 1.

**Correlation (group 2)**: Enter the correlation for group 2.
Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Graphics tab:

Simulation plot: Activate this option if you want to plot different parameters of the test. Two parameters can vary. All remaining parameters are used as they were defined in the General tab.

Y axis: Select the parameter to be used on the Y axis of the simulation plot. You can either choose the power or the sample size.

X axis: Select the parameter to be used on the X axis of the simulation plot. You can either choose the power or the sample size, the type I error (alpha) or the effect size.

Interval size: Enter the minimum, maximum and interval size for the X axis of the simulation plot.

Results

Results: This table displays the parameters of the test and the power or the required number of observations. The parameters obtained by the calculation are in bold format. An explanation is displayed below this table.

Intervals for the simulation plot: This table is composed of two columns: power and sample size or alpha depending on the parameters selected in the dialog box. It helps building the simulation plot.

Simulation plot: This plot shows the evolution of the parameters as defined in the graphics tab of the dialog box.

Example

An example of power calculation based on a test is available on the Addinsoft website at

http://www.xlstat.com/demo-pwr.htm
An example of calculating the required sample size is available on the Addinsoft website at

http://www.xlstat.com/demo-spl.htm

References


Linear regression (XLSTAT-Power)

Use this tool to compute power and necessary sample size in linear regression model.

Description

XLSTAT-Pro offers a tool to apply a linear regression model. XLSTAT-Power estimates the power or calculates the necessary number of observations associated with variations of $R^2$ in the framework of a linear regression.

When testing a hypothesis using a statistical test, there are several decisions to take:

- The null hypothesis $H_0$ and the alternative hypothesis $H_a$.
- The statistical test to use.
- The type I error also known as alpha. It occurs when one rejects the null hypothesis when it is true. It is set a priori for each test and is 5%.

The type II error or beta is less studied but is of great importance. In fact, it represents the probability that one does not reject the null hypothesis when it is false. We can not fix it upfront, but based on other parameters of the model we can try to minimize it. The power of a test is calculated as 1-beta and represents the probability that we reject the null hypothesis when it is false.

We therefore wish to maximize the power of the test. The XLSTAT-Power module calculates the power (and beta) when other parameters are known. For a given power, it also allows to calculate the sample size that is necessary to reach that power.

The statistical power calculations are usually done before the experiment is conducted. The main application of power calculations is to estimate the number of observations necessary to properly conduct an experiment.

XLSTAT allows to compare:

- $R^2$ value to 0.
- Increase in $R^2$ value when new predictors are added to the model to 0.

It means testing the following hypothesis:

- $H_0$: $R^2$ is equal to 0 / $H_a$: $R^2$ is different from 0
- $H_0$: Increase in $R^2$ is equal to 0 / $H_a$: Increase in $R^2$ is different from 0.
**Effect size**

This concept is very important in power calculations. Indeed, Cohen (1988) developed this concept. The effect size is a quantity that will allow to calculate the power of a test without entering any parameters but will tell if the effect to be tested is weak or strong.

In the context of a linear regression, conventions of magnitude of the effect size are:

- $f^2=0.02$, the effect is small.
- $f^2=0.15$, the effect is moderate.
- $f^2=0.35$, the effect is strong.

XLSTAT-Power allows to enter directly the effect size but also allows to enter parameters of the model that will help calculating the effect size. We detail the calculations below:

- Using variances: We can use the variances of the model to define the size of the effect. With $\text{Var}_{\text{expl}}$ being the variance explained by the explanatory variables that we wish to test and $\text{Var}_{\text{Err}}$ being the variance of the error or residual variance, we have:

  \[
  f^2 = \frac{\text{var}_{\text{expl}}}{\text{var}_{\text{Err}}}
  \]

- Using the $R^2$ (in the case $H_0: R^2=0$): We enter the estimated square multiple correlation value (called rho²) to define the size of the effect. We have:

  \[
  f^2 = \frac{\rho^2}{1 - \rho^2}
  \]

- Using the partial $R^2$ (in the case $H_0$: Increase in $R^2=0$): We enter the partial $R^2$ that is the expected difference in $R^2$ when adding predictors to the model to define the size of the effect. We have:

  \[
  f^2 = \frac{R_{\text{part}}^2}{1 - R_{\text{part}}^2}
  \]

- Using the correlations between predictors (in the case $H_0$: $R^2=0$): One must then select a vector containing the correlations between the explanatory variables and the dependent variable $\text{CorrY}$, and a square matrix containing the correlations between the explanatory variables $\text{CorrX}$. We have:

  \[
  f^2 = \frac{\text{Corr}_{\text{Y}}^T (\text{Corr}_{\text{X}})^{-1} \text{Corr}_{\text{Y}}}{1 - \text{Corr}_{\text{Y}}^T (\text{Corr}_{\text{X}})^{-1} \text{Corr}_{\text{Y}}}
  \]

Once the effect size is defined, power and necessary sample size can be computed.

**Methods**

The section of this document dedicated to the linear regression describes in detail the method.
The power of a test is usually obtained by using the associated non-central distribution. For this specific case we will use the Fisher non-central distribution to compute the power.

The power of this test is obtained using the non-central Fisher distribution with degrees of freedom equal to: DF1 is the number of tested variables; DF2 is the sample size from which the total number of explanatory variables included in model plus one is subtracted and the non-centrality parameter is: \( NCP = f^2 N \)

**Calculating sample size**

To calculate the number of observations required, XLSTAT uses an algorithm that searches for the root of a function. It is called the Van Wijngaarden-Dekker-Brent algorithm (Brent, 1973). This algorithm is adapted to the case where the derivatives of the function are not known. It tries to find the root of:

\[ \text{power (N)} - \text{expected_power} \]

We then obtain the size N such that the test has a power as close as possible to the desired power.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload**: Click this button to reload the default options.
- **Delete Data**: Click this button to delete the data selections.
**General** tab:

**Goal:** Choose between computing power and sample size estimation.

**Statistical test:** Select the test you want to apply.

**Alpha:** Enter the value of the type I error (alpha, between 0.001 and 0.999).

**Power** (when sample size estimation has been selected): Enter the value of the power to be reached.

**Sample size** (when power computation has been selected): Enter the size of the first sample.

**Number of tested predictors:** Enter the number of predictors to be tested.

**Total number of predictors** (when testing H0: Increase in R²=0): Enter the total number of predictors included in the model.

**Determine effect size:** Select the way effect size is computed.

**Effect size f²** (when effect size is entered directly): Enter the effect size (see the description part of the help for more details).

**Explained variance** (when effect size is computed from variances): Enter the explained variance by the tested predictors.

**Error variance** (when effect size is computed from variances): Enter the residual variance of the global model.

**Partial R²** (when effect size is computed using the direct approach): Enter the expected increase in R² when new covariates are added to the model.

**rho²** (when effect size is computed using the R²): Enter the expected theoretical value of the R².

**Range:** Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Check this option to display the results in a new worksheet in the active workbook.

**Workbook:** Check this option to display the results in a new workbook.

**Correlations** tab:

This tab appears when the hypothesis to be tested is H0: R²=0 and when effect size is computed with the correlations between predictors.
Correlations with Ys: Select a column corresponding to the correlations between the predictors and the response variable Y. This vector must have a number of lines equal to the number of explanatory variables. Do not select the text of the column but only the numerical values.

Correlations between predictors: Select a Table corresponding to the correlations between the explanatory variables. This table should be symmetrical, have 1 on the diagonal and have a number of rows and columns equal to the number of explanatory variables. Do not select the labels of the columns or of the rows, but only the numerical values.

Graphics tab:

Simulation plot: Activate this option if you want to plot different parameters of the test. Two parameters can vary. All remaining parameters are used as they were defined in the General tab.

Y axis: Select the parameter to be used on the Y axis of the simulation plot. You can either choose the power or the sample size.

X axis: Select the parameter to be used on the X axis of the simulation plot. You can either choose the power or the sample size, the type I error (alpha) or the effect size.

Interval size: Enter the minimum, maximum and interval size for the X axis of the simulation plot.

Results

Inputs: This table displays the parameters used to compute effect size.

Results: This table displays the alpha, the effect size and the power or the required number of observations. The parameters obtained by the calculation are in bold format. An explanation is displayed below this table.

Intervals for the simulation plot: This table is composed of two columns: power and sample size or alpha depending on the parameters selected in the dialog box. It helps building the simulation plot.

Simulation plot: This plot shows the evolution of the parameters as defined in the graphics tab of the dialog box.
Example

An example of power calculation based on a test is available on the Addinsoft website at

http://www.xlstat.com/demo-pwr.htm

An example of calculating the required sample size is available on the Addinsoft website at

http://www.xlstat.com/demo-spl.htm

References


ANOVA/ANCOVA (XLSTAT-Power)

Use this tool to compute power and necessary sample size in analysis of variance, repeated measures analysis of variance or analysis of covariance model.

Description

XLSTAT-Pro offers tools to apply analysis of variance (ANOVA), repeated measures analysis of variance and analysis of covariance (ANCOVA). XLSTAT-Power estimates the power or calculates the necessary number of observations associated with these models.

When testing a hypothesis using a statistical test, there are several decisions to take:

- The null hypothesis H0 and the alternative hypothesis Ha.
- The statistical test to use.
The type I error also known as alpha. It occurs when one rejects the null hypothesis when it is true. It is set a priori for each test and is 5%.

The type II error or beta is less studied but is of great importance. In fact, it represents the probability that one does not reject the null hypothesis when it is false. We can not fix it upfront, but based on other parameters of the model we can try to minimize it. The power of a test is calculated as 1-beta and represents the probability that we reject the null hypothesis when it is false.

We therefore wish to maximize the power of the test. The XLSTAT-Power module calculates the power (and beta) when other parameters are known. For a given power, it also allows to calculate the sample size that is necessary to reach that power.

The statistical power calculations are usually done before the experiment is conducted. The main application of power calculations is to estimate the number of observations necessary to properly conduct an experiment.

XLSTAT can therefore test:
- In the case of a one-way ANOVA or more fixed factors and interactions, as well as in the case of ANCOVA:
  - $H_0$: The means of the groups of the tested factor are equal.
  - $H_a$: At least one of the means is different from another.
- In the case of repeated measures ANOVA for a within-subjects factor:
  - $H_0$: The means of the groups of the within subjects factor are equal.
  - $H_a$: At least one of the means is different from another.
- In the case of repeated measures ANOVA for a between-subjects factor:
  - $H_0$: Les The means of the groups of the between subjects factor are equal.
  - $H_a$: At least one of the means is different from another.
- In the case of repeated measures ANOVA for an interaction between a within-subjects factor and a between-subjects factor:
  - $H_0$: The means of the groups of the within-between subjects interaction are equal.
  - $H_a$: At least one of the means is different from another.

**Effect size**

This concept is very important in power calculations. Indeed, Cohen (1988) developed this concept. The effect size is a quantity that will allow to calculate the power of a test without entering any parameters but will tell if the effect to be tested is weak or strong.

In the context of an ANOVA-type model, conventions of magnitude of the effect size are:
- $f=0.1$, the effect is small.
- $f=0.25$, the effect is moderate.
- $f=0.4$, the effect is strong.
XLSTAT-Power allows to enter directly the effect size but also allows you to enter parameters of the model that will calculate the effect size. We detail the calculations below:

- Using variances: We can use the variances of the model to define the size of the effect. With \( \text{Var}_{\text{Exp}} \) being the variance explained by the explanatory factors that we wish to test and \( \text{Var}_{\text{Err}} \) being the variance of the error or residual variance, we have:

\[
 f = \sqrt{\frac{\text{Var}_{\text{Exp}}}{\text{Var}_{\text{Err}}}}
\]

- Using the direct approach: We enter the estimated value of \( \eta^2 \) which is the ratio between the explained variance by the studied factor and the total variance of the model. For more details on \( \eta^2 \), please refer to Cohen (1988, chap. 8.2). We have:

\[
 f = \sqrt{\frac{\eta^2}{1-\eta^2}}
\]

- Using the means of each group (in the case of one-way ANOVA or within subjects repeated measures ANOVA): We select a vector with the averages for each group. It is also possible to have groups of different sizes, in this case, you must also select a vector with different sizes (the standard option assumes that all groups have equal size). We have:

\[
 f = \sqrt{\frac{\sum_{i}(m_i - m)^2}{k \cdot \text{SD}_{\text{intra}}}}
\]

where \( m_i \) mean of group \( i \), \( m \) mean of the means, \( k \) number of groups and \( \text{SD}_{\text{intra}} \) within-group standard deviation.

- When an ANCOVA is performed, a term has to be added to the model in order to take into account the quantitative predictors. The effect size is then multiplied by \( \sqrt{\frac{1}{1-\rho^2}} \) where \( \rho^2 \) is the theoretical value of the square multiple correlation coefficient associated to the quantitative predictors.

Once the effect size is defined, power and necessary sample size can be computed.

**Methods**

The section of this document dedicated to the different methods describes in detail the methods themselves.

The power of a test is usually obtained by using the associated non-central distribution. For this specific case we will use the Fisher non-central distribution to compute the power.

We first introduce some notations:
- \( \text{NbGroup} \): Number of groups we wish to test.
- \( N \): sample size.
- NumeratorDF: Numerator degrees of freedom for the F distribution (see below for more details).
- NbRep: Number of repetition (measures) for repeated measures ANOVA.
- $\rho$: Correlation between measures for repeated measures ANOVA.
- $\varepsilon$: Geisser-Greenhouse non sphericity correction.
- NbPred: Number of predictors in an ANCOVA model.

For each method, we give the first and second degrees of freedom and the non-centrality parameter:

- One-way ANOVA:
  \[ DF1 = NbGroup - 1 \quad DF2 = N - NbGroup \quad NCP = f^2 N \]
- ANOVA with fixed effects and interactions:
  \[ DF1 = \text{NumeratorDF} \quad DF2 = N - NbGroup \quad NCP = f^2 N \]
- Repeated measures ANOVA within-subjects factor:
  \[ DF1 = NbRep - 1 \quad DF2 = (N - NbGroup)(NbRep - 1) \quad NCP = f^2 \frac{N \cdot NbRep}{1 - \rho} \varepsilon \]
- Repeated measures ANOVA between-subjects factor:
  \[ DF1 = NbGroup - 1 \quad DF2 = N - NbGroup \quad NCP = f^2 \frac{N \cdot NbRep}{1 + \rho(NbRep - 1)} \]
- Repeated measures ANOVA interaction between a within-subject factor and a between-subject factor:
  \[ DF1 = (NbRep - 1)(NbGroup - 1) \quad DF2 = (N - NbGroup)(NbRep - 1) \quad NCP = f^2 \frac{N \cdot NbRep}{1 - \rho} \varepsilon \]
- ANCOVA:
  \[ DF1 = \text{NumeratorDF} \quad DF2 = N - NbGroup - NbPred - 1 \quad NCP = f^2 N \]

**Calculating sample size**

To calculate the number of observations required, XLSTAT uses an algorithm that searches for the root of a function. It is called the Van Wijngaarden-Dekker-Brent algorithm (Brent, 1973). This algorithm is adapted to the case where the derivatives of the function are not known. It tries to find the root of:

\[ \text{power} (N) - \text{expected_power} \]

We then obtain the size $N$ such that the test has a power as close as possible to the desired power.
Numerator degrees of freedom

In the framework of an ANOVA with fixed factor and interactions or an ANCOVA; XLSTAT-Power proposes to enter the number of degrees of freedom for the numerator of the non-central F distribution. This is due to the fact that many different models can be tested and computing numerator degrees of freedom is a simple way to test all kind of models.

Practically, the numerator degrees of freedom is equal to the number of group associated to the factor minus one in the case of a fixed factor. When interactions are studied, it is equal to the product of the degrees of freedom associated to each factor included in the interaction.

Suppose we have a 3-factor model, A (2 groups), B (3 groups), C (3 groups), 3 second order interactions A*B, A*C and B*C and one third-order interaction A*B*C We have 3*3*2=18 groups.

To test the main effects A, we have: NbGroups=18 and NumeratorDF=(2-1)=1.

To test the interactions, eg A*B, we have NbGroups=18 and NumeratorDF=(2-1)(3-1)=2. If you wish to test the third order interaction (A*B*C), we have NbGroups=18 and NumeratorDF=(2-1)(3-1)(3-1)=4.

In the case of an ANCOVA, the calculations will be similar.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

Reload: Click this button to reload the default options.

Delete: Click this button to delete the data selections.
**General** tab:

**Goal:** Choose between computing power and sample size estimation.

**Statistical test:** Select the test you want to apply.

**Alpha:** Enter the value of the type I error (alpha, between 0.001 and 0.999).

**Power** (when sample size estimation has been selected): Enter the value of the power to be reached.

**Sample size** (when power computation has been selected): Enter the size of the first sample.

**Number of groups:** Enter the total number of groups included in the model.

**Number of tested predictors:** Enter the number of predictors to be tested.

**NumDF:** Enter the number of degrees of freedom associated to the tested factor (Number of groups -1 in the case of a first order factor). For more details, see the description part of this help.

**Correlation between measures:** Enter the correlation between measures for repeated measures ANOVA.

**Sphericity correction:** Enter the Geisser-Greenhouse epsilon for correction of non-sphericity for repeated measures ANOVA. If the hypothesis of sphericity is not rejected, then epsilon=1.

**Number of tested predictors:** Enter the number of predictors in the ANCOVA model.

**Determine effect size:** Select the way effect size is computed.

**Effect size f** (when effect size is entered directly): Enter the effect size (see the description part of the help for more details).

**Explained variance** (when effect size is computed from variances): Enter the explained variance by the tested factors.

**Error variance** (when effect size is computed from variances): Enter the residual variance of the global model.

**Within-group variance** (when effect size is computed from variances): Enter the within-group variance of the model.

**Partial eta²** (when effect size is computed using the direct approach): Enter the expected value of eta². For more details, see the description part of this help.

**Within-group standard deviation** (when effect size is computed using the means): Enter the expected within-group standard deviation of the model.
**Range**: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Check this option to display the results in a new worksheet in the active workbook.

**Workbook**: Check this option to display the results in a new workbook.

**Means** tab:

This tab appears when applying a one-way ANOVA or repeated measures ANOVA for a between-subject factor.

**Means**: Select a column corresponding to the means of the groups. This vector must have a number of lines equal to the number of measures (or repetition). Do not select the label of the column but only the numerical values.

**Unequal group size**: Activate this option if the groups have unequal sizes. When activated, select a vector corresponding to the group sizes. This vector must have a number of lines equal to the number of measures (or repetition). Do not select the label of the column but only the numerical values. This option cannot be reached when required sample size is estimated.

**Graphics** tab:

**Simulation plot**: Activate this option if you want to plot different parameters of the test. Two parameters can vary. All remaining parameters are used as they were defined in the General tab.

**Y axis**: Select the parameter to be used on the Y axis of the simulation plot. You can either choose the power or the sample size.

**X axis**: Select the parameter to be used on the X axis of the simulation plot. You can either choose the power or the sample size, the type I error (alpha) or the effect size.

**Interval size**: Enter the minimum, maximum and interval size for the X axis of the simulation plot.

**Results**

**Inputs**: This table displays the parameters used to compute effect size.

**Results**: This table displays the alpha, the effect size and the power or the required number of observations. The parameters obtained by the calculation are in bold format. An explanation is displayed below this table.
**Intervals for the simulation plot:** This table is composed of two columns: power and sample size or alpha depending on the parameters selected in the dialog box. It helps building the simulation plot.

**Simulation plot:** This plot shows the evolution of the parameters as defined in the graphics tab of the dialog box.

---

**Example**

An example of power calculation based on a test is available on the Addinsoft website at

http://www.xlstat.com/demo-pwr.htm

An example of calculating the required sample size is available on the Addinsoft website at

http://www.xlstat.com/demo-spl.htm

---

**References**


Logistic regression (XLSTAT-Power)

Use this tool to compute power and necessary sample size in a logistic regression model.

Description

XLSTAT-Pro offers a tool to apply logistic regression. XLSTAT-Power estimates the power or calculates the necessary number of observations associated with this model.

When testing a hypothesis using a statistical test, there are several decisions to take:

- The null hypothesis H0 and the alternative hypothesis Ha.
- The statistical test to use.
- The type I error also known as alpha. It occurs when one rejects the null hypothesis when it is true. It is set a priori for each test and is 5%.

The type II error or beta is less studied but is of great importance. In fact, it represents the probability that one does not reject the null hypothesis when it is false. We can not fix it upfront, but based on other parameters of the model we can try to minimize it. The power of a test is calculated as 1-beta and represents the probability that we reject the null hypothesis when it is false.

We therefore wish to maximize the power of the test. The XLSTAT-Power module calculates the power (and beta) when other parameters are known. For a given power, it also allows to calculate the sample size that is necessary to reach that power.

The statistical power calculations are usually done before the experiment is conducted. The main application of power calculations is to estimate the number of observations necessary to properly conduct an experiment.

In the general framework of logistic regression model, the goal is to explain and predict the probability P that an event appends (usually Y=1). P is equal to:

\[ P = \frac{\exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k)}{1 + \exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k)} \]

We have: \( \log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k \)

The test used in XLSTAT-Power is based on the null hypothesis that the \( \beta_1 \) coefficient is equal to 0. That means that the X1 explanatory variable has no effect on the model. For more details on logistic regression, please see the associated chapter of this help.
The hypothesis to be tested is:
- H₀ : β₁ = 0
- Ha : β₁ ≠ 0

Power is computed using an approximation which depends on the type of variable.

If X₁ is quantitative and has a normal distribution, the parameters of the approximation are:
- P₀ (baseline probability): The probability that Y=1 when all explanatory variables are set to their mean value.
- P₁ (alternative probability): The probability that X₁ be equal to one standard error above its mean value, all other explanatory variables being at their mean value.
- Odds ratio: The ratio between the probability that Y=1, when X₁ is equal to one standard deviation above its mean and the probability that Y=1 when X₁ is at its mean value.
- The R² obtained with a regression between X₁ and all the other explanatory variables included in the model.

If X₁ is binary and follow a binomial distribution. Parameters of the approximation are:
- P₀ (baseline probability): The probability that Y=1 when X₁=0.
- P₁ (alternative probability): The probability that Y=1 when X₁=1.
- Odds ratio: The ratio between the probability that Y=1, when X₁=1 and the probability that Y=1 when X₁=0.
- The R² obtained with a regression between X₁ and all the other explanatory variables included in the model.
- The percentage of observations with X₁=1.

These approximations depend on the normal distribution.

**Calculating sample size**

To calculate the number of observations required, XLSTAT uses an algorithm that searches for the root of a function. It is called the Van Wijngaarden-Dekker-Brent algorithm (Brent, 1973). This algorithm is adapted to the case where the derivatives of the function are not known. It tries to find the root of:

\[ \text{power (N)} - \text{expected_power} \]

We then obtain the size N such that the test has a power as close as possible to the desired power.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.
Click this button to start the computations.

Click this button to close the dialog box without doing any computation.

Click this button to display the help.

Click this button to reload the default options.

Click this button to delete the data selections.

**General** tab:

**Goal:** Choose between computing power and sample size estimation.

**Alpha:** Enter the value of the type I error (alpha, between 0.001 and 0.999).

**Power** (when sample size estimation has been selected): Enter the value of the power to be reached.

**Sample size** (when power computation has been selected): Enter the size of the first sample.

**Baseline probability (P0):** Enter the probability that Y=1 when all explanatory variables are at their mean value or are equal to 0 when binary.

**Determine effect size:** Select the way effect size is computed.

**Alternative probability (P1):** Enter the probability that Y=1 when X1 is equal to one standard deviation above its mean value or is equal to 0 when binary.

**Odds ratio:** Enter the odds ratio (see the description part of this help).

**R² of X1 with other Xs:** Enter the R² obtained with a regression between X1 and the other explanatory variables of the model.

**Type of variable:** Select the type of variable X1 to be analyzed (quantitative with normal distribution or binary).

**Percent of N with X1=1:** In the case of a binary X1, enter the percentage of observations with X1=1.
**Range:** Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Check this option to display the results in a new worksheet in the active workbook.

**Workbook:** Check this option to display the results in a new workbook.

**Graphics** tab:

**Simulation plot:** Activate this option if you want to plot different parameters of the test. Two parameters can vary. All remaining parameters are used as they were defined in the General tab.

**Y axis:** Select the parameter to be used on the Y axis of the simulation plot. You can either choose the power or the sample size.

**X axis:** Select the parameter to be used on the X axis of the simulation plot. You can either choose the power or the sample size, the type I error (alpha) or the effect size.

**Interval size:** Enter the minimum, maximum and interval size for the X axis of the simulation plot.

**Results**

**Inputs:** This table displays the parameters used to compute power and required sample size.

**Results:** This table displays the alpha and the power or the required number of observations. The parameters obtained by the calculation are in bold format. An explanation is displayed below this table.

**Intervals for the simulation plot:** This table is composed of two columns: power and sample size or alpha depending on the parameters selected in the dialog box. It helps building the simulation plot.

**Simulation plot:** This plot shows the evolution of the parameters as defined in the graphics tab of the dialog box.

**Example**

An example of power calculation based on a test is available on the Addinsoft website at

An example of calculating the required sample size is available on the Addinsoft website at

http://www.xlstat.com/demo-spl.htm

References


**Cox model (XLSTAT-Power)**

Use this tool to compute power and necessary sample size in a Cox proportional hazards ratio model to treat failure time data with covariates.

**Description**

XLSTAT-Life offers a tool to apply the proportional hazards ratio Cox regression model. XLSTAT-Power estimates the power or calculates the necessary number of observations associated with this model.

When testing a hypothesis using a statistical test, there are several decisions to take:

- The null hypothesis $H_0$ and the alternative hypothesis $H_a$.
- The statistical test to use.
- The type I error also known as alpha. It occurs when one rejects the null hypothesis when it is true. It is set a priori for each test and is 5%.

The type II error or beta is less studied but is of great importance. In fact, it represents the probability that one does not reject the null hypothesis when it is false. We can not fix it upfront, but based on other parameters of the model we can try to minimize it. The power of a test is calculated as 1-beta and represents the probability that we reject the null hypothesis when it is false.

We therefore wish to maximize the power of the test. The XLSTAT-Power module calculates the power (and beta) when other parameters are known. For a given power, it also allows to calculate the sample size that is necessary to reach that power.

The statistical power calculations are usually done before the experiment is conducted. The main application of power calculations is to estimate the number of observations necessary to properly conduct an experiment.

The Cox model is based on the hazard function which is the probability that an individual will experience an event (for example, death) within a small time interval, given that the individual has survived up to the beginning of the interval. It can therefore be interpreted as the risk of dying at time $t$. The hazard function (denoted by $\lambda(t,X)$) can be estimated using the following equation:

$$\lambda(t,X) = \lambda_0(t) \exp(\beta_1X_1 + \ldots + \beta_pX_p)$$

The first term depends only on time and the second one depends on $X$. We are only interested by the second term. If all $\beta_i$ are equal to zero then there is no hazard factor. The goal of the Cox model is to focus on the relations between the $\beta_i$s and the hazard function.
The test used in XLSTAT-Power is based on the null hypothesis that the \( \beta_1 \) coefficient is equal to 0. That means that the X1 covariate is not a hazard factor. For more details on Cox model, please see the associated chapter of this help.

The hypothesis to be tested is:
- \( H_0 : \beta_1 = 0 \)
- \( H_a : \beta_1 \neq 0 \)

Power is computed using an approximation which depends on the normal distribution. Other parameters used in this approximation are: the event rate, which is the proportion of uncensored individuals, the standard deviation of X1, the expected value of \( \beta_1 \) known as \( B(\log(\text{hazard ratio})) \) and the \( R^2 \) obtained with the regression between X1 and the other covariates included in the Cox model.

**Calculating sample size**

To calculate the number of observations required, XLSTAT uses an algorithm that searches for the root of a function. It is called the Van Wijngaarden-Dekker-Brent algorithm (Brent, 1973). This algorithm is adapted to the case where the derivatives of the function are not known. It tries to find the root of:

\[
\text{power (N) - expected_power}
\]

We then obtain the size N such that the test has a power as close as possible to the desired power.

**Calculating B**

The \( B(\log(\text{hazard ratio})) \) is an estimation of the coefficient \( \beta_1 \) of the following equation:

\[
\log \left( \frac{\lambda(t|X)}{\lambda_0(t)} \right) = \beta_1 X_1 + ... + \beta_k X_k
\]

\( \beta_1 \) is the change in logarithm of the hazard ratio when X1 is incremented of one unit (all other explanatory variables remaining constant). We can use the hazard ratio instead of the log. For a hazard ratio of 2, we will have \( B=\ln(2)=0.693 \).
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

**OK**: Click this button to start the computations.

**Cancel**: Click this button to close the dialog box without doing any computation.

**Help**: Click this button to display the help.

**Reload**: Click this button to reload the default options.

**Delete**: Click this button to delete the data selections.

**General** tab:

**Goal**: Choose between computing power and sample size estimation.

**Alpha**: Enter the value of the type I error (alpha, between 0.001 and 0.999).

**Power** (when sample size estimation has been selected): Enter the value of the power to be reached.

**Sample size** (when power computation has been selected): Enter the size of the first sample.

**Event rate**: Enter the event rate (uncensored units rate).

**B(log(Hazard ratio)**: Enter the estimation of the parameter B associated to X1 in the Cox model.

**Standard error of X1**: Enter the standard error of X1.

**R² of X1 with other Xs**: Enter the R² obtained with a regression between X1 and the other explanatory variables of the model.

**Range**: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Check this option to display the results in a new worksheet in the active workbook.

**Workbook**: Check this option to display the results in a new workbook.
Graphics tab:

Simulation plot: Activate this option if you want to plot different parameters of the test. Two parameters can vary. All remaining parameters are used as they were defined in the General tab.

Y axis: Select the parameter to be used on the Y axis of the simulation plot. You can either choose the power or the sample size.

X axis: Select the parameter to be used on the X axis of the simulation plot. You can either choose the power or the sample size, the type I error (alpha) or the effect size.

Interval size: Enter the minimum, maximum and interval size for the X axis of the simulation plot.

Results

Inputs: This table displays the parameters used to compute power and required sample size.

Results: This table displays the alpha and the power or the required number of observations. The parameters obtained by the calculation are in bold format. An explanation is displayed below this table.

Intervals for the simulation plot: This table is composed of two columns: power and sample size or alpha depending on the parameters selected in the dialog box. It helps building the simulation plot.

Simulation plot: This plot shows the evolution of the parameters as defined in the graphics tab of the dialog box.

Example

An example of power calculation based on a test is available on the Addinsoft website at http://www.xlstat.com/demo-pwr.htm

An example of calculating the required sample size is available on the Addinsoft website at http://www.xlstat.com/demo-spl.htm
References


Sample size for clinical trials (XLSTAT-Power)

Use this tool to compute sample size and power for different kind of clinical trials: equivalence trial, non-inferiority trial and superiority trial.

Description

XLSTAT-Power enables you to compute the necessary sample size for a clinical trial.

Three types of trials can be studied:

- **Equivalence trials**: An equivalence trial is where you want to demonstrate that a new treatment is no better or worse than an existing treatment.

- **Superiority trials**: A superiority trial is one where you want to demonstrate that one treatment is better than another.

- **Non-inferiority trials**: A non-inferiority trial is one where you want to show that a new treatment is not worse than an existing treatment.

These tests can be applied to a binary outcome or a continuous outcome.

When testing a hypothesis using a statistical test, there are several decisions to take:

- The null hypothesis $H_0$ and the alternative hypothesis $H_a$.
- The statistical test to use.
- The type I error also known as alpha. It occurs when one rejects the null hypothesis when it is true. It is set a priori for each test and is 5%.

The type II error or beta is less studied but is of great importance. In fact, it represents the probability that one does not reject the null hypothesis when it is false. We can not fix it upfront, but based on other parameters of the model we can try to minimize it. The power of a test is calculated as 1-beta and represents the probability that we reject the null hypothesis when it is false.

We therefore wish to maximize the power of the test. The XLSTAT-Power module calculates the power (and beta) when other parameters are known. For a given power, it also allows to calculate the sample size that is necessary to reach that power.

The statistical power calculations are usually done before the experiment is conducted. The main application of power calculations is to estimate the number of observations necessary to properly conduct an experiment.
Methods

The necessary sample size is obtained using simple approximation methods.

Equivalence test for a continuous outcome

The mean outcome is compared between two randomised groups. You must define a difference between these means, \( d \), within which you will accept that the two treatments being compared are equivalent.

The sample size is obtained using:

\[
n = \frac{f(\alpha, \beta/2) \cdot 2 \cdot \sigma^2}{(d)^2}
\]

With \( \sigma^2 \) being the variance of the outcome and:

\[
f(\alpha, \beta) = \left( \Phi^{-1}(\alpha) + \Phi^{-1}(\beta) \right)^2
\]

Equivalence test for a binary outcome

The percentage of patients that “survived” is compared between two randomised groups. You must define a difference between these percentages, \( d \), within which you will accept that the two treatments being compared are equivalent.

The sample size is obtained using:

\[
n = \frac{f(\alpha, \beta/2) \cdot (P(\text{std}) \cdot (100 - P(\text{std})))}{(P(\text{std}) - d)^2}
\]

With \( P(\text{std}) \) being the percentage for the treatments (we suppose these percentage are equivalent for both treatments), \( d \) is defined by the user and:

\[
f(\alpha, \beta) = \left( \Phi^{-1}(\alpha) + \Phi^{-1}(\beta) \right)^2
\]

Non-inferiority test for a continuous outcome

The mean outcome is compared between two randomised groups. The null hypothesis is that the experimental treatment is inferior to the standard treatment. The alternative hypothesis is that the experimental treatment is non-inferior to the standard treatment. You must choose the non-inferiority limit, \( d \), to be the largest difference that is clinically acceptable, so that a difference bigger than this would matter in practice.

The sample size is obtained using:
With sigma^2 being the variance of the outcome and:

\[ f(\alpha, \beta) = \left( \Phi^{-1}(\alpha) + \Phi^{-1}(\beta) \right)^2 \]

Nnon-inferiority test for a binary outcome

The percentage of patients that “survived” is compared between two randomised groups. The null hypothesis is that the percentage for those on the standard treatment is better than the percentage for those on the experimental treatment by an amount d. The alternative hypothesis is that the experimental treatment is better than the standard treatment or only slightly worse (by no more than d). The user must define the non-inferiority limit (d) so that a difference bigger than this would matter in practice. You should normally assume that the percentage 'success' in both standard and experimental treatment groups is the same.

The sample size is obtained using:

\[ n = \frac{f(\alpha, \beta) \cdot 2 \cdot \sigma^2}{(d)^2} \]

With P(std) being the percentage for the standard treatment and P(new) being the percentage for the new treatment, d is defined by the user and:

\[ f(\alpha, \beta) = \left( \Phi^{-1}(\alpha) + \Phi^{-1}(\beta) \right)^2 \]

Superiority test for a continuous outcome

The mean outcome is compared between two randomised groups. We wish to know if the mean associated to a new treatment is higher than the mean with the standard treatment.

The sample size is obtained using:

\[ n = \frac{f(\alpha, \beta) \cdot (P(std) \cdot (100 - P(std)) + P(new) \cdot (100 - P(new)))}{(P(std) - P(new) - d)^2} \]

With sigma^2 being the variance, mu1 and mu2 being the means associated to each group of the outcome and:

\[ f(\alpha, \beta) = \left( \Phi^{-1}(\alpha) + \Phi^{-1}(\beta) \right)^2 \]

When cross-over is present, a formula for adjusting the sample size is used:
Superiority test for a binary outcome

The percentage of patients that “survived” is compared between two randomised groups. We wish to know if the percentage associated to a new treatment is higher than the percentage with the standard treatment.

The sample size is obtained using:

\[
n = \frac{f(\alpha/2, \beta) \cdot (P(\text{std}) \cdot (100 - P(\text{std})) + P(\text{new}) \cdot (100 - P(\text{new})))}{(P(\text{std}) - P(\text{new}))^2}
\]

With \( P(\text{std}) \) being the percentage for the standard treatment and \( P(\text{new}) \) being the percentage for the new treatment and:

\[
f(\alpha, \beta) = \left( \Phi^{-1}(\alpha) + \Phi^{-1}(\beta) \right)^2
\]

When cross-over is present, a formula for adjusting the sample size is used:

\[
n_{\text{adjusted}} = \frac{n \times 10'000}{(100 - c_1 - c_2)}
\]

With \( c_1 \) and \( c_2 \) being the cross-over percentage in each group.

Calculating power

To calculate the power for a fixed sample size, XLSTAT uses an algorithm that searches the beta (1-power) so that:

\[
\text{Sample size (beta) – expected sample size} = 0
\]

We then obtain the power (1-beta) such that the test needs a sample size as close as possible to the desired sample size.
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload**: Click this button to reload the default options.
- **Delete**: Click this button to delete the data selections.

**General** tab:

**Goal**: Choose between computing power and sample size estimation.

**Clinical trial**: Select the type of clinical trial: Equivalence, non-inferiority or superiority trials.

**Outcome variable**: Select the type of outcome variable (continuous or binary).

**Alpha**: Enter the value of the type I error (alpha, between 0.001 and 0.999).

**Power** (when sample size estimation has been selected): Enter the value of the power to be reached.

**Sample size** (when power computation has been selected): Enter the size of the total trial.

The available options will differ with respect to the chosen trial:

**Equivalence trial with continuous outcome**

**Std deviation**: Enter the standard deviation of the outcome.

**Equivalence limit d**: Enter the equivalence limit d.

**Equivalence trial with binary outcome**

**% of success for both groups**: Enter the % of success for both groups.
**Equivalence limit d**: Enter the equivalence limit d.

Non inferiority trial with continuous outcome

**Std deviation**: Enter the standard deviation of the outcome.

**Non inferiority limit d**: Enter the non inferiority limit d.

Non inferiority trial with binary outcome

**% of success for control group**: Enter the % of success for the control group.

**% of success for treatment group**: Enter the % of success for the treatment group.

**Non inferiority limit d**: Enter the non inferiority limit d.

Superiority trial with continuous outcome

**Mean for control group**: Enter the mean for the control group.

**Mean for treatment group**: Enter the mean for the treatment group.

**Std deviation**: Enter the standard deviation of the outcome.

**% cross over for control group**: Enter the percentage of cross-over for the control group.

**% cross over for treatment group**: Enter the percentage of cross-over for the treatment group.

Superiority trial with binary outcome

**% of success for control group**: Enter the % of success for the control group.

**% of success for treatment group**: Enter the % of success for the treatment group.

**% cross over for control group**: Enter the percentage of cross-over for the control group.

**% cross over for treatment group**: Enter the percentage of cross-over for the treatment group.

**Range**: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.
Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Graphics tab:

Simulation plot: Activate this option if you want to plot different parameters of the test. Two parameters can vary. All remaining parameters are used as they were defined in the General tab.

Y axis: Select the parameter to be used on the Y axis of the simulation plot.

X axis: Select the parameter to be used on the X axis of the simulation plot.

Interval size: Enter the minimum, maximum and interval size for the X axis of the simulation plot.

Results

Results: This table displays the parameters of the test and the power or the required number of observations. The parameters obtained by the calculation are in bold format. An explanation is displayed below this table.

Intervals for the simulation plot: This table is composed of two columns: power and sample size or alpha depending on the parameters selected in the dialog box. It helps building the simulation plot.

Simulation plot: This plot shows the evolution of the parameters as defined in the graphics tab of the dialog box.

Example

An example of calculating the required sample size for clinical trials is available on the Addinsoft website at

http://www.xlstat.com/demo-spltrial.htm
References


Subgroup Charts

Use this tool to supervise production quality, in the case where you have a group of measurements for each point in time. The measurements need to be quantitative data. This tool is useful to recap the mean and the variability of the measured production quality.

Integrated in this tool, you will find Box-Cox transformations, calculation of process capability and the application of rules for special causes and Westgard rules (an alternative set of rules to identify special causes) to complete your analysis.

Description

Control charts were first mentioned in a document by Walter Shewhart that he wrote during his time working at Bell Labs in 1924. He described his methods completely in his book (1931).

For a long time, there was no significant innovation in the area of control charts. With the development of CUSUM, UWMA and EWMA charts in 1936, Deming expanded the set of available control charts.

Control charts were originally used in area of goods production. Therefore the wording is still from that domain. Today this approach is being applied to a large number of different fields, for instance services, human resources, and sales. In the following chapters we will use the wording from the production and shop floors.

Subgroup charts

The subgroup charts tool offers you the following chart types alone or in combination:

- $\overline{X}$ (X bar)
- R
- S
- $S^2$

An X bar chart is useful to follow the mean of a production process. Mean shifts are easily visible in the diagrams.

An R chart (Range chart) is useful to analyze the variability of the production. A large difference in production, caused for example by the use of different production lines, will be easily visible.
S and S² charts are also used to analyze the variability of production. The S chart draws the standard deviation of the process and the S² chart draws the variance (which is the square of the standard deviation).

Note 1: If you want to investigate smaller mean shifts, then you can also use CUSUM group charts which are, by the way, often preferred to subgroup control charts.

Note 2: If you have only one measurement for each point in time, then please use the control charts for individuals.

Note 3: If you have measurements in qualitative values (for instance ok, not ok, conform not conform), then use the control charts for attributes.

This tool offers you the following options for the estimation of the standard deviation (sigma) of the data set, given k subgroups and nᵢ (i=1, …k) measurements per subgroup:

- Pooled standard deviation: sigma is computed using the k within-subgroup variances:

\[
\hat{s} = \sqrt{\frac{\sum_{i=1}^{k} (n_i - 1) s_i^2}{\sum_{i=1}^{k} (n_i - 1)}} / c_4 \left(1 + \sum_{i=1}^{k} (n_i - 1)\right)
\]

where \(c_4\) is the control chart constant according to Burr (1969).

- R bar: The estimator for sigma is calculated based on the average range of the k subgroups.

\[
\hat{s} = \frac{\bar{R}}{d_2}
\]

where \(d_2\) is the control chart constant according to Burr (1969).

- S bar: The estimator for sigma is calculated based on the average of the standard deviations of the k subgroups:

\[
\hat{s} = \sqrt{\frac{1}{k} \sum_{i=1}^{k} s_i^2} / c_4,
\]

where \(c_4\) is the control chart constant according to Burr (1969).

**Process capability**

Process capability describes a process and informs if the process is under control and if values taken by the measured variables are inside the specification limits of the process. In the latter case, one says that the process is “capable”.
During the interpretation of the different indicators for the process capability please pay attention to the fact that some indicators suppose normality or at least symmetry of the distribution of the measured values. By the use of a normality test, you can verify these premises (see the Normality Tests in XLSTAT-Pro).

If the data are not normally distributed, you have the following possibilities to obtain results for the process capabilities.

- Use the Box-Cox transformation to improve the normality of the data set. Then verify again the normality using a normality test.
- Use the process capability indicator $C_p$ 5.15.

Let $\hat{s}$ be the estimated standard deviation of the process, USL be the upper specification limit of the process, LSL be the lower specification limit of the process, and target be the selected target. XLSTAT allows to compute the following performance indicators to evaluate the process capability:

- **$C_p$:** The short term process capability is defined as:
  \[
  C_p = \frac{USL - LSL}{6\hat{s}}
  \]

- **$C_{pl}$:** The short term process capability with respect to the lower specification is defined as:
  \[
  C_{pl} = \frac{xbar - LSL}{3\hat{s}}
  \]

- **$C_{pu}$:** The short term process capability with respect to the upper specification is defined as:
  \[
  C_{pu} = \frac{USL - xbar}{3\hat{s}}
  \]

- **$C_{pk}$:** The short term process capability supposing a centered distribution is defined as:
  \[
  C_{pk} = \min(C_{pl}, C_{pu})
  \]

- **$P_p$:** The long term process capability is defined as:
  \[
  P_p = \frac{USL - LSL}{6 \sigma}
  \]

- **$P_{pl}$:** The long term process capability with respect to the lower specification is defined as:
  \[
  P_{pl} = \frac{xbar - LSL}{3 \sigma}
  \]

- **$P_{pu}$:** The long term process capability with respect to the upper specification is defined as:
Ppu = (USL – xbar) / (3 sigma)

- **Ppk**: The long term process capability supposing a centered distribution is defined as:
  
  \[ Ppk = \min(Ppl, Ppu) \]

- **Cpm**: The short term process capability according to Taguchi. This value can be calculated, if the target value has been specified. It is defined as:
  
  \[
  Cpm = \frac{\min(USL-target, target-LSL)}{3\sqrt{\hat{s}^2 + (\bar{X} - \text{target})^2}}
  \]

  where \( \hat{s} \) is the estimated standard deviation using the selected option for the estimation of sigma.

- **Cpm Boyle**: The short term process capability according to Taguchi improved by Boyle. This value can be calculated, if the target value has been specified. It is defined as:
  
  \[
  Cpm \text{ Boyle} = \frac{\min(USL, LSL) / 2}{3\sqrt{(n - 1) \hat{s}^2 / n + (\bar{X} - \text{target})^2}}
  \]

  where \( \hat{s} \) is the estimated standard deviation using the selected option for the estimation of sigma.

- **Cp 5.15**: The short term process capability is defined as:
  
  \[ Cp 5.15 = (USL - LSL) / (5.15 \hat{s}) \]

  where \( \hat{s} \) is the estimated standard deviation using the selected option for the estimation of sigma.

- **Cpk 5.15**: The short term process capability supposing a centered distribution is defined as:
  
  \[
  Cpk 5.15 = \left( \frac{d - |\bar{X} - (USL + LSL) / 2|}{2.575\hat{s}} \right)
  \]

  where \( d = (USL - LSL) / 2 \) and \( \hat{s} \) is the estimated standard deviation using the selected option for the estimation of sigma.

- **Cpmk**: The short term process capability according to Pearn. This value can be calculated, if the target value has been specified. It is defined as:
  
  \[
  Cpmk = \frac{(USL - LSL)^2}{2} - \frac{|\bar{X} - m|}{3\sqrt{\hat{s}^2 + (\bar{X} - \text{target})^2}}
  \]
where \( d = (\text{USL} + \text{LSL}) / 2 \) and \( \sigma \) is the estimated standard deviation using the selected option for the estimation of \( \sigma \).

- **Cs Wright**: The process capability according to Wright. This value can be calculated, if the target value has been specified. It is defined as:

\[
Cs_{\text{Wright}} = \frac{\min(\text{USL} - \bar{X}, \bar{X} - \text{LSL})}{3\sqrt{(n - 1) \frac{s^2}{n} + \frac{(\bar{X} - \text{target})^2}{c_4 s^2 b_3}}}
\]

where \( c_4 \) and \( b_3 \) are from the tables of SPC constants and \( \sigma \) is the estimated standard deviation using the selected option for the estimation of \( \sigma \).

- **Z below**: The amount of standard deviations between the mean and the lower specification limit is defined as:

\[
Z \text{ below} = (\bar{X} - \text{LSL}) / \sigma
\]

- **Z above**: The amount of standard deviations between the mean and the upper specification limit is defined as:

\[
Z \text{ above} = (\text{USL} - \bar{X}) / \sigma
\]

- **Z total**: The amount of standard deviations between the mean and the lower or upper respectively specification limit is defined as:

\[
Z \text{ total} = \Phi^{-1}(p \text{ (not conform) total})
\]

- **p(not conform) below**: The probability of producing a defect product below the lower specification limit is defined as:

\[
p\text{(not conform) below} = \Phi(Z \text{ below})
\]

- **p(not conform) above**: The probability of producing a defect product above the upper specification limit is defined as:

\[
p\text{(not conform) above} = \Phi(Z \text{ above})
\]

- **p(not conform) total**: The probability of producing a defect product below or above the specification limits is defined as:

\[
p\text{(not conform) total} = p\text{(not conform) below} + p\text{(not conform) above}
\]

- **PPM below**: The number of defect products below the lower specification limit during one million items produced is defined as:

\[
PPM \text{ below} = p\text{(not conform) below} \times 10^6
\]
• PPM above: The number of defect products above the upper specification limit during one million items produced is defined as:

$$\text{PPM above} = p(\text{not conform}) \text{ above} \times 10^6$$

• PPM total: The number of defect products below or above the specification limits during one million items produced is defined as:

$$\text{PPM total} = \text{PPM below} + \text{PPM above}$$

**Box-Cox transformation**

Box-Cox transformation is used to improve the normality of the time series; the Box-Cox transformation is defined by the following equation:

$$Y_t = \begin{cases} \frac{X_t^\lambda - 1}{\lambda}, & X_t \geq 0, \lambda > 0 \\ \ln(X_t), & X_t > 0, \lambda = 0 \end{cases}$$

Where the series \( \{X_t\} \) being transformed into series \( \{Y_t\} \), \( t=1,\ldots,n \):

Note: if \( \lambda < 0 \), the first equation is still valid, but \( X_t \) must be strictly positive. XLSTAT accepts a fixed value of \( \lambda \), or it can find the value that maximizes the likelihood value, the model being a simple linear model with the time as sole explanatory variable.

**Chart rules**

XLSTAT offers you the possibility to apply rules for special causes and Westgard rules. Two sets of rules are available in order to interpret control charts. You can activate and deactivate separately the rules in each set.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

Click this button to start the computations.
Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

Reload options: Click this button to reload the default options.

Delete data: Click this button to delete the data selections.

Mode tab:

Chart family: Select the family that you want to use:

- **Subgroup charts**: Activate the option if you have a data set with several measurements for each point in time.
- **Individual charts**: Activate this option if you have a data set with one quantitative measurement for each point in time.
- **Attribute charts**: Activate this option if you have a data set with one qualitative measurement for each point.
- **Time weighted**: Activate this option if you want to use a time weighted chart like UWMA, EWMA or CUSUM.

At this stage, the subgroup charts family should be selected. If not, you should switch to the help corresponding to the selected chart family. The options below correspond to the subgroup charts

Chart type: Select the type of chart you want to use:

- **X bar chart**: Activate this option if you want to calculate the X bar chart to analyze the mean of the process.
- **R chart**: Activate this option if you want to calculate the R chart to analyze variability of the process.
• **S chart:** Activate this option if you want to calculate the S chart to analyze variability of the process.

• **S² chart:** Activate this option if you want to calculate the S² chart to analyze variability of the process.

• **X bar R chart:** Activate this option if you want to calculate the X bar chart together with the R chart to analyze the mean value and variability of the process.

• **X bar S chart:** Activate this option if you want to calculate the X bar chart together with the S chart to analyze the mean value and variability of the process.

• **X bar S² chart:** Activate this option if you want to calculate the X bar chart together with the S² chart to analyze the mean value and variability of the process.

**General** tab:

**Data format:** Select the data format.

- **Columns/Rows:** Activate this option for XLSTAT to take each column (in column mode) or each row (in row mode) as a separate measurement that belongs to the same subgroup.

- **One column/row:** Activate this option if the measurements of the different subgroups are all on the same column (column mode) or one row (row mode). To assign the different measurements to their corresponding subgroup, please enter a constant group size or select a column or row with the group identifier in it.

**Data:** If the data format “One column/row” is selected, please choose the unique column or row that contains all the data. The assignment of the data to their corresponding subgroup must be specified using the Groups field or setting the common subgroup size. If you select the data “Columns/rows” option, please select a data area with one column/row per measurement in a subgroup.

**Groups:** If the data format “One column/row” is selected, then activate this Option to select a column/row that contains the group identifier. Select the data that identify for each element of the data selection the corresponding group.

**Common subgroup size:** If the data format “One column/row” is selected and the subgroup size is constant, then you can deactivate the groups option and enter in this field the common subgroup size.

**Phase:** Activate this option to supply one column/row with the phase identifier.

**Observation labels:** Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels” option is activated you need to include a header in
the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 ...).

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Column/Row labels**: Activate this option if the first row (column mode) or column (row mode) of the data selections contains a label.

**Options** tab:

**Upper control limit**:

- **Bound**: Activate this option, if you want to enter a maximum value to accept for the upper control limit of the process. This value will be used when the calculated upper control limit is greater than the value entered here.

- **Value**: Enter the upper control limit. This value will be used in place of the calculated upper control limit.

**Lower control limit**:

- **Bound**: Activate this option, if you want to enter a minimum value to accept for the lower control limit of the process. This value will be used when the calculated lower control limit is greater than the value entered here.

- **Value**: Enter the lower control limit. This value will be used and overrides the calculated upper control limit.

**Calculate process capabilities**: Activate this option to calculate process capabilities based on the input data (see the description section for more details).

- **USL**: If the calculation of the process capabilities is activated, please enter here the upper specification limit (USL) of the process.

- **LSL**: If the calculation of the process capabilities is activated, please enter here the lower specification limit (LSL) of the process.
• **Target**: If the calculation of the process capabilities is activated, activate this option to add the target value of the process.

• **Confidence interval (%)**: If the “Calculate process capabilities” option is activated, please enter the percentage range of the confidence interval to use for calculating the confidence interval around the process capabilities. Default value: 95.

**Box-Cox**: Activate this option to compute the Box-Cox transformation. You can either fix the value of the Lambda parameter, or decide to let XLSTAT optimize it (see the description section for further details).

**k Sigma**: Activate this option to enter the distance between the upper and the lower control limit and the center line of the control chart. The distance is fixed to k times the factor you enter multiplied by the estimated standard deviation. Corrective factors according to Burr (1969) will be applied.

**alpha**: Activate this option to define the size of the confidence range around the center line of the control chart. 100 - alpha % of the distribution of the control chart is inside the control limits. Corrective factors according to Burr (1969) will be applied.

**Mean**: Activate this option to enter a value for the center line of the control chart. This value should be based on historical data.

**Sigma**: Activate this option to enter a value for the standard deviation of the control chart. This value should be based on historical data. If this option is activated, then you cannot choose an estimation method for the standard deviation in the “Estimation” tab.

**Estimation** tab:

**Method for Sigma**: Select an option to determine the estimation method for the standard deviation of the control chart (see the description section for further details):

• **Pooled standard deviation**

• **R-bar**

• **S-bar**

**Outputs** tab:

**Display zones**: Activate this option to display beside the lower and upper control limit also the limits of the zones A and B.
Normality Tests: Activate this option to check normality of the data. (see the Normality Tests tool for further details).

Significance level (%): Enter the significance level for the tests.

Test special causes: Activate this option to analyze the points of the control chart according to the rules for special causes. You can activate the following rules independently:

- 1 point more than 3s from center line
- 9 points in a row on same side of center line
- 6 points in a row, all increasing or all decreasing
- 14 points in a row, alternating up and down
- 2 out of 3 points > 2s from center line (same side)
- 4 out of 5 points > 1s from center line (same side)
- 15 points in a row within 1s of center line (either side)
- 8 points in a row > 1s from center line (either side)

- **All**: Click this button to select all options.
- **None**: Click this button to deselect all options.

Apply Westgard rules: Activate this option to analyze the points of the control chart according to the Westgard rules. You can activate the following rules independently:

- Rule 1 2s
- Rule 1 3
- Rule 2 2s
- Rule 4s
- Rule 4 1s
- Rule 10 X

- **All**: Click this button to select all options.
- **None**: Click this button to deselect all options.
Charts tab:

Display charts: Activate this option to display the control charts graphically.
- Continuous line: Activate this option to connect the points in the control chart.
- Needles view: Activate this option to display for each point of the control chart, the minimum and maximum of the corresponding subgroup.
- Box view: Activate this option to display the control charts using bars.

Connect through missing: Activate this option to connect the points, even when missing values separate the points.

Normal Q-Q plots: Check this option to display Q-Q plots based on the normal distribution.

Display a distribution: Activate this option to compare histograms of samples selected with a density function.

Run Charts: Activate this option to display a chart of the latest data points. Each individual measurement is displayed.
- Number of observations: Enter the maximum number of the last observations to be displayed in the Run chart.

Results

Estimation:

Estimated mean: This table displays the estimated mean values for the different phases.

Estimated standard deviation: This table displays the estimated standard deviation values for the different phases.

Box-Cox transformation:

Estimates of the parameters of the model: This table is available only if the Lambda parameter has been optimized. It displays the estimator for Lambda.

Series before and after transformation: This table displays the series before and after transformation. If Lambda has been optimized, the transformed series corresponds to the residuals of the model. If it hasn’t then the transformed series is the direct application of the Box-Cox transformation.
**Process capabilities:**

**Process capabilities:** These tables are displayed, if the “process capability” option has been selected. There is one table for each phase. A table contains the following indicators for the process capability and if possible the corresponding confidence intervals: Cp, Cpl, Cpu, Cpk, Pp, Ppl, Ppu, Ppk, Cpm, Cpm (Boyle), Cp 5.5, Cpk 5.5, Cpmk, and Cs (Wright).

For Cp, Cpl, and Cpu, information about the process performance is supplied and for Cp a status information is given to facilitate the interpretation.

Cp values have the following status based on Ekvall and Juran (1974):

- "not adequate" if Cp < 1
- "adequate" if 1 <= Cp <= 1.33
- "more than adequate" if Cp > 1.33

Based on Montgomery (2001), Cp needs to have the following minimal values for the process performance to be as expected:

- 1.33 for existing processes
- 1.50 for new processes or for existing processes when the variable is critical
- 1.67 for new processes when the variable is critical

Based on Montgomery (2001), Cpu and Cpl need to have the following minimal values for process performance to be as expected:

- 1.25 for existing processes
- 1.45 for new processes or for existing processes when the variable is critical
- 1.60 for new processes when the variable is critical

**Capabilities:** This chart contains information about the specification and control limits. A line between the lower and upper limits represents the interval with an additional vertical mark for the center line. The different control limits of each phase are drawn separately.

**Chart information:**
The following results are displayed separately for each requested chart. Charts can be selected alone or in combination with the X bar chart.

**X bar/ R/ S/ S² chart:** This table contains information about the center line and the upper and lower control limits of the selected chart. There will be one column for each phase.

**Observation details:** This table displays detailed information for each subgroup. For each subgroup the corresponding phase, the size, the mean, the minimum and the maximum values, the center line, and the lower and upper control limits are displayed. If the information about the zones A, B and C are activated, then the lower and upper control limits of the zones A and B are displayed as well.

**Rule details:** If the rules options are activated, a detailed table about the rules will be displayed. For each subgroup, there is one row for each rule that applies. "Yes" indicates that the corresponding rule was fired for the corresponding subgroup and “No” indicates that the rule does not apply.

**X bar/ R/ S/ S² chart:** If the charts are activated, then a chart containing the information of the two tables above is displayed. Each subgroup is displayed. The center line and the lower and upper control limits are displayed as well. If the corresponding options have been activated, the lower and upper control limits for the zones A and B are included and there are labels for the subgroups for which rules were fired. A legend with the activated rules and the corresponding rule number is displayed below the chart.

**Normality tests:**

For each of the four tests, the statistics relating to the test are displayed including, in particular, the p-value which is afterwards used in interpreting the test by comparing with the chosen significance threshold.

If requested, a Q-Q plot is then displayed.

**Histograms:** The histograms are displayed. If desired, you can change the color of the lines, scales, titles as with any Excel chart.

**Run chart:** The chart of the last data points is displayed.

**Example**

A tutorial explaining how to use the SPC subgroup charts tool is available on the Addinsoft web site. To consult the tutorial, please go to:


References


Individual Charts

Use this tool to supervise the production quality, in the case where you have a single measurement for each point in time. The measurements need to be quantitative variables.

This tool is useful to recap the moving mean and median and the variability of the production quality that is being measured.

Integrated in this tool, you will find Box-Cox transformations, calculation of process capability and the application of rules for special causes and Westgard rules (an alternative rule set to identify special causes) available to complete your analysis.

Description

Control charts were first mentioned in a document by Walter Shewhart that he wrote during his time working at Bell Labs in 1924. He described his methods completely in his book (1931).

For a long time, there was no significant innovation in the area of control charts. With the development of CUSUM, UWMA and EWMA charts in 1936, Deming expanded the set of available control charts.

Control charts were originally used in area of goods production. Therefore the wording is still from that domain. Today this approach is being applied to a large number of different fields, for instance services, human resources, and sales. In the following lines, we use the wording from the production and shop floors.

Individual charts

The individual charts tool offers you the following chart types alone or in combination:

- X Individual
- MR moving range

An X individual chart is useful to follow the moving average of a production process. Mean shifts are easily visible in the diagrams.

An MR chart (moving range diagram) is useful to analyze the variability of the production. Large difference in production, caused by the use of different production lines, will be easily visible.
Note 1: If you want to investigate smaller mean shifts, then you can also use CUSUM individual charts which are often preferred in comparison with the individual control charts, because they can detect smaller mean shifts.

Note 2: If you have more than one measurement for each point in time, then you should use the control charts for subgroups.

Note 3: If you have measurements in qualitative values (for instance ok, not ok, conform not conform), then use the control charts for attributes.

This tool offers you the following options for the estimation of the standard deviation (sigma) of the data set, given n measurements:

- Average moving range: The estimator for sigma is calculated based on the average moving range using a window length of m measurements.
  \[ \hat{s} = \frac{\bar{m}}{d_2}, \]
  where \(d_2\) is the control chart constant according to Burr (1969).

- Median moving range: The estimator for sigma is calculated based on the median of the moving range using a window length of m measurements.
  \[ \hat{s} = \frac{\text{median}}{d_4}, \]
  where \(d_4\) is the control chart constant according to Burr (1969).

- standard deviation: The estimator for sigma is calculated based on the standard deviation of the n measurements.
  \[ \hat{s} = \frac{s}{c_4} \]
  where \(c_4\) is the control chart constant according to Burr (1969).

**Process capability**

Process capability describes a process and informs if the process is under control and the distribution of the measured variables are inside the specification limits of the process. If the distributions of the measured variables are in the technical specification limits, then the process is called “capable”.

During the interpretation of the different indicators for the process capability please pay attention to the fact that some indicators suppose normality or at least symmetry of the distribution of the measured values. By the use of a normality test, you can verify these premises (see the Normality Tests in XLSTAT-Pro).
If the data are not normally distributed, you have the following possibilities to obtain results for the process capabilities.

- Use the Box-Cox transformation to improve the normality of the data set. Then verify again the normality using a normality test.

- Use the process capability indicator Cp 5.5.

**Box-Cox transformation**

Box-Cox transformation is used to improve the normality of the time series; the Box-Cox transformation is defined by the following equation:

\[
Y_t = \begin{cases} 
\frac{X_t^\lambda - 1}{\lambda}, & X_t \geq 0, \lambda > 0 \\
\ln(X_t), & X_t > 0, \lambda = 0 
\end{cases}
\]

Where the series \{X_t\} being transformed into series \{Y_t\}, (t=1,…,n):

Note: if \( \lambda < 0 \) the first equation is still valid, but \( X_t \) must be strictly positive. XLSTAT accepts a fixed value of \( \lambda \), or it can find the value that maximizes the likelihood value, the model being a simple linear model with the time as sole explanatory variable.

**Chart rules**

XLSTAT offers you the possibility to apply rules for special causes and Westgard rules on the data set. Two sets of rules are available in order to interpret control charts. You can activate and deactivate separately the rules in each set.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
Help: Click this button to display the help.

: Click this button to reload the default options.

: Click this button to delete the data selections.

: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**Mode** tab:

**Chart family:** Select the type of chart family that you want to use:

- **Subgroup charts:** Activate this option if you have a data set with several measurements for each point in time.

- **Individual charts:** Activate this option if you have a data set with one quantitative measurement for each point in time.

- **Attribute charts:** Activate this option if you have a data set with one qualitative measurement for each point.

- **Time weighted:** Activate this option if you want to use a time weighted chart like UWMA, EWMA or CUSUM.

At this stage, the individual charts family is selected. If you want to switch to another chart family, please change the corresponding option and call the help function again if you want to obtain more details on the available options. The options below correspond to the subgroups charts.

**Chart type:** Select the type of chart you want to use:

- **X Individual chart:** Activate this option if you want to calculate the X individual chart to analyze the mean of the process.

- **MR Moving Range chart:** Activate this option if you want to calculate the MR chart to analyze variability of the process.

- **X-MR Individual/Moving Range chart:** Activate this option if you want to calculate the X Individual chart together with the MR chart to analyze the mean value and variability of the process.

**General** tab:
Data: Please choose the unique column or row that contains all the data.

Phase: Activate this option to supply one column/row with the phase identifier.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the “Variable labels” option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column/Row labels: Activate this option if the first row (column mode) or column (row mode) of the data selections contains a label.

Options tab:

Upper control limit:

- **Bound**: Activate this option, if you want to enter a maximum value to accept for the upper control limit of the process. This value will be used when the calculated upper control limit is greater than the value entered here.

- **Value**: Enter the upper control limit. This value will be used and overrides the calculated upper control limit.

Lower control limit:

- **Bound**: Activate this option, if you want to enter a minimum value to accept for the lower control limit of the process. This value will be used when the calculated lower control limit is greater than the value entered here.

- **Value**: Enter the lower control limit. This value will be used in place of the calculated upper control limit.
Calculate Process capabilities: Activate this option to calculate process capabilities based on the input data (see the description section for more details).

USL: If the calculation of the process capabilities is activated, please enter here the upper specification limit (USL) of the process.

LSL: If the calculation of the process capabilities is activated, please enter here the lower specification limit (LSL) of the process.

Target: If the calculation of the process capabilities is activated, activate this option to add the target value of the process.

Confidence interval (%): If the “Calculate Process Capabilities” option is activated, please enter the percentage range of the confidence interval to use for calculating the confidence interval around the parameters. Default value: 95.

Box-Cox: Activate this option to compute the Box-Cox transformation. You can either fix the value of the Lambda parameter, or decide to let XLSTAT optimize it (see the description section for further details).

k Sigma: Activate this option to enter the distance between the upper and the lower control limit and the center line of the control chart. The distance is fixed to k times the factor you enter multiplied by the estimated standard deviation. Corrective factors according to Burr (1969) will be applied.

alpha: Activate this option to enter the size of the confidence range around the center line of the control chart. The alpha is used to compute the upper and lower control limits. 100 – alpha % of the distribution of the control chart is inside the control limits. Corrective factors according to Burr (1969) will be applied.

Mean: Activate this option to enter a value for the center line of the control chart. This value should be based on historical data.

Sigma: Activate this option to enter a value for the standard deviation of the control chart. This value should be based on historical data. If this option is activated, then you cannot choose an estimation method for the standard deviation in the "Estimation" tab.

Estimation tab:

Method for Sigma: Select an option to determine the estimation method for the standard deviation of the control chart (see the description section for further details):

- Average Moving Range
• **Median Moving Range**
  
  o **MR Length**: Change this value to modify the number of observations that are taken into account in the moving range.

• **Standard deviation**: The estimator of sigma is calculated using the standard deviation of the n measurements.

**Outputs** tab:

**Display zones**: Activate this option to display beside the lower and upper control limit also the limits of the zones A and B.

**Normality Tests**: Activate this option to check normality of the data. (see the Normality Tests tool for further details).

**Significance level (%)**: Enter the significance level for the tests.

**Test special causes**: Activate this option to analyze the points of the control chart according to the rules for special causes. You can activate the following rules independently:

• 1 point more than 3s from center line
• 9 points in a row on same side of center line
• 6 points in a row, all increasing or all decreasing
• 14 points in a row, alternating up and down
• 2 out of 3 points > 2s from center line (same side)
• 4 out of 5 points > 1s from center line (same side)
• 15 points in a row within 1s of center line (either side)
• 8 points in a row > 1s from center line (either side)

• **All**: Click this button to select all.
• **None**: Click this button to deselect all.

**Apply Westgard rules**: Activate this option to analyze the points of the control chart according to the Westgard rules. You can activate the following rules independently:

• **Rule 1 2s**
- Rule 13
- Rule 22s
- Rule 4s
- Rule 41s
- Rule 10X

- **All**: Click this button to select all.
- **None**: Click this button to deselect all.

**Charts** tab:

**Display charts**: Activate this option to display the control charts graphically.

- **Continuous line**: Activate this option to connect the points in the control chart.

**Connect through missing**: Activate this option to connect the points in the control charts, even when missing values are between the points.

**Normal Q-Q Charts**: Check this option to display Q-Q plots.

**Display a distribution**: Activate this option to compare histograms of samples selected with a density function.

**Run Charts**: Activate this option to display a chart of the latest data points. Each individual measurement is displayed.

**Number of observations**: Enter the maximal number of the last observations to be displayed in the Run chart.

**Results**

**Estimation**:

**Estimated mean**: This table displays the estimated mean values for the different phases.

**Estimated standard deviation**: This table displays the estimated standard deviation values for the different phases.
**Box-Cox transformation:**

**Estimates of the parameters of the model:** This table is available only if the Lambda parameter has been optimized. It displays the estimator for Lambda.

**Series before and after transformation:** This table displays the series before and after transformation. If Lambda has been optimized, the transformed series corresponds to the residuals of the model. If it hasn’t then the transformed series is the direct application of the Box-Cox transformation.

**Process capability:**

**Process capabilities:** These tables are displayed, if the “process capability” option has been selected. There is one table for each phase. A table contains the following indicators for the process capability and if possible the corresponding confidence intervals: Cp, Cpl, Cpu, Cpk, Pp, Ppl, Ppu, Ppk, Cpm, Cpm (Boyle), Cp 5.5, Cpk 5.5, Cpmk, and Cs (Wright).

For Cp, Cpl, and Cpu, information about the process performance is supplied and for Cp a status information is given to facilitate the interpretation.

Cp values have the following status based on Ekvall and Juran (1974):

- "not adequate" if Cp < 1
- "adequate" if 1 <= Cp <= 1.33
- "more than adequate" if Cp > 1.33

Based on Montgomery (2001), Cp needs to have the following minimal values for the process performance to be as expected:

- 1.33 for existing processes
- 1.50 for new processes or for existing processes when the variable is critical
- 1.67 for new processes when the variable is critical

Based on Montgomery (2001), Cpu and Cpl need to have the following minimal values for process performance to be as expected:

- 1.25 for existing processes
- 1.45 for new processes or for existing processes when the variable is critical
- 1.60 for new processes when the variable is critical
Capabilities: This chart contains information about the specification and control limits. A line between the lower and upper limits represents the interval with an additional vertical mark for the center line. The different control limits of each phase are drawn separately.

Chart information:

The following results are displayed separately for each requested chart. Charts can be selected alone or in combination with the X individual chart.

X Individual / MR moving range chart: This table contains information about the center line and the upper and lower control limits of the selected chart. There will be one column for each phase.

Observation details: This table displays detailed information for each observation. For each observation, the corresponding phase, the mean or median, the center line, the lower and upper control limits are displayed. If the information about the zones A, B and C are activated, then the lower and upper control limits of the zones A and B are displayed as well.

Rule details: If the rules options are activated, a detailed table about the rules will be displayed. For each observation, there is one row for each rule that applies. “Yes” indicates that the corresponding rule was fired, and “No” indicates that the rule does not apply.

X Individual / MR moving range Chart: If the charts are activated, then a chart containing the information of the two tables above is displayed. Each observation is displayed. The center line and the lower and upper control limits are displayed as well. If the corresponding options have been activated, the lower and upper control limits for the zones A and B are included and there are labels for the observations for which rules were fired. A legend with the activated rules and the corresponding rule number is displayed below the chart.

Normality tests:

For each of the four tests, the statistics relating to the test are displayed including, in particular, the p-value which is afterwards used in interpreting the test by comparing with the chosen significance threshold.

If requested, a Q-Q plot is then displayed.

Histograms: The histograms are displayed. If desired, you can change the color of the lines, scales, titles as with any Excel chart.

Run chart: The chart of the last data points is displayed.
Example

A tutorial explaining how to use the SPC subgroup charts tool is available on the Addinsoft web site. To consult the tutorial, please go to:

http://www.xlstat.com/demo-spc2.htm

References


Attribute charts

Use this tool to supervise the production quality, in the case where you have a single measurement for each point in time. The measurements are based on attribute or attribute counts of the process.

This tool is useful to recap the categorical variables of the measured production quality.

Integrated in this tool, you will find Box-Cox transformations, calculation of process capability and the application of rules for special causes and Westgard rules (an alternative rule set to identify special causes) available to complete your analysis.

Description

Control charts were first mentioned in a document by Walter Shewhart that he wrote during his time working at Bell Labs in 1924. He described his methods completely in his book (1931).

For a long time, there was no significant innovation in the area of control charts. With the development of CUSUM, UWMA and EWMA charts in 1936, Deming expanded the set of available control charts.

Control charts were originally used in area of goods production. Therefore the wording is still from that domain. Today this approach is being applied to a large number of different fields, for instance services, human resources, and sales. In the following chapters we will use the wording from the production and shop floors.

Attribute charts

The attribute charts tool offers you the following chart types:

- P chart
- NP chart
- C chart
- U chart

These charts analyze either “nonconforming products” or “nonconformities”. They are usually used to inspect the quality before delivery (outgoing products) or the quality at delivery (incoming products). Not all the products need to be necessarily inspected.

Inspections are done by inspection units having a well defined size. The size can be 1 in the case of the reception of television sets at a warehouse. The size would be 24 in the case of peaches delivered in crates of 24 peaches.
P and NP charts allow to analyze the fraction respectively the absolute number of nonconforming products of a production process. For example, we can count the number of nonconforming television sets, or the number of crates that contain at least one bruised peach.

C and U chart analyze the fraction respectively the absolute number of occurrences of nonconformities in an inspection unit. For example, we can count the number of defect transistors for each inspection unit (there might be more than one transistor not working in one television set), or the number of bruised peaches per crate.

A P chart is useful to follow the fraction of non conforming units of a production process.

An NP chart is useful to follow the absolute number of non conforming units of a production process.

A C chart is useful in the case of a production having a constant size for each inspection unit. It can be used to follow the absolute number of the non conforming items per inspection.

A U chart is useful in the case of a production having a non constant size of each inspection unit. It can be used to follow the fraction of the non conforming items per inspection.

**Process capability**

Process capability describes a process and informs if the process is under control and the distribution of the measured variables are inside the specification limits of the process. If the distributions of the measured variables are in the technical specification limits, then the process is called “capable”.

During the interpretation of the different indicators for the process capability please pay attention to the fact that some indicators suppose normality or at least symmetry of the distribution of the measured values. By the use of a normality test, you can verify these premises (see the Normality Tests in XLSTAT-Pro).

If the data are not normally distributed, you have the following possibilities to obtain results for the process capabilities.

- Use the Box-Cox transformation to improve the normality of the data set. Then verify again the normality using a normality test.
- Use the process capability indicator Cp 5.5.

**Box-Cox transformation**

Box-Cox transformation is used to improve the normality of the time series; the Box-Cox transformation is defined by the following equation:
\[ Y_t = \begin{cases} \frac{X_t^\lambda - 1}{\lambda}, & X_t \geq 0, \quad \lambda > 0 \\ \ln(X_t), & X_t > 0, \quad \lambda = 0 \end{cases} \]

Where the series \( \{X_t\} \) being transformed into series \( \{Y_t\} \), \( t=1,\ldots,n \):

Note: if \( \lambda < 0 \) the first equation is still valid, but \( X_t \) must be strictly positive. XLSTAT accepts a fixed value of \( \lambda \), or it can find the value that maximizes the likelihood value, the model being a simple linear model with the time as sole explanatory variable.

**Chart rules**

XLSTAT offers you the possibility to apply rules for special causes and Westgard rules on the data set. Two sets of rules are available in order to interpret control charts. You can activate and deactivate separately the rules in each set.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- ![Reload options](image) Click this button to reload the default options.
- ![Delete data](image) Click this button to delete the data selections.
- ![Change mode](image) Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**Mode** tab:

**Chart family**: Select the type of chart family that you want to use:
- **Subgroup charts**: Activate this option if you have a data set with several measurements for each point in time.

- **Individual charts**: Activate this option if you have a data set with one quantitative measurement for each point in time.

- **Attribute charts**: Activate this option if you have a data set with one qualitative measurement for each point.

- **Time weighted**: Activate this option if you want to use a time weighted chart like UWMA, EWMA or CUSUM.

At this stage, the attribute charts family is selected. If you want to switch to another chart family, please change the corresponding option and call the help function again if you want to obtain more details on the available options. The options below correspond to the subgroups charts.

**Chart type**: Select the type of chart you want to use (see the description section for more details):

- P chart
- NP chart
- C chart
- U chart

**General** tab:

**Data**: Please choose the unique column or row that contains all the data.

**Phase**: Activate this option to supply one column/row with the phase identifier.

**Observation labels**: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.
**Column/Row labels:** Activate this option if the first row (column mode) or column (row mode) of the data selections contains a label.

**Options** tab:

**Upper control limit:**
- **Bound:** Activate this option, if you want to enter a maximum value to accept for the upper control limit of the process. This value will be used when the calculated upper control limit is greater than the value entered here.
- **Value:** Enter the upper control limit. This value will be used and overrides the calculated upper control limit.

**Lower control limit:**
- **Bound:** Activate this option, if you want to enter a minimum value to accept for the lower control limit of the process. This value will be used when the calculated lower control limit is greater than the value entered here.
- **Value:** Enter the lower control limit. This value will be used and overrides the calculated upper control limit.

**Calculate Process capabilities:** Activate this option to calculate process capabilities based on the input data (see the description section for more details).

**USL:** If the calculation of the process capabilities is activated, please enter here the upper specification limit (USL) of the process.

**LSL:** If the calculation of the process capabilities is activated, please enter here the lower specification limit (LSL) of the process.

**Target:** If the calculation of the process capabilities is activated, activate this option to add the target value of the process.

**Confidence interval (%):** If the “Calculate Process Capabilities” option is activated, please enter the percentage range of the confidence interval to use for calculating the confidence interval around the parameters. Default value: 95.

**Box-Cox:** Activate this option to compute the Box-Cox transformation. You can either fix the value of the Lambda parameter, or decide to let XLSTAT optimize it (see the description section for further details).
**k Sigma**: Activate this option to enter the distance between the upper and the lower control limit and the center line of the control chart. The distance is fixed to k times the factor you enter multiplied by the estimated standard deviation. Corrective factors according to Burr (1969) will be applied.

**alpha**: Activate this option to enter the size of the confidence range around the center line of the control chart. The alpha is used to compute the upper and lower control limits. 100 – alpha % of the distribution of the control chart is inside the control limits. Corrective factors according to Burr (1969) will be applied.

**P bar / C bar / U bar**: Activate this option to enter a value for the center line of the control chart. This value should be based on historical data.

**Outputs** tab:

**Display zones**: Activate this option to display beside the lower and upper control limit also the limits of the zones A and B.

**Normality Tests**: Activate this option to check normality of the data. (see the Normality Tests tool for further details).

**Significance level (%)**: Enter the significance level for the tests.

**Test special causes**: Activate this option to analyze the points of the control chart according to the rules for special causes. You can activate the following rules independently:

- 1 point more than 3s from center line
- 9 points in a row on same side of center line
- 6 points in a row, all increasing or all decreasing
- 14 points in a row, alternating up and down
- 2 out of 3 points > 2s from center line (same side)
- 4 out of 5 points > 1s from center line (same side)
- 15 points in a row within 1s of center line (either side)
- 8 points in a row > 1s from center line (either side)

- **All**: Click this button to select all.
- **None**: Click this button to deselect all.

**Apply Westgard rules**: Activate this option to analyze the points of the control chart according to the Westgard rules. You can activate the following rules independently:

- Rule 1 2s
- Rule 1 3
- Rule 2 2s
- Rule 4s
- Rule 4 1s
- Rule 10 X

- **All**: Click this button to select all.
- **None**: Click this button to deselect all.

**Charts** tab:

**Display charts**: Activate this option to display the control charts graphically.

- **Continuous line**: Activate this option to connect the points in the control chart.

**Connect through missing**: Activate this option to connect the points in the control charts, even when missing values are between the points.

**Normal Q-Q Charts**: Check this option to display Q-Q plots.

**Display a distribution**: Activate this option to compare histograms of samples selected with a density function.

**Run Charts**: Activate this option to display a chart of the latest data points. Each individual measurement is displayed.

**Number of observations**: Enter the maximal number of the last observations to be displayed in the Run chart.
Results

Estimation:

Estimated mean: This table displays the estimated mean values for the different phases.

Estimated standard deviation: This table displays the estimated standard deviation values for the different phases.

Box-Cox transformation:

Estimates of the parameters of the model: This table is available only if the Lambda parameter has been optimized. It displays the estimator for Lambda.

Series before and after transformation: This table displays the series before and after transformation. If Lambda has been optimized, the transformed series corresponds to the residuals of the model. If it hasn't then the transformed series is the direct application of the Box-Cox transformation.

Process capability:

Process capabilities: These tables are displayed, if the “process capability” option has been selected. There is one table for each phase. A table contains the following indicators for the process capability and if possible the corresponding confidence intervals: Cp, Cpl, Cpu, Cpk, Pp, Ppl, Ppu, Ppk, Cpm, Cpm (Boyle), Cp 5.5, Cpk 5.5, Cpmk, and Cs (Wright).

For Cp, Cpl, and Cpu, information about the process performance is supplied and for Cp a status information is given to facilitate the interpretation.

Cp values have the following status based on Ekvall and Juran (1974):

- "not adequate" if Cp < 1
- "adequate" if 1 <= Cp <= 1.33
- "more than adequate" if Cp > 1.33

Based on Montgomery (2001), Cp needs to have the following minimal values for the process performance to be as expected:

- 1.33 for existing processes
- 1.50 for new processes or for existing processes when the variable is critical
- 1.67 for new processes when the variable is critical
Based on Montgomery (2001), Cpu and Cpl need to have the following minimal values for process performance to be as expected:

- 1.25 for existing processes
- 1.45 for new processes or for existing processes when the variable is critical
- 1.60 for new processes when the variable is critical

Capabilities: This chart contains information about the specification and control limits. A line between the lower and upper limits represents the interval with an additional vertical mark for the center line. The different control limits of each phase are drawn separately.

Chart information:

The following results are displayed separately for each requested chart. Charts can be selected alone or in combination with the X attribute chart.

P / NP / C / U chart: This table contains information about the center line and the upper and lower control limits of the selected chart. There will be one column for each phase.

Observation details: This table displays detailed information for each observation. For each observation the corresponding phase, the value for P, NP, C or U, the subgroup size, the center line, the lower and upper control limits are displayed. If the information about the zones A, B and C are activated, then the lower and upper control limits of the zones A and B are displayed as well.

Rule details: If the rules options are activated, a detailed table about the rules will be displayed. For each subgroup there is one row for each rule that applies. “Yes” indicates that the corresponding rule was fired, and “No” indicates that the rule does not apply.

P / NP / C / U Chart: If the charts are activated, then a chart containing the information of the two tables above is displayed. The center line and the lower and upper control limits are displayed as well. If the corresponding options have been activated, the lower and upper control limits for the zones A and B are included and there are labels for the subgroups for which rules were fired. A legend with the activated rules and the corresponding rule number is displayed below the chart.

Normality tests:
For each of the four tests, the statistics relating to the test are displayed including, in particular, the p-value which is afterwards used in interpreting the test by comparing with the chosen significance threshold.

If requested, a Q-Q plot is then displayed.

**Histograms:** The histograms are displayed. If desired, you can change the color of the lines, scales, titles as with any Excel chart.

**Run chart:** The chart of the last data points is displayed.

**Example**

A tutorial explaining how to use the attributes charts tool is available on the Addinsoft web site. To consult the tutorial, please go to:


**References**


**Time Weighted Charts**

Use this tool to supervise production quality, in the case where you have a group of measurements or a single measurement for each point in time. The measurements need to be quantitative variables.

This tool is useful to recap the mean and the variability of the measured production quality.

Integrated in this tool, you will find Box-Cox transformations, calculation of process capability and the application of rules for special causes and Westgard rules (an alternative rule set to identify special causes) available to complete your analysis.

**Description**

Control charts were first mentioned in a document by Walter Shewhart that he wrote during his time working at Bell Labs in 1924. He described his methods completely in his book (1931).

For a long time, there was no significant innovation in the area of control charts. With the development of CUSUM, UWMA and EWMA charts in 1936, Deming expanded the set of available control charts.

Control charts were originally used in area of goods production. Therefore the wording is still from that domain. Today this approach is being applied to a large number of different fields, for instance services, human resources, and sales. In the following chapters we will use the wording from the production and shop floors.

**Time Weighted charts**

The time weighted charts tool offers you the following chart types:

- CUSUM or CUSUM individual
- UWMA or UWMA individual
- EWMA or EWMA individual

A CUSUM, UWMA or EWMA chart is useful to follow the mean of a production process. Mean shifts are easily visible in the diagrams.

**UWMA and EWMA charts**

These charts are not directly based on the raw data. They are based on the smoothed data.
In the case of UWMA charts, data are smoothed using a uniform weighting in a moving window. Then the chart is analyzed like Shewhart charts.

In the case of EWMA charts, the data is smoothed using a exponentially weighting. Then the chart is analyzed like Shewhart charts.

**CUSUM charts**

These charts are not directly based on the raw data. They are based on the normalized data.

These charts help to detect mean shifts of at a user defined granularity. The granularity is defined by the design parameter \( k \). \( k \) is the half of the mean shift to be detected. To detect a 1 sigma shift, \( k \) is set to 0.5.

Two kinds of CUSUM charts can be drawn: one and two sided charts. In the case of a one sided CUSUM chart, upper and lower cumulated sums SH and SL are recursively calculated.

\[
SH_i = \max(0, (z_i - k) + SH_{i-1})
\]

\[
SL_i = \min(0, (z_i + k) + SL_{i-1})
\]

If SH or SL is bigger than the threshold \( h \), then a mean shift is detected. The value of \( h \) can be chosen by the user (\( h \) is usually set to 4 or 5).

The initial value of SH and SL at the beginning of the calculation and after detecting a mean shift is usually 0. Using the option FIR (Fast Initial Response) can change this initial value to a user defined value.

In the case of a two sided CUSUM chart the normalized data are calculated. The upper and lower control limits are called "U mask" or "V mask". These names are related to the shape that the control limits draws on the chart. For a given data point the maximal upper and lower limits for mean shift detection are calculated backwards and drawn in the chart in a U or V mask format. The default data point for the origin of the mask is the last data point. The user can change this by the option origin.

This tool offers you the following options for the estimation of the standard deviation (sigma) of the data set, given \( k \) subgroups and \( n_i (i=1, \ldots k) \) measurements per subgroup:

- **Pooled standard deviation**: sigma is computed using the \( k \) within-subgroup variances:

\[
\hat{s} = \sqrt{\frac{\sum_{i=1}^{k} (n_i - 1) s_i^2}{\sum_{i=1}^{k} (n_i - 1)}} / c_{s}\left(1 + \sum_{i=1}^{k} (n_i - 1)\right)
\]

- **R bar**: The estimator for sigma is calculated based on the average range of the \( k \) subgroups.
\[ \hat{s} = \frac{\bar{R}}{d_2} \]

where \( d_2 \) is the control chart constant according to Burr (1969).

- **S bar**: The estimator for \( \sigma \) is calculated based on the average of the standard deviations of the \( k \) subgroups:

\[ \hat{s} = \sqrt{\frac{1}{k} \sum_{i=1}^{k} s_i^2 / c_4}, \]

In the case of \( n \) Individual measurements:

- **Average moving range**: The estimator for \( \sigma \) is calculated based on the average moving range using a window length of \( m \) measurements.

\[ \hat{s} = \bar{m} / d_2, \]

where \( d_2 \) is the control chart constant according to Burr (1969).

- **Median moving range**: The estimator for \( \sigma \) is calculated based on the median of the moving range using a window length of \( m \) measurements.

\[ \hat{s} = \text{median} / d_4, \]

where \( d_4 \) is the control chart constant according to Burr (1969).

- **standard deviation**: The estimator for \( \sigma \) is calculated based on the standard deviation of the \( n \) measurements.

\[ \hat{s} = s / c_4 \]

where \( c_4 \) is the control chart constant according to Burr (1969).

**Box-Cox transformation**

Box-Cox transformation is used to improve the normality of the time series; the Box-Cox transformation is defined by the following equation:

\[ Y_t = \begin{cases} \frac{X_t^\lambda - 1}{\lambda}, & X_t \geq 0, \lambda > 0 \\ \ln(X_t), & X_t > 0, \lambda = 0 \end{cases} \]

Where the series \{\( X_t \)\} being transformed into series \{\( Y_t \)\}, \( t = 1, \ldots, n \):
Note: if $\lambda < 0$ the first equation is still valid, but $X_t$ must be strictly positive. XLSTAT accepts a fixed value of $\lambda$, or it can find the value that maximizes the likelihood value, the model being a simple linear model with the time as sole explanatory variable.

**Process capability**

Process capability describes a process and informs if the process is under control and the distribution of the measured variables are inside the specification limits of the process. If the distributions of the measured variables are in the technical specification limits, then the process is called “capable”.

During the interpretation of the different indicators for the process capability please pay attention to the fact that some indicators suppose normality or at least symmetry of the distribution of the measured values. By the use of a normality test, you can verify these premises (see the Normality Tests in XLSTAT-Pro).

If the data are not normally distributed, you have the following possibilities to obtain results for the process capabilities.

- Use the Box-Cox transformation to improve the normality of the data set. Then verify again the normality using a normality test.

- Use the process capability indicator $C_p$ 5.5.

**Chart rules**

XLSTAT offers you the possibility to apply rules for special causes and Westgard rules on the data set. Two sets of rules are available in order to interpret control charts. You can activate and deactivate separately the rules in each set.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload default options**: Click this button to reload the default options.
Click this button to delete the data selections.

Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**Mode** tab:

**Chart family:** Select the type of chart family that you want to use:

- **Subgroup charts:** Activate this option if you have a data set with several measurements for each point in time.
- **Individual charts:** Activate this option if you have a data set with one quantitative measurement for each point in time.
- **Attribute charts:** Activate this option if you have a data set with one qualitative measurement for each point.
- **Time weighted:** Activate this option if you want to use a time weighted chart like UWMA, EWMA or CUSUM.

At this stage, the time weighted charts family is selected. If you want to switch to another chart family, please change the corresponding option and call the help function again if you want to obtain more details on the available options. The options below correspond to the subgroups charts

**Chart type:** Select the type of chart you want to use (see the description section for more details):

- CUSUM chart
- CUSUM individual chart
- UWMA chart
- UWMA individual chart
- EWMA chart
- EWMA individual chart

**General** tab:

**Data format:** Select the data format.
• **Columns/Rows**: Activate this option for XLSTAT to take each column (in column mode) or each row (in row mode) as a separate measurement that belongs to the same subgroup.

• **One column/Row**: Activate this option, if the measurements of subgroups continuously follow one after the other in one column or one row. To assign the different measurements to their corresponding subgroup, please enter a constant group size or select a column or row with the group identifier in it.

**Data**: If the data format « One column/row » is selected, please choose the unique column or row that contains all the data. The assignment of the data to their corresponding subgroup must be specified using the Groups field or setting the common subgroup size. If you select the data « Columns/rows » option, please select a data area with one column/row per measurement in a subgroup.

**Groups**: If the data format « one column/row » is selected, then activate this Option to select a column/row that contains the group identifier. Select the data that identifies for each element of the data selection the corresponding group.

**Common subgroup size**: If the data format « One column/row » is selected and the subgroup size is constant, then you can deactivate the groups option and enter in this field the common subgroup size.

**Phase**: Activate this option to supply one column/row with the phase identifier.

**Observation labels**: Activate this option if observations labels are available. Then select the corresponding data. If the “Variable labels” option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Column/Row labels**: Activate this option if the first row (column mode) or column (row mode) of the data selections contains a label.

**Standardize**: In the case of a CUSUM chart, please activate this option to display the cumulated sums and the control limits normalized.
**Target:** In the case of a CUSUM chart, please activate this option to enter the target value that will be used during the normalization of the data. Default value is the estimated mean.

**Weight:** In the case of a EWMA chart, please activate this option to enter the weight factor of the exponential smoothing.

**MA Length:** In the case of a UWMA chart, please activate this option to enter the length of the window of the moving average.

**Options** tab:

**Upper control limit:**

- **Bound:** Activate this option, if you want to enter a maximum value to accept for the upper control limit of the process. This value will be used when the calculated upper control limit is greater than the value entered here.

- **Value:** Enter the upper control limit. This value will be used and overrides the calculated upper control limit.

**Lower control limit:**

- **Bound:** Activate this option, if you want to enter a minimum value to accept for the lower control limit of the process. This value will be used when the calculated lower control limit is greater than the value entered here.

- **Value:** Enter the lower control limit. This value will be used and overrides the calculated upper control limit.

**Calculate Process capabilities:** Activate this option to calculate process capabilities based on the input data (see the description section for more details).

**USL:** If the calculation of the process capabilities is activated, please enter here the upper specification limit (USL) of the process.

**LSL:** If the calculation of the process capabilities is activated, please enter here the lower specification limit (LSL) of the process.

**Target:** If the calculation of the process capabilities is activated, activate this option to add the target value of the process.

**Confidence interval (%):** If the “Calculate Process Capabilities” option is activated, please enter the percentage range of the confidence interval to use for calculating the confidence interval around the parameters. Default value: 95.
**Box-Cox:** Activate this option to compute the Box-Cox transformation. You can either fix the value of the Lambda parameter, or decide to let XLSTAT optimize it (see the description section for further details).

**k Sigma:** Activate this option to enter the distance between the upper and the lower control limit and the center line of the control chart. The distance is fixed to k times the factor you enter multiplied by the estimated standard deviation. Corrective factors according to Burr (1969) will be applied.

**alpha:** Activate this option to enter the size of the confidence range around the center line of the control chart. The alpha is used to compute the upper and lower control limits. 100 – alpha % of the distribution of the control chart is inside the control limits. Corrective factors according to Burr (1969) will be applied.

**Mean:** Activate this option to enter a value for the center line of the control chart. This value should be based on historical data.

**Sigma:** Activate this option to enter a value for the standard deviation of the control chart. This value should be based on historical data. If this option is activated, then you cannot choose an estimation method for the standard deviation in the "Estimation" tab.

**Estimation tab:**

**Method for Sigma:** Select an option to determine the estimation method for the standard deviation of the control chart (see the description section for further details):

- **Pooled standard deviation:** The standard deviation is calculated using all available measurements. That means having n subgroups with k measurements for each subgroup, all the n * k measurements will be weighted equally to calculate the standard deviation.

- **R bar:** The estimator of sigma is calculated using the average range of the n subgroups.

- **S bar:** The estimator of sigma is calculated using the average standard deviation of the n subgroups.

- **Average Moving Range:** The estimator of sigma is calculated using the average moving range using a window length of m measurements.

- **Median Moving Range:** The estimator of sigma is calculated using the median of the moving range using a window length of m measurements.
• **MR Length**: Activate this option to change the window length of the moving range.

• **Standard deviation**: The estimator of sigma is calculated using the standard deviation of the n measurements.

**Design tab:**

This tab is only active, if CUSUM charts are selected.

**Scheme**: Chose one of the following options depending on the kind of chart that you want (see the description section for further details):

• **One sided (LCL/UCL)**: The upper and lower cumulated sum are calculated separately for each point.
  
  o **FIR**: Activate this option to change the initial value of the upper and lower cumulated sum. Default value is 0.

• **Two sided (U-Mask)**: The normalized values are displayed. Starting from the origin point the upper and lower limits for the mean shift detection a displayed backwards in form of a mask.
  
  o **Origin**: Activate this option to change the origin of the mask. Default value is the last data point.

**Design**: In this section you can determine the Parameter of the mean-shift detection (see the description section for further details):

• **h**: Enter the threshold for the upper and lower cumulated sum or mask from above which a mean shift is detected.

• **k**: Enter the granularity of the mean shift detection. K is the half of the mean shift to be detected. Default value is 0.5 to detect 1 sigma mean shifts.

**Outputs tab:**

**Display zones**: Activate this option to display beside the lower and upper control limit also the limits of the zones A and B.

**Normality Tests**: Activate this option to check normality of the data. (see the Normality Tests tool for further details).
Significance level (%): Enter the significance level for the tests.

Test special causes: Activate this option to analyze the points of the control chart according to the rules for special causes. You can activate the following rules independently:

- 1 point more than 3s from center line
- 9 points in a row on same side of center line
- 6 points in a row, all increasing or all decreasing
- 14 points in a row, alternating up and down
- 2 out of 3 points > 2s from center line (same side)
- 4 out of 5 points > 1s from center line (same side)
- 15 points in a row within 1s of center line (either side)
- 8 points in a row > 1s from center line (either side)

- All: Click this button to select all.
- None: Click this button to deselect all.

Apply Westgard rules: Activate this option to analyze the points of the control chart according to the Westgard rules. You can activate the following rules independently:

- Rule 1 2s
- Rule 1 3
- Rule 2 2s
- Rule 4s
- Rule 4 1s
- Rule 10 X

- All: Click this button to select all.
- None: Click this button to deselect all.

Charts tab:
Display charts: Activate this option to display the control charts graphically.

- Continuous line: Activate this option to connect the points in the control chart.
- Box view: Activate this option to display the control charts using bars.

Connect through missing: Activate this option to connect the points in the control charts, even when missing values are between the points.

Normal Q-Q Charts: Check this option to display Q-Q plots.

Display a distribution: Activate this option to compare histograms of samples selected with a density function.

Run Charts: Activate this option to display a chart of the latest data points. Each individual measurement is displayed.

- Number of observations: Enter the maximal number of the last observations to be displayed in the Run chart.

Results

Estimation:

Estimated mean: This table displays the estimated mean values for the different phases.

Estimated standard deviation: This table displays the estimated standard deviation values for the different phases.

Box-Cox transformation:

Estimates of the parameters of the model: This table is available only if the Lambda parameter has been optimized. It displays the estimator for Lambda.

Series before and after transformation: This table displays the series before and after transformation. If Lambda has been optimized, the transformed series corresponds to the residuals of the model. If it hasn't then the transformed series is the direct application of the Box-Cox transformation.

Process capabilities:

Process capabilities: These tables are displayed, if the “process capability” option has been selected. There is one table for each phase. A table contains the following indicators for the
process capability and if possible the corresponding confidence intervals: Cp, Cpl, Cpu, Cpk, Pp, Ppl, Ppu, Ppk, Cpm, Cpm (Boyle), C5 5.5, Cpk 5.5, Cpmk, and Cs (Wright).

For Cp, Cpl, and Cpu, information about the process performance is supplied and for Cp a status information is given to facilitate the interpretation.

Cp values have the following status based on Ekvall and Juran (1974):

- "not adequate" if Cp < 1
- "adequate" if 1 <= Cp <= 1.33
- "more than adequate" if Cp > 1.33

Based on Montgomery (2001), Cp needs to have the following minimal values for the process performance to be as expected:

- 1.33 for existing processes
- 1.50 for new processes or for existing processes when the variable is critical
- 1.67 for new processes when the variable is critical

Based on Montgomery (2001), Cpu and Cpl need to have the following minimal values for process performance to be as expected:

- 1.25 for existing processes
- 1.45 for new processes or for existing processes when the variable is critical
- 1.60 for new processes when the variable is critical

Capabilities: This chart contains information about the specification and control limits. A line between the lower and upper limits represents the interval with an additional vertical mark for the center line. The different control limits of each phase are drawn separately.

Chart information:

The following results are displayed separately for the requested chart.

**UWMA / EWMA / CUSUM chart:** This table contains information about the center line and the upper and lower control limits of the selected chart. There will be one column for each phase.
**Observation details:** This table displays detailed information for each subgroup. For each subgroup the corresponding phase, the values according to the selected diagram type, the center line, the lower and upper control limits are displayed. If the information about the zones A, B and C are activated, then the lower and upper control limits of the zones A and B are displayed as well.

**Rule details:** If the rules options are activated, a detailed table about the rules will be displayed. For each subgroup there is one row for each rule that applies. “Yes” indicates that the corresponding rule was fired, and “No” indicates that the rule does not apply.

**UWMA / EWMA / CUSUM Chart:** If the charts are activated, then a chart containing the information of the two tables above is displayed. The center line and the lower and upper control limits are displayed as well. If the corresponding options have been activated, the lower and upper control limits for the zones A and B are included and there are labels for the subgroups for which rules were fired. A legend with the activated rules and the corresponding rule number is displayed below the chart.

**Normality tests:**

For each of the four tests, the statistics relating to the test are displayed including, in particular, the p-value which is afterwards used in interpreting the test by comparing with the chosen significance threshold.

If requested, a Q-Q plot is then displayed.

**Histograms:** The histograms are displayed. If desired, you can change the color of the lines, scales, titles as with any Excel chart.

**Run chart:** The chart of the last data points is displayed.

**Example**

A tutorial explaining how to use the SPC time weighted charts tool is available on the Addinsoft web site. To consult the tutorial, please go to:

References


Pareto plots

Use this tool to calculate descriptive statistics and display Pareto plots (bar and pie charts) for a set of qualitative variables.

Description

A Pareto chart draws its name from an Italian economist, but J. M. Juran is credited with being the first to apply it to industrial problems.

The causes that should be investigated (e.g., nonconforming items) are listed and percentages assigned to each one so that the total is 100%. The percentages are then used to construct the diagram that is essentially a bar or pie chart. Pareto analysis uses the ranking causes to determine which of them should be pursued first.

XLSTAT offers you a large number of descriptive statistics and charts which give you a useful and relevant insight of your data.

Although you can select several variables (or samples) at the same time, XLSTAT calculates all the descriptive statistics for each of the samples independently.

Descriptive statistics for qualitative data:

For a sample made up of N qualitative values, we define:

- **Number of observations**: The number N of values in the selected sample.
- **Number of missing values**: The number of missing values in the sample analyzed. In the subsequent statistical calculations, values identified as missing are ignored. We define $n$ to be the number of non-missing values, and \{w_1, w_2, \ldots w_n\} to be the sub-sample of weights for the non-missing values.
- **Sum of weights***: The sum of the weights, $Sw$. When all the weights are 1, $Sw=n$.
- **Mode***: The mode of the sample analyzed. In other words, the most frequent category.
- **Frequency of mode***: The frequency of the category to which the mode corresponds.
- **Category**: The names of the various categories present in the sample.
- **Frequency by category***: The frequency of each of the categories.
- **Relative frequency by category***: The relative frequency of each of the categories.
• **Cumulated relative frequency by category**: The cumulated relative frequency of each of the categories.

(*) Statistics followed by an asterisk take the weight of observations into account.

Several types of chart are available for qualitative data:

**Charts for qualitative data:**

• **Bar charts**: Check this option to represent the frequencies or relative frequencies of the various categories of qualitative variables as bars.

• **Pie charts**: Check this option to represent the frequencies or relative frequencies of the various categories of qualitative variables as pie charts.

• **Double pie charts**: These charts are used to compare the frequencies or relative frequencies of sub-samples with those of the complete sample.

• **Doughnuts**: this option is only checked if a column of sub-samples has been selected. These charts are used to compare the frequencies or relative frequencies of sub-samples with those of the complete sample.

• **Stacked bars**: this option is only checked if a column of sub-samples has been selected. These charts are used to compare the frequencies or relative frequencies of sub-samples with those of the complete sample.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.

- **Cancel**: Click this button to close the dialog box without doing any computation.

- **Help**: Click this button to display the help.

- ![Reload Default Options](image): Click this button to reload the default options.

- ![Delete Data Selections](image): Click this button to delete the data selections.
Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

**Causes:** Select a column (or a row in row mode) of qualitative data that represent the list of causes you want to calculate descriptive statistics for.

**Frequencies:** Check this option, if your data is already aggregated in a list of causes and a corresponding list of frequencies of these causes. Select here the list of frequencies that correspond to the selected list of causes.

**Sub-sample:** Check this option to select a column showing the names or indexes of the sub-samples for each of the observations.

**Range:** Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Check this option to display the results in a new worksheet in the active workbook.

**Workbook:** Check this option to display the results in a new workbook.

**Sample labels:** Check this option if the first line of the selections (qualitative date, sub-samples, and weights) contains a label.

**Weights:** Check this option if the observations are weighted. If you do not check this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Sample labels" option is activated.

- **Standardize the weights:** if you check this option, the weights are standardized such that their sum equals the number of observations.

**Options** tab:

**Descriptive statistics:** Check this option to calculate and display descriptive statistics.

**Charts:** Check this option to display the charts.
**Compare to total sample:** this option is only checked if a column of sub-samples has been selected. Check this option so that the descriptive statistics and charts are also displayed for the total sample.

**Sort up:** Check this option to sort the data upwards.

**Combine categories:** Select the option that determine if and how categories of the qualitative data should be combined.

- **None:** Choose this option to not combine any categories.
- **Frequency less than:** Choose this option to combine categories having a frequency smaller that the user defined value.
- **% smaller than:** Choose this option to combine categories having a % smaller that the user defined value.
- **Smallest categories:** Choose this option to combine the m smallest categories. The value m is defined by the user.
- **Cumulated %:** Choose this option to combine all categories, as soon as the cumulative % of the Pareto plot is bigger than the user defined value.

**Outputs** tab:

**Qualitative data:** Activate the options for the descriptive statistics you want to calculate. The various statistics are described in the description section.

- **All:** Click this button to select all.
- **None:** Click this button to deselect all.
- **Display vertically:** Check this option so that the table of descriptive statistics is displayed vertically (one line per descriptive statistic).

**Charts** tab:

**Bar charts:** Check this option to represent the frequencies or relative frequencies of the various categories of qualitative variables as bars.

**Pie charts:** Check this option to represent the frequencies or relative frequencies of the various categories of qualitative variables as pie charts.
• **Doubles**: this option is only checked if a column of sub-samples has been selected. These charts are used to compare the frequencies or relative frequencies of sub-samples with those of the complete sample.

**Doughnuts**: this option is only checked if a column of sub-samples has been selected. These charts are used to compare the frequencies or relative frequencies of sub-samples with those of the complete sample.

**Stacked bars**: this option is only checked if a column of sub-samples has been selected. These charts are used to compare the frequencies or relative frequencies of sub-samples with those of the complete sample.

**Values used**: choose the type of data to be displayed:

- **Frequencies**: choose this option to make the scale of the plots correspond to the frequencies of the categories.

- **Relative frequencies**: choose this option to make the scale of the plots correspond to the relative frequencies of the categories.

**Example**

An example showing how to create Pareto charts is available on the Addinsoft website:


**References**


Gage R&R for quantitative variables (Measurement System Analysis)

Use this tool to control and validate your measurement method and measurement systems, in the case where you have several quantitative measures taken by one or more operators on several parts.

Description

Measurement System Analysis (MSA) or Gage R&R (Gage Repeatability and Reproducibility) is a method to control and judge a measurement process. It is useful to determine which sources are responsible for the variation of the measurement data. Variability can be caused by the measurement system, the operator or the parts. Gage R&R applied to quantitative measurements is based on two common methods: ANOVA and R control charts.

The word “gage” (or gauge) refers to the fact that the methodology is aimed at validating instruments or measurement methods.

A measurement is “repeatable” if the measures taken by a given operator for the same object (product, unit, part, or sample, depending of the field of application) repeatedly, do not vary above a given threshold. If the repeatability of a measurement system is not satisfactory, one should question the quality of the measurement system, or train the operators that do not obtain repeatable results if the measurement system does not appear to be responsible for the high variability.

A measurement is “reproducible” if the measures obtained for a given object (product, unit, part, or sample, depending of the field of application) by several operators do not vary above a given threshold. If the reproducibility of a measurement system is not satisfactory, one should train the operators so that their results are more homogeneous.

The goal of a Gage R&R analysis is to identify the sources of variability and to take the necessary actions to reduce them if necessary.

When the measures are quantitative data, two alternative methods are available for Gage R&R analysis. This first is based on analysis of variance (ANOVA) and on R control charts (Range and average).

In the descriptions below, $\hat{\sigma}^2_{\text{Repeatability}}$ stands for the variance corresponding to repeatability. The lower it is, the more repeatable the measurement (an operator gives coherent results for a given part). Its computation is different for the ANOVA and for the R control charts.

$\hat{\sigma}^2_{\text{Reproducibility}}$ is the fraction of the total variance that corresponds to reproducibility. The lower it is, the more reproducible the measurement (the various operators given consistent measurements for a given part). Its computation is different for the ANOVA and for the R control charts.
\( \sigma_{R&R}^2 \) is the variance of the gage R&R. The computation is always the sum of the two previous variances \( \sigma_{R&R}^2 = \sigma_{\text{Repeatability}}^2 + \sigma_{\text{Reproducibility}}^2 \).

**ANOVA**

When the ANOVA model is used in R&R analysis, one can statistically test whether the variability of the measures is related to the operators, and/or to the parts being measured themselves, and/or to an interaction between both (some operators might give for some parts significantly higher or lower measures), or not. Two designs are available when doing gage R&R analysis: the crossed design (balanced) and the nested design. XLSTAT includes both.

**Crossed design:**

A balanced ANOVA with the two factors Operator and Part is carried out. You can choose between a reduced ANOVA model that involves only the main factors, or a full model that includes the interaction term as well (Part*Operator).

For a crossed ANOVA, the data must satisfy the needs of a balanced ANOVA. That means that for a given factor, you have equal frequencies for all categories, and each operator must have measured each part. In the case of a full ANOVA, the F statistics are calculated as follows:

\[
F_{\text{operator}} = \frac{MSE_{\text{operator}}}{MSE_{\text{part} \times \text{operator}}}
\]

\[
F_{\text{part}} = \frac{MSE_{\text{part}}}{MSE_{\text{part} \times \text{operator}}}
\]

where MSE stands for mean squared error.

If the p-Value of the interaction Operator*Part is bigger or equal to the user defined threshold (usually 25 %), the interaction term is removed from the model. We then have a reduced model.

In the case of a crossed ANOVA with interaction, the variances are defined as follows:

\[
\hat{\sigma}^2 = MSE_{\text{Error}}
\]

\[
\hat{\sigma}_{\text{part} \times \text{operator}}^2 = \frac{(MSE_{\text{part} \times \text{operator}} - MSE_{\text{Error}})}{n\text{Rep}}
\]

\[
\hat{\sigma}_{\text{Operator}}^2 = \frac{(MSE_{\text{Operator}} - MSE_{\text{part} \times \text{operator}})}{(n\text{Part} \cdot n\text{Rep})}
\]

\[
\hat{\sigma}_{\text{Part}}^2 = \frac{(MSE_{\text{Part}} - MSE_{\text{part} \times \text{operator}})}{(n\text{Operator} \cdot n\text{Rep})}
\]
\[ \hat{\sigma}^2_{\text{Repeatability}} = \hat{\sigma}^2 \]
\[ \hat{\sigma}^2_{\text{Reproducibility}} = \hat{\sigma}^2_{\text{Operator}} + \hat{\sigma}^2_{\text{part*operator}} \]
\[ \hat{\sigma}^2_{\text{R&R}} = \hat{\sigma}^2_{\text{Reproducibility}} + \hat{\sigma}^2_{\text{Repeatability}} \]

In the case of a reduced model (without interaction), the variances are defined as follows:

\[ \hat{\sigma}^2 = \text{MSE}_{\text{Error}} \]
\[ \hat{\sigma}^2_{\text{part*operator}} = 0 \]
\[ \hat{\sigma}^2_{\text{Operator}} = \left( \text{MSE}_{\text{Operator}} / (n\text{Part} \cdot n\text{Rep}) \right) \]
\[ \hat{\sigma}^2_{\text{Part}} = \left( \text{MSE}_{\text{Part}} / (n\text{Operator} \cdot n\text{Rep}) \right) \]
\[ \hat{\sigma}^2_{\text{Repeatability}} = \hat{\sigma}^2 \]
\[ \hat{\sigma}^2_{\text{Reproducibility}} = \hat{\sigma}^2_{\text{Operator}} + \hat{\sigma}^2_{\text{part*operator}} \]
\[ \hat{\sigma}^2_{\text{R&R}} = \hat{\sigma}^2_{\text{Repeatability}} + \hat{\sigma}^2_{\text{Reproducibility}} \]

where MSE stands for mean squared error, nRep is the number of repetitions, nPart is the number of parts, and nOperator is the number of operators.

**Nested design:**

A nested ANOVA with the two factors Operator and Part(Operator) is carried out.

For a nested ANOVA, the data must satisfy the following prerequisites: for a given factor, you must have equal frequencies for all categories, and a part is checked by only one operator. The F statistics are calculated as follows:

\[ F_{\text{operator}} = \frac{\text{MSE}_{\text{operator}}}{\text{MSE}_{\text{part(operator)}}} \]
\[ F_{\text{part(operator)}} = \frac{\text{MSE}_{\text{part(operator)}}}{\text{MSE}_{\text{Error}}} \]

where MSE stands for mean squared error.

\[ \hat{\sigma}^2 = \text{MSE}_{\text{Error}} \]
\[ \hat{\sigma}^2_{\text{Repeatability}} = \hat{\sigma}^2 \]
\[
\hat{\sigma}_{\text{Repeatability}}^2 = \left( \text{MSE}_{\text{Operator}} - \text{MSE}_{\text{part(operator)}} \right) / (n\text{Part} \cdot n\text{Rep})
\]
\[
\hat{\sigma}_{R&R}^2 = \hat{\sigma}_{\text{Repeatability}}^2 + \hat{\sigma}_{\text{Reproducibility}}^2
\]

where MSE stands for mean squared error, nRep is the number of repetitions, nPart is the number of parts, and nOperator is the number of operators.

**R charts**

While less powerful than the ANOVA method, the Gage R&R analysis based on Range and Average analysis, is easy to compute and produces control charts (R charts). As the ANOVA method, it allows to compute the repeatability and the reproducibility of the measurement process. To use this method you need to have several parts, operators and repetitions (typically 10 parts, 3 operators, and 2 repetitions).

Based on the R chart, the different variances can be calculated as follows:

\[
\hat{\sigma}_{\text{Repeatability}}^2 = \frac{\bar{R}}{d_2^* (n\text{Rep}, n\text{Part} \cdot n\text{Operator})}
\]
\[
\hat{\sigma}_{\text{Reproducibility}}^2 = \left( \frac{\text{Max} (\mu_{\text{Part}}) - \text{Min} (\mu_{\text{Part}})}{d_2^* (n\text{Operator}, 1)} \right)^2 - \frac{\hat{\sigma}_{\text{Repeatability}}^2}{(n\text{Part} \cdot n\text{Operator})}
\]
\[
\hat{\sigma}_{R&R}^2 = \hat{\sigma}_{\text{Repeatability}}^2 + \hat{\sigma}_{\text{Reproducibility}}^2
\]
\[
\hat{\sigma}_{\text{Part}}^2 = \left( \frac{\text{Max} (\mu_{\text{Operator}}) - \text{Min} (\mu_{\text{Operator}})}{d_2^* (n\text{Part}, 1)} \right)^2
\]
\[
\hat{\sigma}^2 = \hat{\sigma}_{R&R}^2 + \hat{\sigma}_{\text{Part}}^2
\]

where Max(\mu Part respectively Operator)-Min(\mu Part respectively Operator) is the difference between the maximum and the minimum across operators (respectively parts) of the averages for each part (respectively operators), nRep is the number of repetitions, nPart is the number of parts, nOperator is the number of operators and \(d_2^* (m, k)\) is the control chart constant according to Burr (1969).

During the computation of the repeatability, we see that the mean amplitude of the Range chart is used. The variability of the parts and the reproducibility are based on the mean values of the X bar chart.
Indicators

XLSTAT offers several indicators derived from the variances to describe the measurement system.

The study variation for the different sources is calculated as product of the corresponding standard deviation of the source and the used defined factor $k$ Sigma:

$$\text{Study variation} = k \cdot \hat{\sigma}$$

The tolerance in percent is defined as the ratio of the variance in the study and the user defined tolerance:

$$\% \text{ tolerance} = \frac{\text{Study variation}}{\text{tolerance}}$$

The process sigma in percent is defined as ratio of the standard deviation of the source and the user defined historic process sigma:

$$\% \text{ process} = \frac{\text{standard deviation of the source}}{\text{process sigma}}$$

Precision to tolerance ratio (P/T):

$$\frac{P}{T} = \frac{k \cdot \hat{\sigma}_{R&R}^2}{\text{tolerance}}$$

Rho P (Rho Part):

$$\rho_{\text{Part}} = \frac{\hat{\sigma}_{\text{Part}}^2}{\hat{\sigma}^2}$$

Rho M:

$$\rho_{\text{M}} = \frac{\hat{\sigma}_{R&R}^2}{\hat{\sigma}^2}$$

Signal to noise ratio (SNR):

$$\text{SNR} = \sqrt{\frac{2 \rho_{\text{Part}}}{1 - \rho_{\text{Part}}}}$$

Discrimination ratio (DR):

$$\text{DR} = \frac{1 + \rho_{\text{Part}}}{1 - \rho_{\text{Part}}}$$

Bias:

$$\text{Bias} = \mu_{\text{Measurements}} - \text{target}$$
Bias in percent:

\[
\text{Bias \%} = \left( \frac{\mu_{\text{Measurements}} - \text{target}}{\text{tolerance}} \right)
\]

Resolution:

\[
\text{Resolution} = \text{Bias} + 3 \times \hat{\sigma}_{R&R}^2
\]

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.

- Click this button to reload the default options.
- Click this button to delete the data selections.

Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab:**

**Y / Measurement**: Choose the unique column or row that contains all the data. The assignment of the data to their corresponding subgroup must be specified using the Operator and the Parts field.

**X / Operator**: Select the data that identify for each element of the data selection the corresponding operator.

**Parts**: Select the data that identify for each element of the data selection the corresponding part.

**Method**: Choose the method to be used:
- **ANOVA**: Activate this option, to calculate variances based on an ANOVA analysis.
- **R chart**: Activate this option, to calculate variances based on an R chart.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Column/Row labels**: Activate this option if the first row (column mode) or column (row mode) of the data selections contains a label.

**Variable-category labels**: Activate this option to display in the results the categories in the form of variable name – category name.

**Sort categories alphabetically**: Activate this option to sort the categories of the variables in alphabetic order.

**Options** tab:

- **k Sigma**: Enter the user defined dispersion. Default value is 6.

- **Tolerance interval**: Activate this option to define the amplitude of the tolerance interval (also USL – LSL).

- **Sigma**: Activate this option to enter a value for the standard deviation of the control chart. This value should be based on historical data.

- **Target**: Activate this option to add the reference value of the measurements.

**ANOVA**: Choose the ANOVA model that should be used for the analysis:

- **reduced**
- **crossed**
  - **Significance level (%)**: Enter the threshold below which the interaction of the crossed model should be taken into account. Default value is 5.
- **nested**
**Estimation** tab:

**Method for Sigma:** Select the method for estimating the standard deviation of the control chart (see the description for further details):

- Pooled standard deviation
- R-bar
- S-bar

**Outputs** tab:

**Variance components:** Activate this option to show the table that displays the various variance components.

**Status indicator:** Activate this option to display the status indicators for the assessment of the measurement system.

**Analysis of variance:** Activate this option to display the variance analysis table.

**Display zones:** Activate this option to display, beside the lower and upper control limit, the limits of the A and B zones.

**Charts** tab:

**Display charts:** Activate this option to display the control charts graphically.

**Continuous line:** Activate this option to connect the points on the control chart.

**Needles view:** Activate this option to display for each point of the control chart, the minimum and maximum of the corresponding subgroup.

**Box view:** Activate this option to display the control charts using bars.

**Connect through missing:** Activate this option to connect the points, even when missing values separate the points.

**Box plots:** Check this option to display box plots (or box-and-whisker plots). See the description section of the univariate plots for more details.

**Scattergrams:** Check this option to display scattergrams. The mean (red +) and the median (red line) are always displayed.

**Means charts:** Activate this option to display the charts used to display the means of the various categories of the various factors.
• **Minimum/Maximum**: Check this option to systematically display the points corresponding to the minimum and maximum (box plots).

• **Outliers**: Check this option to display the points corresponding to outliers (box plots) with a hollowed-out circle.

• **Label position**: Select the position where the labels have to be placed on the box plots and scattergrams plots.

## Results

**Variance components:**

The first table and the corresponding chart display the variance split into its different sources. The contributions to the total variance and to the variance in the study, which is calculated using the user defined dispersion value, are given afterwards.

If a tolerance interval was defined, then the distribution of the variance by the variance according to the tolerance interval is displayed as well.

If a process sigma has been defined, then the distribution of the variance by the variance according to the process sigma is displayed as well.

The next table shows a detailed distribution of the variance by the different sources. Absolute values of the variance components and the percentage of the total variance are displayed.

The third table shows the distribution of the standard deviation for the different sources. It displays the absolute values of the variance components, the study variation that is calculated as the product of the standard deviation and the dispersion, the percentage of the study variation, the tolerance variability, which is defined as the ratio between variability of the study and the process sigma, and the percentage of the process variability.

**Status indicator:**

The first table shows information for the assessment of the measurement system. The Precision to tolerance ratio (P/T), Rho P, Rho M, Signal to noise ratio (SNR), Discrimination ratio (DR), absolute bias, and percentage and the resolution are displayed. The definition of the different indicators is given in the section description.

P/T values have the following status:
"more than adequate" if P/T <= 0.1
"adequate" if 0.1 < P/T <= 0.3
"not adequate" if P/T > 0.3

SNR values have the following status:

"not acceptable" if SNR < 2
"not adequate" if 2 <= SNR <= 5
"adequate" if SNR > 5

**Goodness of fit statistics:** The statistics relating to the fitting of the regression model are shown in this table:

- **Observations:** The number of observations used in the calculations. In the formulas shown below, \( n \) is the number of observations.

- **Sum of weights:** The sum of the weights of the observations used in the calculations. In the formulas shown below, \( W \) is the sum of the weights.

- **DF:** The number of degrees of freedom for the chosen model (corresponding to the error part).

- **R²:** The determination coefficient for the model. This coefficient, whose value is between 0 and 1, is only displayed if the constant of the model has not been fixed by the user. Its value is defined by:

  \[
  R^2 = 1 - \frac{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} w_i (y_i - \bar{y})^2}, \quad \text{where} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} w_i y_i,
  \]

  The \( R^2 \) is interpreted as the proportion of the variability of the dependent variable explained by the model. The nearer \( R^2 \) is to 1, the better is the model. The problem with the \( R^2 \) is that it does not take into account the number of variables used to fit the model.

- **Adjusted R²:** The adjusted determination coefficient for the model. The adjusted \( R^2 \) can be negative if the \( R^2 \) is near to zero. This coefficient is only calculated if the constant of the model has not been fixed by the user. Its value is defined by:

  \[
  \hat{R}^2 = 1 - (1 - R^2) \frac{W - 1}{W - p - 1}
  \]
The adjusted $R^2$ is a correction to the $R^2$ which takes into account the number of variables used in the model.

- **MSE**: The mean squared error (MSE) is defined by:

$$MSE = \frac{1}{W - p} \sum_{i=1}^{a} w_i \left( y_i - \hat{y}_i \right)^2$$

- **RMSE**: The root mean square of the errors (RMSE) is the square root of the MSE.

**Analysis of variance:**

The variance analysis table is used to evaluate the explanatory power of the explanatory variables. The explanatory power is evaluated by comparing the fit (as regards least squares) of the final model with the fit of the rudimentary model whose independent variable would be a constant equal to the mean.

**Chart information:**

The following results are displayed separately for each requested chart. Charts can be selected alone or in combination with the X bar chart.

**X bar/ R chart**: This table contains information about the center line and the upper and lower control limits of the selected chart. There will be one column for each phase.

**Observation details**: This table displays detailed information for each subgroup (a subgroup corresponds to a pair of Operator*Part). For each subgroup the corresponding phase, the size, the mean, the minimum and the maximum values, the center line, and the lower and upper control limits are displayed. If the information about the zones A, B and C are activated, then the lower and upper control limits of the zones A and B are displayed as well.

**X bar/ R chart**: If the charts are activated, then a chart containing the information of the two tables above is displayed. Each subgroup is displayed. The center line and the lower and upper control limits are displayed as well. If the corresponding options have been activated, the lower and upper control limits for the zones A and B are included and there are labels for the subgroups for which rules were fired. A legend with the activated rules and the corresponding rule number is displayed below the chart.

Finally the mean charts for each operator, for each part and for the interaction Operator*Part are displayed.
Example

A tutorial explaining how to use the Gage R&R tool is available on the Addinsoft web site. To consult the tutorial, please go to:

http://www.xlstat.com/demo-rrx.htm

References


Gage R&R for Attributes (Measurement System Analysis)

Use this tool to control and validate your measurement method and measurement systems, in the case where you have qualitative measurements (attributes) or ordinal quantitative measurements taken by one or more operators on several parts.

Description

Measurement System Analysis (MSA) or Gage R&R (Gage Repeatability and Reproducibility) is a method to control and judge a measurement process. It is useful to determine which sources are responsible for the variation of the measurement data. The word "gage" (or gauge) refers to the fact that the methodology is aimed at validating instruments or measurement methods.

In contrast to the Gage R&R for quantitative measurements, the analysis based on attributes gives information on the "agreement" and on the "correctness". The concepts of variance, repeatability and reproducibility are not relevant in this case.

A high "agreement" of the measures taken by a given operator for the same object (product, unit, part, or sample, depending of the field of application) repeatedly, shows that the operator is consistent. If the agreement of a measurement system is low, one should question the quality of the measurement system or protocol, or train the operators that do not obtain a high agreement, if the measurement system does not appear to be responsible for the lack of agreement.

A high "correctness" of the measures taken by an operator for the same object (product, unit, part, or sample, depending of the field of application) in comparison to the given reference or standard value shows that the operator comes to correct results. If the correctness of a measurement system is low, one should train the operators so that their results are more correct.

Correctness can be computed using the Kappa or the Kendall statistics. Kappa coefficients can be used in the case of qualitative and ordinal quantitative measurements. Kendall coefficients can be used in the case of ordinal measurements with at least 3 categories.

The two concepts "agreement" and "correctness" can be computed for a given operator, for a given operator compared to the standard, between two operators and for all operators compared to the standard.

The goal of a Gage R&R analysis for attributes is to identify the sources of low agreement and low correctness, and to take the necessary actions if necessary.

When the measures are qualitative or ordinal quantitative data, the Gage R&R analysis for attributes is based on the following statistics to evaluate the agreement and correctness:
Agreement statistics

Disagreement statistics

Kappa coefficients

Kendall coefficients

If possible, the following comparisons are performed:

- Intra rater
- Operator vs. standard
- Inter rater
- All Operators vs. standard

The standard corresponds to the measurements reported by an expert or a method that is considered as highly reliable.

**Agreement statistics**

It is possible to calculate these statistics in all of the sections.

In the intra rater section, XLSTAT computes for each operator the number of cases where he agrees with himself for a given part across repetitions. Additionally the ratio of the number of cases and the total number of inspections of the operator is computed.

In the Operator vs. standard section, XLSTAT gives the number of cases where an operator agrees with the standard across repetitions. Additionally the ratio of the number of cases and the total number of inspections of the operator is computed.

In the inter rater section, XLSTAT computes the number of cases where all operators agree for a given part and across repetitions. Additionally the ratio of the number of cases and the total number of inspections of all the operators is computed.

In the all operators vs. standard section, XLSTAT computes the number of cases where all operators agree with the standard, across all repetitions. Additionally the ratio of the number of cases and the total number of inspections of all the operators is computed.

In addition, confidence intervals are calculated. For proportions, XLSTAT allows you to use the simple (Wald, 1939) or adjusted (Agresti and Coull, 1998) Wald intervals, a calculation based on the Wilson score (Wilson, 1927), possibly with a correction of continuity, or the Clopper-Pearson (1934) intervals. Agresti and Caffo recommend using the adjusted Wald interval or the Wilson score intervals.
**Disagreement statistics**

This statistic is only calculated in the Operator vs. standard section in the case where the measurement variable is binary (for example, success or failure). Three different kinds of disagreements statistics are calculated for each operator:

- **False Negatives**: This statistic counts the number of cases where a given operator systematically evaluates a part as category 0 while the standard evaluates it as category 1. Additionally, the proportion of false negatives across all parts of category 0 is displayed.

- **False Positives**: This statistic counts the number of cases where a given operator systematically evaluates a part as category 1 while the standard evaluates it as category 0. Additionally, the proportion of false positive across all parts of category 1 is displayed.

- **Mixed**: This statistic counts the number of cases where an operator will be inconsistent in the rating of a given part across repetitions. The proportion of such cases computed as the ratio between Mixed and the total number of parts is displayed.

**Kappa coefficients**

Cohen's and Fleiss Kappa are well suited for qualitative variables. These coefficients are calculated on contingency tables obtained from paired samples. The Fleiss’ kappa is a generalization of the Cohen’s kappa. The kappa coefficient varies between -1 and 1. The closer the kappa is to 1, the higher the association.

In the case of an intra rater analysis, it is necessary that 2 or more measures have been taken by an operator for a given part.

In the case of operator vs. standard, the number of the measures for each operator must be the same as the number of measures for the standard.

In the case of inter rater, the number of the investigations for the two operators being compared must be the same.

In the case of all operators vs. standard the number of investigations for each operator for a given part has to be the same.

**Kendall coefficients**

These indicators are available for ordinal quantitative variables with at least 3 categories.

Kendall’s tau: This coefficient, also referred to as tau-b, allows to measure on a -1 to 1 scale the degree of concordance between two ordinal variables.

The Kendall’s coefficient of concordance: This coefficient measures on a 0 (no agreement) to 1 (perfect agreement) scale the degree of concordance between two ordinal variables.

The coefficients are computed to evaluate the measurement system by comparing each operator to the standard, operators between each other, and all operators vs. standard
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

: Click this button to reload the default options.

: Click this button to delete the data selections.

: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Y / Measurement: Choose the unique column or row that contains all the data. The assignment of the data to their corresponding subgroup must be specified using the Operator and the Parts field.

Data Type: Choose the data type:

- **Ordinal**: Activate this option if the measurement data is ordinal.
- **Nominal**: Activate this option if the measurement data is nominal.

X / Operator: Select the data that identify for each element of the data selection the corresponding operator.

Parts: Select the data that identify for each element of the data selection the corresponding part.
**Reference**: Activate this option, if reference or standard values are available. Select the data that indicate for each measurement the reference values.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Column/Row labels**: Activate this option if the first row (column mode) or column (row mode) of the data selections contains a label.

**Variable-category labels**: Activate this option to display in the results the categories in the form of variable name – category name.

**Sort categories alphabetically**: Activate this option to sort the categories of the variables in alphabetic order.

**Options** tab:

**Confidence intervals**:

- **Size (%)**: Enter the size of the confidence interval in % (default value: 95).

- **Wald**: Activate this option if you want to calculate confidence intervals on the various indexes using the approximation of the binomial distribution by the normal distribution. Activate "Adjusted" to use the adjustment of Agresti and Coull.

- **Wilson score**: Activate this option if you want to calculate confidence intervals on the various indexes using the Wilson score approximation.

- **Clopper-Pearson**: Activate this option if you want to calculate confidence intervals on the various indexes using the Clopper-Pearson approximation.

- **Continuity correction**: Activate this option if you want to apply the continuity correction to the Wilson score and to the interval on ratios.

**Kappa**:

- **Fleiss' Kappa**

- **Cohen's Kappa**
**Outputs** tab:

**Agreement:** Activate this option to display the tables with the agreement statistics.

**Disagreement:** Activate this option to display the tables with the disagreement statistics.

**Kappa:** Activate this option to display the tables with the Kappa statistics.

**Kendall:** Activate this option to display the tables with the Kendall statistics.

**Charts** tab:

**Charts:** Activate this option to display the charts that show the mean values and their corresponding confidence intervals for the agreement statistics.

**Results**

The tables with the selected statistics will be displayed.

The results are divided into the following four sections:

- Intra rater
- Operator vs. standard
- Inter rater and
- All Operators vs. standard

Within each section, the following indicators are displayed, as far as the calculation is wanted and possible:

- agreement statistics
- disagreement statistics
- Kappa statistics
- Kendall statistics
References


Screening designs

Use this module to generate a design to analyze the effect of 2 to 35 factors on one or more responses. This family of screening design is used to find the most influencing factors out of all the studied factors.

Description

The family of screening designs aims for the study of the effect of two or more factors. In general, factorial designs are the most efficient for this type of study. But the number of necessary tests is often too large when using factorial designs. There are other possible types of designs in order to take into account the limited number of experiments that can be carried out.

This tool integrates a large base of several hundred orthogonal design tables. Orthogonal design tables are preferred, as the ANOVA analysis will be based on a balanced design. Designs that are close to the design described by user input will be available for selection without having to calculate for an optimal design. All existing orthogonal designs are available for up to 35 factors having each between 2 and 7 categories. Most common families like full factorial designs, Latin square and Placket and Burman designs are included.

If the existing orthogonal designs in the knowledge base do not satisfy your needs, it is possible to search for D-Optimal designs. However, these designs might not be orthogonal.

Model

This tool generates designs that can be analyzed using an additive model without interactions for the estimation of the mean factor effects. If \( p \) is the number of factors, the ANOVA model is written as follows

\[
y_i = \beta_0 + \sum_{j=1}^{p} \beta_{k(i,j),j} + \epsilon_i
\]  

Common designs

When starting the creation of an experimental design, the internal knowledge base is searched for common orthogonal designs that are close to the problem. A distance measure \( d \) between your problem and each common design is calculated in the following way:

\( p_i = \text{number of factors with } i \text{ categories in the problem} \)

\( c_i = \text{number of factors with } i \text{ categories in the common design} \)

\( p_{\text{exp}} = \text{number of experiments in the problem} \)
\( c_{\text{exp}} = \text{number of experiments in the common design} \)

\[
d(c, p) = \sum_{i=2}^{2} |c_i - p_i| + c_{\text{exp}} - p_{\text{exp}} \quad (1)
\]

All common designs having the same number of factors as the problem and having a distance \( d \) smaller than 20 are proposed in a selection list.

The formal name for common designs is written in the two following ways:

\( \text{Ln} \ (p_1^{c_1} .. p_m^{c_m}) \) or \( \text{Ln} \ (p_1^{c_1} .. p_m^{c_m}) \)

Where

\( n = \text{number of experiments} \)
\( c_i = \text{number of categories of the group of factors } p_i \)
\( p_i = \text{number of factors having } c_i \text{ categories} \)

A common name for each design is displayed in the list if available.

**Optimization**

This tool implements an exchange algorithm with 3 excursions to search for \( d \)-optimal designs.

The internal representation of the design matrix uses the following encoding. For a factor \( f_i \) having \( c_i \) categories, \( c_i - 1 \) columns \( k_1 .. k_{c_i-1} \) are added in the design matrix \( X \) in the following way for the different category values of \( f_i \):

\[
\begin{array}{c|ccc}
\text{f}_i & k_{c_i,1} & .. & k_2 & k_1 \\
1 & -1 & .. & -1 & -1 \\
2 & 0 & .. & 0 & 1 \\
3 & 0 & .. & 1 & 0 \\
\end{array}
\]

\( c_i \)

The complete design matrix \( X \) is composed of \( n \) lines, where \( n \) is the number of experiences. The matrix contains a first column with 1 in each line and \( c_i - 1 \) columns for each factor \( f_i \) in the design, where \( c_i \) is the number of categories of the corresponding factor \( f_i \).

\( X \) is the encoded design matrix, where every line represents the encoded experiment corresponding to the experimental design.
The criterion used for the optimization is defined as:

\[ c = \log_{10}(\det(X'X)) \]  

(2)

With

\[ X'X = \text{information matrix} \]

\[ X = \text{encoded design matrix} \]

This criterion is named in the results as follows:

\[ c = \log(|I|) \]

The following common used criterion is also displayed in the results:

\[ \log(|I|^{1/p}) \]

When comparing experimental designs that have a different number of experiences, the normalized log is used to be able to compare the different criteria values:

\[ \text{Norm. log} = \log_{10}\left(\frac{1}{N} \left(\det \left(\frac{1}{N} (X'X)\right)^{1/p}\right)\right) \]  

(3)

This criterion is named in the results as follows:

\[ \text{Norm. log} = \log(\left|\frac{1}{n}*I\right|^{1/p}) \]

This measure allows comparing the optimality of different experimental designs, even if the number of experiences is different.

The implemented algorithm offers 3 different starting options:

**Random:** A valid initial partition is generated using random numbers.

**Simultaneous:** A small number of experiences \((n = 5)\) is generated by random. The rest of the initial partition is added maximizing the optimization criteria of the exchange algorithm.

**User defined:** The user selects the initial partition to be used.

In the first two cases a number of repetitions should be selected in order to find a good local optimum.

**Output**

This tool will provide a new design for testing. Optional experiment sheets for each individual test might be generated on separated Excel sheets for printing. After having carried out the
experiments, complete the corresponding cells in the created experimental design in the corresponding Excel sheet.

A hidden sheet with important information about the design is included in your Excel file in order to have all necessary information for the XLSTAT analysis for screening designs ready. In this way incorrect analysis of an experimental design is inhibited. Therefore please carry out your analysis of your experimental design in the same Excel workbook where you created the design itself.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- : Click this button to reload the default options.
- : Click this button to delete the data selections.
- : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab:**

- **Model name**: Choose a short model name for the design. This name will be used for the name of the Excel sheets and during the selections of the analysis to create the link between the design and the analysis of the model.

- **Number of factors**: Choose the number of factors to be studied in the design. The possible range is between 2 and 35 factors.
**Minimum number of experiments**: Enter the minimum number of experiments to be carried out during the experimental design.

**Maximum number of experiments**: Enter the maximum number of experiments to be carried out during the experimental design.

**Number of responses**: Enter the number of responses that you want to analyze with the design.

**Repetitions**: Activate this option to choose the number of repetitions of the design.

**Randomize**: Activate this option to change the order of the lines of the design into a random order.

**Print experiment sheets**: Activate this option in order to generate for each individual experiment a separate Excel sheet with information about the experiment. This can be useful when printed out for the realization of the experiment.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Options** tab:

**Method**: Choose the method you want to use to generate the design.

- **Automatic**: This method allows to search for an optimal design.
  - **Initial partition**: Choose how the initial partition is generated. The available methods are random and simultaneous, and user defined. In the latter case, you must select the design of experiments that will be used to start the search for the optimal design.
  - **Repetitions**: In the case of a random initial partition, enter the number of the repetitions to perform.
  - **Initial design**: In the case of a user defined initial partition, select the range in the Excel sheet that contains the initial design. The header line with the factor names have to be included in the selection.
  - **Stop conditions**:
• **Iterations**: Enter the maximum number of iterations for the algorithm. The calculations are stopped when the maximum number of iterations has been exceeded. Default value: 50.

• **Convergence**: Enter the maximum value of the evolution in the criterion from one iteration to another which, when reached, means that the algorithm is considered to have converged. Default value: 0.0001.

Note: this method can cause long calculation times as the total number of models explored is equal to the number of combinations \( C(n,k) = \frac{n!}{(n-k)!k!} \), where \( n \) is the number of experiments of the full experimental design and \( k \) the maximum number of experiments to include in the design. It is recommended to gradually increase the value of \( k \), the maximum number of experiments in the design.

• **Common designs**: Choose this option to select one of the available common designs.

**Factors** tab:

**Selection**: Select one of the two following options to determine the selection mode for this window:

• **Manual selection**: All information about the factors will be inserted directly into the text fields of the window.

• **Sheet selection**: All information about the factors will be selected as ranges in the Excel sheet. In this case a column with as much entries as the number of factors is expected.

**Short name**: Enter a short name for the factors composed of some characters. If manual selection is activated, there is a text field for each factor. If sheet selection is activated, please choose a range in the Excel sheet that contains one field with a value for each factor. The order of the different factors must be the same for all the selections in this window. Header lines must not be included in the selection. The first row of the selected range must contain data values.

**Long name**: Enter a long name for the factors composed of some characters. If manual selection is activated, there is a text field for each factor. If sheet selection is activated, please choose a range in the Excel sheet that contains one field with a value for each factor. The order of the different factors must be the same for all the selections in this window. Header lines must not be included in the selection. The first row of the selected range must contain data values.

**Unit**: Enter a description of the unit of the factors. If manual selection is activated, there is a text field for each factor. If sheet selection is activated, please choose a range in the Excel sheet that contains one field with a value for each factor. The order of the different factors must be the same for all the selections in this window. Header lines must not be included in the selection. The first row of the selected range must contain data values.
be the same for all the selections in this window. Header lines must not be included in the
selection. The first row of the selected range must contain data values.

**Unit (symbol):** Enter the physical unit of the factors. If manual selection is activated, there is a
text field for each factor. If sheet selection is activated, please choose a range in the Excel
sheet that contains one field with a value for each factor. The order of the different factors must
be the same for all the selections in this window. Header lines must not be included in the
selection. The first row of the selected range must contain data values.

**Number of categories:** Enter the number of categories of the factors. If manual selection is
activated, there is a text field for each factor. If sheet selection is activated, please choose a
range in the Excel sheet that contains one field with a value for each factor. The order of the
different factors must be the same for all the selections in this window. Header lines must not
be included in the selection. The first row of the selected range must contain data values.

**Category labels:** Activate this option, if you have labels of the categories available. Select
columns with a list of labels of the categories in the Excel sheet. If manual selection is
activated, there is a text field for each factor. If sheet selection is activated, please choose a
range in the Excel sheet that contains one field with a value for each factor. The order of the
different factors must be the same for all the selections in this window. Header lines must not
be included in the selection. The first row of the selected range must contain data values.

**Responses tab:**

**Selection:** Select one of the two following options to determine the selection mode for this
window:

- **Manual selection:** All information about the responses will be inserted directly into the
text fields of the window.

- **Sheet selection:** All information about the responses will be selected as ranges in the
Excel sheet. In this case a column with as much entries as the number of factors is
expected.

**Short name:** Enter a short name for the responses composed of some characters. If manual
selection is activated, there is a text field for each response. If sheet selection is activated,
please choose a range in the Excel sheet that contains one field with a value for each
response. The order of the different responses must be the same for all the selections in this
window. Header lines must not be included in the selection. The first row of the selected range
must contain data values.

**Long name:** Enter a long name for the responses composed of some characters. If manual
selection is activated, there is a text field for each response. If sheet selection is activated,
please choose a range in the Excel sheet that contains one field with a value for each
response. The order of the different responses must be the same for all the selections in this
window. Header lines must not be included in the selection. The first row of the selected range must contain data values.

**Unit:** Enter a description of the unit of the responses. If manual selection is activated, there is a text field for each response. If sheet selection is activated, please choose a range in the Excel sheet that contains one field with a value for each response. The order of the different response must be the same for all the selections in this window. Header lines must not be included in the selection. The first row of the selected range must contain data values.

**Unit (symbol):** Enter the physical unit of the responses. If manual selection is activated, there is a text field for each response. If sheet selection is activated, please choose a range in the Excel sheet that contains one field with a value for each response. The order of the different response must be the same for all the selections in this window. Header lines must not be included in the selection. The first row of the selected range must contain data values.

**Outputs** tab:

- **Optimization summary:** Activate this option to display the optimization summary.
- **Details of iterations:** Activate this option to display the details of iterations.
- **Burt table:** Activate this option to display the Burt table of the experimental design.
- **Encoded design:** Activate this option to display the encoded experimental design in the case of a d-optimal design.

- **Sort up:** Activate this option to sort the categories in increasing order, the sort criterion being the value of the category. If this option is activated, the sort is ascending.
- **Sort the categories alphabetically:** Activate this option so that the categories of all the variables are sorted alphabetically.
- **Variable-Category labels:** Activate this option to use variable-category labels when displaying outputs. Variable-Category labels include the variable name as a prefix and the category name as a suffix.

**Charts** tab:

- **Evolution of the criterion:** Activate this option for the evolution chart of the chosen criterion.
- **3D view of the Burt table:** Activate this option to display a 3D visualization of the Burt table.

**Screening designs / Common designs** dialog box:
**Selection of experimental design:** This dialog box lets you select the design of experiment you want to use. Thus, a list of fractional factorial designs is presented with their respective distance to the design that was to be generated. If you select a design and you click Select, then the selected design will appear. If no design fits your needs, click on the "optimize" button, and an algorithm will give you a design corresponding exactly to the selected factors.

**Screening designs / optimal** dialog box:

**Selection of experimental design:** This dialog box lets you select the design of experiment you want to use. This dialog box is displayed, if the option "optimize" was selected, and if the minimum number of experiments is strictly less than the maximum number of experiments. Thus, a list of fractional factorial designs is presented with an optimal design for each number of experiments. The list contains for each design the number of experiments, the logarithm of the determinant of the information matrix and the normalized logarithm of that determinant. The histogram on the right displays the normalized logarithm for the designs, which are sorted in an ascending number of experiments from the left to the right. The selected design in the list on the left will appear red in the histogram on the right. If you select a design and you click Select, then the selected design will appear in your analysis.

**Results**

If an Optimization was selected, then the following sections are displayed:

The start and end time, and the duration of the optimization are displayed.

**Optimization summary:** If the minimum number of experiments is strictly inferior to the maximum number of experiments, then a table with information for each number of experiments is displayed. This table displays for each optimization run the number of experiments, the criterion log(determinant), the criterion norm. log(determinant) and the criterion Log(| I |^1/p). The best result is displayed in bold in the first line. The criterion norm. log(determinant) is shown in a chart.

**Statistics for each iteration:** This table shows for the selected experimental design the evolution of the criterion during the iterations of the optimization. If the corresponding option is activated in the Charts tab, a chart showing the evolution of the criterion is displayed.

Then a second table is displayed, if the minimum number of experiments is strictly inferior to the maximum number of experiments. This table displays for each optimization run the number of experiments, the number of iteration steps during the optimization, the criterion log(determinant) and the criterion norm. log(determinant). The best result is displayed in bold in the first line.
**Burt table**: The Burt table is displayed only if the corresponding option is activated in the dialog box. The 3D bar chart that follows is the graphical visualization of this table.

**Variables information**: This table shows the information about the factors. For each factor the short name, long name, unit and physical unit are displayed.

Then the **Model name** is displayed, in order to select this field as identification when performing the analysis of the generated design.

**Experimental design**: This table displays the complete experimental design. Additional columns include information on the factors and on the responses, a label for each experiment, the sort order, the run order and the repetition.

**Endoded design**: This table shows the encoded experimental design. This table is only displayed in the case of a d-optimal experimental design.

If the generation of **experiment sheets** was activated in the dialog box and if there are less than 200 experiments to be carried out, an experiment sheet is generated for each line of the experimental design on separate Excel sheets.

These sheets start with the report header of the experimental design and the model name to simplify the identification of the experimental design that this sheet belongs to. Then the running number of the experiment and the total number of experiments are displayed. The values of the additional columns of the experimental design, i.e. sort order, run order, and repetition are given for the experiment.

Last, the information on the experimental conditions of the factors is displayed with fields so that the user can enter the results obtained for the various responses. Short names, long names, units, physical units and values are displayed for each factor.

These sheets can be printed out or can be used in electronic format to assist during the realization of the experiments.

**Example**

A tutorial on the generation and analysis of a screening design is available on the Addinsoft website:

[http://www.xlstat.com/demo-doe1.htm](http://www.xlstat.com/demo-doe1.htm)

**References**


Analysis of a screening design

Use this tool to analyze a screening design of 2 to 35 factors and a user defined number of results. A linear model with or without interactions will be used for the analysis.

Description

Analysis of a screening design uses the same conceptual framework as linear regression and variance (ANOVA). The main difference comes from the nature of the underlying model. In ANOVA, explanatory variables are often called factors.

If \( p \) is the number of factors, the ANOVA model is written as follows:

\[
y_i = \beta_0 + \sum_{j=1}^{p} \beta_{k(i,j),j} + \varepsilon_i
\]  

(1)

where \( y_i \) is the value observed for the dependent variable for observation \( i \), \( k(i,j) \) is the index of the category of factor \( j \) for observation \( i \), and \( \varepsilon_i \) is the error of the model.

The hypotheses used in ANOVA are identical to those used in linear regression: the errors \( \varepsilon_i \) follow the same normal distribution \( N(0,\sigma) \) and are independent.

The way the model with this hypothesis added is written means that, within the framework of the linear regression model, the \( y_i \)s are the expression of random variables with mean \( \mu_i \) and variance \( \sigma^2 \), where

\[
\mu_i = \beta_0 + \sum_{j=1}^{p} \beta_{k(i,j),j}
\]

To use the various tests proposed in the results of linear regression, it is recommended to check retrospectively that the underlying hypotheses have been correctly verified. The normality of the residues can be checked by analyzing certain charts or by using a normality test. The independence of the residues can be checked by analyzing certain charts or by using the Durbin Watson test.

For more information on ANOVA and linear regression please consider the corresponding sections in the online help.

Balanced and unbalanced ANOVA
We talk of balanced ANOVA when for each factor (and interaction if available) the number of observations within each category is the same. When this is not true, the ANOVA is said to be unbalanced. XLSTAT can handle both cases.

If you are in a balanced or an unbalanced case of ANOVA depend on the experimental design you have chosen.

**Constraints**

During the calculations, each factor is broken down into a sub-matrix containing as many columns as there are categories in the factor. Typically, this is a full disjunctive table. Nevertheless, the breakdown poses a problem: if there are g categories, the rank of this sub-matrix is not g but g-1. This leads to the requirement to delete one of the columns of the sub-matrix and possibly to transform the other columns. Several strategies are available depending on the interpretation we want to make afterwards:

a1=0: the parameter for the first category is null. This choice allows us force the effect of the first category as a standard. In this case, the constant of the model is equal to the mean of the dependent variable for group 1.

Note: even if the choice of constraint influences the values of the parameters, it has no effect on the predicted values and on the different fitting statistics.

Note: The option a1=0 is always applied when using this module you cannot change this option.

**Multi-response and desirability**

In the case of many response values y1, ..., ym it is possible to optimize each response value individually and to create a combined desirability function and analyze its values. Proposed by Derringer and Suich (1980), this approach is to first convert each response yi into an individual desirability function di that varies over the range 0 <= di <= 1.

When yi has reached its target, then di = 1. If yi is outside an acceptable region around the target, then di = 0. Between these two extreme cases, intermediate values of di exist as shown below.

The 3 different optimization cases for di are present with the following definitions:

L = lower value. Every value smaller than L has di = 0

U = upper value. Every value bigger than U has di = 0.

T(L) = left target value.

T(R) = right target value. Every value between T(L) and T(R) has di = 1.

s, t = weighting parameters that define the shape of the optimization function between L and T(L) and T(R) and U.
The following equation has to be respected when defining $L$, $U$, $T(L)$ and $T(R)$:

$$L \leq T(L) \leq T(R) \leq U$$

**Maximize** the value of $y_i$:

$$d_i = \begin{cases} 
0 & y_i < L \\
\left(\frac{y_i - L}{T(L) - L}\right)^s & L \leq y_i \leq T(L) \\
1 & y_i > T(L)
\end{cases}$$

**Minimize** the value of $y_i$:

$$d_i = \begin{cases} 
1 & y_i < T(R) \\
\left(\frac{U - y_i}{U - T(R)}\right)^t & T(R) \leq y_i \leq U \\
0 & y_i > U
\end{cases}$$

Two sided desirability function as shown below to target a certain interval of $y_i$:
The design variables are chosen to maximize the overall desirability $D$

$$ D = \left( d_1^{w_1} \cdot d_2^{w_2} \cdots \cdot d_m^{w_m} \right)^{1/(w_1 + w_2 + \cdots + w_m)} $$

Where $1 \leq w_i \leq 10$ are weightings of the individual desirability functions. The bitter $w_i$, the more important is $d_i$ taken into account during the optimization.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
Click this button to reload the default options.

Click this button to delete the data selections.

Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

**Model name**: Select the corresponding cell in the Excel sheet with the generated design that you want to analyze. The Model name is used as part of the names of Excel sheets and during the selection of the analysis in order to make the link between the design and the analysis of the results of the design.

**Y / results**: Select the columns of the experimental design that contain the results. These columns should now hold the results of the experiments carried out. If several result variables have been selected, XLSTAT carries out calculations for each of the variables separately, and then an analysis of the desirability is carried out. If a column header has been selected, check that the "Variable labels" option has been activated.

**Experimental design**: Activate this option, if you made changes to the values of the generated experimental design. Then the changes will be shown in the results.

If you have the possibility to select the additional columns (the columns on the left of the factor columns of the generated experimental design) and the columns with factors of the experimental design and you want to select them for comparison with the original experimental design. It is important include into the selection the column with the sort order information. Using this option includes changes to the experimental design in the factor columns into the analysis. If this option is not activated, the experimental design at the moment of its generation is used for the analysis. The selected data has to be numerical. If a column header has been selected, check that the "Variable labels" option has been activated.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.
Variable labels: This option is always activated. The first row of the selected data (data and observation labels) must contain a label.

Sort up: Activate this option to sort the categories in increasing order, the sort criterion being the value of the category. If this option is activated, the sort is ascending.

Responses tab:

Selection: Select one of the two following options to determine the selection mode for this window:

- **Manual selection:** All information about the responses will be inserted directly into the text fields of the window.
- **Sheet selection:** All information about the responses will be selected as ranges in the Excel sheet. In this case a column with as much entries as the number of factors is expected.

Short name: Enter a short name for the responses composed of some characters. If manual selection is activated, there is a text field for each response. If sheet selection is activated, please choose a range in the Excel sheet that contains one field with a value for each response. The order of the different responses must be the same for all the selections in this window. Header lines must not be included in the selection. The first row of the selected range must contain data values.

Long name: Enter a long name for the responses composed of some characters. If manual selection is activated, there is a text field for each response. If sheet selection is activated, please choose a range in the Excel sheet that contains one field with a value for each response. The order of the different responses must be the same for all the selections in this window. Header lines must not be included in the selection. The first row of the selected range must contain data values.

Aim: Choose the aim of the optimization. You have the choice between Minimum, Optimum and Maximum. If manual selection is activated, there is a text field for each response. If sheet selection is activated, please choose a range in the Excel sheet that contains one field with a value for each response. The order of the different responses must be the same for all the selections in this window. Header lines must not be included in the selection. The first row of the selected range must contain data values.

If the selected aim is Optimum or Maximum, then the following two fields are activated.

Lower: Enter the value of the lower bound, below which the desirability is 0. If manual selection is activated, there is a text field for each response. If sheet selection is activated, please choose a range in the Excel sheet that contains one field with a value for each response. The order of the different responses must be the same for all the selections in this window. Header lines must not be included in the selection. The first row of the selected range must contain data values.
window. Header lines must not be included in the selection. The first row of the selected range must contain data values.

**Target (left):** Enter the value of the lower bound, above which the desirability is 1. The desirability function increases monotonously from 0 to 1 between the lower bound and the left target. If manual selection is activated, there is a text field for each response. If sheet selection is activated, please choose a range in the Excel sheet that contains one field with a value for each response. The order of the different responses must be the same for all the selections in this window. Header lines must not be included in the selection. The first row of the selected range must contain data values.

If the selected aim is Minimum or Optimum, then the following two fields are activated.

**Target (right):** Enter the value of the upper bound, below which the desirability is 1. The desirability function decreases monotonously from 1 to 0 between the right target and the upper bound. If manual selection is activated, there is a text field for each response. If sheet selection is activated, please choose a range in the Excel sheet that contains one field with a value for each response. The order of the different responses must be the same for all the selections in this window. Header lines must not be included in the selection. The first row of the selected range must contain data values.

**Upper:** Enter the value of the upper bound, above which the desirability is 0. If manual selection is activated, there is a text field for each response. If sheet selection is activated, please choose a range in the Excel sheet that contains one field with a value for each response. The order of the different responses must be the same for all the selections in this window. Header lines must not be included in the selection. The first row of the selected range must contain data values.

**s:** Activate this option, if the increasing desirability function should have a non linear shape. Enter the value of the shape parameter, which should be a value between 0.01 and 100. Examples are shown in the graphic in the window at the right. If manual selection is activated, there is a text field for each response. If sheet selection is activated, please choose a range in the Excel sheet that contains one field with a value for each response. The order of the different responses must be the same for all the selections in this window. Header lines must not be included in the selection. The first row of the selected range must contain data values.

**t:** Activate this option, if the decreasing desirability function should have a non linear shape. Enter the value of the shape parameter, which should be a value between 0.01 and 100. Examples are shown in the graphic in the window at the right. If manual selection is activated, there is a text field for each response. If sheet selection is activated, please choose a range in the Excel sheet that contains one field with a value for each response. The order of the different responses must be the same for all the selections in this window. Header lines must not be included in the selection. The first row of the selected range must contain data values.

**Weight:** Activate this option, if the responses should have an exponent different from 1 during the calculation of the desirability function. Enter the value of the shape parameter, which should be a value between 0.01 and 100. If manual selection is activated, there is a text field for each response. If sheet selection is activated, please choose a range in the Excel sheet
that contains one field with a value for each response. The order of the different responses
must be the same for all the selections in this window. Header lines must not be included in the
selection. The first row of the selected range must contain data values.

**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the variables
selected.

**Correlations**: Activate this option to display the correlation matrix for quantitative variables
(dependent or explanatory).

**Experimental design**: Activate this option to display the table with the experimental design.

**Analysis of variance**: Activate this option to display the analysis of variance table.

**Goodness of fit statistics**: Activate this option to display the table of goodness of fit statistics
for the model.

**Contribution**: Activate this option to display the contribution of the factors to the model.

**Standardized coefficients**: Activate this option if you want the standardized coefficients (beta
coefficients) for the model to be displayed.

**Predictions and residuals**: Activate this option to display the predictions and residuals for all
the observations.

- **Adjusted predictions**: Activate this option to calculate and display adjusted predictions
  in the table of predictions and residuals.

- **Cook’s D**: Activate this option to calculate and display Cook’s distances in the table of
  predictions and residuals.

**Charts** tab:

**Regression charts**: Activate this option to display regression chart:

- **Standardized coefficients**: Activate this option to display the standardized parameters
  for the model with their confidence interval on a chart.

- **Predictions and residuals**: Activate this option to display the following charts.

  (1) Line of regression: This chart is only displayed if there is only one explanatory
  variable and this variable is quantitative.
(2) Explanatory variable versus standardized residuals: This chart is only displayed if there is only one explanatory variable and this variable is quantitative.

(3) Dependent variable versus standardized residuals.

(4) Predictions for the dependent variable versus the dependent variable.

(5) Bar chart of standardized residuals.

- **Confidence intervals**: Activate this option to have confidence intervals displayed on charts (1) and (4).

**Pareto plots**: Activate this option, to display the chart that represents the contribution of the factors to the response in a Pareto plot.

**Means charts**: Activate this option to display the charts used to display the means of the various categories of the various factors.

### Results

**Descriptive statistics**: These tables show the simple statistics for all the variables selected. The number of observations, missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed. For qualitative explanatory variables the names of the various categories are displayed together with their respective frequencies.

**Variables information**: This table shows the information about the factors. For each factor the short name, long name, unit and physical unit are displayed.

Then the **Model name** is displayed, in order to select this field as identification later on during the analysis of the generated design.

**Experimental design**: In this table the complete experimental design is shown. There are additional columns, columns for the factors and columns for the responses displayed. The additional columns contain a label for each experiment, the sort order, the run order, the bloc number and the point type.

If changes were made to the values between the generation of the experimental design and the analysis, these values are displayed in bold.

After that the parameters of the desirability function are displayed, if there is more than one response present in the design. The table shows for each response the short name, long name, unit, physical unit, aim, lower bound, left target value, right target value, upper bound, shape parameters s and t and the weight parameter.

If **mean charts** have been requested, the corresponding results are then displayed.
Then for each response and the global desirability function, the following tables and charts are displayed.

**Goodness of fit statistics:** The statistics relating to the fitting of the regression model are shown in this table:

- **Observations**: The number of observations used in the calculations. In the formulas shown below, \( n \) is the number of observations.
- **Sum of weights**: The sum of the weights of the observations used in the calculations. In the formulas shown below, \( W \) is the sum of the weights.
- **DF**: The number of degrees of freedom for the chosen model (corresponding to the error part).
- **\( R^2 \)**: The determination coefficient for the model. This coefficient, whose value is between 0 and 1, is only displayed if the constant of the model has not been fixed by the user. Its value is defined by:

\[
R^2 = 1 - \frac{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} w_i (y_i - \bar{y})^2}, \quad \text{where } \bar{y} = \frac{1}{n} \sum_{i=1}^{n} w_i y_i,
\]

The \( R^2 \) is interpreted as the proportion of the variability of the dependent variable explained by the model. The nearer \( R^2 \) is to 1, the better is the model. The problem with the \( R^2 \) is that it does not take into account the number of variables used to fit the model.

- **Adjusted \( R^2 \)**: The adjusted determination coefficient for the model. The adjusted \( R^2 \) can be negative if the \( R^2 \) is near to zero. This coefficient is only calculated if the constant of the model has not been fixed by the user. Its value is defined by:

\[
\hat{R}^2 = 1 - (1 - R^2) \frac{W - 1}{W - p - 1}
\]

The adjusted \( R^2 \) is a correction to the \( R^2 \) which takes into account the number of variables used in the model.

- **MSE**: The mean squared error (MSE) is defined by:

\[
MSE = \frac{1}{W - p} \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2
\]

- **RMSE**: The root mean square of the errors (RMSE) is the square root of the MSE.

- **MAPE**: The *Mean Absolute Percentage Error* is calculated as follows:
\[ \text{MAPE} = \frac{100}{W} \sum_{i=1}^{n} w_i \left| \frac{y_i - \hat{y}_i}{y_i} \right| \]

- **DW**: The Durbin-Watson statistic is defined by:

\[
DW = \frac{\sum_{i=2}^{n} \left[ (y_i - \hat{y}_i) - (y_{i-1} - \hat{y}_{i-1}) \right]^2}{\sum_{i=1}^{n} w_i \left( y_i - \hat{y}_i \right)^2}
\]

This coefficient is the order 1 autocorrelation coefficient and is used to check that the residuals of the model are not autocorrelated, given that the independence of the residuals is one of the basic hypotheses of linear regression. The user can refer to a table of Durbin-Watson statistics to check if the independence hypothesis for the residuals is acceptable.

- **Cp**: Mallows Cp coefficient is defined by:

\[
Cp = \frac{\text{SSE}}{\hat{\sigma}} + 2p^* - W
\]

where SSE is the sum of the squares of the errors for the model with p explanatory variables and \( \hat{\sigma} \) is the estimator of the variance of the residuals for the model comprising all the explanatory variables. The nearer the Cp coefficient is to p*, the less the model is biased.

- **AIC**: Akaike’s Information Criterion is defined by:

\[
AIC = W \ln \left( \frac{\text{SSE}}{W} \right) + 2p^*
\]

This criterion, proposed by Akaike (1973) is derived from the information theory and uses Kullback and Leibler’s measurement (1951). It is a model selection criterion which penalizes models for which adding new explanatory variables does not supply sufficient information to the model, the information being measured through the MSE. The aim is to minimize the AIC criterion.

- **SBC**: Schwarz’s Bayesian Criterion is defined by:

\[
SBC = W \ln \left( \frac{\text{SSE}}{W} \right) + \ln(W) p^*
\]

This criterion, proposed by Schwarz (1978) is similar to the AIC, and the aim is to minimize it.

- **PC**: Amemiya’s Prediction Criterion is defined by:
\[ PC = \frac{(1 - R^2)(W + p \ast)}{W - p \ast} \]

This criterion, proposed by Amemiya (1980) is used, like the adjusted R² to take account of the parsimony of the model.

- **Press RMSE**: Press' statistic is only displayed if the corresponding option has been activated in the dialog box. It is defined by:

\[ \text{Press} = \sum_{i=1}^{n} w_i \left( y_i - \hat{y}_{i(-i)} \right)^2 \]

where \( \hat{y}_{i(-i)} \) is the prediction for observation i when the latter is not used for estimating parameters. We then get:

\[ \text{Press RMSE} = \sqrt{\frac{\text{Press}}{W - p \ast}} \]

Press's RMSE can then be compared to the RMSE. A large difference between the two shows that the model is sensitive to the presence or absence of certain observations in the model.

- **Q²**: The Q² statistic is displayed. It is defined as

\[ Q^2 = 1 - \frac{\text{PressRMSE}}{SSE} \]

The closer Q² is to 1, the better and more robust is the model.

The analysis of variance table is used to evaluate the explanatory power of the explanatory variables. Where the constant of the model is not set to a given value, the explanatory power is evaluated by comparing the fit (as regards least squares) of the final model with the fit of the rudimentary model including only a constant equal to the mean of the dependent variable. Where the constant of the model is set, the comparison is made with respect to the model for which the dependent variable is equal to the constant which has been set.

Then the contributions and the corresponding Pareto plot are displayed, if the corresponding option has been activated and all the factors are binary.

The equation of the model is then displayed to make it easier to read or re-use the model.

The table of standardized coefficients (also called beta coefficients) are used to compare the relative weights of the variables. The higher the absolute value of a coefficient, the more important the weight of the corresponding variable. When the confidence interval around standardized coefficients has value 0 (this can be easily seen on the chart of standardized coefficients), the weight of a variable in the model is not significant.
The **predictions and residuals** table shows, for each observation, its weight, the value of the qualitative explanatory variable, if there is only one, the observed value of the dependent variable, the model's prediction, the residuals, the confidence intervals together with the fitted prediction and Cook's D if the corresponding options have been activated in the dialog box. Two types of confidence interval are displayed: a confidence interval around the mean (corresponding to the case where the prediction would be made for an infinite number of observations with a set of given values for the explanatory variables) and an interval around the isolated prediction (corresponding to the case of an isolated prediction for the values given for the explanatory variables). The second interval is always greater than the first, the random values being larger.

The **charts** which follow show the results mentioned above. If there is only one explanatory variable in the model, the first chart displayed shows the observed values, the regression line and both types of confidence interval around the predictions. The second chart shows the standardized residuals as a function of the explanatory variable. In principle, the residuals should be distributed randomly around the X-axis. If there is a trend or a shape, this shows a problem with the model.

The **three charts** displayed next respectively show the evolution of the standardized residuals as a function of the dependent variable, the distance between the predictions and the observations (for an ideal model, the points would all be on the bisector), and the standardized residuals on a bar chart. The last chart quickly shows if an abnormal number of values are outside the interval \([-2, 2]\) given that the latter, assuming that the sample is normally distributed, should contain about 95% of the data.

**Example**

A tutorial on the generation and the analysis of a screening design is available on the Addinsoft website:

http://www.xlstat.com/doe1.htm

**References**


Surface response designs

Use this module to generate a design to analyze the surface response for 2 to 6 factors and one or more responses.

Description

The family of surface response design is used for modeling and analysis of problems in which a response of interest is influenced by several variables and the objective is to optimize this response.

Remark: In contrast to this, screening designs aim to study the input factors, not the response value.

For example, suppose that an engineer wants to find the optimal levels of the pressure ($x_1$) and the temperature ($x_2$) of an industrial process to produce concrete, which should have a maximum hardness $y$.

\[ y = f(x_1, x_2) + \varepsilon_i \]  

Model

This tool supposes a second-order model. If $k$ is the number of factors, the quadratic model is written as follows:

\[ Y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_i x_i^2 + \sum_{i<j} \beta_{ij} x_i x_j + \varepsilon \]

Design

The tool offers the following design approaches for surface modeling:

Full factorial design with 3 levels: All combinations of 3 values for each factor (minimum, mean and maximum) are generated in the design. The number of experiments $n$ for $k$ factors is given as:

\[ n = 3^k \]

Central composite design: Proposed by Box G.E.P. and Wilson K.B. (1951), the points of experiments are generated on a sphere around the center point. The number of different factor
levels is minimized. The center point is repeated in order to maximize the prediction precision around the supposed optimum. The number of repetitions \( n_0 \) of the center point is calculated by the following formulas for \( k \) factors based on the uniform precision:

\[
\gamma = \frac{(k + 3) + \sqrt{9k^2 + 14k - 7}}{4(k + 2)}
\]

\[
n_0 = \text{floor}(\gamma(\sqrt{2^k + 2})^2 - 2^k - 2k)
\]

, where floor designates the biggest integer value smaller than the argument. The number of experiments \( n \) for \( k \) factors is given as:

\[
n = 2^k + 2k + 1
\]

**Box-Behnken**: This design was proposed by Box G.E.P. and Behnken D.W (1960) and is based on the same principles as the central composite design, but with a smaller number of experiments. The number of experiments \( n \) for \( k \) factors is given as:

\[
n = 2k^2 - 2k + 1
\]

**Doehlert**: This design was proposed by Doehlert D.H. (1970) and is based on the same principles as the central composite and Box-Behnken design, but with a smaller number of experiments. This design has a larger amount of different factor levels for several factors of the design and might therefore be difficult to use. The number of experiments \( n \) for \( k \) factors is given as:

\[
n = k^2 + k + 1
\]

The following table displays the number of different experiments for each of the 4 design choices and a given number of factors \( k \) to be analyzed. In this calculation, the center point is only present one time.

<table>
<thead>
<tr>
<th>( k )</th>
<th>full fact.</th>
<th>Cent. comp.</th>
<th>Box-Behnken</th>
<th>Doehlert</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>9</td>
<td>9</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>27</td>
<td>15</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>4</td>
<td>81</td>
<td>25</td>
<td>25</td>
<td>21</td>
</tr>
<tr>
<td>5</td>
<td>243</td>
<td>43</td>
<td>41</td>
<td>31</td>
</tr>
<tr>
<td>6</td>
<td>729</td>
<td>77</td>
<td>61</td>
<td>43</td>
</tr>
<tr>
<td>7</td>
<td>2187</td>
<td>143</td>
<td>86</td>
<td>57</td>
</tr>
</tbody>
</table>

**Output**

This tool will provide a new design for testing. Optional experiment sheets for each individual test might be generated on separated Excel sheets for printing. After having carried out the
experiments, complete the corresponding cells in the created experimental design in the corresponding Excel sheet.

A hidden sheet with important information about the design is included in your Excel file in order to have all necessary information for the XLSTAT analysis for response surface designs is ready. In this way incorrect analysis of an experimental design is inhibited. Therefore please carry out your analysis of your experimental design in the same Excel workbook where you created the design itself.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- ![Reload Default Options](image): Click this button to reload the default options.
- ![Delete Data Selections](image): Click this button to delete the data selections.

**General** tab:

**Model name**: Choose a short model name for the design. This name is used for the name of the Excel sheets and to relate the design to the analysis of the model.

**Number of factors**: Choose the number of factors to be studied in the design. The possible range is between 2 and 6 factors.
**Experimental design:** Choose the design that you want to use. Depending on the number of factors several alternative designs are suggested among which the “central composite design” and the “full factorial design with 3 levels”.

**Force the number of repetitions of the central point:** In the case of a central composite design, you have the possibility to change the number of the repetitions of the central point. Activate this option to force the number of repetitions of the central point.

**Number of responses:** Enter the number of responses that you want to analyze with the design.

**Repetitions:** Activate this option to choose the number of repetitions of the design.

**Randomize:** Activate this option to change the order of the lines of the design into a random order.

**Display experiment sheets:** Activate this option in order to generate for each individual experiment a separate Excel sheet with information about the experiment. This can be useful when printed out for the realization of the experiment.

**Range:** Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Activate this option to display the results in a new worksheet of the active workbook.

**Workbook:** Activate this option to display the results in a new workbook.

**Factors** tab:

**Information on factors:** Select one of the two following options to determine how the information on the factors is entered:

- **Enter manually:** All information on the factors is directly entered in the text fields of the dialog box.

- **Select on a sheet:** All information on the factors is selected in the Excel sheet. In this case you must select columns with as many rows as there are factors.

**Format:** Select one of the two following options to determine the way the factor intervals are entered:
- **Range**: Select this option, if you want to enter for each factor the minimum and maximum value of the interval to be studied.

- **Center + Step**: Select this option, if you want to enter for each factor the center and the maximum step size between two values.

**Short name**: Enter a few letters name for each factor. If manual selection has been chosen, enter the name in the corresponding field for each factor. If sheet selection is activated, select on the Excel sheet a range that contains the short name for each factor. The order of the different factors must be the same for all the selections in this window. Headers must not be included in the selection.

**Long name**: Enter the full name for each factor. If manual selection has been chosen, enter the name in the corresponding field for each factor. If sheet selection is activated, select on the Excel sheet a range that contains the long name for each factor. The order of the different factors must be the same for all the selections in this window. Headers must not be included in the selection.

**Unit**: Enter a description that corresponds to the unit of each factor (for example “degrees Celcius”). If manual selection has been chosen, enter the name in the corresponding field for each factor. If sheet selection is activated, select on the Excel sheet a range that contains the description of the unit for each factor. The order of the different factors must be the same for all the selections in this window. Headers must not be included in the selection.

**Unit (symbol)**: Enter the physical unit of the factors (for example “°C”). If manual selection has been chosen, enter the name in the corresponding field for each factor. If sheet selection is activated, select on the Excel sheet a range that contains the physical unit for each factor. The order of the different factors must be the same for all the selections in this window. Headers must not be included in the selection.

If the format “Amplitude” option is activated, the following two fields are visible and must be filled in.

- **Minimum**: Enter the minimum of the range to be studied for each factor. If manual selection has been chosen, enter the name in the corresponding field for each factor. If sheet selection is activated, select on the Excel sheet a range that contains the minimum of each factor. The order of the different factors must be the same for all the selections in this window. Headers must not be included in the selection.

- **Maximum**: Enter the maximum of the range to be studied for each factor. If manual selection has been chosen, enter the name in the corresponding field for each factor. If sheet selection is activated, select on the Excel sheet a range that contains the maximum of each factor. The order of the different factors must be the same for all the selections in this window. Headers must not be included in the selection.

If the “Center + Step” option is activated, the following two fields are visible.

- **Center**: Enter the central value of the range to be studied for each factor. If manual selection has been chosen, enter the name in the corresponding field for each factor. If sheet selection
is activated, select on the Excel sheet a range that contains the central value for each factor. The order of the different factors must be the same for all the selections in this window. Headers must not be included in the selection.

**Step:** Enter the step size between two successive values of the range to be studied for each factor. If manual selection has been chosen, enter the name in the corresponding field for each factor. If sheet selection is activated, select on the Excel sheet a range that contains the step size between two successive values for each factor. The order of the different factors must be the same for all the selections in this window. Headers must not be included in the selection.

**Responses** tab:

**Information on responses:** Select one of the two following options to determine how the information on the responses is entered:

- **Enter manually:** All information on the responses is directly entered in the text fields of the dialog box.
- **Select on a sheet:** All information on the responses is selected in the Excel sheet. In this case you must select columns with as many rows as there are responses.

**Short name:** Enter a few letters name for each response. If manual selection has been chosen, enter the name in the corresponding field for each response. If sheet selection is activated, select on the Excel sheet a range that contains the short name for each response. The order of the different responses must be the same for all the selections in this window. Headers must not be included in the selection.

**Long name:** Enter the full name for each response. If manual selection has been chosen, enter the name in the corresponding field for each response. If sheet selection is activated, select on the Excel sheet a range that contains the long name for each response. The order of the different responses must be the same for all the selections in this window. Headers must not be included in the selection.

**Unit:** Enter a description of the unit of the responses. If manual selection is activated, there is a text field for each response. If sheet selection is activated, please choose a range in the Excel sheet that contains one field with a value for each response. The order of the different response must be the same for all the selections in this window. Header lines must not be included in the selection. The first row of the selected range must contain data values.

**Unit (symbol):** Enter the physical unit of the responses. If manual selection is activated, there is a text field for each response. If sheet selection is activated, please choose a range in the Excel sheet that contains one field with a value for each response. The order of the different response must be the same for all the selections in this window. Header lines must not be included in the selection. The first row of the selected range must contain data values.
Results

Variables information: This table shows the information about the factors. For each factor the short name, long name, unit and physical unit are displayed.

Then the Model name is displayed, in order to select this field as identification when performing the analysis of the generated design.

Experimental design: This table displays the complete experimental design. Additional columns include information on the factors and on the responses, a label for each experiment, the sort order, the run order and the repetition.

If the generation of experiment sheets was activated in the dialog box and if there are less than 200 experiments to be carried out, an experiment sheet is generated for each line of the experimental design on separate Excel sheets.

These sheets start with the report header of the experimental design and the model name to simplify the identification of the experimental design that this sheet belongs to. Then the running number of the experiment and the total number of experiments are displayed. The values of the additional columns of the experimental design, i.e. sort order, run order and repetition are given for the experiment.

Last, the information on the experimental conditions of the factors is displayed with fields so that the user can enter the results obtained for the various responses. Short names, long names, units, physical units and values are displayed for each factor.

These sheets can be printed out or can be used in electronic format to assist during the realization of the experiments.

Example

A tutorial on the generation of a surface response design is available on the Addinsoft website: 
http://www.xlstat.com/demo-doe2.htm

References

Box G. E. P. and Behnken D. W. (1960). Some new three level designs for the study of quantitative variables, Technometrics, 2, Number 4, 455-475.


Analysis of a Surface response design

Use this tool to analyze a surface response design for 2 to 6 factors and a user defined number of results. A second order model is used for the analysis.

Description

The analysis of a surface response design uses the same statistical and conceptual framework as linear regression. The main difference comes from the model that is used. A quadratic form is used as a model

If k is the number of factors, the quadratic model is written as follows:

\[ Y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_{i<j} \sum_{j} \beta_{ij} x_i x_j + \epsilon \]  

(1)

For more information on ANOVA and linear Regression please consider the corresponding sections in the online help.

Multi-response and desirability

In the case of many response values \( y_1, \ldots, y_m \), it is possible to optimize each response value individually and to create a combined desirability function and analyze its values. Proposed by Derringer and Suich (1980), this approach consists of converting each response \( y_i \) into an individual desirability function \( d_i \) that varies over the range \( 0 \leq d_i \leq 1 \).

When \( y_i \) has reached its target, then \( d_i = 1 \). If \( y_i \) is outside an acceptable region around the target, \( d_i = 0 \). Between these two extreme cases, intermediate values of \( d_i \) exist as shown below.

The 3 different optimization cases for \( d_i \) are present with the following definitions:

L = lower value. Every value smaller than L has \( d_i = 0 \)

U = upper value. Every value bigger than U has \( d_i = 0 \).

T(L) = left target value.

T(R) = right target value. Every value between T(L) and T(R) has \( d_i = 1 \).

s, t = weighting parameters that define the shape of the optimization function between L and T(L) and T(R) and U.

The following equation has to be respected when defining L, U, T(L) and T(R):
L \leq T(L) \leq T(R) \leq U

**Maximize** the value of y:

\[
d_i = \begin{cases} 
0 & y_i < L \\
\left( \frac{y_i - L}{T(L) - L} \right)^s & L \leq y_i \leq T(L) \\
1 & y_i > T(L)
\end{cases}
\]

**Minimize** the value of y:

\[
d_i = \begin{cases} 
1 & y_i < T(R) \\
\left( \frac{U - y_i}{U - T(R)} \right)^t & T(R) \leq y_i \leq U \\
0 & y_i > U
\end{cases}
\]

Two sided desirability function as shown below to **target** a certain interval of y:
The design variables are chosen to maximize the overall desirability $D$

$$D = \left( d_1^{w_1} \cdot d_2^{w_2} \cdot \ldots \cdot d_m^{w_m} \right)^{1/(w_1 + w_2 + \ldots + w_m)}$$

Where $1 \leq w_i \leq 10$ are weightings of the individual desirability functions. The bitter $w_i$, the more important is $d_i$ taken into account during the optimization.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
: Click this button to reload the default options.

: Click this button to delete the data selections.

: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

**Model name:** Select the corresponding cell in the Excel sheet with the generated design that you want to analyze. The Model name is used as part of the names of Excel sheets and during the selection of the analysis in order to make the link between the design and the analysis of the results of the design.

**Y / results:** Select the columns of the experimental design that contain the results. These columns should now hold the results of the experiments carried out. If several result variables have been selected, XLSTAT carries out calculations for each of the variables separately, and then an analysis of the desirability is carried out. If a column header has been selected, check that the "Variable labels" option has been activated.

**Experimental design:** Activate this option, if you made changes to the values of the generated experimental design. Then the changes will be shown in the results.

If you have the possibility to select the additional columns (the columns on the left of the factor columns of the generated experimental design) and the columns with factors of the experimental design and you want to select them for comparison with the original experimental design. It is important include into the selection the column with the sort order information. Using this option includes changes to the experimental design in the factor columns into the analysis. If this option is not activated, the experimental design at the moment of its generation is used for the analysis. The selected data has to be numerical. If a column header has been selected, check that the "Variable labels" option has been activated.

**Range:** Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Activate this option to display the results in a new worksheet of the active workbook.

**Workbook:** Activate this option to display the results in a new workbook.
Variable labels: This option is always activated. The first row of the selected data (data and observation labels) must contain a label.

Responses tab:

Selection: Select one of the two following options to determine the selection mode for this window:

- Manual selection: All information about the responses will be inserted directly into the text fields of the window.
- Sheet selection: All information about the responses will be selected as ranges in the Excel sheet. In this case a column with as much entries as the number of factors is expected.

Short name: Enter a short name for the responses composed of some characters. If manual selection is activated, there is a text field for each response. If sheet selection is activated, please choose a range in the Excel sheet that contains one field with a value for each response. The order of the different responses must be the same for all the selections in this window. Header lines must not be included in the selection. The first row of the selected range must contain data values.

Long name: Enter a long name for the responses composed of some characters. If manual selection is activated, there is a text field for each response. If sheet selection is activated, please choose a range in the Excel sheet that contains one field with a value for each response. The order of the different responses must be the same for all the selections in this window. Header lines must not be included in the selection. The first row of the selected range must contain data values.

Aim: Choose the aim of the optimization. You have the choice between Minimum, Optimum and Maximum. If manual selection is activated, there is a text field for each response. If sheet selection is activated, please choose a range in the Excel sheet that contains one field with a value for each response. The order of the different responses must be the same for all the selections in this window. Header lines must not be included in the selection. The first row of the selected range must contain data values.

If the selected aim is Optimum or Maximum, then the following two fields are activated.

Lower: Enter the value of the lower bound, below which the desirability is 0. If manual selection is activated, there is a text field for each response. If sheet selection is activated, please choose a range in the Excel sheet that contains one field with a value for each response. The order of the different responses must be the same for all the selections in this window. Header lines must not be included in the selection. The first row of the selected range must contain data values.
**Target (left):** Enter the value of the lower bound, above which the desirability is 1. The desirability function increases monotonously from 0 to 1 between the lower bound and the left target. If manual selection is activated, there is a text field for each response. If sheet selection is activated, please choose a range in the Excel sheet that contains one field with a value for each response. The order of the different responses must be the same for all the selections in this window. Header lines must not be included in the selection. The first row of the selected range must contain data values.

If the selected aim is Minimum or Optimum, then the following two fields are activated.

**Target (right):** Enter the value of the upper bound, below which the desirability is 1. The desirability function decreases monotonously from 1 to 0 between the right target and the upper bound. If manual selection is activated, there is a text field for each response. If sheet selection is activated, please choose a range in the Excel sheet that contains one field with a value for each response. The order of the different responses must be the same for all the selections in this window. Header lines must not be included in the selection. The first row of the selected range must contain data values.

**Upper:** Enter the value of the upper bound, above which the desirability is 0. If manual selection is activated, there is a text field for each response. If sheet selection is activated, please choose a range in the Excel sheet that contains one field with a value for each response. The order of the different responses must be the same for all the selections in this window. Header lines must not be included in the selection. The first row of the selected range must contain data values.

**s:** Activate this option, if the increasing desirability function should have a non linear shape. Enter the value of the shape parameter, which should be a value between 0.01 and 100. Examples are shown in the graphic in the window at the right. If manual selection is activated, there is a text field for each response. If sheet selection is activated, please choose a range in the Excel sheet that contains one field with a value for each response. The order of the different responses must be the same for all the selections in this window. Header lines must not be included in the selection. The first row of the selected range must contain data values.

**t:** Activate this option, if the decreasing desirability function should have a non linear shape. Enter the value of the shape parameter, which should be a value between 0.01 and 100. Examples are shown in the graphic in the window at the right. If manual selection is activated, there is a text field for each response. If sheet selection is activated, please choose a range in the Excel sheet that contains one field with a value for each response. The order of the different responses must be the same for all the selections in this window. Header lines must not be included in the selection. The first row of the selected range must contain data values.

**Weight:** Activate this option, if the responses should have an exponent different from 1 during the calculation of the desirability function. Enter the value of the shape parameter, which should be a value between 0.01 and 100. Examples are shown in the graphic in the window at the right. If manual selection is activated, there is a text field for each response. If sheet selection is activated, please choose a range in the Excel sheet that contains one field with a value for each response. The order of the different responses must be the same for all the
selections in this window. Header lines must not be included in the selection. The first row of the selected range must contain data values.

**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the variables selected.

**Correlations**: Activate this option to display the correlation matrix for quantitatitive variables (dependent or explanatory).

**Experimental design**: Activate this option to display the table with the experimental design.

**Analysis of variance**: Activate this option to display the analysis of variance table.

**Goodness of fit statistics**: Activate this option to display the table of goodness of fit statistics for the model.

**Standardized coefficients**: Activate this option if you want the standardized coefficients (beta coefficients) for the model to be displayed.

**Predictions and residuals**: Activate this option to display the predictions and residuals for all the observations.

- **Adjusted predictions**: Activate this option to calculate and display adjusted predictions in the table of predictions and residuals.
- **Studendized residuals**: Activate this option to calculate and display studentized residuals in the table of predictions and residuals
- **Cook’s D**: Activate this option to calculate and display Cook’s distances in the table of predictions and residuals.

**Charts** tab:

**Regression charts**: Activate this option to display regression chart:

- **Standardized coefficients**: Activate this option to display the standardized parameters for the model with their confidence interval on a chart.
- **Predictions and residuals**: Activate this option to display the following charts.

  1. **Line of regression**: This chart is only displayed if there is only one explanatory variable and this variable is quantititative.
  2. **Explanatory variable versus standardized residuals**: This chart is only displayed if there is only one explanatory variable and this variable is quantititative.
(3) Dependent variable versus standardized residuals.

(4) Predictions for the dependent variable versus the dependent variable.

(5) Bar chart of standardized residuals.

- **Confidence intervals**: Activate this option to have confidence intervals displayed on charts (1) and (4).

**Contour plot**: Activate this option to display charts that represent the desirability function in contour plots in the case of a model with 2 factors.

**Trace plot**: Activate this option to display charts that represent the trace of the desirability function for each of the factors, with the other factors set to the mean value.

**Results**

**Descriptive statistics**: The tables of descriptive statistics show the simple statistics for all the variables selected. The number of observations, missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed. For qualitative explanatory variables, the names of the various categories are displayed together with their respective frequencies.

**Variables information**: This table shows the information about the factors. For each factor, the short name, long name, unit and physical unit are displayed.

Then the **Model name** is displayed, in order to select this field as identification later on during the analysis of the generated design.

**Experimental design**: In this table, the complete experimental design is shown. There are additional columns, columns for the factors and columns for the responses displayed. The additional columns contain a label for each experiment, the sort order, the run order, the block number, and the point type.

If changes were made to the values between the generation of the experimental design and the analysis, these values are displayed in bold.

After that, the parameters of the desirability function are displayed, if there is more than one response present in the design. The table shows, for each response, the short name, long name, unit, physical unit, aim, lower bound, left target value, right target value, upper bound, shape parameters s and t, and the weight parameter.

**Correlation matrix**: This table is displayed to give you a view of the correlations between the various variables selected.

Then for each response and the global desirability function, the following tables and charts are displayed.
Goodness of fit statistics: The statistics relating to the fitting of the regression model are shown in this table:

- **Observations**: The number of observations used in the calculations. In the formulas shown below, \( n \) is the number of observations.

- **Sum of weights**: The sum of the weights of the observations used in the calculations. In the formulas shown below, \( W \) is the sum of the weights.

- **DF**: The number of degrees of freedom for the chosen model (corresponding to the error part).

- **\( R^2 \)**: The determination coefficient for the model. This coefficient, whose value is between 0 and 1, is only displayed if the constant of the model has not been fixed by the user. Its value is defined by:

\[
R^2 = 1 - \frac{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} w_i (y_i - \bar{y})^2}, \quad \text{where} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} w_i y_i,
\]

The \( R^2 \) is interpreted as the proportion of the variability of the dependent variable explained by the model. The nearer \( R^2 \) is to 1, the better is the model. The problem with the \( R^2 \) is that it does not take into account the number of variables used to fit the model.

- **Adjusted \( R^2 \)**: The adjusted determination coefficient for the model. The adjusted \( R^2 \) can be negative if the \( R^2 \) is near to zero. This coefficient is only calculated if the constant of the model has not been fixed by the user. Its value is defined by:

\[
\hat{R}^2 = 1 - \left(1 - R^2\right) \frac{W - 1}{W - p - 1}
\]

The adjusted \( R^2 \) is a correction to the \( R^2 \) which takes into account the number of variables used in the model.

- **MSE**: The mean squared error (MSE) is defined by:

\[
MSE = \frac{1}{W - p} \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2
\]

- **RMSE**: The root mean square of the errors (RMSE) is the square root of the MSE.

- **MAPE**: The Mean Absolute Percentage Error is calculated as follows:

\[
MAPE = \frac{100}{W} \sum_{i=1}^{n} w_i \left| \frac{y_i - \hat{y}_i}{y_i} \right|
\]

- **DW**: The Durbin-Watson statistic is defined by:
This coefficient is the order 1 autocorrelation coefficient and is used to check that the residuals of the model are not autocorrelated, given that the independence of the residuals is one of the basic hypotheses of linear regression. The user can refer to a table of Durbin-Watson statistics to check if the independence hypothesis for the residuals is acceptable.

- **Cp**: Mallows Cp coefficient is defined by:

\[
Cp = \frac{SSE}{\hat{\sigma}^2} + 2p^* - W
\]

where SSE is the sum of the squares of the errors for the model with p explanatory variables and \(\hat{\sigma}\) is the estimator of the variance of the residuals for the model comprising all the explanatory variables. The nearer the Cp coefficient is to \(p^*\), the less the model is biased.

- **AIC**: Akaike’s Information Criterion is defined by:

\[
AIC = W \ln \left( \frac{SSE}{W} \right) + 2p^*
\]

This criterion, proposed by Akaike (1973) is derived from the information theory and uses Kullback and Leibler’s measurement (1951). It is a model selection criterion which penalizes models for which adding new explanatory variables does not supply sufficient information to the model, the information being measured through the MSE. The aim is to minimize the AIC criterion.

- **SBC**: Schwarz’s Bayesian Criterion is defined by:

\[
SBC = W \ln \left( \frac{SSE}{W} \right) + \ln(W) p^*
\]

This criterion, proposed by Schwarz (1978) is similar to the AIC, and the aim is to minimize it.

- **PC**: Amemiya’s Prediction Criterion is defined by:

\[
PC = \frac{(1 - R^2)(W + p^*)}{W - p^*}
\]

This criterion, proposed by Amemiya (1980) is used, like the adjusted \(R^2\) to take account of the parsimony of the model.
• **Press RMSE**: Press' statistic is only displayed if the corresponding option has been activated in the dialog box. It is defined by:

\[
\text{Press} = \sum_{i=1}^{n} w_i \left( y_i - \hat{y}_{i(-i)} \right)^2
\]

where \( \hat{y}_{i(-i)} \) is the prediction for observation \( i \) when the latter is not used for estimating parameters. We then get:

\[
\text{Press RMSE} = \sqrt{\frac{\text{Press}}{W - p^*}}
\]

Press's RMSE can then be compared to the RMSE. A large difference between the two shows that the model is sensitive to the presence or absence of certain observations in the model.

• **Q²**: The Q² statistic is displayed. It is defined as

\[
Q^2 = 1 - \frac{\text{PressRMSE}}{\text{SSE}}
\]

The closer Q² is to 1, the better and more robust is the model.

The analysis of variance table is used to evaluate the explanatory power of the explanatory variables. Where the constant of the model is not set to a given value, the explanatory power is evaluated by comparing the fit (as regards least squares) of the final model with the fit of the rudimentary model including only a constant equal to the mean of the dependent variable. Where the constant of the model is set, the comparison is made with respect to the model for which the dependent variable is equal to the constant which has been set.

The equation of the model is then displayed to make it easier to read or re-use the model.

The table of standardized coefficients (also called beta coefficients) are used to compare the relative weights of the variables. The higher the absolute value of a coefficient, the more important the weight of the corresponding variable. When the confidence interval around standardized coefficients has value 0 (this can be easily seen on the chart of standardized coefficients), the weight of a variable in the model is not significant.

The predictions and residuals table shows, for each observation, its weight, the value of the qualitative explanatory variable, if there is only one, the observed value of the dependent variable, the model's prediction, the residuals, the confidence intervals together with the fitted prediction and Cook's D if the corresponding options have been activated in the dialog box. Two types of confidence interval are displayed: a confidence interval around the mean (corresponding to the case where the prediction would be made for an infinite number of
observations with a set of given values for the explanatory variables) and an interval around
the isolated prediction (corresponding to the case of an isolated prediction for the values given
for the explanatory variables). The second interval is always greater than the first, the random
values being larger.

The charts which follow show the results mentioned above. If there is only one explanatory
variable in the model, the first chart displayed shows the observed values, the regression line
and both types of confidence interval around the predictions. The second chart shows the
standardized residuals as a function of the explanatory variable. In principle, the residuals
should be distributed randomly around the X-axis. If there is a trend or a shape, this shows a
problem with the model.

The three charts displayed next show respectively the evolution of the standardized residuals
as a function of the dependent variable, the distance between the predictions and the
observations (for an ideal model, the points would all be on the bisector), and the standardized
residuals on a bar chart. The last chart quickly shows if an abnormal number of values are
outside the interval \([-2, 2]\) given that the latter, assuming that the sample is normally
distributed, should contain about 95% of the data.

Then the contour plot is displayed, if the design has two factors and the corresponding option
is activated. The contour plot is shown as a two dimensional projection and as a 3D chart.
Using these charts it is possible to analyze the dependence of the two factors simultaneously.

Then the trace plots are displayed, if the corresponding option is activated. The trace plots
show for each factor the response variable as a function of the factor. All other factors are set
to their mean value. These charts are shown in two options: with the standardized factors and
with the factors in original values. Using these plots the dependence of a response on a given
factor can be analyzed.

Example

A tutorial on the generation and the analysis of a surface response design is available on the
Addinsoft website:

http://www.xlstat.com/demo-doe2.htm

References

Derringer R. and Suich R. (1980). Simultaneous optimization of several response variables,

d'expériences : une approche pragmatique et illustrée, Alpha Graphic, Olivet, 2005.

Mixture designs

Use this module to generate a mixture design for 2 to 6 factors.

Description

Mixture designs are used to model the results of experiments where these relate to the optimization of formulations. The resulting model is called "mixture distribution"

Mixture designs differ from factorial designs by the following characteristics:

- The factors studied are proportions which sum is equal to 1
- Construction of the design of experiments is subjected to constraints because the factors may not evolve independently of each other (the sum of the proportions being 1).

Experimental space of a mixture

When the concentrations of the n components are not submitted to any constraint, the experimental design is a simplex, that is to say, a regular polyhedron with n vertices in a space of dimension n-1. For example, for a mixture of three components, the experimental field is an equilateral triangle; for 4 constituents it is a regular tetrahedron.

Creating mixture designs therefore consist of positioning regularly the experiences in the simplex to optimize the accuracy of the model. The most conventional designs are Scheffé’s designs, Scheffé-centroid designs, and augmented designs.

If constraints on the components of the model are introduced by defining a minimum amount or a maximum amount not to exceed, then, the experimental domain can be a simplex, an inverted simplex (also called simplex B) or a any convex polyhedron. In the latter case, the simplex designs are no longer usable.

To treat irregular domains, algorithmic experimental designs are used: the optimality criterion used in XLSTAT is the D-optimality.

Warning: if the number of components is important and there are many constraints on the components, it is possible that the experimental domain does not exist.

The Scheffé simplex networks are the easiest designs to build. They allow to build models of any degree m. These matrices are related to a canonical model having a high number of coefficients (Full Canonical Model).
To improve the sequentiality of the experiments, Scheffé proposed to add points to the center of experimental space. These experimental designs are known Simplex Centroid Designs.

These mixture designs allow to construct a reduced polynomial model, which comprises only product terms of the components. The number of experiments thus increases less rapidly than in the case of a Scheffé’s simplex. Centered simplexes add additional mixtures in the center of the experimental space compared to conventional simplexes. This has the effect of improving the quality of predictions in the center of the field.

**Output**

This tool will provide a new design for testing. Optional experiment sheets for each individual test might be generated on separated Excel sheets for printing. After having carried out the experiments, complete the corresponding cells in the created experimental design in the corresponding Excel sheet.

A hidden sheet with important information about the design is included in your Excel file in order to have all necessary information for the XLSTAT analysis for response surface designs is ready. In this way incorrect analysis of an experimental design is inhibited. Therefore please carry out your analysis of your experimental design in the same Excel workbook where you created the design itself.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box. 

<table>
<thead>
<tr>
<th>Degree of the model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constituents</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>10</td>
</tr>
</tbody>
</table>
General tab:

**Model name:** Choose a short model name for the design. This name is used for the name of the Excel sheets and to relate the design to the analysis of the model.

**Number of factors:** Choose the number of factors to be studied in the design. The possible range is between 2 and 6 factors.

**Experimental design:** Choose the design that you want to use among Scheffé’s simplex, centered Scheffé’s simplex, augmented simplex.

**Degree of the model:** In the case of a Scheffé’s simplex, it is possible to choose the number of degrees of the model (from 1 to 4). The higher the degree of the model, the more the number of experiments increases.

**Number of responses:** Enter the number of responses that you want to analyze with the design.

**Repetitions:** Activate this option to choose the number of repetitions of the design.

**Randomize:** Activate this option to change the order of the lines of the design into a random order.

**Display experiment sheets:** Activate this option in order to generate for each individual experiment a separate Excel sheet with information about the experiment. This can be useful when printed out for the realization of the experiment.

**Range:** Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Activate this option to display the results in a new worksheet of the active workbook.
Workbook: Activate this option to display the results in a new workbook.

Factors tab:

Information on factors: Select one of the two following options to determine how the information on the factors is entered:

- Enter manually: All information on the factors is directly entered in the text fields of the dialog box.
- Select on a sheet: All information on the factors is selected in the Excel sheet. In this case you must select columns with as many rows as there are factors.

Format: Select one of the two following options to determine the way the factor intervals are entered:

- Range: Select this option, if you want to enter for each factor the minimum and maximum value of the interval to be studied.
- Center + Step: Select this option, if you want to enter for each factor the center and the maximum step size between two values.

Short name: Enter a few letters name for each factor. If manual selection has been chosen, enter the name in the corresponding field for each factor. If sheet selection is activated, select on the Excel sheet a range that contains the short name for each factor. The order of the different factors must be the same for all the selections in this window. Headers must not be included in the selection.

Long name: Enter the full name for each factor. If manual selection has been chosen, enter the name in the corresponding field for each factor. If sheet selection is activated, select on the Excel sheet a range that contains the long name for each factor. The order of the different factors must be the same for all the selections in this window. Headers must not be included in the selection.

Unit: Enter a description that corresponds to the unit of each factor (for example “degrees Celcius”). If manual selection has been chosen, enter the name in the corresponding field for each factor. If sheet selection is activated, select on the Excel sheet a range that contains the description of the unit for each factor. The order of the different factors must be the same for all the selections in this window. Headers must not be included in the selection.

Unit (symbol): Enter the physical unit of the factors (for example “°C”). If manual selection has been chosen, enter the name in the corresponding field for each factor. If sheet selection is activated, select on the Excel sheet a range that contains the physical unit for each factor. The order of the different factors must be the same for all the selections in this window. Headers must not be included in the selection.
If the format “Amplitude” option is activated, the following two fields are visible and must be filled in.

**Minimum:** Enter the minimum of the range to be studied for each factor. If manual selection has been chosen, enter the name in the corresponding field for each factor. If sheet selection is activated, select on the Excel sheet a range that contains the minimum of each factor. The order of the different factors must be the same for all the selections in this window. Headers must not be included in the selection.

**Maximum:** Enter the maximum of the range to be studied for each factor. If manual selection has been chosen, enter the name in the corresponding field for each factor. If sheet selection is activated, select on the Excel sheet a range that contains the maximum of each factor. The order of the different factors must be the same for all the selections in this window. Headers must not be included in the selection.

If the “Center + Step” option is activated, the following two fields are visible.

**Center:** Enter the central value of the range to be studied for each factor. If manual selection has been chosen, enter the name in the corresponding field for each factor. If sheet selection is activated, select on the Excel sheet a range that contains the central value for each factor. The order of the different factors must be the same for all the selections in this window. Headers must not be included in the selection.

**Step:** Enter the step size between two successive values of the range to be studied for each factor. If manual selection has been chosen, enter the name in the corresponding field for each factor. If sheet selection is activated, select on the Excel sheet a range that contains the step size between two successive values for each factor. The order of the different factors must be the same for all the selections in this window. Headers must not be included in the selection.

**Responses** tab:

**Information on responses:** Select one of the two following options to determine how the information on the responses is entered:

- **Enter manually:** All information on the responses is directly entered in the text fields of the dialog box.

- **Select on a sheet:** All information on the responses is selected in the Excel sheet. In this case you must select columns with as many rows as there are responses.

**Short name:** Enter a few letters name for each response. If manual selection has been chosen, enter the name in the corresponding field for each response. If sheet selection is activated, select on the Excel sheet a range that contains the short name for each response. The order of the different responses must be the same for all the selections in this window. Headers must not be included in the selection.
**Long name:** Enter the full name for each response. If manual selection has been chosen, enter the name in the corresponding field for each response. If sheet selection is activated, select on the Excel sheet a range that contains the long name for each response. The order of the different responses must be the same for all the selections in this window. Headers must not be included in the selection.

**Unit:** Enter a description of the unit of the responses. If manual selection is activated, there is a text field for each response. If sheet selection is activated, please choose a range in the Excel sheet that contains one field with a value for each response. The order of the different response must be the same for all the selections in this window. Header lines must not be included in the selection. The first row of the selected range must contain data values.

**Unit (symbol):** Enter the physical unit of the responses. If manual selection is activated, there is a text field for each response. If sheet selection is activated, please choose a range in the Excel sheet that contains one field with a value for each response. The order of the different response must be the same for all the selections in this window. Header lines must not be included in the selection. The first row of the selected range must contain data values.

---

**Results**

**Variables information:** This table shows the information about the factors. For each factor the short name, long name, unit and physical unit are displayed.

Then the **Model name** is displayed, in order to select this field as identification when performing the analysis of the generated design.

**Experimental design:** This table displays the complete experimental design. Additional columns include information on the factors and on the responses, a label for each experiment, the sort order, the run order and the repetition.

If the generation of **experiment sheets** was activated in the dialog box and if there are less than 200 experiments to be carried out, an experiment sheet is generated for each line of the experimental design on separate Excel sheets.

These sheets start with the report header of the experimental design and the model name to simplify the identification of the experimental design that this sheet belongs to. Then the running number of the experiment and the total number of experiments are displayed. The values of the additional columns of the experimental design, i.e. sort order, run order and repetition are given for the experiment.

Last, the information on the experimental conditions of the factors is displayed with fields so that the user can enter the results obtained for the various responses. Short names, long names, units, physical units and values are displayed for each factor.
These sheets can be printed out or can be used in electronic format to assist during the realization of the experiments.

**Example**

A tutorial on the generation and analysis of a mixture design is available on the Addinsoft website:
http://www.xlstat.com/demo-mixture.htm

**References**


Analysis of a mixture design

Use this tool to analyze a mixture design for 2 to 6 factors.

Description

The analysis of a mixture design is based on the same principle as linear regression. The major difference comes from the model that is used. Several models are available.

By default, XLSTAT associates a reduced model (Simplified Canonical Model) to centroïd simplexes. However, it is possible to change the model if the number of degrees of freedom is sufficient (by increasing the number of repetitions of the experiments). Otherwise, an error message will be displayed informing you that the number of experiments is too small for all the model coefficients to be estimated.

To fulfill the constraint associated to a mixture design, a polynomial model with no intercept is used. We distinguish two types of models, simplified (special) models and full models (from level 3).

The model equations are:

- Linear model (level 1):
  \[ Y = \sum_{i} \beta_i x_i \]

- Quadratic model (level 2):
  \[ Y = \sum_{i} \beta_i x_i + \sum_{i < j} \beta_{ij} x_i x_j \]

- Cubic model (level 3):
  \[ Y = \sum_{i} \beta_i x_i + \sum_{i < j} \beta_{ij} x_i x_j + \sum_{j} \delta_j x_j (x_i - x_j) + \sum_{k} \sum_{j < k} \sum_{i < j} \beta_{ijk} x_i x_j x_k \]

- Simplified cubic model (special):
  \[ Y = \sum_{i} \beta_i x_i + \sum_{i < j} \beta_{ij} x_i x_j + \sum_{k} \sum_{j < k} \sum_{i < j} \beta_{ijk} x_i x_j x_k \]

XLSTAT allows to apply models up to level 4.
Estimation of these models is done with classical regression. For more details on ANOVA and linear regression, please refer to the chapters of this help associated to these methods.

**Multi-response and desirability**

In the case of many response values $y_1, \ldots, y_m$, it is possible to optimize each response value individually and to create a combined desirability function and analyze its values. Proposed by Derringer and Suich (1980), this approach consists of converting each response $y_i$ into an individual desirability function $d_i$ that varies over the range $0 \leq d_i \leq 1$.

When $y_i$ has reached its target, then $d_i = 1$. If $y_i$ is outside an acceptable region around the target, $d_i = 0$. Between these two extreme cases, intermediate values of $d_i$ exist as shown below.

The 3 different optimization cases for $d_i$ are present with the following definitions:

$L = \text{lower value}. \quad \text{Every value smaller than } L \text{ has } d_i = 0$

$U = \text{upper value}. \quad \text{Every value bigger than } U \text{ has } d_i = 0.$

$T(L) = \text{left target value}.$

$T(R) = \text{right target value}. \quad \text{Every value between } T(L) \text{ and } T(R) \text{ has } d_i = 1.$

$s, t = \text{weighting parameters that define the shape of the optimization function between } L \text{ and } T(L) \text{ and } T(R) \text{ and } U.$

The following equation has to be respected when defining $L, U, T(L)$ and $T(R)$:

$L \leq T(L) \leq T(R) \leq U$

**Maximize** the value of $y_i$: 

\[
    d_i = \begin{cases} 
        0 & y_i < L \\
        \left(\frac{y_i - L}{T(L) - L}\right)^s & L \leq y_i \leq T(L) \\
        1 & y_i > T(L) 
    \end{cases}
\]
Minimize the value of $y_i$:

Two sided desirability function as shown below to target a certain interval of $y_i$:

$$d_i = \begin{cases} 
1 & y_i < T(R) \\
\left(\frac{U - y_i}{U - T(R)}\right)^{s} & T(R) \leq y_i \leq U \\
0 & y_i > U 
\end{cases}$$

The design variables are chosen to maximize the overall desirability $D$

$$D = \left( d_1^{w_1} \cdot d_2^{w_2} \cdots d_m^{w_m} \right)^{-\frac{1}{w_1 + w_2 + \cdots + w_m}}$$
Where $1 \leq w_i \leq 10$ are weightings of the individual desirability functions. The bitter $w_i$, the more important is $d_i$ taken into account during the optimization.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload**: Click this button to reload the default options.
- **Delete**: Click this button to delete the data selections.
- **Switch**: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab:**

**Model name**: Select the corresponding cell in the Excel sheet with the generated design that you want to analyze. The Model name is used as part of the names of Excel sheets and during the selection of the analysis in order to make the link between the design and the analysis of the results of the design.

**Y / results**: Select the columns of the experimental design that contain the results. These columns should now hold the results of the experiments carried out. If several result variables have been selected, XLSTAT carries out calculations for each of the variables separately, and then an analysis of the desirability is carried out. If a column header has been selected, check that the "Variable labels" option has been activated.
**Experimental design**: Activate this option, if you made changes to the values of the generated experimental design. Then the changes will be shown in the results.

If you have the possibility to select the additional columns (the columns on the left of the factor columns of the generated experimental design) and the columns with factors of the experimental design and you want to select them for comparison with the original experimental design. It is important include into the selection the column with the sort order information. Using this option includes changes to the experimental design in the factor columns into the analysis. If this option is not activated, the experimental design at the moment of its generation is used for the analysis. The selected data has to be numerical. If a column header has been selected, check that the "Variable labels" option has been activated.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Variable labels**: This option is always activated. The first row of the selected data (data and observation labels) must contain a label.

**Responses** tab:

**Selection**: Select one of the two following options to determine the selection mode for this window:

- **Manual selection**: All information about the responses will be inserted directly into the text fields of the window.

- **Sheet selection**: All information about the responses will be selected as ranges in the Excel sheet. In this case a column with as much entries as the number of factors is expected.

**Short name**: Enter a short name for the responses composed of some characters. If manual selection is activated, there is a text field for each response. If sheet selection is activated, please choose a range in the Excel sheet that contains one field with a value for each response. The order of the different responses must be the same for all the selections in this window. Header lines must not be included in the selection. The first row of the selected range must contain data values.
**Long name:** Enter a long name for the responses composed of some characters. If manual selection is activated, there is a text field for each response. If sheet selection is activated, please choose a range in the Excel sheet that contains one field with a value for each response. The order of the different responses must be the same for all the selections in this window. Header lines must not be included in the selection. The first row of the selected range must contain data values.

**Aim:** Choose the aim of the optimization. You have the choice between Minimum, Optimum and Maximum. If manual selection is activated, there is a text field for each response. If sheet selection is activated, please choose a range in the Excel sheet that contains one field with a value for each response. The order of the different responses must be the same for all the selections in this window. Header lines must not be included in the selection. The first row of the selected range must contain data values.

If the selected aim is Optimum or Maximum, then the following two fields are activated.

**Lower:** Enter the value of the lower bound, below which the desirability is 0. If manual selection is activated, there is a text field for each response. If sheet selection is activated, please choose a range in the Excel sheet that contains one field with a value for each response. The order of the different responses must be the same for all the selections in this window. Header lines must not be included in the selection. The first row of the selected range must contain data values.

**Target (left):** Enter the value of the lower bound, above which the desirability is 1. The desirability function increases monotonously from 0 to 1 between the lower bound and the left target. If manual selection is activated, there is a text field for each response. If sheet selection is activated, please choose a range in the Excel sheet that contains one field with a value for each response. The order of the different responses must be the same for all the selections in this window. Header lines must not be included in the selection. The first row of the selected range must contain data values.

If the selected aim is Minimum or Optimum, then the following two fields are activated.

**Target (right):** Enter the value of the upper bound, below which the desirability is 1. The desirability function decreases monotonously from 1 to 0 between the right target and the upper bound. If manual selection is activated, there is a text field for each response. If sheet selection is activated, please choose a range in the Excel sheet that contains one field with a value for each response. The order of the different responses must be the same for all the selections in this window. Header lines must not be included in the selection. The first row of the selected range must contain data values.

**Upper:** Enter the value of the upper bound, above which the desirability is 0. If manual selection is activated, there is a text field for each response. If sheet selection is activated, please choose a range in the Excel sheet that contains one field with a value for each response. The order of the different responses must be the same for all the selections in this window. Header lines must not be included in the selection. The first row of the selected range must contain data values.
s: Activate this option, if the increasing desirability function should have a non linear shape. Enter the value of the shape parameter, which should be a value between 0.01 and 100. Examples are shown in the graphic in the window at the right. If manual selection is activated, there is a text field for each response. If sheet selection is activated, please choose a range in the Excel sheet that contains one field with a value for each response. The order of the different responses must be the same for all the selections in this window. Header lines must not be included in the selection. The first row of the selected range must contain data values.

t: Activate this option, if the decreasing desirability function should have a non linear shape. Enter the value of the shape parameter, which should be a value between 0.01 and 100. Examples are shown in the graphic in the window at the right. If manual selection is activated, there is a text field for each response. If sheet selection is activated, please choose a range in the Excel sheet that contains one field with a value for each response. The order of the different responses must be the same for all the selections in this window. Header lines must not be included in the selection. The first row of the selected range must contain data values.

Weight: Activate this option, if the responses should have an exponent different from 1 during the calculation of the desirability function. Enter the value of the shape parameter, which should be a value between 0.01 and 100. Examples are shown in the graphic in the window at the right. If manual selection is activated, there is a text field for each response. If sheet selection is activated, please choose a range in the Excel sheet that contains one field with a value for each response. The order of the different responses must be the same for all the selections in this window. Header lines must not be included in the selection. The first row of the selected range must contain data values.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Correlations: Activate this option to display the correlation matrix for quantitative variables (dependent or explanatory).

Experimental design: Activate this option to display the table with the experimental design.

Analysis of variance: Activate this option to display the analysis of variance table.

Goodness of fit statistics: Activate this option to display the table of goodness of fit statistics for the model.

Standardized coefficients: Activate this option if you want the standardized coefficients (beta coefficients) for the model to be displayed.

Predictions and residuals: Activate this option to display the predictions and residuals for all the observations.
• **Adjusted predictions**: Activate this option to calculate and display adjusted predictions in the table of predictions and residuals.

• **Studendized residuals**: Activate this option to calculate and display studentized residuals in the table of predictions and residuals.

• **Cook's D**: Activate this option to calculate and display Cook's distances in the table of predictions and residuals.

**Charts** tab:

**Regression charts**: Activate this option to display regression chart:

• **Standardized coefficients**: Activate this option to display the standardized parameters for the model with their confidence interval on a chart.

• **Predictions and residuals**: Activate this option to display the following charts.
  
  (1) Line of regression: This chart is only displayed if there is only one explanatory variable and this variable is quantitative.

  (2) Explanatory variable versus standardized residuals: This chart is only displayed if there is only one explanatory variable and this variable is quantitative.

  (3) Dependent variable versus standardized residuals.

  (4) Predictions for the dependent variable versus the dependent variable.

  (5) Bar chart of standardized residuals.

  o **Confidence intervals**: Activate this option to have confidence intervals displayed on charts (1) and (4).

**Ternary diagram**: Activate this option to display a ternary diagram.
# $ + @ Results

**Descriptive statistics:** The tables of descriptive statistics show the simple statistics for all the variables selected. The number of observations, missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed. For qualitative explanatory variables the names of the various categories are displayed together with their respective frequencies.

**Variables information:** This table shows the information about the factors. For each factor the short name, long name, unit and physical unit are displayed.

Then the **Model name** is displayed, in order to select this field as identification later on during the analysis of the generated design.

**Experimental design:** In this table the complete experimental design is shown. There are additional columns, columns for the factors and columns for the responses displayed. The additional columns contain a label for each experiment, the sort order, the run order, the block number and the point type.

If changes were made to the values between the generation of the experimental design and the analysis, these values are displayed in bold.

After that the parameters of the desirability function are displayed, if there is more than one response present in the design. The table shows for each response the short name, long name, unit, physical unit, aim, lower bound, left target value, right target value, upper bound, shape parameters s and t and the weight parameter.

**Correlation matrix:** This table is displayed to give you a view of the correlations between the various variables selected.

Then for each response und the global desirability function, the following tables and charts are displayed.
**Goodness of fit statistics**: The statistics relating to the fitting of the regression model are shown in this table:

- **Observations**: The number of observations used in the calculations. In the formulas shown below, \( n \) is the number of observations.

- **Sum of weights**: The sum of the weights of the observations used in the calculations. In the formulas shown below, \( W \) is the sum of the weights.

- **DF**: The number of degrees of freedom for the chosen model (corresponding to the error part).

- **\( R^2 \)**: The determination coefficient for the model. This coefficient, whose value is between 0 and 1, is only displayed if the constant of the model has not been fixed by the user. Its value is defined by:

\[
R^2 = 1 - \frac{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} w_i (y_i - \bar{y})^2}
\]

The \( R^2 \) is interpreted as the proportion of the variability of the dependent variable explained by the model. The nearer \( R^2 \) is to 1, the better is the model. The problem with the \( R^2 \) is that it does not take into account the number of variables used to fit the model.

- **Adjusted \( R^2 \)**: The adjusted determination coefficient for the model. The adjusted \( R^2 \) can be negative if the \( R^2 \) is near to zero. This coefficient is only calculated if the constant of the model has not been fixed by the user. Its value is defined by:

\[
\hat{R}^2 = 1 - \left(1 - R^2\right) \frac{W - 1}{W - p - 1}
\]

The adjusted \( R^2 \) is a correction to the \( R^2 \) which takes into account the number of variables used in the model.

- **MSE**: The mean squared error (MSE) is defined by:

\[
MSE = \frac{1}{W - p} \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2
\]

- **RMSE**: The root mean square of the errors (RMSE) is the square root of the MSE.

- **MAPE**: The *Mean Absolute Percentage Error* is calculated as follows:

\[
MAPE = \frac{100}{W} \sum_{i=1}^{n} w_i \left| \frac{y_i - \hat{y}_i}{y_i} \right|
\]
• **DW**: The Durbin-Watson statistic is defined by:

\[
DW = \frac{\sum_{i=2}^{n} [(y_i - \hat{y}_i) - (y_{i-1} - \hat{y}_{i-1})]^2}{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}
\]

This coefficient is the order 1 autocorrelation coefficient and is used to check that the residuals of the model are not autocorrelated, given that the independence of the residuals is one of the basic hypotheses of linear regression. The user can refer to a table of Durbin-Watson statistics to check if the independence hypothesis for the residuals is acceptable.

• **Cp**: Mallows Cp coefficient is defined by:

\[
C_p = \frac{SSE}{\hat{\sigma}} + 2p^* - W
\]

where \(SSE\) is the sum of the squares of the errors for the model with \(p\) explanatory variables and \(\hat{\sigma}\) is the estimator of the variance of the residuals for the model comprising all the explanatory variables. The nearer the Cp coefficient is to \(p^*\), the less the model is biased.

• **AIC**: Akaike’s Information Criterion is defined by:

\[
AIC = W \ln \left( \frac{SSE}{W} \right) + 2p^*
\]

This criterion, proposed by Akaike (1973) is derived from the information theory and uses Kullback and Leibler’s measurement (1951). It is a model selection criterion which penalizes models for which adding new explanatory variables does not supply sufficient information to the model, the information being measured through the MSE. The aim is to minimize the AIC criterion.

• **SBC**: Schwarz’s Bayesian Criterion is defined by:

\[
SBC = W \ln \left( \frac{SSE}{W} \right) + \ln (W)p^*
\]

This criterion, proposed by Schwarz (1978) is similar to the AIC, and the aim is to minimize it.

• **PC**: Amemiya’s Prediction Criterion is defined by:

\[
PC = \frac{(1 - R^2)(W + p^*)}{W - p^*}
\]
This criterion, proposed by Amemiya (1980) is used, like the adjusted R² to take account of the parsimony of the model.

- **Press RMSE**: Press' statistic is only displayed if the corresponding option has been activated in the dialog box. It is defined by:

\[
\text{Press} = \sum_{i=1}^{n} \nu_i \left( y_i - \hat{y}_{i(-i)} \right)^2
\]

where \( \hat{y}_{i(-i)} \) is the prediction for observation i when the latter is not used for estimating parameters. We then get:

\[
\text{Press RMSE} = \sqrt{\frac{\text{Press}}{n - p}}
\]

Press's RMSE can then be compared to the RMSE. A large difference between the two shows that the model is sensitive to the presence or absence of certain observations in the model.

- **Q²**: The Q² statistic is displayed. It is defined as

\[
Q^2 = 1 - \frac{\text{PressRMSE}}{\text{SSE}}
\]

The closer Q² is to 1, the better and more robust is the model.

The **analysis of variance table** is used to evaluate the explanatory power of the explanatory variables. Where the constant of the model is not set to a given value, the explanatory power is evaluated by comparing the fit (as regards least squares) of the final model with the fit of the rudimentary model including only a constant equal to the mean of the dependent variable. Where the constant of the model is set, the comparison is made with respect to the model for which the dependent variable is equal to the constant which has been set.

The **equation of the model** is then displayed to make it easier to read or re-use the model.

The table of **standardized coefficients** (also called beta coefficients) are used to compare the relative weights of the variables. The higher the absolute value of a coefficient, the more important the weight of the corresponding variable. When the confidence interval around standardized coefficients has value 0 (this can be easily seen on the chart of standardized coefficients), the weight of a variable in the model is not significant.

The **predictions and residuals** table shows, for each observation, its weight, the value of the qualitative explanatory variable, if there is only one, the observed value of the dependent variable, the model's prediction, the residuals, the confidence intervals together with the fitted
prediction and Cook’s D if the corresponding options have been activated in the dialog box. Two types of confidence interval are displayed: a confidence interval around the mean (corresponding to the case where the prediction would be made for an infinite number of observations with a set of given values for the explanatory variables) and an interval around the isolated prediction (corresponding to the case of an isolated prediction for the values given for the explanatory variables). The second interval is always greater than the first, the random values being larger.

The charts which follow show the results mentioned above. If there is only one explanatory variable in the model, the first chart displayed shows the observed values, the regression line and both types of confidence interval around the predictions. The second chart shows the standardized residuals as a function of the explanatory variable. In principle, the residuals should be distributed randomly around the X-axis. If there is a trend or a shape, this shows a problem with the model.

The three charts displayed next show respectively the evolution of the standardized residuals as a function of the dependent variable, the distance between the predictions and the observations (for an ideal model, the points would all be on the bisector), and the standardized residuals on a bar chart. The last chart quickly shows if an abnormal number of values are outside the interval ]-2, 2[ given that the latter, assuming that the sample is normally distributed, should contain about 95% of the data.

For each combination of factors, we draw a ternary diagram. This graph shows a response surface on one of the faces of the polyhedron to which the experimental space corresponds. These graphs facilitate the interpretation of the model and allow to identify the optimal configurations.

Example

A tutorial on the generation and the analysis of a mixture design is available on the Addinsoft website:

http://www.xlstat.com/demo-mixture.htm

References


Kaplan-Meier analysis

Use this tool to build a population survival curve, and to obtain essential statistics such as the median survival time. Kaplan-Meier analysis, which main result is the Kaplan-Meier table, is based on irregular time intervals, contrary to the Life table analysis, where the time intervals are regular.

Description

The Kaplan Meier method (also called product-limit) analysis belongs to the descriptive methods of survival analysis, as does Life table analysis. The life table analysis method was developed first, but the Kaplan-Meier method has been shown to be superior in many cases.

Kaplan-Meier analysis allows to quickly obtain a population survival curve and essential statistics such as the median survival time. Kaplan-Meier analysis, which main result is the Kaplan-Meier table, is based on irregular time intervals, contrary to the Life table analysis, where the time intervals are regular.

Kaplan-Meier analysis is used to analyze how a given population evolves with time. This technique is mostly applied to survival data and product quality data. There are three main reasons why a population of individuals or products may evolve: some individuals die (products fail), some other go out of the surveyed population because they get healed (repaired) or because their trace is lost (individuals move from location, the study is terminated, ...). The first type of data is usually called "failure data", or "event data", while the second is called "censored data".

There are several types of censoring of survival data:

Left censoring: when an event is reported at time t=t(i), we know that the event occurred at t < t(i).

Right censoring: when an event is reported at time t=t(i), we know that the event occurred at t > t(i), if it ever occurred.

Interval censoring: when an event is reported at time t=t(i), we know that the event occurred during [t(i-1); t(i)].

Exact censoring: when an event is reported at time t=t(i), we know that the event occurred exactly at t=t(i).

The Kaplan Meier method requires that the observations are independent. Second, the censoring must be independent: if you consider two random individuals in the study at time t-1, if one of the individuals is censored at time t, and if the other survives, then both must have equal chances to survive at time t. There are four different types of independent censoring:
Simple type I: all individuals are censored at the same time or equivalently individuals are followed during a fixed time interval.

Progressive type I: all individuals are censored at the same date (for example, when the study terminates).

Type II: the study is continued until \( n \) events have been recorded.

Random: the time when a censoring occurs is independent of the survival time.

The Kaplan Meier analysis allows to compare populations, through their survival curves. For example, it can be of interest to compare the survival times of two samples of the same product produced in two different locations. Tests can be performed to check if the survival curves have arisen from identical survival functions. These results can later be used to model the survival curves and to predict probabilities of failure.

### Confidence interval

Computing confidence intervals for the survival function can be done using three different methods:

Greenwood’s method: \( S(T) \pm z_{1-\alpha/2} \sqrt{\frac{\text{var}(S(T))}{S^2(T)}} \)

Exponential Greenwood’s method: \( \exp\left(-\exp\left(-\log(\log(S(T)))\right) \pm z_{1-\alpha/2} \sqrt{\text{var}(S(T))}\right) \)

Log-transformed method: \( S(T)^{1/\theta}, S(T)^{\theta} \text{ avec } \theta = \exp\left(\frac{z_{1-\alpha/2} \sqrt{\text{var}(S(T))}}{\log(S(T))}\right) \)

These three approaches give similar results, but the last ones will be preferred when samples are small.

### Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.
Date data: Select the data that correspond to the times or the dates when the events or the censoring are recorded. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

Weighted data: Activate this option if for a given time, several events are recorded on the same row (for example, at time t=218, 10 failures and 2 censored data have been observed). If you activate this option, the "Event indicator" field replaces the "Status variable" field, and the "Censoring indicator" field replaces the "Event code" and "Censored code" boxes.

Status indicator: Select the data that correspond to an event or censoring data. This field is not available if the "Weighted data" option is checked. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

Event code: Enter the code used to identify an event data within the Status variable. Default value is 1.

Censored code: Enter the code used to identify an event data within the Status variable. Default value is 1.

Censored code: Enter the code used to identify a censored data within the Status variable. Default value is 0.

Event indicator: Select the data that correspond to the counts of events recorded at each time. Note: this option is available only if the "weighted data" option is selected. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

Censoring indicator: Select the data that correspond to the counts of right-censored data recorded at a given time. Note: this option is available only if the "weighted data" option is selected. If a column header has been selected on the first row, check that the "Column labels" option has been activated.
**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Labels included**: Activate this option if the row and column labels have been selected.

**Options** tab:

**Significance level (%)**: Enter the significance level for the comparison tests (default value 5%). This value is also used to determine the confidence intervals around the estimated statistics.

**Confidence interval**: Choose the method to use to compute the confidence interval to be displayed in the outputted table.

**Data options** tab:

**Missing data**:

**Do not accept missing data**: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

**Remove observations**: Activate this option to remove the observations with missing data.

**Groups**:

**By group analysis**: Activate this option and select the data that describe to which group each observation belongs, if you want that XLSTAT performs the analysis on each group separately.

- **Compare**: Activate this option if want to compare the survival curves, and perform the comparison tests.

**Filter**: Activate this option and select the data that describe to which group each observation belongs, if you want that XLSTAT performs the analysis for some groups that you will be able to select in a separate dialog box during the computations. If the “By group analysis” option is also activated, XLSTAT will perform the analysis for each group separately, only for the selected subset of groups.

**Charts** tab:
Survival distribution function: Activate this option to display the charts corresponding to the survival distribution function.

-Log(SDF): Activate this option to display the –Log() of the survival distribution function (SDF).

Log(-Log(SDF)): Activate this option to display the Log(–Log()) of the survival distribution function.

Censored data: Activate this option to identify on the charts the times when censored data have been recorded (the identifier is a hollowed circle "o").

Results

Basic statistics: This table displays the total number of observations, the number of events, and the number of censored data.

Kaplan-Meier table: This table displays the various results obtained from the analysis, including:

- Interval start time: lower bound of the time interval.
- At risk: number of individuals that were at risk.
- Events: number of events recorded.
- Censored: number of censored data recorded.
- Proportion failed: proportion of individuals who "failed" (the event did occur).
- Survival rate: proportion of individuals who "survived" (the event did not occur).
- Survival distribution function (SDF): Probability of an individual to survive until at least the time of interest. Also called cumulative survival distribution function, or survival curve.
- Survival distribution function standard error: standard error of the previous
- Survival distribution function confidence interval: confidence interval of the previous.

Mean and Median residual lifetime: A first table displays the mean residual lifetime, the standard error, and a confidence range. A second table displays statistics (estimator, and confidence range) for the 3 quartiles including the median residual lifetime (50%). The median residual lifetime is one of the key results of the Kaplan-Meier analysis as it allows to evaluate the time remaining for half of the population to "fail".

Charts: Depending on the selected options, up to three charts are displayed: Survival distribution function (SDF), -Log(SDF) and Log(-Log(SDF)).
If the "Compare" option has been activated in the dialog box, XLSTAT displays the following results:

**Test of equality of the survival functions**: This table displays the statistics for three different tests: the Log-rank test, the Wilcoxon test, and the Tarone Ware test. These tests are based on a Chi-square test. The lower the corresponding p-value, the more significant the differences between the groups.

**Charts**: Depending on the selected options, up to three charts with one curve for each group are displayed: Survival distribution function (SDF), -Log(SDF), Log(-Log(SDF)).

**Example**

An example of survival analysis based on the Kaplan-Meier method is available on the Addinsoft website:

[http://www.xlstat.com/demo-km.htm](http://www.xlstat.com/demo-km.htm)

**References**


Life tables

Use this tool to build a survival curve for a given population, and to obtain essential statistics such as the median survival time. Life table analysis, which main result is the life table (also named actuarial table) works on regular time intervals, contrary to the Kaplan Meier analysis, where the time intervals are taken as they are in the data set. XLSTAT enables you to take into account censored data, and grouping information.

Description

Life table analysis belongs to the descriptive methods of survival analysis, as well as Kaplan Meier analysis. The life table analysis method was developed first, but the Kaplan-Meier method has been shown to be superior in many cases.

Life table analysis allows to quickly obtain a population survival curve and essential statistics such as the median survival time. Life table analysis, which main result is the life table (also called actuarial table) works on regular time intervals, contrary to the Kaplan Meier analysis, where the time intervals are taken as they are in the data set.

Life table analysis allows to analyze how a given population evolves with time. This technique is mostly applied to survival data and product quality data. There are three main reasons why a population of individuals or products may evolve: some individuals die (products fail), some other go out of the surveyed population because they get healed (repaired) or because their trace is lost (individuals move from location, the study is terminated, …). The first type of data is usually called "failure data", or "event data", while the second is called "censored data".

There are several types of censoring of survival data:

Left censoring: when an event is reported at time \( t=t(i) \), we know that the event occurred at \( t \leq t(i) \).

Right censoring: when an event is reported at time \( t=t(i) \), we know that the event occurred at \( t \geq t(i) \), if it ever occurred.

Interval censoring: when an event is reported at time \( t=t(i) \), we know that the event occurred during \([t((i-1)); t(i)]\).

Exact censoring: when an event is reported at time \( t=t(i) \), we know that the event occurred exactly at \( t=t(i) \).

The life table method requires that the observations are independent. Second, the censoring must be independent: if you consider two random individuals in the study at time \( t-1 \), if one of the individuals is censored at time \( t \), and if the other survives, then both must have equal chances to survive at time \( t \). There are four different types of independent censoring:
Simple type I: all individuals are censored at the same time or equivalently individuals are followed during a fixed time interval.

Progressive type I: all individuals are censored at the same date (for example, when the study terminates).

Type II: the study is continued until n events have been recorded.

Random: the time when a censoring occurs is independent of the survival time.

The life table method allows to compare populations, through their survival curves. For example, it can be of interest to compare the survival times of two samples of the same product produced in two different locations. Tests can be performed to check if the survival curves have arisen from identical survival functions. These results can later be used to model the survival curves and to predict probabilities of failure.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload**: Click this button to reload the default options.
- **Delete**: Click this button to delete the data selections.

**General** tab:

**Date data**: Select the data that correspond to the times or the dates when the events or the censoring are recorded. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

**Weighted data**: Activate this option if for a given time, several events are recorded on the same row (for example, at time t=218, 10 failures and 2 censored data have been observed). If
you activate this option, the "Event indicator" field replaces the "Status variable" field, and the "Censoring indicator" field replaces the "Event code" and "Censured code" boxes.

**Status indicator**: Select the data that correspond to an event or censoring data. This field is not available if the "Weighted data" option is checked. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

**Event code**: Enter the code used to identify an event data within the Status variable. Default value is 1.

**Censored code Event code**: Enter the code used to identify an event data within the Status variable. Default value is 1.

**Censored code**: Enter the code used to identify a censored data within the Status variable. Default value is 0.

**Event indicator**: Select the data that correspond to the counts of events recorded at each time. Note: this option is available only if the "weighted data" option is selected. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

**Censoring indicator**: Select the data that correspond to the counts of right-censored data recorded at a given time. Note: this option is available only if the "weighted data" option is selected. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Labels included**: Activate this option if the row and column labels have been selected.

**Options** tab:

**Significance level (%)**: Enter the significance level for the comparison tests (default value 5%). This value is also used to determine the confidence intervals around the estimated statistics.

Time intervals:
• **Constant width**: Activate this option if you want to enter the constant interval width. In this case, the lower bound is automatically set to 0.

• **User defined**: Activate this option to define the intervals that should be used to perform the life table analysis. Then select the data that correspond to the lower bound of the first interval and to the upper bounds of all the intervals.

**Data options** tab:

**Missing data**:

**Do not accept missing data**: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

**Remove observations**: Activate this option to remove the observations with missing data.

**Groups**:

**By group analysis**: Activate this option and select the data that describe to which group each observation belongs, if you want that XLSTAT performs the analysis on each group separately.

• **Compare**: Activate this option if you want to compare the survival curves, and perform the comparison tests.

**Filter**: Activate this option and select the data that describe to which group each observation belongs, if you want that XLSTAT performs the analysis for some groups that you will be able to select in a separate dialog box during the computations. If the “By group analysis” option is also activated, XLSTAT will perform the analysis for each group separately, only for the selected subset of groups.

**Charts** tab:

**Survival distribution function**: Activate this option to display the charts corresponding to the survival distribution function.

- **-Log(SDF)**: Activate this option to display the –Log() of the survival distribution function (SDF).

- **Log(-Log(SDF))**: Activate this option to display the Log(–Log()) of the survival distribution function.

**Censored data**: Activate this option to identify on the charts the times when censored data have been recorded (the identifier is a hollowed circle "o").
Results

Basic statistics: This table displays the total number of observations, the number of events, and the number of censored data.

Life table: This table displays the various results obtained from the analysis, including:

- Interval: Time interval.
- At risk: Number of individuals that were at risk during the time interval.
- Events: Number of events recorded during the time interval.
- Censored: Number of censored data recorded during the time interval.
- Effective at risk: Number of individuals that were at risk at the beginning of the interval minus half of the individuals who have been censored during the time interval.
- Survival rate: Proportion of individuals who "survived" (the event did not occur) during the time interval. Ratio of individuals who survived over the individuals who were "effective at risk".
- Conditional probability of failure: Ratio of individuals who failed over the individuals who were "effective at risk".
- Standard error of the conditional probability: Standard error of the previous.
- Survival distribution function (SDF): Probability of an individual to survive until at least the time interval of interest. Also called survivor function.
- Standard error of the survival function: standard error of the previous.
- Probability density function: estimated density function at the midpoint of the interval.
- Standard error of the probability density: standard error of the previous.
- Hazard rate: estimated hazard rate function at the midpoint of the interval. Also called failure rate. Corresponds to the failure rate for the survivors.
- Standard error of the hazard rate: Standard error of the previous.
- Median residual lifetime: Amount of time remaining to reduce the surviving population (individuals at risk) by one half. Also called median future lifetime.
- Median residual lifetime standard error: Standard error of the previous.
**Median residual lifetime**: Table displaying the median residual lifetime at the beginning of the experiment, and its standard error. This statistic is one of the key results of the life table analysis as it allows to evaluate the time remaining for half of the population to "fail".

**Charts**: Depending on the selected options, up to five charts are displayed: Survival distribution function (SDF), Probability density function, Hazard rate function, -Log(SDF), Log(-Log(SDF)).

If the "Compare" option has been activated in the dialog box, XLSTAT displays the following results:

**Test of equality of the survival functions**: This table displays the statistics for three different tests: the Log-rank test, the Wilcoxon test, and the Tarone Ware test. These tests are based on a Chi-square test. The lower the corresponding p-value, the more significant the differences between the groups.

**Charts**: Depending on the selected options, up to five charts with one curve for each group are displayed: Survival distribution function (SDF), Probability density function, Hazard rate function, -Log(SDF), Log(-Log(SDF)).

**Example**

An example of survival analysis by the mean of life tables is available on the Addinsoft website:


**References**


Nelson-Aalen analysis

Use this tool to build cumulative hazard curves using the Nelson-Aalen method. The Nelson-Aalen method allows to estimate the hazard functions based on irregular time intervals, contrary to the Life table analysis, where the time intervals are regular.

Description

The Nelson-Aalen analysis method belongs to the descriptive methods for survival analysis. With the Nelson-Aalen approach you can quickly obtain a curve of cumulative hazard. The Nelson-Aalen method enables to estimate the hazard functions based on irregular time intervals.

Nelson-Aalen analysis is used to analyze how a given population evolves with time. This technique is mostly applied to survival data and product quality data. There are three main reasons why a population of individuals or products may evolve: some individuals die (products fail), some other go out of the surveyed population because they get healed (repaired) or because their trace is lost (individuals move from location, the study is terminated, ...). The first type of data is usually called "failure data", or "event data", while the second is called "censored data".

There are several types of censoring of survival data:

Left censoring: when an event is reported at time \( t=t(i) \), we know that the event occurred at \( t < t(i) \).

Right censoring: when an event is reported at time \( t=t(i) \), we know that the event occurred at \( t > t(i) \), if it ever occurred.

Interval censoring: when an event is reported at time \( t=t(i) \), we know that the event occurred during \( [t(i-1); t(i)] \).

Exact censoring: when an event is reported at time \( t=t(i) \), we know that the event occurred exactly at \( t=t(i) \).

The Nelson-Aalen method requires that the observations are independent. Second, the censoring must be independent: if you consider two random individuals in the study at time \( t-1 \), if one of the individuals is censored at time \( t \), and if the other survives, then both must have equal chances to survive at time \( t \). There are four different types of independent censoring:

Simple type I: all individuals are censored at the same time or equivalently individuals are followed during a fixed time interval.

Progressive type I: all individuals are censored at the same date (for example, when the study terminates).
Type II: the study is continued until n events have been recorded.

Random: the time when a censoring occurs is independent of the survival time.

The Nelson-Aalen analysis allows to compare populations, through their hazards curves. Nelson-Aalen estimator should be preferred to Kaplan-Meier estimator when analyzing cumulative hazard functions. When analyzing cumulative survival functions, Kaplan-Meier estimator should be preferred.

The cumulative hazard function is: \( H(T) = \sum_{T_i \leq T} \frac{d_i}{r_i} \) with \( d_i \) being the number of observation falling at time \( T_i \) and \( r_i \) the number of observation at risk (still in the study) at time \( T_i \).

Several different variance estimators are available:

- Simple: \( \text{var}(H(T)) = \sum_{T_i \leq T} \frac{d_i}{r_i^2} \)

- Plug-in: \( \text{var}(H(T)) = \sum_{T_i \leq T} \frac{d_i (r_i - d_i)}{r_i^3} \)

- Binomial: \( \text{var}(H(T)) = \sum_{T_i \leq T} \frac{d_i (r_i - d_i)}{r_i^2 (r_i - 1)} \)

Confidence intervals can also be obtained:

- Greenwood’s method: \( H(T) \pm z_{1-\alpha/2} \sqrt{\text{var}(H(T))} \)

- Log-transformed method: \( H(T) / \phi, H(T) \cdot \phi \) with \( \phi = \exp \left( \frac{z_{1-\alpha/2} \sqrt{\text{var}(H(T))}}{H(T)} \right) \)

The second one will be preferred with small samples.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

**OK** : Click this button to start the computations.
General tab:

**Date data:** Select the data that correspond to the times or the dates when the events or the censoring are recorded. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

**Weighted data:** Activate this option if for a given time, several events are recorded on the same row (for example, at time t=218, 10 failures and 2 censored data have been observed). If you activate this option, the "Event indicator" field replaces the "Status variable" field, and the "Censoring indicator" field replaces the "Event code" and "Censored code" boxes.

**Status indicator:** Select the data that correspond to an event or censoring data. This field is not available if the "Weighted data" option is checked. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

**Event code:** Enter the code used to identify an event data within the Status variable. Default value is 1.

**Censored code Event code:** Enter the code used to identify an event data within the Status variable. Default value is 1.

**Censored code:** Enter the code used to identify a censored data within the Status variable. Default value is 0.

**Event indicator:** Select the data that correspond to the counts of events recorded at each time. Note: this option is available only if the "weighted data" option is selected. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

**Censoring indicator:** Select the data that correspond to the counts of right-censored data recorded at a given time. Note: this option is available only if the "weighted data" option is selected. If a column header has been selected on the first row, check that the "Column labels" option has been activated.
Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Labels included: Activate this option if the column labels have been selected.

Options tab:

Significance level (%): Enter the significance level for the comparison tests (default value 5%). This value is also used to determine the confidence intervals around the estimated statistics.

Variance: Choose the method to use to compute the variance to be displayed in the outputted table.

Confidence interval: Choose the method to use to compute the confidence interval to be displayed in the outputted table.

Data options tab:

Missing data:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Groups:

By group analysis: Activate this option and select the data that describe to which group each observation belongs, if you want that XLSTAT performs the analysis on each group separately.

• Compare: Activate this option if want to compare the survival curves, and perform the comparison tests.

Filter: Activate this option and select the data that describe to which group each observation belongs, if you want that XLSTAT performs the analysis for some groups that you will be able to select in a separate dialog box during the computations. If the “By group analysis” option is also activated, XLSTAT will perform the analysis for each group separately, only for the selected subset of groups.
**Charts** tab:

**Cumulative hazard function**: Activate this option to display the charts corresponding to the cumulative hazard function.

**Survival distribution function**: Activate this option to display the charts corresponding to the survival distribution function.

**Log(Cumulative hazard function)**: Activate this option to display the Log() of the cumulative hazard function.

**Censored data**: Activate this option to identify on the charts the times when censored data have been recorded (the identifier is a hollowed circle "o").

**Results**

**Basic statistics**: This table displays the total number of observations, the number of events, and the number of censored data.

**Nelson-Aalen table**: This table displays the various results obtained from the analysis, including:

- **Interval start time**: lower bound of the time interval.
- **At risk**: number of individuals that were at risk.
- **Events**: number of events recorded.
- **Censored**: number of censored data recorded.
- **Cumulative hazard function**: hazard associated with an individual at the considered time.
- **Cumulative hazard function error**: standard error of the previous
- **Cumulative hazard function confidence interval**: confidence interval of the previous
- **Survival distribution function**: probability for an individual to survive until the considered time (calculated as $S(T) = \exp(-H(T))$).

**Charts**: Depending on the selected options, up to three charts are displayed: Cumulative hazard function, survival distribution function, and Log(Hazard function).
If the "Compare" option has been activated in the dialog box, XLSTAT displays the following results:

**Test of equality of the survival functions:** This table displays the statistics for three different tests: the Log-rank test, the Wilcoxon test, and the Tarone Ware test. These tests are based on a Chi-square test. The lower the corresponding p-value, the more significant the differences between the groups.

**Charts:** Depending on the selected options, up to three charts with one curve for each group are displayed: Cumulative hazard function, survival distribution function, and Log(Hazard function).

**Example**

An example of survival analysis based on the Nelson-Aalen method is available on the Addinsoft website:


**References**


Cumulative incidence

Use this tool to analyze survival data when competing risks are present. The cumulative incidence allows to estimate the impact of an event when several competitive events may occur. The time intervals should not necessarily be regular. XLSTAT allows the treatment of censored data with competing risks and to compare different groups within the population.

Description

The cumulative incidence allows estimating the impact when several competitive events may occur. It is usually called competing risks case. The time intervals should not necessarily be regular. XLSTAT allows the treatment of censored data in competing risks and to compare different groups within the population.

For a given period, the cumulative incidence is the probability that an observation still included in the analysis at the beginning of this period will be affected by an event during the period. It is especially appropriate in the case of competing risks, that is to say, when several types of events may occur.

This technique is used for the analysis of survival data, whether individuals (cancer research, for example) or products (resistance time of a production tool, for example): some individuals die (in this case we will have 2 causes of death: from the disease or an other cause), the products break (in this case we can model different breaking points), but others leave the study because they heal, you lose track of them (moving for example) or because the study was discontinued. The first type of data is usually called "failure data", or "event data", while the second is called "censored data".

There are several types of censoring of survival data:

Left censoring: when an event is reported at time \( t = t(i) \), we know that the event occurred at \( t * t(i) \).

Right censoring: when an event is reported at time \( t = t(i) \), we know that the event occurred at \( t * t(i) \), if it ever occurred.

Interval censoring: when an event is reported at time \( t = t(i) \), we know that the event occurred during \([t(i-1); t(i)]\).

Exact censoring: when an event is reported at time \( t = t(i) \), we know that the event occurred exactly at \( t = t(i) \).

The cumulative incidence method requires that the observations are independent. Second, the censoring must be independent: if you consider two random individuals in the study at time \( t-1 \), if one of the individuals is censored at time \( t \), and if the other survives, then both must have equal chances to survive at time \( t \). There are four different types of independent censoring:
Simple type I: all individuals are censored at the same time or equivalently individuals are followed during a fixed time interval.

Progressive type I: all individuals are censored at the same date (for example, when the study terminates).

Type II: the study is continued until n events have been recorded.

Random: the time when a censoring occurs is independent of the survival time.

When working with competing risks, the different types of events can happen only once, after the event has occurred, the observation is withdrawn from the analysis. We can calculate the risk of occurrence of an event in the presence of competitive events. XLSTAT allows you to compare the types of events but also to take account of groups of observations (depending on the treatment administered, for example).

The cumulative incidence function is: \( I_k(T) = \sum_{T_j \leq T} \hat{S}(T_{j-1}) \frac{d_{kj}}{n_j} \) for event k at time T. With \( \hat{S}(T_{j-1}) \) being the survival distribution function obtained using the Kaplan-Meier estimator, \( d_{kj} \) being the number of observation failing with event k at time \( T_j \) and \( n_j \), the number of observation at risk (still in the study) at time \( T_j \).

Variance estimator is:

\[
\text{var}(I_k(T)) = \sum_{T_j \leq T} \left[ (I_k(T) - I_k(T_j))^2 \frac{d_j}{n_j(n_j - d_j)} \right] + \sum_{T_j \leq T} \left[ \left( \hat{S}(T_{j-1}) \right)^2 \frac{(n_j - d_j)}{n_j} \frac{d_{kj}}{n_j^2} \right] - 2 \sum_{T_j \leq T} \left[ (I_k(T) - I_k(T_j)) \hat{S}(T_{j-1}) \frac{d_j}{n_j^2} \right].
\]

Confidence intervals are obtained using: \( I_k(T) \exp \left( \pm z_{\alpha/2} \sqrt{\text{var}(I_k(T))} \right) / I_k(T) \log(I_k(T)) \).

**Gray test for group comparison**

Gray test is used to compare groups in a cumulative incidence framework. When competing risks are present, a classic comparison of groups test cannot be applied. Gray developed a test for that case. It is based on a k-sample test that compares the cumulative incidence of a particular type of failure among different groups. For a complete presentation of that test, see Gray (1988).

A p-value for each failure type is obtained for all the groups being studied.
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload default options**: Click this button to reload the default options.
- **Delete data selections**: Click this button to delete the data selections.

**General** tab:

- **Date data**: Select the data that correspond to the times or the dates when the events or the censoring are recorded. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

- **Status indicator**: Select the data that correspond to an event or censoring data. This field is not available if the "Weighted data" option is checked. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

- **Censored code**: Enter the code used to identify a censored data within the Status variable. Default value is 0.

- **Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

- **Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

- **Workbook**: Activate this option to display the results in a new workbook.

- **Labels included**: Activate this option if the column labels have been selected.
**Groups**: Activate this option if you want to group the data. Then select the data that correspond to the group to which each observation belongs.

**Gray test**: Activate this option if you want to perform a Gray test to compare cumulative incidence associated to groups of observations for each failure type.

**Options** tab:

**Significance level (%)**: Enter the significance level for the comparison tests (default value 5%). This value is also used to determine the confidence intervals around the estimated statistics.

**Missing data** tab:

**Do not accept missing data**: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

**Remove observations**: Activate this option to remove the observations with missing data.

**Charts** tab:

**Cumulative incidence function**: Activate this option to display the charts corresponding to the cumulative incidence function.

**Survival distribution function**: Activate this option to display the charts corresponding to the survival distribution function.

**Censored data**: Activate this option to identify on the charts the times when censored data have been recorded (the identifier is a hollowed circle "o").

**Results**

**Basic statistics**: This table displays the total number of observations, the number of events, and the number of censored data.

Each table and plots are displayed for each event type.

**Cumulative incidence**: This table displays the various results obtained from the analysis, including:

- **Interval start time**: lower bound of the time interval.
At risk: number of individuals that were at risk.

Events i: number of events of type i recorded.

All types of events: number of events of all types recorded.

Censored: number of censored data recorded.

Cumulative incidence: Cumulative incidence obtained for event I at the considered time.

Cumulative incidence standard error: standard error of the previous

Cumulative incidence confidence interval: confidence interval of the previous

Cumulative Survival function: This table displays the various results obtained from the analysis, including:

Interval start time: lower bound of the time interval.

At risk: number of individuals that were at risk.

Events i: number of events of type i recorded.

All types of events: number of events of all types recorded.

Censored: number of censored data recorded.

Cumulative survival function: Cumulative survival function obtained for event i at the considered time.

Cumulative survival function standard error: standard error of the previous

Cumulative survival function confidence interval: confidence interval of the previous

Charts: Depending on the selected options, up to three charts are displayed: Cumulative incidence and cumulative survival function.

Gray test: For each failure type the Gray test statistic and the associated degrees of freedom and p-values are displayed.

Example

An example of survival analysis based on the cumulative incidence method is available on the Addinsoft website:
References


Cox Proportional Hazards Model

Use Cox proportional hazards (also known as Cox regression) to model a survival time using quantitative and/or qualitative covariates.

Description

Cox proportional hazards model is a frequently used method in the medical domain (when a patient will get well or not).

The principle of the proportional hazards model is to link the survival time of an individual to covariates. For example, in the medical domain, we are seeking to find out which covariate has the most important impact on the survival time of a patient.

Models

A Cox model is a well-recognized statistical technique for exploring the relationship between the survival of a patient and several explanatory variables. A Cox model provides an estimate of the treatment effect on survival after adjustment for other explanatory variables. It allows us to estimate the hazard (or risk) of death, or other event of interest, for individuals, given their prognostic variables.

Interpreting a Cox model involves examining the coefficients for each explanatory variable. A positive regression coefficient for an explanatory variable means that the hazard is higher. Conversely, a negative regression coefficient implies a better prognosis for patients with higher values of that variable.

Cox’s method does not assume any particular distribution for the survival times, but it rather assumes that the effects of the different variables on survival are constant over time and are additive in a particular scale.

The hazard function is the probability that an individual will experience an event (for example, death) within a small time interval, given that the individual has survived up to the beginning of the interval. It can therefore be interpreted as the risk of dying at time $t$. The hazard function (denoted by $\lambda(t,X)$) can be estimated using the following equation:

$$\lambda(t,X) = \lambda_0(t) \exp(\beta X)$$
The first term depends only on time and the second one depends on X. We are only interested on the second term.

If we only estimate the second term, a very important hypothesis has to be verified: the proportional hazards hypothesis. It means that the hazard ratio between two different observations does not depend on time.

Cox developed a modification of the likelihood function called partial likelihood to estimate the coefficients $\beta$ not taking into account the time dependent term of the hazard function:

$$
\log(L(\beta)) = \sum_{i=1}^{n} \beta X_i - \log\left(\sum_{j=1}^{n} \exp(\beta X_j)\right)
$$

To estimate the $\beta$ parameters of the model (the coefficients of the linear function), we try to maximize the partial likelihood function. Contrary to linear regression, an exact analytical solution does not exist. So an iterative algorithm has to be used. XLSTAT uses a Newton-Raphson algorithm. The user can change the maximum number of iterations and the convergence threshold if desired.

**Strata**

When the proportional hazards hypothesis does not hold, the model can be stratified. If the hypothesis holds on sub-samples, then the partial likelihood is estimated on each sub-sample and these partial likelihoods are summed in order to obtain the estimated partial likelihood. In XLSTAT, strata are defined using a qualitative variable.

**Qualitative variables**

Qualitative covariates are treated using a complete disjunctive table. In order to have independent variables in the model, the binary variable associated to the first modality of each qualitative variable has to be removed from the model. In XLSTAT, the first modality is always selected and, thus, its effect corresponds to a standard. The impact of the other categories are obtained relatively to the omitted modality.

**Ties handling**

The proportional hazards model has been developed by Cox (1972) in order to treat continuous time survival data. However, frequently in practical applications, some observations occur at the same time. The classical partial likelihood cannot be applied. With XLSTAT, you can use two alternative approaches in order to handle ties:

- Breslow's method (1974) (default method): The partial likelihood has the following form:
\[ \log(L(\beta)) = \sum_{i=1}^{T} \beta \sum_{t=1}^{d_i} X_i - d_i \log \left( \sum_{j \in j(i)} \exp(\beta X_j) \right), \]

where \( T \) is the number of times and \( d_i \) is the number of observations associated to time \( t(i) \).

- Efron’s method (1977): The partial likelihood has the following form:

\[ \log(L(\beta)) = \sum_{i=1}^{T} \beta \sum_{t=1}^{d_i} X_i - \sum_{r=0}^{d_i-1} \log \left( \sum_{j \in j(i)} \exp(\beta X_j) - \frac{r}{d_i} \sum_{j=1}^{d_i} \exp(\beta X_j) \right) \]

where \( T \) is the number of times and \( d_i \) is the number of observations associated to time \( t(i) \).

If there are no ties, partial likelihoods are equivalent to Cox partial likelihood.

### Indices to validate the model

XLSTAT-Life allows you to display indices that help validating the model. They are obtained through bootstrapping. As a consequence, for each index you obtain the mean, the standard error, as well as a confidence interval.

The available indices are:

- \( R^2(\text{Cox and Snell}) \): This coefficient, as the classical \( R^2 \), takes values between 0 and 1, and measure the goodness of fit of the model. It equals 1 minus the likelihood ratio that compares the likelihood of the model of interest and the likelihood of the independent model;

- \( R^2(\text{Nagelkerke}) \): This coefficient, as the classical \( R^2 \), takes values between 0 and 1, and measure the goodness of fit of the model. It is equal to the ratio of the Cox and Snell \( R^2 \), divided by 1 minus the likelihood of the independent model;

- Shrinkage index: This index allows quantifying the overfitting of the model. When it is lower than 0.85, on can say that there is some overfitting in the model, and that one should reduce the number of parameters in the model.

- The c index: The concordance index (or general discrimination index) allows evaluating the predictive quality of the model. When it is close to 1, the quality is good, and when it is close to 0, it is bad.

- Sommer’s D: This index is directly related to the c index, as we have \( D=2\ast(c-0.5) \). As a correlation, it takes values between -1 and 1.

These indices make it easier for the user to validate the Cox model that has been obtained. For a detailed description on the bootstrap and validation for the Cox model, please refer to Harrell et al. (1996).
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

Reload default options: Click this button to reload the default options.

Delete data selections: Click this button to delete the data selections.

Change data handling: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Date data: Select the data that correspond to the times or the dates when the events or the censoring are recorded. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

Status indicator: Select the data that correspond to an event or censoring data. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

Event code: Enter the code used to identify an event data within the Status variable. Default value is 1.

Censored code: Enter the code used to identify a censored data within the Status variable. Default value is 0.

Explanatory variables:

Quantitative: Activate this option if you want to include one or more quantitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The
data selected may be of the numerical type. If the variable header has been selected, check that the "Column labels" option has been activated.

**Qualitative**: Activate this option if you want to include one or more qualitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Column labels" option has been activated (see *description*).

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Column labels**: Activate this option if the first row of the data selections (time, status and explanatory variables labels) includes a header.

**Options** tab:

**Significance level (%)**: Enter the significance level for the comparison tests (default value 5%). This value is also used to determine the confidence intervals around the estimated statistics.

**Ties handling**: Select the method to be used when there is more than one observation for one time (see *description*). Default method: Breslow’s method.

**Stop conditions**:

- **Iterations**: Enter the maximum number of iterations for the Newton-Raphson algorithm. The calculations are stopped when the maximum number if iterations has been exceeded. Default value: 100.

- **Convergence**: Enter the maximum value of the evolution of the log of the likelihood from one iteration to another which, when reached, means that the algorithm is considered to have converged. Default value: 0.000001.

**Model selection**: Activate this option if you want to use one of the two selection methods provided:
- **Forward**: The selection process starts by adding the variable with the largest contribution to the model. If a second variable is such that its entry probability is greater than the **entry threshold value**, then it is added to the model. This process is iterated until no new variable can be entered in the model.

- **Backward**: This method is similar to the previous one but starts from a complete model.

**Resampled statistics**: Activate this option in order to display the validation indexes that have been obtained using the bootstrap method (see the description section).

- **Resamplings**: If the previous option has been activated, enter the number of samples to generate when bootstrapping.

**Missing data** tab:

**Do not accept missing data**: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

**Remove observations**: Activate this option to remove the observations with missing data.

**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the variables selected.

**Goodness of fit statistics**: Activate this option to display the table of goodness of fit statistics for the model.

**Test of the null hypothesis H0: beta=0**: Activate this option to display the table of statistics associated to the test of the null hypothesis H0 (likelihood ratio, Wald statistic and score statistic)

**Model coefficients**: Activate this option to display the table of coefficients for the model. The last columns display the hazard ratios and their confidence intervals (the hazard ratio is calculated as the exponential of the estimated coefficient).

**Residuals**: Activate this option to display the residuals for all the observations (deviance residuals, martingale residuals, Schoenfeld residuals and score residuals).

**Charts** tab:

**Survival distribution function**: Activate this option to display the charts corresponding to the cumulative survival distribution function.
-Log(SDF): Activate this option to display the –Log() of the survival distribution function (SDF).

Log(-Log(SDF)): Activate this option to display the Log(–Log()) of the survival distribution function.

Hazard function: Activate this option to display the hazard function when all covariates are at their mean value.

Residuals: Activate this option to display all the residual charts.

Results

XLSTAT displays a large number of tables and charts to help in analyzing and interpreting the results.

Summary statistics: This table displays descriptive statistics for all the variables selected. For the quantitative variables, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed. For qualitative variables, the categories with their respective frequencies and percentages are displayed.

Summary of the variables selection: When a selection method has been chosen, XLSTAT displays the selection summary.

Goodness of fit coefficients: This table displays a series of statistics for the independent model (corresponding to the case where there is no impact of covariates, beta=0) and for the adjusted model.

- Observations: The total number of observations taken into;
- DF: Degrees of freedom;
- -2 Log(Like.): The logarithm of the likelihood function associated with the model;
- AIC: Akaike’s Information Criterion;
- SBC: Schwarz’s Bayesian Criterion;
- Iterations: Number of iterations until convergence.

Test of the null hypothesis H0: beta=0: The H0 hypothesis corresponds to the independent model (no impact of the covariates). We seek to check if the adjusted model is significantly more powerful than this model. Three tests are available: the likelihood ratio test (-2 Log(Like.)), the Score test and the Wald test. The three statistics follow a Chi² distribution whose degrees of freedom are shown.
Model parameters: The parameter estimate, corresponding standard deviation, Wald's Chi², the corresponding p-value and the confidence interval are displayed for each variable of the model. The hazard ratios for each variable with confidence intervals are also displayed.

The residual table shows, for each observation, the time variable, the censoring variable and the value of the residuals (deviance, martingale, Schoenfeld and score).

Charts: Depending on the selected options, charts are displayed: Cumulative Survival distribution function (SDF), -Log(SDF) and Log(-Log(SDF)), hazard function at mean of covariates, residuals.

Example

A tutorial on how to use Cox regression is available on the Addinsoft website:

http://www.xlstat.com/demo-cox.htm

References


Parametric survival models

Use the parametric survival model, also known as Weibull model, to model a survival time using a given probability distribution and, if necessary, quantitative and/or qualitative covariates. These models fit into the framework of the methods for survival data analysis.

Description

The parametric survival model is a method that applies in the context of the analysis of survival data. It allows modelling survival time with right-censored data. It is widely used in medicine (survival time or cure of a patient).

The principle of the parametric survival model is to link the survival time of an individual to a probability distribution (the Weibull distribution is often used) and, when necessary, covariates. For example, in the medical domain, we are seeking to find out which covariate has the most important impact on the survival time of a patient based on a defined distribution.

XLSTAT-Life offers two tools for parametric survival models:

- The **parametric survival regression**, which lets you apply a regression model and analyze the impact of explanatory variables on survival time (assuming an underlying distribution).

- The **parametric survival curve** uses a chosen distribution to model the survival time.

These two methods are exactly equivalent to a methodological standpoint, the difference lies in the fact that, in the first case, explanatory variables are included.

Models

The parametric survival model is similar to the classical regression models in the sense that one tries to link an event (modelled by a date) to a number of explanatory variables.

The parametric survival model is a parametric model. It is based on the assumption that survival times follow a distribution. This assumes a structure for the hazard function that is associated with the chosen distribution.

The parametric survival model is applicable to any situation where one which to study the time of occurrence of an event. This event may be the recurrence of a disease, the response to a treatment, the death, etc. For each subject, we know the date of the latest event (censored or not).

The subjects for which we do not know the status are censored data. The explanatory variables are noted Xj and do not vary along the study.
The \( T \) variable is the time until the event. The parametric survival model can express the risk of occurrence of the event as a function of time \( t \) and of the explanatory variables \( X_j \). These variables may represent risk factors, prognostic factors, treatment, about the intrinsic characteristics, ...

The survival function, noted \( S(t) \), is defined depending on the selected distribution. XLSTAT-Life offers different distributions, among others, the exponential distribution (the survival rate is constant, \( h(t) = l \)), the Weibull distribution (often called Weibull model), the distributions of extreme values...

The exponential and Weibull models are very interesting because they are simultaneously proportional hazards models (such as the Cox model) and accelerated failure time models (for all individuals \( i \) and \( j \) with survival time \( S_i() \) and \( S_j() \), there exists a constant \( \phi \) such that \( S_i(t) = S_j(t^\phi) \) for all \( t \)).

The estimation of such models is done with the maximum likelihood method. Generally \( Y = \log(T) \) is used as dependent variable (for Weibull and exponential models).

Unlike linear regression, an exact analytical solution does not exist. It is therefore necessary to use an iterative algorithm. XLSTAT uses a Newton-Raphson algorithm. The user can change if desired maximum number of iterations and the convergence threshold.

Interpretation of results is done both by studying the graphs associated with cumulative survival functions and studying the tables of coefficients and goodness of fit indices.

**Qualitative variables**

Qualitative covariates are treated using a complete disjunctive table. In order to have independent variables in the model, the binary variable associated to the first modality of each qualitative variable has to be removed from the model. In XLSTAT, the first or the last modality can be selected and, thus, its effect corresponds to a standard. The impacts of the other modalities are obtained relatively to the omitted modality.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

[OK] Click this button to start the computations.
Click this button to close the dialog box without doing any computation.

Click this button to display the help.

Click this button to reload the default options.

Click this button to delete the data selections.

Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

**Date data**: Select the data that correspond to the times or the dates when the events or the censoring are recorded. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

**Status indicator**: Select the data that correspond to an event or censoring data. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

**Event code**: Enter the code used to identify an event data within the Status variable. Default value is 1.

**Censored code**: Enter the code used to identify a censored data within the Status variable. Default value is 0.

**Explanatory variables (in the case of a parametric survival regression)**:

**Quantitative**: Activate this option if you want to include one or more quantitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The data selected may be of the numerical type. If the variable header has been selected, check that the "Column labels" option has been activated.

**Qualitative**: Activate this option if you want to include one or more qualitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Column labels" option has been activated (see *description*).
Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column labels: Activate this option if the first row of the data selections (time, status and explanatory variables labels) includes a header.

Distribution: Select the distribution to be used to fit your model. XLSTAT-Life offers different distributions including Weibull, exponential, extreme value…

Regression weights: Activate this option if you want to carry out a weighted least squares regression. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

Options tab:

Significance level (%): Enter the significance level for the comparison tests (default value 5%). This value is also used to determine the confidence intervals around the estimated statistics.

Initial parameters: Activate this option if you want to take initial parameters into account. If you do not activate this option, the initial parameters are automatically obtained. If a column header has been selected, check that the "Variable labels" option is activated.

Fixed constant: Activate this option to fix the constant of the regression model to a value you then enter (0 by default).

Tolerance: Activate this option to prevent the initial regression calculation algorithm taking into account variables which might be either constant or too correlated with other variables already used in the model (0.0001 by default).

Constraints: Details on the various options are available in the description section.

a1 = 0: Choose this option so that the parameter of the first category of each factor is set to 0.

an = 0: Choose this option so that the parameter of the last category of each factor is set to 0.

Stop conditions:

- Iterations: Enter the maximum number of iterations for the Newton-Raphson algorithm. The calculations are stopped when the maximum number if iterations has been exceeded. Default value: 100.
- **Convergence**: Enter the maximum value of the evolution of the log of the likelihood from one iteration to another which, when reached, means that the algorithm is considered to have converged. Default value: 0.000001.

**Model selection**: Activate this option if you want to use one of the two selection methods provided:

- **Forward**: The selection process starts by adding the variable with the largest contribution to the model. If a second variable is such that its entry probability is greater than the **entry threshold value**, then it is added to the model. This process is iterated until no new variable can be entered in the model.

- **Backward**: This method is similar to the previous one but starts from a complete model.

**Missing data** tab:

- **Do not accept missing data**: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

- **Remove observations**: Activate this option to remove the observations with missing data.

**Outputs** tab:

- **Descriptive statistics**: Activate this option to display descriptive statistics for the variables selected.

- **Goodness of fit statistics**: Activate this option to display the table of goodness of fit statistics for the model.

- **Test of the null hypothesis H0: beta=0**: Activate this option to display the table of statistics associated to the test of the null hypothesis H0 (likelihood ratio, Wald statistic and score statistic)

- **Model coefficients**: Activate this option to display the table of coefficients for the model. The last columns display the hazard ratios and their confidence intervals (the hazard ratio is calculated as the exponential of the estimated coefficient).

- **Residuals and predictions**: Activate this option to display the residuals for all the observations (standardized residuals, Cox-Snell residuals). The value of the estimated cumulative distribution function, the hazard function and the cumulative survival function for each observation are displayed.
**Quantiles**: Activate this option to display the quantiles for each observation (in the case of a parametric survival regression) and for different values of the percentiles (1, 5, 10, 25, 50, 75, 90, 95 and 99 %).

**Charts** tab:

**Survival distribution function**: Activate this option to display the charts corresponding to the cumulative survival distribution function.

**-Log(SDF)**: Activate this option to display the –Log() of the survival distribution function (SDF).

**Log(-Log(SDF))**: Activate this option to display the Log(–Log()) of the survival distribution function.

**Hazard function**: Activate this option to display the hazard function when all covariates are at their mean value.

**Residuals**: Activate this option to display all the residual charts.

**Results**

XLSTAT displays a large number of tables and charts to help in analyzing and interpreting the results.

**Summary statistics**: This table displays descriptive statistics for all the variables selected. For the quantitative variables, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed. For qualitative variables, the categories with their respective frequencies and percentages are displayed.

**Summary of the variables selection**: When a selection method has been chosen, XLSTAT displays the selection summary.

**Goodness of fit coefficients**: This table displays a series of statistics for the independent model (corresponding to the case where there is no impact of covariates, beta=0) and for the adjusted model.

- **Observations**: The total number of observations taken into;
- **DF**: Degrees of freedom;
- **-2 Log(Like.)**: The logarithm of the likelihood function associated with the model;
- **AIC**: Akaike’s Information Criterion;
- **SBC**: Schwarz’s Bayesian Criterion;
• **Iterations**: Number of iterations until convergence.

**Test of the null hypothesis H0: beta=0**: The H0 hypothesis corresponds to the independent model (no impact of the covariates). We seek to check if the adjusted model is significantly more powerful than this model. Three tests are available: the likelihood ratio test (-2 Log(Like.)), the Score test and the Wald test. The three statistics follow a Chi² distribution whose degrees of freedom are shown.

**Model parameters**: The parameter estimate, corresponding standard deviation, Wald's Chi², the corresponding p-value and the confidence interval are displayed for each variable of the model. Confidence intervals are also displayed.

The **residual and predictions** table shows, for each observation, the time variable, the censoring variable, the value of the residuals, the cumulative distribution function, the cumulative survival function and the hazard function.

**Charts**: Depending on the selected options, charts are displayed: Cumulative Survival distribution function (SDF), -Log(SDF) and Log(-Log(SDF)), hazard function, residuals.

**Example**

A tutorial on how to use parametric survival regression is available on the Addinsoft website:

http://www.xlstat.com/demo-survreg.htm

A tutorial on how to use parametric survival curve is available on the Addinsoft website:

http://www.xlstat.com/demo-survcurve.htm

**References**


Sensitivity and Specificity

Use this tool to compute, among others, the sensitivity, specificity, odds ratio, predictive values, and likelihood ratios associated with a test or a detection method. These indices can be used to assess the performance of a test. For example in medicine it can be used to evaluate the efficiency of a test used to diagnose a disease or in quality control to detect the presence of a defect in a manufactured product.

Description

This method was first developed during World War II to develop effective means of detecting Japanese aircrafts. It was then applied more generally to signal detection and medicine where it is now widely used.

The problem is as follows: we study a phenomenon, often binary (for example, the presence or absence of a disease) and we want to develop a test to detect effectively the occurrence of a precise event (for example, the presence of the disease).

Let V be the binary or multinominal variable that describes the phenomenon for N individuals that are being followed. We note by + the individuals for which the event occurs and by - those for which it does not. Let T be a test which goal is to detect if the event occurred or not. T can be a binary (presence/absence), a qualitative (for example the color), or a quantitative variable (for example a concentration). For binary or qualitative variables, let t1 be the category corresponding to the occurrence of the event of interest. For a quantitative variable, let t1 be the threshold value under or above which the event is assumed to happen.

Once the test has been applied to the N individuals, we obtain an individuals/variables table in which for each individual you find if the event occurred or not, and the result of the test.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individu</td>
<td>Maladie</td>
<td>Test</td>
</tr>
<tr>
<td>11</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>12</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>13</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>14</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>15</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>16</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>17</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>18</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>19</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>110</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>111</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>L</th>
<th>M</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individu</td>
<td>Maladie</td>
<td>T1</td>
</tr>
<tr>
<td>11</td>
<td>+</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>+</td>
<td>0.1</td>
</tr>
<tr>
<td>13</td>
<td>+</td>
<td>0.2</td>
</tr>
<tr>
<td>14</td>
<td>+</td>
<td>0.3</td>
</tr>
<tr>
<td>15</td>
<td>-</td>
<td>0.4</td>
</tr>
<tr>
<td>16</td>
<td>+</td>
<td>0.5</td>
</tr>
<tr>
<td>17</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>19</td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td>110</td>
<td>-</td>
<td>4</td>
</tr>
</tbody>
</table>

Case of binary test  Case of a quantitative test

These tables can be summarized in a 2x2 contingency table:
In the example above, there are 25 individuals for whom the test has detected the presence of the disease and 13 for which it has detected its absence. However, for 20 individuals diagnosis is bad because for 8 of them the test contends the absence of the disease while the patients are sick, and for 12 of them, it concludes that they are sick while they are not.

The following vocabulary is being used:

**True positive (TP):** Number of cases that the test declares positive and that are truly positive.

**False positive (FP):** Number of cases that the test declares positive and that in reality are negative.

**True negative (VN):** Number of cases that the test declares negative and that are truly negative.

**False negative (FN):** Number of cases that the test declares negative and that in reality are positive.

Several indices have been developed to evaluate the performance of a test:

**Sensitivity** (equivalent to the True Positive Rate): Proportion of positive cases that are well detected by the test. In other words, the sensitivity measures how the test is effective when used on positive individuals. The test is perfect for positive individuals when sensitivity is 1, equivalent to a random draw when sensitivity is 0.5. If it is below 0.5, the test is counter-performing and it would be useful to reverse the rule so that sensitivity is higher than 0.5 (provided that this does not affect the specificity). The mathematical definition is given by: Sensitivity = TP/(TP + FN).

**Specificity** (also called True Negative Rate): proportion of negative cases that are well detected by the test. In other words, specificity measures how the test is effective when used on negative individuals. The test is perfect for negative individuals when the specificity is 1, equivalent to a random draw when the specificity is 0.5. If it is below 0.5, the test is counter-performing-and it would be useful to reverse the rule so that specificity is higher than 0.5 (provided that this does not affect the sensitivity). The mathematical definition is given by: Specificity = TN/(TN + FP).

**False Positive Rate (FPR):** Proportion of negative cases that the test detects as positive (FPR = 1-Specificity).

**False Negative Rate (FNR):** Proportion of positive cases that the test detects as negative (FNR = 1-Sensitivity)

**Prevalence:** relative frequency of the event of interest in the total sample (TP+FN)/N.

**Positive Predictive Value (PPV):** Proportion of truly positive cases among the positive cases detected by the test. We have PPV = TP / (TP + FP), or PPV = Sensitivity x Prevalence /
\[(\text{Sensitivity} \times \text{Prevalence} + (1 - \text{Specificity})(1 - \text{Prevalence})].\] It is a fundamental value that depends on the prevalence, an index that is independent of the quality of the test.

**Negative Predictive Value (NPV):** Proportion of truly negative cases among the negative cases detected by the test. We have \(\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}},\) or \(\text{PPV} = \frac{\text{Specificity} \times (1 - \text{Prevalence})}{[(\text{Specificity} - 1) \times \text{Prevalence}] + (1 - \text{Sensibility}) \times \text{Prevalence}}\). This index depends also on the prevalence that is independent of the quality of the test.

**Positive Likelihood Ratio (LR+):** This ratio indicates to which point an individual has more chances to be positive in reality when the test is telling it is positive. We have \(\text{LR+} = \frac{\text{Sensitivity}}{1 - \text{Specificity}}\). The LR+ is a positive or null value.

**Negative Likelihood Ratio (LR-):** This ratio indicates to which point an individual has more chances to be negative in reality when the test is telling it is positive. We have \(\text{LR-} = \frac{(1 - \text{Sensitivity})}{\text{Specificity}}\). The LR- is a positive or null value.

**Odds ratio:** The odds ratio indicates how much an individual is more likely to be positive if the test is positive, compared to cases where the test is negative. For example, an odds ratio of 2 means that the chance that the positive event occurs is twice higher if the test is positive than if it is negative. The odds ratio is a positive or null value. We have \(\text{Odds ratio} = \frac{\text{TP} \times \text{TN}}{\text{FP} \times \text{FN}}\).

**Relative risk:** The relative risk is a ratio that measures how better the test behaves when it is a positive report than when it is negative. For example, a relative risk of 2 means that the test is twice more powerful when it is positive that when it is negative. A value close to 1 corresponds to a case of independence between the rows and columns, and to a test that performs as well when it is positive as when it is negative. Relative risk is a null or positive value given by: \(\text{Relative risk} = \frac{\text{TP} / (\text{TP} + \text{FP})}{\text{FN} / (\text{FN} + \text{TN})}\).

**Confidence intervals**

For the various presented above, several methods of calculating their variance and, therefore their confidence intervals, have been proposed. There are two families: the first concerns proportions, such as sensitivity and specificity, and the second ratios, such as LR+, LR- the odds ratio and the relative risk.

For proportions, XLSTAT allows you to use the simple (Wald, 1939) or adjusted (Agresti and Coull, 1998) Wald intervals, a calculation based on the Wilson score (Wilson, 1927), possibly with a correction of continuity, or the Clopper-Pearson (1934) intervals. Agresti and Caffo recommend using the adjusted Wald interval or the Wilson score intervals.

For ratios, the variances are calculated using a single method, with or without correction of continuity.
Once the variance of the above statistics is calculated, we assume their asymptotic normality (or of their logarithm for ratios) to determine the corresponding confidence intervals. Many of the statistics are proportions and should lie between 0 and 1. If the intervals fall partly outside these limits, XLSTAT automatically corrects the bounds of the interval.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

**OK**  Click this button to start the computations.

**Cancel**  Click this button to close the dialog box without doing any computation.

**Help**  Click this button to display the help.

****: Click this button to reload the default options.

**: Click this button to delete the data selections.

**: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

**Data format:**

**2x2 table (Test/Event):** Choose this option if your data are available in a 2x2 contingency table with the tests results in rows and the positive and negative events in columns. You can then specify in which column of the table are located the positive events, and on which row are located the cases detected as positive by the test. The option "Labels included" must be activated if the labels of the rows and columns were selected with the data.

**Individual data:** Choose this option if your data are recorded in a individuals/variables table. You must then select the **event data** that correspond to the phenomenon of interest (for example, the presence or absence of a disease) and specify which **code** is associated with
positive events (for example + when a disease is diagnosed). You must also select the **test data** corresponding to the value of the diagnostic test. This test may be quantitative (concentration), binary (positive or negative) or qualitative (color). If the test is quantitative, you must specify if XLSTAT should consider it as positive when the test is above or below a given threshold value. If the test is qualitative or binary, you must select the value corresponding to a positive test.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Labels included**: Activate this option if the row and column labels are selected. This option is available if you selected the “2x2 table” format.

**Variable labels**: Activate this option if, in column mode, the first row of the selected data contains a header, or in row mode, if the first column of the selected data contains a header. This option is available if you selected the “individual data” format.

**Weights**: Activate this option if the observations are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

**Options** tab:

**Confidence intervals**:

- **Size (%)**: Enter the size of the confidence interval in % (default value: 95).

- **Wald**: Activate this option if you want to calculate confidence intervals on the various indexes using the approximation of the binomial distribution by the normal distribution. Activate "Adjusted" to use the adjustment of Agresti and Coull.

- **Wilson score**: Activate this option if you want to calculate confidence intervals on the various indexes using the Wilson score approximation.

- **Clopper-Pearson**: Activate this option if you want to calculate confidence intervals on the various indexes using the Clopper-Pearson approximation.

- **Continuity correction**: Activate this option if you want to apply the continuity correction to the Wilson score and to the interval on ratios.
A priori prevalence: If you know that the disease involves a certain proportion of individuals in the total population, you can use this information to adjust predictive values calculated from your sample.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

Results

The results are made of the contingency table followed by the table that displays the various indices described in the description section.

Example

An example showing how to compute sensitivity and specificity is available on the Addinsoft website:

http://www.xlstat.com/demo-sens.htm

References


**ROC curves**

Use this tool to generate an ROC curve that allows to represent the evolution of the proportion of true positive cases (also called sensitivity) as a function of the proportion of false positives cases (corresponding to 1 minus specificity), and to evaluate a binary classifier such as a test to diagnose a disease, or to control the presence of defects on a manufactured product.

**Description**

ROC curves have first been developed during World War II to develop effective means of detecting Japanese aircrafts. This methodology was then applied more generally to signal detection and medicine where it is now widely used.

The problem is as follows: we study a phenomenon, often binary (for example, the presence or absence of a disease) and we want to develop a test to detect effectively the occurrence of a precise event (for example, the presence of the disease).

Let $V$ be the binary or multinomial variable that describes the phenomenon for $N$ individuals that are being followed. We note by $+$ the individuals for which the event occurs and by $-$those for which it does not. Let $T$ be a test which goal is to detect if the event occurred or not. $T$ is most of the time continuous (for example, a concentration) but it can also be ordinal (to represent levels).

We want to set the threshold value below or beyond which the event occurs. To do so, we examine a set of possible threshold values for each we calculate various statistics among which the simplest are:

- **True positive (TP):** Number of cases that the test declares positive and that are truly positive.
- **False positive (FP):** Number of cases that the test declares positive and that in reality are negative.
- **True negative (VN):** Number of cases that the test declares negative and that are truly negative.
- **False negative (FN):** Number of cases that the test declares negative and that in reality are positive.
- **Prevalence:** Relative frequency of the event of interest in the total sample $(TP+FN)/N$.

Several indices have been developed to evaluate the performance of a test at a given threshold value:
**Sensitivity** (equivalent to the True Positive Rate): Proportion of positive cases that are well detected by the test. In other words, the sensitivity measures how the test is effective when used on positive individuals. The test is perfect for positive individuals when sensitivity is 1, equivalent to a random draw when sensitivity is 0.5. If it is below 0.5, the test is counterperforming and it would be useful to reverse the rule so that sensitivity is higher than 0.5 (provided that this does not affect the specificity). The mathematical definition is given by: \( \text{Sensitivity} = \frac{TP}{TP + FN} \).

**Specificity** (also called True Negative Rate): proportion of negative cases that are well detected by the test. In other words, specificity measures how the test is effective when used on negative individuals. The test is perfect for negative individuals when the specificity is 1, equivalent to a random draw when the specificity is 0.5. If it is below 0.5, the test is counterperforming-and it would be useful to reverse the rule so that specificity is higher than 0.5 (provided that this does not affect the sensitivity). The mathematical definition is given by: \( \text{Specificity} = \frac{TN}{TN + FP} \).

**False Positive Rate (FPR):** Proportion of negative cases that the test detects as positive (FPR = 1-Specificity).

**False Negative Rate (FNR):** Proportion of positive cases that the test detects as negative (FNR = 1-Sensitivity)

**Prevalence:** relative frequency of the event of interest in the total sample \( (TP+FN)/N \).

**Positive Predictive Value (PPV):** Proportion of truly positive cases among the positive cases detected by the test. We have \( \text{PPV} = \frac{TP}{TP + FP} \), or \( \text{PPV} = \frac{\text{Sensitivity} \times \text{Prevalence}}{\text{Sensitivity} \times \text{Prevalence} + (1-\text{Specificity})(1-\text{Prevalence})} \). It is a fundamental value that depends on the prevalence, an index that is independent of the quality of the test.

**Negative Predictive Value (NPV):** Proportion of truly negative cases among the negative cases detected by the test. We have \( \text{NPV} = \frac{TN}{TN + FN} \), or \( \text{NPV} = \frac{\text{Specificity} \times (1 \text{- Prevalence})}{\text{Specificity} \times (1 \text{- Prevalence}) + (1-\text{Sensibility}) \times \text{Prevalence}} \). This index depends also on the prevalence that is independent of the quality of the test.

**Positive Likelihood Ratio (LR+):** This ratio indicates to which point an individual has more chances to be positive in reality when the test is telling it is positive. We have \( \text{LR+} = \frac{\text{Sensitivity}}{1-\text{Specificity}} \). The LR+ is a positive or null value.

**Negative Likelihood Ratio (LR-):** This ratio indicates to which point an individual has more chances to be negative in reality when the test is telling it is positive. We have \( \text{LR-} = \frac{(1-\text{Sensitivity})}{\text{Specificity}} \). The LR- is a positive or null value.

**Odds ratio:** The odds ratio indicates how much an individual is more likely to be positive if the test is positive, compared to cases where the test is negative. For example, an odds ratio of 2 means that the chance that the positive event occurs is twice higher if the test is positive than if
it is negative. The odds ratio is a positive or null value. We have \( \text{Odds ratio} = \frac{TP \times TN}{FP \times FN} \).

**Relative risk:** The relative risk is a ratio that measures how better the test behaves when it is a positive report than when it is negative. For example, a relative risk of 2 means that the test is twice more powerful when it is positive than when it is negative. A value close to 1 corresponds to a case of independence between the rows and columns, and to a test that performs as well when it is positive as when it is negative. Relative risk is a null or positive value given by: \( \text{Relative risk} = \frac{TP}{(TP+FP)} / \frac{FN}{(FN+TN)} \).

**Confidence intervals**

For the various presented above, several methods of calculating their variance and, therefore their confidence intervals, have been proposed. There are two families: the first concerns proportions, such as sensitivity and specificity, and the second ratios, such as \( \text{LR}^+ \), \( \text{LR}^- \) the odds ratio and the relative risk.

For proportions, XLSTAT allows you to use the simple (Wald, 1939) or adjusted (Agresti and Coull, 1998) Wald intervals, a calculation based on the Wilson score (Wilson, 1927), possibly with a correction of continuity, or the Clopper-Pearson (1934) intervals. Agresti and Caffo recommend using the adjusted Wald interval or the Wilson score intervals.

For ratios, the variances are calculated using a single method, with or without correction of continuity.

Once the variance of the above statistics is calculated, we assume their asymptotic normality (or of their logarithm for ratios) to determine the corresponding confidence intervals. Many of the statistics are proportions and should lie between 0 and 1. If the intervals fall partly outside these limits, XLSTAT automatically corrects the bounds of the interval.

**ROC curve**

The ROC curve corresponds to the graphical representation of the couple \((1 - \text{specificity}, \text{sensitivity})\) for the various possible threshold values.
The area under the curve (AUC) is a synthetic index calculated for ROC curves. The AUC is the probability that a positive event is classified as positive by the test given all possible values of the test. For an ideal model we have AUC = 1 (above in blue), where for a random pattern we have AUC = 0.5 (above in red). One usually considers that the model is good when the value of the AUC is higher than 0.7. A well discriminating model should have an AUC between 0.87 and 0.9. A model with an AUC above 0.9 is excellent.

Sen (1960), Bamber (1975) and Hanley and McNeil (1982) have proposed different methods to calculate the variance of the AUC. All are available in XLSTAT. XLSTAT offers as well a comparison test of the AUC to 0.5, the value 0.5 corresponding to a random classifier. This test is based on the difference between the AUC and 0.5 divided by the variance calculated according to one of the three proposed methods. The statistic obtained is supposed to follow a standard normal distribution, which allows the calculation of the p-value.

The AUC can also be used to compare different tests between them. If the different tests have been applied to different groups of individuals, samples are independent. In this case, XLSTAT uses a Student test to compare the AUCs (which requires assuming the normality of the AUC, which is acceptable if the samples are not too small). If different tests were applied to the same individuals, the samples are paired. In this case, XLSTAT calculates the covariance matrix of the AUCs as described by Delong and Delong (1988) on the basis of Sen’s work (1960), to then calculate the variance of the difference between two AUCs, and to calculate the p-value assuming the normality.
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- ****: Click this button to reload the default options.
- ****: Click this button to delete the data selections.
- ****: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab:**

**Event data**: Select the data that correspond to the phenomenon being studied (for example, the presence or absence of a disease) and specify which code is associated to the positive event (for example D or + for a diseased individual).

**Test data**: Select the data that correspond to test value of the diagnostic. The data must be quantitative. If the data are ordinal, they must be recoded as quantitative data (for example 0,1,2,3,4). You must then specify if one should consider it as positive when the test value is greater or lower than a threshold value determined during the computations.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.
**Variable labels:** Activate this option if, in column mode, the first row of the selected data contains a header, or in row mode, if the first column of the selected data contains a header.

**Weights:** Activate this option if the observations are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

**Options** tab:

**Confidence intervals:**

- **Size (%):** Enter the size of the confidence interval in % (default value: 95).
- **Wald:** Activate this option if you want to calculate confidence intervals on the various indexes using the approximation of the binomial distribution by the normal distribution. Activate "Adjusted" to use the adjustment of Agresti and Coull.
- **Wilson score:** Activate this option if you want to calculate confidence intervals on the various indexes using the Wilson score approximation.
- **Clopper-Pearson:** Activate this option if you want to calculate confidence intervals on the various indexes using the Clopper-Pearson approximation.
- **Continuity correction:** Activate this option if you want to apply the continuity correction to the Wilson score and to the interval on ratios.

**A priori prevalence:** If you know that the disease involves a certain proportion of individuals in the total population, you can use this information to adjust predictive values calculated from your sample.

**Test on AUC:** You can compare the AUC (Area Under the Curve) to 0.5, the value it would have if the test variable were purely random. This test is conducted using the method of calculating the variance chosen above.

**Costs:** Activate this option if you want to evaluate the cost associated with the various possible decisions based on the threshold values of the test variable. You need to enter the costs that correspond to the different situations: TP (true positive), FP (false positive), FN (true negative), TN (true negative).

**Data options** tab:
Missing data:

**Do not accept missing data**: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

**Remove observations**: Activate this option to remove the observations with missing data.

**Ignore missing data**: Activate this option to ignore missing data.

Groups:

**By group analysis**: Activate this option and select the data that describe to which group each observation belongs, if you want that XLSTAT performs the analysis on each group separately.

- **Compare**: Activate this option if want to compare the ROC curves, and perform the comparison tests.

**Filter**: Activate this option and select the data that describe to which group each observation belongs, if you want that XLSTAT performs the analysis for some groups that you will be able to select in a separate dialog box during the computations. If the “By group analysis” option is also activated, XLSTAT will perform the analysis for each group separately, only for the selected subset of groups.

**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the selected variables.

**ROC analysis**: Activate this option to display the table that lists the various indices calculated for each value of the test variable. You can choose to show or not show predictive values, likelihood ratios and of true/false positive and negative counts.

**Test on the AUC**: Activate this option if you want to display the results of the comparison of the AUC to 0.5, the value that corresponds to a random classifier.

**Comparison of the AUCs**: If you have selected several test variables or a group variable, activate this option to compare the AUCs obtained for the different variables or different groups.

**Charts** tab:

**ROC curve**: Activate this option to display the ROC curve.

**True/False +/-**: Activate this option to display the stacked bars chart that shows the % of the TP/TN/FP/FN for the different values of the test variable.
**Decision plot:** Activate this option to display the decision plot of your choice. This plot will help you to decide what level of the test variable is best.

**Comparison of the ROC curves:** Activate this option to display on a single plot the ROC curves that correspond to the various test variables or to the different groups. This option is only available if you select two or more test variables or if a group variable has been selected.

**Results**

**Summary statistics:** In this first table you can find statistics for the selected test(s), followed by a table recalling, for the phenomenon of interest, for the number of occurrences of each event and the prevalence of the positive event in the sample. The row displayed in bold corresponds to the positive event.

**ROC curve:** The ROC curve is then displayed. The straight dotted line that goes from (0;0) to (1;1) corresponds to the curve of a random test with no discrimination. The colored line corresponds to the ROC curve. Small squares correspond to observations (one square per observed value of the test variable).

**ROC analysis:** This table displays for each possible threshold value of the test variable, the various indices presented in the description section. On the line below the table you'll find a reminder of the rule set out in the dialog box to identify positive cases compared to the threshold value. Below the table you will find a stacked bars chart showing the evolution of the TP, TN, FP, FN depending on the value of the threshold value. If the corresponding option was activated, the decision plot is then displayed (for example, changes in the cost depending on the threshold value).

**Area under the curve (AUC):** This table displays the AUC, its standard error and a confidence interval.

**Comparison of the AUC to 0.5:** These results allow to compare the test to a random classifier. The confidence interval corresponds to the difference. Various statistics are then displayed including the p-value, followed by the interpretation of the comparison test.

**Comparison of the AUCs:** If you selected several test variables, once the above results are displayed for each variable, you will find the covariance matrix of the AUC, followed by the table of differences for each pair of AUCs with as comments the confidence interval, and then the table of the p-values. Values in bold correspond to significant differences. Last, a graph that compares the ROC curves displayed.
Example

An example showing how to compute ROC curves is available on the Addinsoft website:

http://www.xlstat.com/demo-roc.htm

An example showing how to compute ROC curves and compare them is available on the Addinsoft website:

http://www.xlstat.com/demo-roccompare.htm

References


Bamber D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. Journal of Mathematical Psychology, 12, 387-415.


Method comparison

Use this tool to compare a method to a reference method or to a comparative method. Tests, confidence intervals are computed, and several plots are displayed to visualize differences, including the Bland Altman plot and the Difference plot. With this tool you are able to meet the recommendations of the Clinical and Laboratory Standards Institute (CLSI).

Description

When developing a new method to measure the concentration or the quantity of an element (molecule, micro organism, …) you might want to check whether it gives results that are similar to a reference or comparative method or not. If there is a difference, you might be interested in knowing if this is due to a bias that depends on where you are on the scale variation. If a new measurement method is cheaper than a standard, but if there is a known and fixed bias, you might take into account the bias while reporting the results.

XLSTAT provides a series of tools to evaluate the performance of a method compared to another.

Repeatability analysis

Repeatability and reproducibility analysis of measurement systems is available in the XLSTAT-SPC module (see gage R&R). The repeatability analysis provided here is a lighter version that is aimed at analyzing the repeatability of each method separately and to compare the repeatability of the methods. To evaluate the repeatability of a method, one needs to have several replicates. Replicates can be specified using the “Groups” field of the dialog box (replicates must have the same identifier). This corresponds to the case where several measures are taken on a given sample. If the method is repeatable, the variance within the replicates is low. XLSTAT computes the repeatability as a standard deviation and displays a confidence interval. Ideally, the confidence interval should contain 0. Repeatability plots are displayed for each method and show for each subject the standard deviation versus the mean computed across replicates.

Paired t-test

Among the comparison methods, a paired t-test can be computed. The paired t-test allows to test the null hypothesis $H_0$ that the mean of the differences between the results of the two methods is not different from 0, against an alternative hypothesis $H_a$ that it is.
**Scatter plots**

First, you can draw a scatter plot to compare the reference or comparative method against the method being tested. If the data are on both sides of the identity line (bisector) and close to it, the two methods give close and consistent results. If the data are above the identify line, the new method overestimates the value of interest. If the data are under the line, the new method underestimates the value of interest, at least compared to the comparative or reference method. If the data are crossing the identify line, there is a bias that depends on where you are on the scale of variation. If the data are randomly scattered around the identity line with some observations that are far from it, the new method is not performing well.

1. Consistent methods
2. Positive constant bias
3. Negative constant bias
4. Linear bias
5. Inconsistent methods

**Bias**

The bias is estimated as the mean of the differences (or differences %, or ratio) between the two methods. If replicates are available, a first step computes the mean of the replicates. The standard deviation is computed as well as a confidence interval. Ideally, the confidence interval should contain 0.

Note: The bias is computed for the criterion that has been chosen for the Bland Altman analysis (difference, difference % or ratio).

**Bland Altman and related comparison methods**

Bland and Altman recommend plotting the difference (T-S) between the test (T) and comparative or reference method (S) against the average (T+S)/2 of the results obtained from the two methods. In the ideal case, there should not be any correlation between the difference and the average whether there is a bias or not. XLSTAT tests whether the correlation is significantly different from 0 or not. Alternative possibilities are available for the ordinates of the plot: you can choose between the difference (T-S), the difference as a % of the sum (T-S)/(T+S), and the ratio (T/S). On the Bland Altman plot, XLSTAT displays the bias line, the confidence lines around the bias, and the confidence lines around the individual differences (or the difference % or the ratio).

**Histogram and box plot**

Histogram and box plots of the differences (or difference % or ratio) are plotted to validate the hypothesis that the difference (or difference % or ratio) is normally distributed, which is used to compute confidence intervals around the bias and the individual differences. When the size of the samples is small, the histogram is of little interest and one should only consider the box
plot. If the distribution does not seem to be normal, one might want to verify that point with a normality test, and one should consider with caution the confidence intervals.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- ![Reload](image)
  - Click this button to reload the default options.
- ![Delete](image)
  - Click this button to delete the data selections.
- ![Change Data](image)
  - Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General tab:**

**Data (Method 1):** Select the data that correspond to the first method, or to the reference method. If the name of the method is available in the first position of the data, make sure you activate the “Variable labels” option.

**Data (Method 2):** Select the data that correspond to the second method. If the name of the method is available in the first position of the data, make sure you activate the “Variable labels” option.

**Groups:** If replicates are available, select in this field the identifier of the measures. Two measures with the same group identifier are considered as replicates. XLSTAT uses the mean of the replicates for the analysis, and will provide you with repeatability results.
Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if, in column mode, the first row of the selected data contains a header, or in row mode, if the first column of the selected data contains a header.

Options tab:

Bland Altman analysis: Activate this option if you want to run a Bland Altman analysis and/or display a Bland Altman plot. Then, you need to specify the variable to use for the ordinates.

Difference analysis: Activate this option if you want to run a Difference analysis and/or display a Difference plot. Then, you need to specify the variable to use for the abscissa.

Significance level (%): Enter the size value of the significance level that is used to determine the critical value of the Student’s t test and to generate the conclusion of the test.

Confidence intervals (%): Enter the size of the confidence interval in % (default value: 95).

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Ignore missing data: Activate this option to ignore missing data. This option is only visible if the “Groups” option is active.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the two methods.

Paired t-test: Activate this option to display the results corresponding to a paired Student’s t test to test whether the difference between the two methods is significant or not.

Bland Altman analysis: Activate this option to compute the Bias statistic and the corresponding confidence interval.
Charts tab:

**Scatter plot:** Activate this option to display the scatter plot showing on the abscissa the reference or comparative method, and on the ordinates the test method.

**Bland Altman plot:** Activate this option to display the Bland Altman plot.

**Histogram:** Activate this option to display the histogram of the differences (or differences % or ratios).

**Box plot:** Activate this option to display the box plot of the differences (or differences % or ratios).

**Difference plot:** Activate this option to display the difference plot.

Results

**Summary statistics:** In this first table you can find the basic descriptive statistics for each method.

**t-test for two paired samples:** These results correspond to the test of the null hypothesis that the two methods are not different versus the alternative hypothesis that they are. Note: this test is made using the assumption that the samples obtained with both methods are normally distributed.

**A scatter plot** is then displayed to allow comparing the two methods visually. The identity line is displayed on the plot. It corresponds to the ideal case where the samples on which the two methods are applied are identical and where the two methods would give exactly the same results.

The **Bland Altman analysis** is starts with an estimate of the bias, using the criterion that has been chosen (difference, difference in %, or ratio), the standard error and a confidence interval being as well displayed. The Bland Altman plot is displayed so that the difference between the two methods can be visualized. XLSTAT displays the correlation between the abscissa and the ordinates. One would expect it to be non-significantly different from 0, which means the confidence interval around the correlation should include 0.

The **histogram and the box plot** allow to visualize how the difference (or the difference % or the ratio) is distributed. A normality assumption is used when computing the confidence interval around the differences.

The **Difference plot** shows the difference between the two methods against the average of both methods, or against the reference method with an estimate of the bias, using the criterion
that has been chosen (difference, difference in %, or ratio), the standard error and a confidence interval being as well displayed.

Example

An example showing how to compare two methods is available on the Addinsoft website:

http://www.xlstat.com/demo-bland.htm

References


Passing and Bablok regression

Use this tool to compare two methods of measurement by a minimum of assumptions about their distribution.

Description

Passing and Bablok (1983) developed a regression method that allows comparing two measurement methods (for example, two techniques for measuring concentration of an analyte), which overcomes the assumptions of the classical linear regression single that are inappropriate for this application. As a reminder the assumptions of the OLS regression are:

- The explanatory variable, $X$ in the model $y(i) = a + b.x(i) + \varepsilon(i)$, is deterministic (no measurement error),
- The dependent variable $Y$ follows a normal distribution with expectation $aX$
- The variance of the measurement error is constant.

Furthermore, extreme values can highly influence the model.

Passing and Bablok proposed a method which overcomes these assumptions: the two variables are assumed to have a random part (representing the measurement error and the distribution of the element being measured in medium) without needing to make assumption about their distribution, except that they both have the same distribution. We then define:

- $y(i) = a + b.x(i) + \eta(i)$
- $x(i) = A + B.y(i) + \xi(i)$

Where $\eta$ and $\xi$ follow the same distribution. The Passing and Bablok method allows calculating the $a$ and $b$ coefficients (from which we deduce $A$ and $B$ using $B = 1/b$ and $A = -a/b$) as well as a confidence interval around these values. The study of these values helps comparing the methods. If they are very close, $b$ is close to 1 and $a$ is close to 0.

Passing and Bablok also suggested a linearity test to verify that the relationship between the two measurement methods is stable over the interval of interest. This test is based on a CUSUM statistic that follows a Kolmogorov distribution. XLSTAT provides the statistic, the critical value for the significance level chosen by the user, and the p-value associated with the statistic.
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- OK: Click this button to start the computations.
- Cancel: Click this button to close the dialog box without doing any computation.
- Help: Click this button to display the help.
- : Click this button to reload the default options.
- : Click this button to delete the data selections.
- : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

X: Select the data that correspond to the method that will be displayed on the abscissa axis. If the name of the variable is available in the first position of the data, make sure you activate the “Variable labels” option.

Y: Select the data that correspond to the method that will be displayed on the ordinates axis. If the name of the variable is available in the first position of the data, make sure you activate the “Variable labels” option.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if, in column mode, the first row of the selected data contains a header, or in row mode, if the first column of the selected data contains a header.
**Options** tab:

**Confidence intervals (%)**: Enter the size of the confidence interval in % (default value: 95).

**Missing data** tab:

**Do not accept missing data**: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

**Remove observations**: Activate this option to remove the observations with missing data.

**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the two methods.

**Charts** tab:

**Predictions and residuals**: Activate this option to display table corresponding to the input data, the predictions, the residuals and the perpendicular residuals.

**Results**

**Summary statistics**: In this first table you can find the basic descriptive statistics for each method.

**Coefficients of the model**: In this table are shown the coefficients a and b of the model and their respective confidence intervals.

**Predictions and residuals**: This table displays for each observation, the value of X, the value of Y, the model prediction, the residual and the perpendicular residual (the distance to the regression line by orthogonal projection).

The charts allow to visualize the regression line, the observations and the model $Y = X$ (corresponding to the bisector of the plane) and the corresponding confidence interval calculated using the RMSE obtained from the model of Passing and Bablok but with the usual method for linear regression. This chart allows to visually check if the model is far from the model that would correspond to the hypothesis that the methods are identical.
Example

An example showing how to compare two methods using the Passing and Bablok regression is available on the Addinsoft website:

http://www.xlstat.com/demo-passing.htm

References

Deming regression

Use this tool to compare two methods of measurement with error on both measurements using Deming regression.

Description

Deming (1943) developed a regression method, that allows comparing two measurement methods (for example, two techniques for measuring concentration of an analyte), which supposes that measurement error are present in both X and Y. It overcomes the assumptions of the classical linear regression that are inappropriate for this application. As a reminder the assumptions of the OLS regression are

- The explanatory variable, X in the model $y(i)=a+b.x(i)+\varepsilon(i)$, is deterministic (no measurement error),
- The dependent variable Y follows a normal distribution with expectation $aX$
- The variance of the measurement error is constant.

Furthermore, extreme values can highly influence the model.

Deming proposed a method which overcomes these assumptions: the two variables are assumed to have a random part (representing the measurement). The distribution has to be normal. We then define:

- $y(i)=y(i)^*+\varepsilon(i)$
- $x(i)=x(i)^*+\eta(i)$

Assume that the available data $(y_i, x_i)$ are mismeasured observations of the “true” values $(y(i)^*, x(i)^*)$ where errors $\varepsilon$ and $\eta$ are independent. The ratio of their variances is assumed to be known:

$$\delta=\sigma^2(\varepsilon)/\sigma^2(\eta)$$

In practice, the variance of the x and y is often unknown which complicates the estimate of $\delta$ but when the measurement methods for x and y are the same they are likely to be equal so that $\delta=1$ for this case. XLSTAT-Life allows you to define $\delta$.

We seek to find the line of “best fit” $y^* = a + b x^*$, such that the weighted sum of squared residuals of the model is minimized.

Where $\eta$ and $\varepsilon$ follow a normal distribution. The Deming method allows calculating the a and b coefficients as well as a confidence interval around these values. The study of these values helps comparing the methods. If they are very close, then b is close to 1 and a is close to 0.
The Deming regression has two forms:

- Simple Deming regression: The error terms are constant and the ratio between variances has to be chosen (with default value being 1). The estimation is very simple using a direct formula (Deming, 1943).

- Weighted Deming regression: In the case where replicates of the experiments are present, the weighted Deming regression supposes that the error terms are not constant but only proportional. Within each replication, you can take into account the mean or the first experiment to estimate the coefficients. In that case, a direct estimation is not possible. An iterative method is used (Linnet, 1990).

Confidence interval of the intercept and slope coefficient are complex to compute. XLSTAT-Life uses a jackknife approach to compute confidence intervals, as stated in Linnet (1993).

A linearity test to verify that the relationship between the two measurement methods is stable over the interval of interest is also displayed. This test is based on a CUSUM statistic that follows a Kolmogorov distribution. XLSTAT provides the statistic, the critical value for the significance level chosen by the user, and the p-value associated with the statistic.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload**: Click this button to reload the default options.
- **Delete**: Click this button to delete the data selections.
- **Data Handling**: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.
General tab:

X: Select the data that correspond to the method that will be displayed on the abscissa axis. If the name of the variable is available in the first position of the data, make sure you activate the “Variable labels” option.

Y: Select the data that correspond to the method that will be displayed on the ordinates axis. If the name of the variable is available in the first position of the data, make sure you activate the “Variable labels” option.

Replicates: Activate this option if more than one replicate has been measured. Select the data that associate the replicates of the experiments to the observations. If the name of the variable is available in the first position of the data, make sure you activate the “Variable labels” option.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if, in column mode, the first row of the selected data contains a header, or in row mode, if the first column of the selected data contains a header.

Constant error: Activate this option if the errors of both X and Y are supposed to be constant.

Proportional error: Activate this option if the errors of both X and Y are supposed to be proportional. This option is available only if replicates have been selected.

Options tab:

Confidence intervals (%): Enter the size of the confidence interval in % (default value: 95).

Variance ratio: If the constant error option is selected. Enter the variance ratio (delta parameter). See the description part of this chapter).

Replicates: If the replicates have been selected with proportional error. Select the method to estimate the parameter. In the weighted Deming regression, within each replicate, you can use the mean or the first replicate in the iterative algorithm. Four options are available, the default one being mean versus mean.
**Missing data** tab:

**Do not accept missing data**: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

**Remove observations**: Activate this option to remove the observations with missing data.

**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the two methods.

**Charts** tab:

**Predictions and residuals**: Activate this option to display the table corresponding to the input data, the predictions and the residuals.

**Results**

**Summary statistics**: In this first table you can find the basic descriptive statistics for each method.

**Coefficients of the model**: In this table are shown the coefficients $a$ and $b$ of the model and their respective confidence intervals.

**Predictions and residuals**: This table displays for each observation, the value of $X$, the value of $Y$, the model prediction and the residuals.

The charts allow to visualize the regression line, the observations and the model $Y = X$ (corresponding to the bisector of the plane) and the corresponding confidence interval calculated using the RMSE obtained from the model of Deming but with the usual method for linear regression. This chart enables to visually check if the model is far from the model that would correspond to the hypothesis that the methods are identical.
Example

An example showing how to compare two methods using the Deming regression is available on the Addinsoft website:

http://www.xlstat.com/demo-deming.htm

References


Differential expression

Use this tool to detect the most differentially expressed elements according to explanatory variables within a features/individuals data matrix that may be very large.

Description

Differential expression allows identifying features (genes, proteins, metabolites…) that are significantly affected by explanatory variables. For example, we might be interested in identifying proteins that are differentially expressed between healthy and diseased individuals. In this kind of studies, data often have a very important size ( = high-throughput data). At this stage, we may talk about omics data analyses, in reference to analyses performed over the genome (genomics) or the transcriptome (transcriptomics) or the proteome (proteomics) or the metabolome (metabolomics), etc.

In order to test if features are differentially expressed, we often use traditional statistical tests. However, the size of the data may cause problems in terms of computation time as well as readability and statistical reliability of results. Those tools must therefore be slightly adapted in order to overcome these problems.

Statistical tests

The statistical tests proposed in the differential expression tool in XLSTAT are traditional parametric or non-parametric tests: Student t-test, ANOVA, Mann-Whitney, Kruskal-Wallis).

Post-hoc corrections

The p-value represents the risk that we take to be wrong when stating that an effect is statistically significant. Running a test several times increases the number of computed p-values, and subsequently the risk of detecting significant effects which are not significant in reality. Considering a significance level alpha of 5%, we would likely find 5 significant p-values by chance over 100 computed p-values. When working with high-throughput data, we often test the effect of an explanatory variable on the expression of thousands of genes, thus generating thousands of p-values. Consequently, p-values should be corrected ( = increased = penalized) as their number grow. XLSTAT proposes three common p-value correction methods:

Benjamini-Hochberg: this procedure makes sure that p-values increase both with their number and the proportion of non-significant p-values. It is part of the FDR (False Discovery Rate) correction procedure family. The Benjamini-Hochberg correction is poorly conservative ( = not very severe). It is therefore adapted to situations where we are looking for a large number of genes which are likely affected by the explanatory variables. It is widely used in differential expression studies.
The corrected p-value according to the Benjamini-Hochberg procedure is defined by:
\[ p_{\text{Benjamini-Hochberg}} = \min( p^* \frac{nbp}{j}, 1) \]
where \( p \) is the original (uncorrected) p-value, \( nbp \) is the number of computed p-values in total and \( j \) is the rank of the original p-value when p-values are sorted in ascending order.

Benjamini-Yekutieli: this procedure makes sure that p-values increase both with their number and the proportion of non-significant p-values. It is part of the FDR (False Discovery Rate) correction procedure family. In addition to Benjamini-Hochberg’s approach, it takes into account a possible dependence between the tested features, making it more conservative than this procedure. However, it is far less stringent than the Bonferroni approach which we describe just after.

The corrected p-value according to the Benjamini-Yekutieli procedure is defined by:
\[ p_{\text{Benjamini-Yekutieli}} = \min\left[ ( p^* \frac{nbp}{\sum_{i=1}^{nbp} \frac{1}{i}}) / j, 1 \right] \]
where \( p \) is the original p-value, \( nbp \) is the number of computed p-values in total and \( j \) is the rank of the original p-value when p-values are sorted in ascending order.

Bonferroni: p-values increase only with their number. This procedure is very conservative. It is part of the FWER (Familywise error rate) correction procedure family. It is rarely used in differential expression analyses. It is useful when the goal of the study is to select a very low number of differentially expressed features.

The corrected p-value according to the Bonferroni procedure is defined by:
\[ p_{\text{Bonferroni}} = \min( p^* \frac{nbp}{1}, 1) \]
where \( p \) is the original p-value and \( nbp \) is the number of computed p-values in total.

**Multiple pairwise comparisons**

After one-way ANOVAs or Kruskal-Wallis tests, it is possible to perform multiple pairwise comparisons for each feature taken separately. XLSTAT provides different options including:

- Tukey’s HSD test: this test is the most used (HSD: Honestly Significant Difference).
- Fisher’s LSD test: this is Student’s test that tests the hypothesis that all the means for the various categories are equal (LSD: Least Significant Difference).
- Bonferroni’s t* test: this test is derived from Student’s test and is less reliable as it takes into account the fact that several comparisons are carried out simultaneously. Consequently, the significance level of the test is modified according to the following formula:
where \( g \) is the number of categories of the factor whose categories are being compared.

- Dunn-Sidak's test: This test is derived from Bonferroni's test. It is more reliable in some situations.

\[
\alpha^* = 1 - \left(1 - \alpha \right)^{2/(g(g-1))}
\]

**Non-specific filtering**

Before launching the analyses, it is interesting to filter out features with very poor variability across individuals. Non-specific filtering has two major advantages:

- It allows computations to focus less on features which are very likely to be not differentially expressed thus saving computation time.

- It limits post-hoc penalizations, as fewer p-values are computed.

Two methods are available in XLSTAT:

- The user specifies a variability threshold (interquartile range or standard deviation), and features with lower variability are eliminated prior to analyses.

- The user specifies a percentage of features with low variability (interquartile range or standard deviation) to be removed prior to analyses.

**Biological effects and statistical effects: the volcano plot**

A statistically significant effect is not necessarily interesting at the biological scale. An experiment involving very precise measurements with a high number of replicates may provide low p-values associated to very weak biological differences. It is thus recommended to keep an eye on biological effects and not to rely only on p-values. The volcano plot is a scatter chart that combines statistical effects on the y-axis and biological effects on the x-axis for a whole features/individuals data matrix. The only constraint is that it can only be executed to examine the difference between the levels of two-level qualitative explanatory variables.

The y axis coordinates are -log10(p-values) making the chart easier to read: high values reflect the most significant effects whereas low values correspond to effects which are less significant.

XLSTAT provides two ways of building the x axis coordinates:
- Difference between the mean of the first level and the mean of the second, for each feature. Generally, we use this format when handling data on a transformed scale such as log or square root.

- Log2 of fold change between the two means: log2(mean1/mean2). This format should preferably be used with untransformed data.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- ![Reload Default Options](image)
  - Click this button to reload the default options.
- ![Delete Data Selections](image)
  - Click this button to delete the data selections.

**General** tab:

**Features/individuals table**: Select the features/individuals data matrix in the Excel worksheet. The data selected must be of type numeric.

**Data format**:

**Features in rows**: activate this option if features are stored in lines and individuals (or samples) are stored in columns.

**Features in columns**: activate this option if features are stored in columns and individuals (or samples) are stored in lines.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.
Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Labels included: Activate this option if feature and individual labels are included in the selection.

Cluster features: Activate this option if you wish the heat map to include clustering on features

Cluster individuals: Activate this option if you wish the heat map to include clustering on individuals (or samples).

Options tab:

Center: Activate this option to center each row separately.

Reduce: Activate this option to reduce each row separately.

Non-specific filtering: Activate this option to filter out features with low variability prior to computations.

Criterion and threshold: Select the non-specific filtering criterion.

- Standard deviation<: all features with a standard deviation lower than the selected threshold are removed.
- Interquartile range<: all features with an interquartile range lower than the selected threshold are removed.
- % (Std. dev.): a percentage of features with low standard deviation are removed. The percentage should be indicated in the threshold box
- % (IQR): a percentage of features with low interquartile range are removed. The percentage should be indicated in the threshold box.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.
**Estimate missing data:** Activate this option to estimate missing data before starting the computations.

- **Mean or mode:** Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

- **Nearest neighbour:** Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.

**Outputs** tab:

**Descriptive statistics:** Activate this option to display descriptive statistics for the selected variables.

**Charts** tab:

**Color scale:** select the color range to use in the heat map (red to green through black; red to blue through white; red to yellow).

**Color calibration:**

- **Automatic:** Activate this option if you want XLSTAT to automatically choose boundary values that will delimit the heatmap color range.

- **User defined:** Activate this option if you want to manually choose the minimum (Min) and maximum (Max) values that will delimit the heatmap color range.

**Width** and **height:** select a magnification factor for the heat map’s width or height.

**Results**

**Summary statistics:** The tables of descriptive statistics show the simple statistics for all individuals. The number of observations, missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed.

**Heat map:** The features dendrogram is displayed vertically (rows) and the individuals dendrogram is displayed horizontally (columns). A heat map is added to the chart, reflecting data values.

Similarly expressed features are characterized by horizontal rectangles of homogeneous color along the map.
Similar individuals are characterized by vertical rectangles of homogeneous color along the map.

Clusters of similar individuals characterized by clusters of similarly expressed features can be detected by examining rectangles or squares of homogeneous color at the intersection between feature clusters and individual clusters inside the map.

Example

A tutorial on differential expression analysis is available on the Addinsoft website:

http://www.xlstat.com/demo-omicsdiff.htm

References


Heat maps

Use this tool to perform clustering on both columns and rows of a features/individuals data matrix, and to draw heat maps.

Description

While exploring features/individuals matrices, it is interesting to examine how correlated features (i.e. genes, proteins, metabolites) correspond to similar individuals (i.e. samples). For example, a cluster of diseased kidney tissue samples may be characterized by a high expression of a group of genes, compared to other samples. The heat maps tool in XLSTAT allows performing such explorations.

How it works in XLSTAT

Both features and individuals are clustered independently using ascendant hierarchical clustering based on Euclidian distances, optionally preceded by the k-means algorithm depending on the matrix’s size. The data matrix’s rows and columns are then permuted according to corresponding clusterings, which brings similar columns closer to each other and similar lines closer to each other. A heat map is then displayed, reflecting data in the permuted matrix (data values are replaced by corresponding color intensities).

Non-specific filtering

Before launching the analyses, it is interesting to filter out features with very poor variability across individuals. In heat map analysis, non-specific filtering has two major advantages:

- It allows computations to focus less on features which are very likely to be not differentially expressed thus saving computation time.

- It improves the readability of the heat map chart.

Two methods are available in XLSTAT:

- The user specifies a variability threshold (interquartile range or standard deviation), and features with lower variability are eliminated prior to analyses.

- The user specifies a percentage of features with low variability (interquartile range or standard deviation) to be removed prior to analyses.
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

Reload: Click this button to reload the default options.

Delete: Click this button to delete the data selections.

General tab:

Features/individuals table: Select the features/individuals data matrix in the Excel worksheet. The data selected must be of type numeric.

Data format:

Features in rows: activate this option if features are stored in lines and individuals (or samples) are stored in columns.

Features in columns: activate this option if features are stored in columns and individuals (or samples) are stored in lines.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Labels included: Activate this option if feature and individual labels are included in the selection.
Cluster features: Activate this option if you wish the heat map to include clustering on features.

Cluster individuals: Activate this option if you wish the heat map to include clustering on individuals (or samples).

Options tab:

Center: Activate this option to center each row separately.

Reduce: Activate this option to reduce each row separately.

Non-specific filtering: Activate this option to filter out features with low variability prior to computations.

Criterion and threshold: Select the non-specific filtering criterion.

- Standard deviation: all features with a standard deviation lower than the selected threshold are removed.

- Interquartile range: all features with an interquartile range lower than the selected threshold are removed.

- %(Std. dev.): a percentage of features with low standard deviation are removed. The percentage should be indicated in the threshold box.

- %(IQR): a percentage of features with low interquartile range are removed. The percentage should be indicated in the threshold box.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- Mean or mode: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

- Nearest neighbour: Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.
**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the selected variables.

**Charts** tab:

**Color scale**: select the color range to use in the heat map (red to green through black; red to blue through white; red to yellow).

**Width** and **height**: select a magnification factor for the heat map’s width or height.

**Color calibration**:

- **Automatic**: Activate this option if you want XLSTAT to automatically choose boundary values that will delimit the heat map color range.

- **User defined**: Activate this option if you want to manually choose the minimum (Min) and maximum (Max) values that will delimit the heat map color range.

**Results**

**Summary statistics**: The tables of descriptive statistics show the simple statistics for all individuals. The number of observations, missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed.

**Heat map**: The features dendrogram is displayed vertically (rows) and the individuals dendrogram is displayed horizontally (columns). A heat map is added to the chart, reflecting data values.

Similarly expressed features are characterized by horizontal rectangles of homogeneous color along the map.

Similar individuals are characterized by vertical rectangles of homogeneous color along the map.

Clusters of similar individuals characterized by clusters of similarly expressed features can be detected by examining rectangles or squares of homogeneous color at the intersection between feature clusters and individual clusters inside the map.
Example

A tutorial on two-way clustering is available on the Addinsoft website:

http://www.xlstat.com/demo-omicsheat.htm

References

Canonical Correlation Analysis (CCorA)

Use Canonical Correlation Analysis (CCorA, sometimes CCA, but we prefer to use CCA for Canonical Correspondence Analysis) to study the correlation between two sets of variables and to extract from these tables a set of canonical variables that are as much as possible correlated with both tables and orthogonal to each other.

Description

Canonical Correlation Analysis (CCorA, sometimes CCA, but we prefer to use CCA for Canonical Correspondence Analysis) is one of the many methods that allow to study the relationship between two sets of variables. Discovered by Hotelling (1936) this method is used a lot in ecology but is has been supplanted by RDA (Redundancy Analysis) and by CCA (Canonical Correspondence Analysis).

This method is symmetrical, contrary to RDA, and is not oriented towards prediction. Let Y1 and Y2 be two tables, with respectively p and q variables. CCorA aims at obtaining two vectors a(i) and b(i) such that

$$\rho(i) = \text{cor}(Y_1a(i), Y_2b(i)) = \frac{\text{cov}(Y_1a(i), Y_2b(i))}{\text{var}(Y_1a(i)) \cdot \text{var}(Y_2b(i))}$$

is maximized. Constraints must be introduced so that the solution for a(i) et b(i) is unique. As we are in the end trying to maximize the covariance between Y1a(i) and Y2b(i) and to minimize their respective variance, we might obtain components that are well correlated among each other, but that are not explaining well Y1 and Y2. Once the solution has been obtained for i=1, we look for the solution for i=2 where a(2) and b(2) must respectively be orthogonal to a(1) and b(2), and so on. The number of vectors that can be extracted is to the maximum equal to min(p, q).

Note: The inter-batteries analysis of Tucker (1958) is an alternative where one wants to maximize the covariance between the Y1a(i) and Y2b(i) components.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

Click this button to start the computations.
Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

Reload default options: Click this button to reload the default options.

Delete data selections: Click this button to delete the data selections.

Column/Row labels: Click these buttons to change the way XLSTAT handles the data. If the arrow points down (column mode), XLSTAT considers that rows correspond to sites and columns to objects/variables. If the arrow points to the right (row mode), XLSTAT considers that rows correspond to objects/variables and columns to sites.

General tab:

Y1: Select the data that corresponds to the first table. If the “Column labels” option is activated (column mode) you need to include a header on the first row of the selection. If the "Row labels" option is activated (row mode) you need to include a header in the first column of the selection in the selection.

Y2: Select the data that corresponds to the second table. If the "Column labels" option is activated (column mode) you need to include a header on the first row of the selection. If the "Row labels" option is activated (row mode) you need to include a header in the first column of the selection in the selection.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column/Row labels: Activate this option if, in column mode, the first row of the selected data contains a header, or in row mode, if the first column of the selected data contains a header.

Observation labels: Activate this option if observation labels are available. Then select the corresponding data. If the "Column labels" option is activated you need to include a header in the selection. If this option is not activated, the sites labels are automatically generated by XLSTAT (Obs1, Obs2 …).
**Options** tab:

**Type of analysis**: Select from which type of matrix the canonical analysis should be performed.

**Y1**:
- **Center**: Activate this option to center the variables of table Y1.
- **Reduce**: Activate this option to standardize the variables of table Y1.

**Y2**:
- **Center**: Activate this option to center the variables of table Y2.
- **Reduce**: Activate this option to standardize the variables of table Y2.

**Filter factors**: You can activate one of the following two options in order to reduce the number of factors for which results are displayed.

- **Minimum %**: Activate this option then enter the minimum percentage of the total variability that the chosen factors must represent.
- **Maximum Number**: Activate this option to set the number of factors to take into account.

**Missing data** tab:

**Do not accept missing data**: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

**Remove observations**: Activate this option to remove the observations with missing data.

**Estimate missing data**: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbour**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.

**Outputs** tab:
**Descriptive statistics:** Activate this option to display descriptive statistics for the selected variables.

**Covariance/Correlations/[Y1Y2][Y1Y2]:** Activate this option to display the similarity matrix that is being used.

**Eigenvalues:** Activate this option to display the table and the scree plot of the eigenvalues.

**Wilks Lambda test:** Activate this option to display the results of the Wilks lambda test.

**Canonical correlations:** Activate this option to display the canonical correlations.

**Redundancy coefficients:** Activate this option to display the redundancy coefficients.

**Canonical coefficients:** Activate this option to display the canonical coefficients.

**Variables/Factors correlations:** Activate this option to display the correlations between the initial variables of Y1 and Y2 with the canonical variables.

**Canonical variables adequacy coefficients:** Activate this option to display canonical variables adequacy coefficients.

**Squared cosines:** Activate this option to display the squared cosines of the initial variables in the canonical space.

**Scores:** Activate this option to display the coordinates of the observations in the space of the canonical variables.

**Charts** tab:

**Correlation charts:** Activate this option to display the charts involving correlations between the components and the variables.

- **Vectors:** Activate this option to display the variables with vectors.

- **Colored labels:** Activate this option to display the labels with the same color as the corresponding points. If this option is not activated the labels are displayed in black.

**Results**

**Summary statistics:** This table displays the descriptive statistics for the objects and the explanatory variables.
**Similarity matrix**: The matrix that corresponds to the “type of analysis” chosen in the dialog box is displayed.

**Eigenvalues and percentages of inertia**: In this table are displayed the eigenvalues, the corresponding inertia, and the corresponding percentages. Note: in some software, the eigenvalues that are displayed are equal to L / (1-L), where L is the eigenvalues given by XLSTAT.

**Wilks Lambda test**: This test allows to determine if the two tables Y1 and Y2 are significantly related to each canonical variable.

**Canonical correlations**: The canonical correlations, bounded by 0 and 1, are higher when the correlation between Y1 and Y2 is high. However, they do not tell to what extent the canonical variables are related to Y1 and Y2. The squared canonical correlations are equal to the eigenvalues and, as a matter of fact, correspond to the percentage of variability carried by the canonical variable.

The results listed below are computed separately for each of the two groups of input variables.

**Redundancy coefficients**: These coefficients allow to measure for each set of input variables what proportion of the variability of the input variables is predicted by the canonical variables.

**Canonical coefficients**: These coefficients (also called *Canonical weights*, or *Canonical function coefficients*) indicate how the canonical variables were constructed, as they correspond to the coefficients in the linear combine that generates the canonical variables from the input variables. They are standardized if the input variables have been standardized. In that case, the relative weights of the input variables can be compared.

**Correlations between input variables and canonical variables** (also called *Structure correlation coefficients*, or *Canonical factor loadings*) allow understanding how the canonical variables are related to the input variables.

The **canonical variable adequacy coefficients** correspond, for a given canonical variable, to the sum of the squared correlations between the input variables and canonical variables, divided by the number of input variables. They give the percentage of variability taken into account by the canonical variable of interest.

**Square cosines**: The squared cosines of the input variables in the space of canonical variables allow to know if an input variable is well represented in the space of the canonical variables. The squared cosines for a given input variable sum to 1. The sum over a reduced number of canonical axes gives the communality.

**Scores**: The scores correspond to the coordinates of the observations in the space of the canonical variables.
Example

An example of Canonical Correlation Analysis is available on the Addinsoft website:

http://www.xlstat.com/demo-ccora.htm

References


Redundancy Analysis (RDA)

Use Redundancy Analysis (RDA) to analyze a table of response variables using the information provided by a set of explanatory variables, and visualize on the same plot the two sets of variables, and the observations.

Description

Redundancy Analysis (RDA) was developed by Van den Wollenberg (1977) as an alternative to Canonical Correlation Analysis (CCorA). RDA allows studying the relationship between two tables of variables Y and X. While the CCorA is a symmetric method, RDA is non-symmetric. In CCorA, the components extracted from both tables are such that their correlation is maximized. In RDA, the components extracted from X are such that they are as much as possible correlated with the variables of Y. Then the components of Y are extracted so that they are as much as possible correlated with the components extracted from X.

Principles of RDA

Let Y be a table of response variables with n observations and p variables. This table can be analyzed using Principal Component Analysis (PCA) to obtain a simultaneous map of the observations and the variables in two or three dimensions.

Let X be a table that contains the measures recorded for the same n observations on q quantitative and/or qualitative variables.

Redundancy Analysis allows to obtain a simultaneous representation of the observations, the Y variables, and the X variables in two or three dimensions, that is optimal for a covariance criterion (Ter Braak 1986).

Redundancy Analysis can be divided into two parts:

- A constrained analysis in a space which number of dimensions is equal to min(n-1, p, q). This part is the one of main interest as it corresponds to the analysis of the relation between the two tables.

- An unconstrained part, which corresponds to the analysis of the residuals. The number of dimensions for the unconstrained RDA is equal to min(n-1, p).

Partial RDA

Partial RDA adds a preliminary step. The X table is subdivided into two groups. The first group X(1) contains conditioning variables which effect we want to remove, as it is either known or
without interest for the study. Regressions are run on the Y and X(2) tables and the residuals of the regressions are used for the RDA step. Partial RDA allows to analyze the effect of the second group of variables, after the effect of the first group has been removed.

The terminology Response variables/Observations/Explanatory Variables is used in XLSTAT. When the method is used in ecology, "Species" could be used instead of "Response variables", "Sites" could be used instead of "observations", and "Environmental variables" instead of "Explanatory variables".

**Biplot scaling**

XLSTAT offers three different types of scaling. The type of scaling changes the way the scores of the response variables and the observations are computed, and as a matter of fact, their respective position on the plot. Let \( u(ik) \) be the normalized score of variable i on the kth axis, \( v(ik) \) the normalized score of observation i on the kth axis, \( L(k) \) the eigenvalue corresponding to axis k, and \( T \) the total inertia (the sum of the \( L(k) \) for the constrained and unconstrained RDA). The three scalings available in XLSTAT are identical to those of vegan (a package for the R software, Oksanen, 2007). The \( u(ik) \) are multiplied by \( c \), and the \( v(ik) \) by \( d \), and \( r \) is a constant equal to \( \frac{4}{\sqrt{(n-1)T}} \), where \( n \) is the number of observations.

Scaling 1: \[ c = r \sqrt{\frac{L(k)}{T}} \quad d = r \]

Scaling 2: \[ c = r \quad d = r \sqrt{\frac{L(k)}{T}} \]

Scaling 3: \[ c = r \frac{4}{\sqrt{T}} \sqrt{\frac{L(k)}{T}} \quad d = r \frac{4}{\sqrt{T}} \sqrt{\frac{L(k)}{T}} \]

In addition to the observations and the response variables, the explanatory variables can be displayed. The coordinates of the latter are obtained by computing the correlations between the X table and the observation scores.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.
Click this button to start the computations.

Click this button to close the dialog box without doing any computation.

Click this button to display the help.

Click this button to reload the default options.

Click this button to delete the data selections.

Click these buttons to change the way XLSTAT handles the data. If the arrow points down (column mode), XLSTAT considers that rows correspond to sites and columns to objects/variables. If the arrow points to the right (row mode), XLSTAT considers that rows correspond to objects/variables and columns to sites.

**General** tab:

**Response variables Y:** Select the table that corresponds to response variables. If the "Column labels" option is activated (column mode) you need to include a header on the first row of the selection. If the "Row labels" option is activated (row mode) you need to include a header in the first column of the selection in the selection.

**Explanatory variables X:** Select the data that correspond to the various explanatory variables that have been measured for the same observations as for table Y.

- **Quantitative:** Activate this option if you want to use quantitative variables and then select these variables.

- **Qualitative:** Activate this option if you want to use qualitative variables and then select these variables.

**Range:** Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Activate this option to display the results in a new worksheet of the active workbook.

**Workbook:** Activate this option to display the results in a new workbook.

**Partial RDA:** Activate this option to run a partial RDA. If you activate this option, a dialog box will be displayed during the analysis, so that you can select the conditioning variables (see the description section for further details).
**Column/Row labels**: Activate this option if, in column mode, the first row of the selected data contains a header, or in row mode, if the first column of the selected data contains a header.

**Observation labels**: Activate this option if observation labels are available. Then select the corresponding data. If the "Column labels" option is activated you need to include a header in the selection. If this option is not activated, the sites labels are automatically generated by XLSTAT (Obs1, Obs2 …).

**Options** tab:

**Filter factors**: You can activate one of the following two options in order to reduce the number of factors for which results are displayed.

- **Minimum %**: Activate this option then enter the minimum percentage of the total variability that the chosen factors must represent.
- **Maximum Number**: Activate this option to set the number of factors to take into account.

**Permutation test**: Activate this option if you want to use a permutation test to test the independence of the two tables.

- **Number of permutations**: Enter the number of permutations to perform for the test (Default value: 500)
- **Significance level (%)**: Enter the significance level for the test.

**Response variables**:

- **Center**: Activate this option to center the variables before running the RDA.
- **Reduce**: Activate this option to standardize the variables before running the RDA

**Explanatory variables X**:

- **Center**: Activate this option to center the variables before running the RDA.
- **Reduce**: Activate this option to standardize the variables before running the RDA.

**Biplot type**: Select the type of biplot you want to display. The type changes the way the scores of the response variables and the observations are scaled (see the description section for further details).
**Missing data** tab:

**Do not accept missing data:** Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

**Remove observations:** Activate this option to remove the observations with missing data.

**Estimate missing data:** Activate this option to estimate missing data before starting the computations.

- **Mean or mode:** Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

- **Nearest neighbour:** Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.

**Outputs** tab:

**Descriptive statistics:** Activate this option to display descriptive statistics for the selected variables.

**RDA results:** Activate this option to display the RDA results.

**Unconstrained RDA results:** Activate this option to display the results of the unconstrained RDA.

**Eigenvalues:** Activate this option to display the table and the scree plot of the eigenvalues.

**Scores (Observations):** Activate this option to display the scores of the observations.

**Scores (Response variables):** Activate this option to display the scores of the response variables.

- **WA scores:** Activate this option to compute and display the weighted average scores.

- **LC scores:** Activate this option to compute and display the linear combinations scores.

**Contributions:** Activate this option to display the contributions of the observations and the response variables.
Squared cosines: Activate this option to display the squared cosines of the observations and the response variables.

Scores (Explanatory variables): Activate this option to display the scores of the explanatory variables.

Charts tab:
Select the information you want to display on the plot/biplot/triplot.

- Observations: Activate this option to display the observations on the chart.
- Response variables: Activate this option to display the response variables on the chart.
- Explanatory variables: Activate this option to display the explanatory variables on the chart.

Labels: Activate this option to display the labels of the sites on the charts.
- Colored labels: Activate this option to display the labels with the same color as the corresponding points. If this option is not activated the labels are displayed in black.

Vectors: Activate this option to display the vectors for the standard coordinates on the asymmetric charts.
- Length factor: Activate this option to modulate the length of the vectors.

Results

Summary statistics: This table displays the descriptive statistics for the objects and the explanatory variables.

If a permutation test was requested, its results are first displayed so that we can check if the relationship between the tables is significant or not.

Eigenvalues and percentages of inertia: In these tables are displayed for the constrained RDA and the unconstrained RDA the eigenvalues, the corresponding inertia, and the corresponding percentages, either in terms of constrained inertia (or unconstrained inertia), or in terms of total inertia.
The scores of the observations, response variables and explanatory variables are displayed. These coordinates are used to produce the plot.

The charts allow to visualize the relationship between the observations, the response variables and the explanatory variables. When qualitative variables have been included, the corresponding categories are displayed with a hollowed red circle.

**Example**

An example of Redundancy Analysis is available on the Addinsoft website:


**References**


Canonical Correspondence Analysis (CCA)

Use Canonical Correspondence Analysis (CCA) to analyze a contingency table (typically with sites as rows and species in columns) while taking into account the information provided by a set of explanatory variables contained in a second table and measured on the same sites.

Description

Canonical Correspondence Analysis (CCA) has been developed to allow ecologists to relate the abundance of species to environmental variables (Ter Braak, 1986). However, this method can be used in other domains. Geomarketing and demographic analyses should be able to take advantage of it.

Principles of CCA

Let T1 be a contingency table corresponding to the counts on n sites of p objects. This table can be analyzed using Correspondence Analysis (CA) to obtain a simultaneous map of the sites and objects in two or three dimensions.

Let T2 be a table that contains the measures recorded on the same n sites of corresponding to q quantitative and/or qualitative variables.

Canonical Correspondence Analysis allows to obtain a simultaneous representation of the sites, the objects, and the variables in two or three dimensions, that is optimal for a variance criterion (Ter Braak 1986, Chessel 1987).

Canonical Correspondence Analysis can be divided into two parts:

- A constrained analysis in a space which number of dimensions is equal to q. This part is the one of main interest as it corresponds to the analysis of the relation between the two tables.

- An unconstrained part, which corresponds to the analysis of the residuals. The number of dimensions for the unconstrained CCA is equal to \( \min(n-1-q, p-1) \).

Partial CCA

Partial CCA adds a preliminary step. The T2 table is subdivided into two groups of variables: the first group contains conditioning variables which effect we want to remove, as it is either known or without interest for the study. A CCA is run using these variables. A second CCA is run using the second group of variables which effect we want to analyze. Partial CCA allows to
analyze the effect of the second group of variables, after the effect of the first group has been removed.

**PLS-CCA**

Tenenhaus (1998) has shown that it is possible to relate discriminant PLS to CCA. Addinsoft is the first software editor to propose a comprehensive and effective integration between the two methods. Using a restructuring of data based on the proposal Tenenhaus, a PLS step is applied to the data, either to create orthogonal PLS components that are optimally designed for the CCA to avoid the constraints in terms of number of variables that can be used, or to select the most influential variables before running the CCA. As calculations of the CCA step and results are identical to what is done with the classical CCA, users can see this approach as a selection method that identifies the variables of higher interest, either because they are selected in the model, or by looking at the chart of the VIPs (see the section on PLS regression for more information). In the case of a partial CCA, the preliminary step is unchanged.

The terminology Sites/Objects/Variables is used in XLSTAT. “Individuals” or “observations” could be used instead of “sites”, and “species” instead of “objects” when the method is used in ecology.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload**: Click this button to reload the default options.
- **Delete**: Click this button to delete the data selections.
Click these buttons to change the way XLSTAT handles the data. If the arrow points down (column mode), XLSTAT considers that rows correspond to sites and columns to objects/variables. If the arrow points to the right (row mode), XLSTAT considers that rows correspond to objects/variables and columns to sites.

**General** tab:

**Sites/Objects data**: Select the contingency table that corresponds to the counts of the various objects recorded on each different site. If the "Column labels" option is activated (column mode) you need to include a header on the first row of the selection. If the "Row labels" option is activated (row mode) you need to include a header in the first column of the selection in the selection.

**Sites/Variables data**: Select the data that correspond to the various explanatory variables that have been measured on the various sites and that you want to use in the analysis.

- **Quantitative**: Activate this option if you want to use quantitative variables and then select these variables.
- **Qualitative**: Activate this option if you want to use qualitative variables and then select these variables.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Partial CCA**: Activate this option to run a partial CCA. If you activate this option, a dialog box will be displayed during the analysis, so that you can select the conditioning variables (see the description for additional details).

**Column/Row labels**: Activate this option if, in column mode, the first row of the selected data contains a header, or in row mode, if the first column of the selected data contains a header.

**Sites labels**: Activate this option if sites labels are available. Then select the corresponding data. If the "Column labels" option is activated you need to include a header in the selection. If this option is not activated, the sites labels are automatically generated by XLSTAT (Obs1, Obs2 ...).
CCA: Activate this option if you want to run a classical CCA.

PLS-CCA: Activate this option if you want to run a PLS-CCA (see the description section for additional details).

Options tab:

Filter factors: You can activate one of the following two options in order to reduce the number of factors for which results are displayed.

- Minimum %: Activate this option then enter the minimum percentage of the total variability that the chosen factors must represent.

- Maximum Number: Activate this option to set the number of factors to take into account.

Permutation test: Activate this option if you want to use a permutation test to test the independence of the two tables.

- Number of permutations: Enter the number of permutations to perform for the test (Default value: 500)

- Significance level (%): Enter the significance level for the test.

PLS-CCA: If you choose to run a PLS-CCA the following options are available.

- Automatic: Select this option if you want XLSTAT to automatically determine how many PLS components should be used for the CCA step.

- User defined:
  - Max components: Activate this option to define the number of components to extract in the PLS step. If this option is not activated, the number of components is automatically determined by XLSTAT.

  - Number of variables: Activate this option to define the number of variables that should enter the CCA step. The variables with the higher VIPs are selected. The VIPs that are used are those corresponding to the PLS model with the number of components set in “Max components”.

Missing data tab:
**Do not accept missing data:** Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

**Remove observations:** Activate this option to remove the observations with missing data.

**Estimate missing data:** Activate this option to estimate missing data before starting the computations.

- **Mean or mode:** Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

- **Nearest neighbour:** Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.

**Outputs** tab:

**Descriptive statistics:** Activate this option to display descriptive statistics for the selected variables.

**Row and column profiles:** Activate this option to display the row and column profiles.

**CCA results:** Activate this option to display the CCA results.

**Unconstrained CCA results:** Activate this option to display the results of the unconstrained CCA.

**Eigenvalues:** Activate this option to display the table and the scree plot of the eigenvalues.

**Principal coordinates:** Activate this option to display the principal coordinates of the sites, objects and variables.

**Standard coordinates:** Activate this option to display the standard coordinates of the sites, objects and variables.

**Contributions:** Activate this option to display the contributions of the sites, objects and variables.

**Squared cosines:** Activate this option to display the squared cosines of the sites, objects and variables.

**Weighted averages:** Activate this option to display the weighted averages that correspond to the variables of the sites/variables table.
Regression coefficients: Activate this option to display regression coefficients that correspond to the various variables in the factor space.

Charts tab:

Sites and objects:

- Sites and objects / Symmetric: Activate this option to display a symmetric chart that includes both the sites and the objects. For both the sites and the objects, the principal coordinates of are used.

- Sites / Asymmetric: Activate this option to display the asymmetric chart of the sites. The principal coordinates are used for the sites, and the standard coordinates are used for the objects.

- Objects / Asymmetric: Activate this option to display the asymmetric chart of the objects. The principal coordinates are used for the objects, and the standard coordinates are used for the sites.

- Sites: Activate this option to display a chart on which only the sites are displayed. The principal coordinates are used.

- Objects: Activate this option to display a chart on which only the objects are displayed. The principal coordinates are used.

Variables:

- Correlations: Activate this option to display the quantitative and qualitative variables on the charts, using as coordinates their correlations (equal to their standard coordinates).

- Regression coefficients: Activate this option to display the quantitative and qualitative variables on the charts, using the regression coefficients as coordinates.

Labels: Activate this option to display the labels of the sites on the charts.

- Colored labels: Activate this option to display the labels with the same color as the corresponding points. If this option is not activated the labels are displayed in black.

Vectors: Activate this option to display the vectors for the standard coordinates on the asymmetric charts.

- Length factor: Activate this option to modulate the length of the vectors.
Results

Summary statistics: This table displays the descriptive statistics for the objects and the explanatory variables.

Inertia: This table displays the distribution of the inertia between the constrained CCA and the unconstrained CCA.

Eigenvalues and percentages of inertia: In these tables are displayed for the CCA and the unconstrained CCA the eigenvalues, the corresponding inertia, and the corresponding percentages, either in terms of constrained inertia (or unconstrained inertia), or in terms of total inertia.

Weighted averages: This table displays the weighted means as well the global weighted means.

The principal coordinates and standard coordinates of the sites, the objects and the variables are then displayed. These coordinates are used to produce the various charts.

Regression coefficients: This table displays the regression coefficients of the variables in the factor space.

The charts allow to visualize the relationship between the sites, the objects and the variables. When qualitative variables have been included, the corresponding categories are displayed with a hollowed red circle.

Example

An example of Canonical Correspondence Analysis is available on the Addinsoft website:

http://www.xlstat.com/demo-cca.htm

References


Principal Coordinate Analysis (PCoA)

Use Principal Coordinate Analysis to graphically visualize a square matrix that describes the similarity or the dissimilarity between p elements (individuals, variables, objects, …).

Description

Principal Coordinate Analysis (often referred to as PCoA) is aimed at graphically representing a resemblance matrix between p elements (individuals, variables, objects, …).

If the input matrix is a similarity matrix, XLSTAT transforms it into a dissimilarity matrix before applying the calculations described by Gower (1966), or before any of the changes suggested by various authors and summarized in the Numerical Ecology book by Legendre and Legendre (1998).

Concept

Let D be a p x p symmetric matrix that contains the distances between p elements: we compute the A matrix which elements a(ij), corresponding to the ith row and to the jth column, are given by

\[ a(ij) = d^2(ij) / 2 \]

We then center the A matrix by rows and by columns to obtain the \( \Delta 1 \) matrix which elements \( \delta 1(ij) \) are given by

\[ \delta 1(ij) = a(ij) - \bar{a}(i) - \bar{a}(j) - \bar{a} \]

where \( \bar{a}(i) \) is the mean of the a(ij) for row i, \( \bar{a}(j) \) is the mean of the a(ij) for column j and \( \bar{a} \) is the mean of all the elements.

Last, we compute the eigen-decomposition of \( \Delta 1 \). The eigenvectors are sorted by decreasing order of eigenvalues and transformed so that, if \( u(k) \) is the eigenvector associated to the \( \lambda(k) \) eigenvalue, we have:

\[ u'(k)u(k) = \lambda(k) \]

The rescaled eigenvectors correspond to the principal coordinates that can be used to display the p objects in a space with 1, 2, p-1 dimensions.

As with PCA (Principal Component Analysis) eigenvalues can be interpreted in terms of percentage of total variability that is being represented in a reduced space.
Note: because $\Delta 1$ is centered, we obtain at most, $p-1$ non null eigenvalues. In the case where the initial matrix $D$ is an Euclidean matrix distance, we can easily understand that $p-1$ axes are enough to fully describe $p$ objects (by two points passes one line, three points are contained in a plane, …). In the case where the points are confounded in a sub-space, we can obtain several null eigenvalues (for example, three points can be aligned on a line).

**Negative eigenvalues**

When the $D$ matrix is not metric, or if missing values were present in the data that were used to compute the distances, the eigen-decomposition can lead to negative eigenvalues. This can especially happen with semi-metric or non metric distances. This problem is described in the article by Gower and Legendre (1986).

XLSTAT suggests two transformations to solve the problem of negative eigenvalues. The first solution consists in taking as input distances the square root of the input distances. The second, inspired by the results of Lingoes (1971), consists in adding a constant to the $D$ matrix (except to the diagonal elements) such that there is no negative eigenvalue. This constant is equal to the opposite of the largest negative eigenvalue.

**PCA, MDS and PCoA**

PCA and PCoA are quite similar in that PCA can also represent observations in a space with less dimensions, the later being optimal in terms of variability carried. A PCoA applied to matrix of Euclidean distances between observations (calculated after standardization of the columns using the unbiased standard deviation) leads to the same results as a PCA based on the correlation matrix. The eigenvalues obtained with the PCoA are equal to $(p-1)$ times those obtained with the PCA.

PCoA and MDS (Multidimensional Scaling) share the same goal of representing objects for which we have a proximity matrix.

MDS has two drawbacks compared with PCoA:

- The algorithm is much more complex and performs slower.
- Axes obtained with MDS cannot be interpreted in terms of variability.

MDS has two advantages compared with PCoA:

- The algorithm allows having missing data in the proximity matrix.
- The non-metric version of MDS provides a simpler and clear way to handle matrices where only the ranking of the distances is important.
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload**: Click this button to reload the default options.
- **Delete**: Click this button to delete the data selections.
- **Switch Data Handling**: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

**Data**: Select a similarity or dissimilarity matrix. If only the lower or upper triangle is available, the table is accepted. If differences are detected between the lower and upper parts of the selected matrix, XLSTAT warns you and offers to change the data (by calculating the average of the two parts) to continue with the calculations.

**Dissimilarities / Similarities**: Choose the option that corresponds to the type of your data.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet in the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.
Labels included: Activate this option if you have included row and column labels in the selection.

Options tab:

Correction for negative eigenvalues: Activate the option that corresponds to the strategy to apply if eigenvalues are detected during the eigen-decomposition:

- **None**: Nothing is done when negative eigenvalues are found.
- **Square root**: The elements of the distance matrix D are replaced by their square root.
- **Lingoes**: A transformation is applied so that that eigen-decomposition does not lead to any negative eigenvalue.

Filter factors: You can activate one of the following two options in order to reduce the number of factors for which results are displayed.

- **Minimum %**: Activate this option then enter the minimum percentage of the total variability that the chosen factors must represent.
- **Maximum Number**: Activate this option to set the number of factors to take into account.

Outputs tab:

Delta1 matrix: Activate this option to display the Delta1 matrix that is used to compute the eigenvalues and the eigenvectors.

Eigenvalues: Activate this option to display the table and the chart (scree plot) of the eigenvalues.

Principal coordinates: Activate this option to display the principal coordinates.

Contributions: Activate this option to display the contributions.

Squared cosines: Activate this option to display the squared cosines.

Charts tab:

Chart: Activate this option to display the chart.
Results

**Delta1 matrix:** This matrix corresponds to the $\Delta_1$ matrix of Gower, used to compute the eigen decomposition.

**Eigenvalues and percentage of inertia:** this table displays the eigenvalues and the corresponding percentage of inertia.

**Principal coordinates:** This table displays of the principal coordinates of the objects that are used to create the chart where the proximities between the charts can be interpreted.

**Contributions:** This table displays the contributions that help evaluate how much an object contributes to a given axis.

**Squared cosines:** This table displays the contributions that help evaluate how close an object is to a given axis.

Example

An example showing how to run a Principal Coordinate Analysis is available on the Addinsoft website at:

http://www.xlstat.com/demo-pcoa.htm

References


**Multiple Factor Analysis (MFA)**

Use the Multiple Factor Analysis (MFA) to simultaneously analyze several tables of variables, and to obtain results, particularly charts, that allow to study the relationship between the observations, the variables and the tables. Within a table, the variables must be of the same type (quantitative or qualitative), but the tables can be of different types.

**Description**

Multiple Factor Analysis (MFA) makes it possible to analyze several tables of variables simultaneously, and to obtain results, in particular charts, that allow studying the relationship between the observations, the variables and tables (Escofier and Pagès, 1984). Inside a table the variables must be of the same type (quantitative or qualitative), but the tables can be of different types.

The MFA is a synthesis of PCA (Principal Component Analysis) and MCA (Multiple Correspondence Analysis), that it generalizes to enable the use of quantitative and qualitative variables. The methodology of the MFA breaks up into two phases:

1. We successively carry out for each table a PCA or an MCA according to the type of the variables of the table. One stores the value of the first eigenvalue of each analysis to then weigh the various tables in the second part of the analysis.

2. One carries out a weighted PCA on the columns of all the tables, knowing that the tables of qualitative variables are transformed into a complete disjunctive table, each indicator variable having a weight that is a function of the frequency of the corresponding category. The weighting of the tables makes it possible to prevent that the tables that include more variables do not weigh too much in the analysis.

This method can be very useful to analyze surveys for which one can identify several groups of variables, or for which the same questions are asked at several time intervals.

The authors that developed the method (Escofier and Pagès, 1984) particularly insisted on the use of the results which are obtained from the MFA. The originality of method is that it allows visualizing in a two or three dimensional space, the tables (each table being represented by a point), the variables, the principal axes of the analyses of the first phase, and the individuals. In addition, one can study the impact of the other tables on an observation by simultaneously visualizing the observation described by the all the variables and the projected observations described by the variables of only one table.

**Note 1:** as for PCA, the qualitative variables are represented by the centroids of the categories on the charts of the observations.

**Note 2:** an MFA performed on K tables that contain each one qualitative variable is equivalent to an MCA performed on the K variables.
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload default options**: Click this button to reload the default options.
- **Delete data selections**: Click this button to delete the data selections.
- **Change data handling direction**: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

**Observations/variables table**: Select the data that correspond to N observations described by P quantitative variables and grouped into K tables. If column headers have been selected, check that the “Variable labels” option has been activated.

**Number of tables**: Enter the number K of tables in which the selected data are subdivided.

**Table labels**: Activate this option if you want to use labels for the K tables. If this option is not activated, the name of the tables are automatically generated (Table1, Table2, ...). If column headers have been selected, check that the “Variable labels” option has been activated.

**Number of variables per table**:

- **Equal**: Choose this option if the number of variables is identical for all the tables. In that case XLSTAT determines automatically the number of variables in each table.
• **User defined**: Choose this option to select a column that contains the number of variables contained in each table. If the "Variable labels" option has been activated, the first row must correspond to a header.

**Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet**: Activate this option to display the results in a new worksheet of the active workbook.

**Workbook**: Activate this option to display the results in a new workbook.

**Variable labels**: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header. Where the selection is a correlation or covariance matrix, if this option is activated, the first column must also include the variable labels.

**Observation labels**: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

**Weights**: Activate this option if the observations are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

**Options** tab:

**PCA type**: Choose the type of matrix to be used for PCA. The difference between the Pearson (n) and the Pearson (n-1) options, only influences the way the variables are standardized, and the difference can only be noticed on the coordinates of the observations.

**Data type**: Specify which is the type of data contained in the various tables, knowing that the type must be the same within a given table. In the case where the "Mixed type" is selected, you need to select a column that indicates the type of data in each table. Use 0 for a table that contains quantitative variables, and 1 for a table that contains qualitative variables.

**Filter factors**: You can activate one of the following two options in order to reduce the number of factors for which results are displayed.

• **Minimum %**: Activate this option then enter the minimum percentage of the total variability that the chosen factors must represent.
• Maximum Number: Activate this option to set the number of factors to take into account.

Display charts on two axes: Activate this option if you want that the numerous graphical representations displayed after the PCA, MCA and MFA are only displayed on the first two axes, without your being prompted after each analysis.

Supplementary data tab:

Supplementary observations: Activate this option if you want to calculate the coordinates and represent additional observations. These observations are not taken into account for the factor axis calculations (passive observations as opposed to active observations). Several methods for selecting supplementary observations are provided:

• Random: The observations are randomly selected. The “Number of observations” N to display must then be specified.

• N last rows: The last N observations are selected for validation. The “Number of observations” N to display must then be specified.

• N first rows: The first N observations are selected for validation. The “Number of observations” N to display must then be specified.

• Group variable: If you choose this option, you must then select an indicator variable set to 0 for active observations and 1 for passive observations.

Supplementary tables: Activate this option if you want to use some tables as supplementary tables. The variables of these tables will not be taken into account for the computation of the factors of the MFA. However, the separate analyses of the first phase of the MFA will be run on these tables. Select a column that contains the indicators (0/1) that let XLSTAT know which are among the K tables the active ones (1) and the supplementary ones (0).

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Remove observations: Activate this option to ignore the observations that contain missing data.

Adapted strategies: Activate this option to choose strategies that are adapted to the data type.
• Quantitative variables:

  o **Pairwise deletion**: Activate this option to remove observations with missing data only when the variables involved in the calculations have missing data. For example, when calculating the correlation between two variables, an observation will only be ignored if the data corresponding to one of the two variables is missing.

  o **Mean**: Activate this option to estimate the missing data of an observation by the mean of the corresponding variable.

  o **Nearest neighbour**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.

• Qualitative variables:

  o **New category**: Choose this option to group missing data into a new category of the corresponding variable.

  o **Nearest neighbour**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.

**Outputs** tab:

The outputs tab is divided into four sub-tabs:

**General**:

These outputs concern all the analyses:

**Descriptive statistics**: Activate this option to display the descriptive statistics for all the selected variables.

**Correlations**: Activate this option to display correlation matrix for the selected quantitative variables.

**Eigenvalues**: Activate this option to display the table and chart (scree plot) of eigenvalues.

**Contributions**: Activate this option to display the contribution tables.

**Squared cosines**: Activate this option to display the tables of squared cosines.

**PCA**:

These outputs only concern the PCA:
**Factor loadings:** Activate this option to display the coordinates of the variables in the factor space.

**Variables/Factors correlations:** Activate this option to display correlations between factors and variables.

**Factor scores:** Activate to display the coordinates of the observations (factor scores) in the new space created by PCA.

**MCA:**

These outputs only concern the MCA:

**Disjunctive table:** Activate this option to display the full disjunctive table that corresponds to the selected qualitative variables.

**Burt table:** Activate this option to display the Burt table.

**Display results for:**

- **Observations:** Activate this option to display the results that concern the observations.
- **Variables:** Activate this option to display the results that concern the variables.

**Principal coordinates:** Activate this option to display the principal coordinates.

**Standard coordinates:** Activate this option to display the standard coordinates.

**Test-values:** Activate this option to display the test-values for the variables.

- **Significance level (%)** Enter the significance level used to determine if the test values are significant or not.

**MFA:**

These results correspond to the second phase of the MFA:

**Tables:**

- **Coordinates:** Activate this option to display the coordinates of the tables in the MFA space. Note: the contributions and the squared cosines are also displayed if the corresponding options are checked in the Outputs/General tab.
- **Lg coefficients**: Activate this option to display the Lg coefficients.
- **RV coefficients**: Activate this option to display the RV coefficients.

**Variables:**
- **Factor loadings**: Activate this option to display the factor loadings in the MFA space.
- **Variables/Factors correlations**: Activate this option to display the correlations between factors and variables in the MFA space.

**Partial axes:**
- **Maximum number**: Enter the maximum number of factors to keep from the analyses of the first phase that you then want to analyze in the MFA space.
- **Coordinates**: Activate this option to display the coordinates of the partial axes in the space obtained from the MFA.
- **Correlations**: Activate this option to display the correlations between the factors of the MFA and the partial axes.
- **Correlations between axes**: Activate this option to display the correlation between the partial axes.

**Observations:**
- **Factor scores**: Activate this option to display the factor scores in the MFA space.
- **Coordinates of the projected points**: Activate this option to display the coordinates of the projected points in the MFA space. The projected points correspond to the projections of the observations in spaces reduced to the number of dimensions of each table.

**Charts** tab:

The charts tab is divided into four sub-tabs:

**General:**

These options are for all the analyses:
- **Colored labels**: Activate this option to show labels in the same color as the points.
Filter: Activate this option to modulate the number of observations displayed:

- **Random**: The observations to display are randomly selected. The “Number of observations” N to display must then be specified.

- **N first rows**: The N first observations are displayed on the chart. The “Number of observations” N to display must then be specified.

- **N last rows**: The N last observations are displayed on the chart. The “Number of observations” N to display must then be specified.

- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to display.

PCA:

These options concern only the PCA:

**Correlation charts**: Activate this option to display the charts involving correlations between the components and the variables.

- **Vectors**: Activate this option to display the variables with vectors.

**Observations charts**: activate this option to display the charts that allow visualizing the observations in the new space.

- **Labels**: Activate this option to display the observations labels on the charts.

**Biplots**: Activate this option to display the charts where the input variables and the observations are simultaneously displayed.

- **Vectors**: Activate this option to display the input variables with vectors.

- **Labels**: Activate this option to display the observations labels on the biplots.

**Type of biplot**: Choose the type of biplot you want to display. See the description section of the PCA for more details.

- **Correlation biplot**: Activate this option to display correlation biplots.

- **Distance biplot**: Activate this option to display distance biplots.

- **Symmetric biplot**: Activate this option to display symmetric biplots.

- **Coefficient**: Choose the coefficient whose square root is to be multiplied by the coordinates of the variables. This coefficient lets you to adjust the position of the variable points in the biplot in order to make it more readable. If set to other than 1, the length of the variable vectors can no longer be interpreted as standard deviation (correlation biplot) or contribution (distance biplot).
MCA:

These options concern only the MCA:

**Symmetric plots:** Activate this option to display the symmetric observations and variables plots.

- **Observations and variables:** Activate this option to display a plot that shows both the observations and variables.
- **Observations:** Activate this option to display a plot that shows only the observations.
- **Variables:** Activate this option to display a plot that shows only the variables.

**Asymmetric plots:** Activate this option to display plots for which observations and variables play an asymmetrical role. These plots are based on the principal coordinates for the observations and the standard coordinates for the variables.

- **Observations:** Activate this option to display an asymmetric plot where the observations are displayed using their principal coordinates, and where the variables are displayed using their standard coordinates.
- **Variables:** Activate this option to display an asymmetric plot where the variables are displayed using their principal coordinates, and where the observations are displayed using their standard coordinates.

**Labels:** Activate this option to display the labels of the categories on the charts.

**Vectors:** Activate this option to display the vectors for the standard coordinates on the asymmetric charts.

- **Length factor:** Activate this option to modulate the length of the vectors.

MFA:

These options concern only the results of the second phase of the MCA:

**Table charts:** Activate this option to display the charts that allow to visualize the tables in the MFA space.

**Correlation charts:** Activate this option to display the charts involving correlations between the components and the quantitative variables used in the MFA.
Observations charts: Activate this option to display the chart of the observations in the MFA space.

- Color observations: Activate this option so that the observations are displayed using different colors, depending on the value of the first qualitative supplementary variable.

- Display the centroids: Activate this option to display the centroids that correspond to the categories of the qualitative variables of the supplementary tables.

Correlation charts (partial axes): Activate this option to display the correlation chart for the partial axes obtained from the first phase of the MFA.

Charts of the projected points: Activate this option to display the chart that shows at the same time the observations in the MFA space, and the observations projected in the sub-space of each table.

- Observation labels: Activate this option to display the observations labels on the charts.

- Projected points labels: Activate this option to display the labels of the projected points.

Results

Descriptive statistics: The table of descriptive statistics shows the simple statistics for all the variables selected. This includes the number of observations, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased).

Correlation/Covariance matrix: This table shows the correlations between all the quantitative variables. The type of coefficient depends on what has been chosen in the dialog box.

The results of the analyses performed on each individual table (PCA or MCA) are then displayed. These results are identical to those you would obtain after running the PCA or MCA function of XLSTAT.

Afterwards, the results of the second phase of the MFA are displayed.

Eigenvalues: The eigenvalues and corresponding chart (scree plot) are displayed. The number of eigenvalues is equal to the number of non-null eigenvalues.

Eigenvectors: This table shows the eigenvectors obtained from the spectral decomposition. These vectors take into account the variable weights used in the MFA.
The coordinates of the tables are then displayed and used to create the plots of the tables. The latter allow to visualize the distance between the tables. The coordinates of the supplementary tables are displayed in the second part of the table.

Contributions (%): Contributions are an interpretation aid. The tables which had the highest influence in building the axes are those whose contributions are highest.

Squared cosines: As in other factor methods, squared cosine analysis is used to avoid interpretation errors due to projection effects. If the squared cosines associated with the axes used on a chart are low, the position of the observation or the variable in question should not be interpreted.

Lg coefficients: The Lg coefficients of relationship between the tables allow to measure to what extent the tables are related two by two. The more variables of a first table are related to the variables of the second table, the higher the Lg coefficient.

RV coefficients: The RV coefficients of relationship between the tables are another measure derived from the Lg coefficients. The value of the RV coefficients varies between 0 and 1.

The results that follow concern the quantitative variables. As for a PCA, the coordinates of the variables (factor loadings), their correlation with the axes, the contributions and the squared cosines are displayed.

The coordinates of the partial axes, and even more their correlations, allow to visualize in the new space the link between the factors obtained from the first phase of the MFA, and those obtained from the second phase.

Les results that concern the observations are then displayed as they are after a PCA (coordinates, contributions in %, and squared cosines).

Last, the coordinates of the projected points in the space resulting from the MFA are displayed. The projected points correspond to projections of the observations in the spaces reduced to the dimensions of each table. The representation of the projected points superimposed with those of the complete observations makes it possible to visualize at the same time the diversity of the information brought by the various tables for a given observation, and to visualize the relative distances from two observations according to the various tables.
Example

An example of Multiple Factor Analysis is available on the Addinsoft website:
http://www.xlstat.com/demo-mfa.htm

References


Latent class clustering

This tool is part of the XLSTAT-LG module. Use this tool to classify cases into meaningful clusters (latent classes) that differ on one or more parameters from latent class (LC) Cluster models. LC Cluster models classify based on combinations of continuous and/or categorical (nominal or ordinal) variables.

Description

This the latent class clustering feature of XLSTAT is part of the XLSTAT-LG module, a powerful clustering tool based on Latent GOLD® 5.0:

Latent class analysis (LCA) involves the construction of latent classes (LC) which are unobserved (latent) subgroups or segments of cases. The latent classes are constructed based on the observed (manifest) responses of the cases on a set of indicator variables. Cases within the same latent class are homogeneous with respect to their responses on these indicators, while cases in different latent classes differ in their response patterns. Formally, latent classes are represented by K distinct categories of a nominal latent variable X. Since the latent variable is categorical, LC modeling differs from more traditional latent variable approaches such as factor analysis, structural equation models, and random-effects regression models since these approaches are based on continuous latent variables.

XLSTAT-LG contains separate modules for estimating two different model structures - LC Cluster models and LC Regression models - which are useful in somewhat different application areas. To better distinguish the output across modules, latent classes are labeled 'clusters' for LC Cluster models and 'classes' for LC Regression models. In this manual we also refer to latent classes using the term 'segments'.

The LC Cluster Model:

- Includes a nominal latent variable X with K categories, each category representing a cluster.
- Each cluster contains a homogeneous group of persons (cases) who share common interests, values, characteristics, and/or behavior (i.e., share common model parameters).
- These interest, values, characteristics, and/or behavior constitute the observed variables (indicators) Y upon which the latent clusters are derived.
Advantages over more traditional ad-hoc types of cluster analysis methods include model selection criteria and probability-based classification. Posterior membership probabilities are estimated directly from the model parameters and used to assign cases to the modal class - the class for which the posterior probability is highest.

A special feature of LC cluster models is the ability to obtain an equation for calculating these posterior membership probabilities directly from the observed variables (indicators). This equation can be used to score new cases based on a LC cluster model estimated previously. That is, the equation can be used to classify new cases into their most likely latent class as a function of the observed variables. This feature is unique to LC models – it is not available with any other clustering technique.

The scoring equation is obtained as a special case of the more general Step3 methodology for LC cluster models (Vermunt, 2010). In Step1, model parameter estimates are obtained. In Step2, cases are assigned to classes based on their posterior membership probabilities. In Step3, the latent classes are used as predictors or dependent variables in further analyses. For further details, see Section 2.3 (Step3 Scoring) in Vermunt and Magidson (2013b).

Copyright ©2014 Statistical Innovations Inc. All rights reserved.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
- **Reload**: Click this button to reload the default options.
- **Delete**: Click this button to delete the data selections.
Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

**Observations/variables table:**

**Continuous:** Select the continuous variable(s). The data must be continuous. If the ‘Column labels’ option is activated make sure that the headers of the variable(s) have also been selected.

**Nominal:** Select the nominal variable(s). The data must be nominal. If the ‘Column labels’ option is activated make sure that the headers of the variable(s) have also been selected.

**Ordinal:** Select the ordinal variable(s). The data must be numeric. If the ‘Column labels’ option is activated make sure that the headers of the variable(s) have also been selected.

**Direct effects:** Activate this option if you want to specify a direct effect to be included in the model. After specifying your model and clicking “OK” from the dialog box, an interactions box will pop up. All pairs of variables eligible for a direct effect parameter appear. To include a direct effect, click in the check-box and a check appears. Direct effect parameters will be estimated for the pairs of variables that have been so selected (direct effect check-box equals on). The inclusion of direct effects is one way to relax the assumption of local dependence.

**Range:** Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Activate this option to display the results in a new worksheet of the active workbook.

**Workbook:** Activate this option to display the results in a new workbook.

**Column labels:** Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

**Observation labels:** Activate this option if labels are available for the N observations. Then select the corresponding data. If the ‘Column labels’ option is activated you need to include a header in the selection.

With repeated measures data (multiple records per case) the Observation labels variable serves as a case ID variable, which groups the records from each case together so that they are assigned to the same fold during cross-validation. If this option is not activated, labels for
the observations are automatically generated by XLSTAT (Obs1, Obs2 …), so that each case contains a single record.

**Case weights:** Activate this option if you want to weight the observations. If you do not activate this option, all weights are set to 1. The weights must be non-negative values. Setting a case weight to 2 is equivalent to repeating the same observation twice. If the 'Variable labels' option is activated, make sure that the header (first row) has also been selected.

**Number of clusters:**

**from:** Enter a number between 1-25.

**to:** Enter a number between 1-25.

Note: To specify a fixed number of clusters $K$: use from $K$ to $K$. For example, to estimate a 2 class model: from 2 to 2.

**Use separate sheets:** Activate this option if you want the program to produce separate sheets for each cluster model estimated. A separate sheet with summary statistics for all models estimated will also be produced.

**Options** tab:

Parameter estimation occurs using an iterative algorithm which begins using the Expectation-Maximization (EM) algorithm until either the maximum number of EM iterations (Iterations EM) or the EM convergence criterion (Tolerance(EM)) is reached. Then, the program switches to perform Newton Raphson (NR) iterations which continue until the maximum number of NR iterations (Iterations Newton-Raphson) or the overall converge criterion (Tolerance) is reached. The program also stops iterating when the change in the log-posterior is negligible (smaller than $10^{-12}$). A warning is given if one of the elements of the gradient is larger than $10^{-3}$.

Sometimes, for example in the case of models with many parameters, it is more efficient to use only the EM algorithm. This is accomplished by setting Iterations Newton-Raphson to 0. With very large models, one may also consider suppressing the computation of standard errors (and associated Wald statistics) in the Output tab.

**Convergence:**

**Tolerance(EM):** Expectation-Maximization (EM) Tolerance is the sum of absolute relative changes of parameter values in a single iteration as long as the EM algorithm is used. It determines when the program switches from EM to Newton-Raphson (if the NR iteration limit has been set to > 0). Increasing the EM Tolerance will switch faster from EM to NR. To change this option, click the value to highlight it, then type in a new value. You may enter any non-
negative real number. The default is 0.01. Values between 0.01 and 0.1 (1% and 10%) are reasonable.

**Tolerance**: Overall Tolerance (Tolerance) is the sum of absolute relative changes of parameter values in a single iteration. It determines when the program stops its iteration. The default is $1.0 \times 10^{-8}$ which specifies a tight convergence criterion. To change this option, click the value to highlight it, then type in a new value. You may enter any non-negative real number.

Note: when only EM iterations are used, the effective tolerance is the maximum of Tolerance(EM) and Overall Tolerance.

**Iterations:**

**EM**: Maximum number of EM iterations. The default is 250. If the model does not converge after 250 iterations, this value should be increased. You also may want to increase this value if you set Newton-Raphson iterations = 0. To change this option, click the value to highlight it, then type in a new value. You may enter any non-negative integer.

**Newton-Raphson**: Maximum number of NR iterations. The default is 50. If the model does not converge after 50 iterations, this value should be increased. To change this option, click the value to highlight it, then type in a new value. You may enter any non-negative integer. A value of 0 is entered to direct XLSTAT-LG to use only EM, which may produce faster convergence in models with many parameters or in models that contain continuous indicators.

**Start values:**

The best way to prevent ending up with a local solution is the use of multiple sets of starting values since different sets of starting values may yield solutions with different log-posterior values. The use of such multiple sets of random starting values is automated. This procedure increases considerably the probability of finding the global solution, but in general does not guarantee that it will be found in a single run. To reduce the likelihood of obtaining a local solution, the following options can be used to either increasing the number of start sets, the number of iterations per set, or both.

**Random sets**: The default is 16 for the number of random sets of starting values to be used to start the iterative estimation algorithm. Increasing the number of sets of random starting values for the model parameters reduces the likelihood of converging to a local (rather than global) solution. To change this option, click the value to highlight it, then type in a new value. You may enter any non-negative integer. Using either the value 0 or 1 results in the use of a single set of starting values.

**Iterations**: This option allows specification of the number of iterations to be performed per set of start values. XLSTAT-LG first performs this number of iterations within each set and
subsequently twice this number within the best 10% of the start sets. For some models, many more than 50 iterations per set may need to be performed to avoid local solutions.

**Seed** (random numbers): The default value of 123456789 means that the Seed is obtained during estimation using a pseudo random number generator based on clock time. Specifying a non-negative integer different from 0, yields the same result each time.

To specify a particular numeric seed (such as the Best Start Seed reported in the Model Summary Output for a previously estimated model), click the value to highlight it, then type in a non-negative integer. When using the Best Start Seed, be sure to deactivate the Random Sets option (using Random Sets = 0).

**Tolerance**: Indicates the convergence criterion to be used when running the model of interest with the various start sets. The definition of this tolerance is the same as the one that is used for the EM and Newton-Raphson Iterations.

**Bayes Constants**:

The Bayes options can be used to eliminate the possibility of obtaining boundary solutions. You may enter any non-negative real value. Separate Bayes constants can be specified for three different situations:

**Latent**: The default is 1. Increase the value to increase the weight allocated to the Dirichlet prior which is used to prevent the occurrence of boundary zeroes in estimating the latent distribution. The number can be interpreted as a total number of added cases that is equally distributed among the classes (and the covariate patterns). To change this option, click the value to highlight it, then type in a new value.

**Categorical**: The default is 1. Increase the value to increase the weight allocated to the Dirichlet prior which is used in estimating multinomial models with variables specified as Ordinal or Nominal. This number can be interpreted as a total number of added cases to the cells in the models for the indicators to prevent the occurrence of boundary solutions. To change this option, click the value to highlight it, then type in a new value.

**Error variance**: The default is 1. Increase the value to increase the weight allocated to the inverse-Wishart prior which is used in estimating the error variance-covariance matrix in models for continuous dependent variables or indicators. The number can be interpreted as the number of pseudo-cases added to the data, each pseudo-case having a squared error equal to the total variance of the indicator concerned. Such a prior prevents variances of zero from occurring. To change this option, click the value to highlight it, then type in a new value.

For technical details, see section 7.3 of Vermunt and Magidson (2013a).

**Cluster Independent**:
**Error (Co)variances**: This option indicates that the error covariances are restricted to be equal across classes (class independent). Note that this option only applies to pairs of continuous indicators for which direct effects have been included in the model (see the Direct Effects option in the General tab).

**Missing data** tab:

**Do not accept missing data**: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

**Remove observations**: Activate this option to remove the observations with missing data.

**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the variables selected.

**Statistics**: Activate this option to display the following statistics about the model(s).

**Chi-squared**: Activate this option to display various chi-square based statistics related to model fit.

**Log-likelihood**: Activate this option to display log-likelihood statistics.

**Classification**: Activate this option to display the Classification Table, which cross-tabulates modal and probabilistic class assignment.

**Profile**: Activate this option to display the probabilities or means associated with each Indicator.

- The first row of numbers shows how large each cluster is.
- The body of the table contains (marginal) conditional probabilities that show how the clusters are related to the Nominal or Ordinal variables. These probabilities sum to 1 within each cluster (column).
- For indicators specified as Continuous, the body of the table contains means (rates) instead of probabilities. For indicators specified as Ordinal, means are displayed in addition to the conditional probabilities.

**Standard Errors**: Activate this option to display the standard errors (and associated Wald statistics). The standard (Hessian) computation method makes use of the second-order derivatives of the log-likelihood function called the Hessian matrix.
**Bivariate Residuals:** Activate this option to display the bivariate residuals for the model.

**Frequencies / Residuals:** Activate this option to display the observed and expected frequencies along with the standardized residuals for a model. This output is not available if at least one indicator is continuous. This output is not reported in the case 1 or more continuous indicators.

**Iteration Details:** Activate this option to display technical information associated with the performance of the estimation algorithm, such as log-posterior and log-likelihood values at convergence:

- EM algorithm,
- Newton-Raphson algorithm.

When applicable, this file also contains warning messages concerning non-convergence, unidentified parameters and boundary solutions.

**Scoring Equation:** Activate this option to display the scoring equation, consisting of regression coefficients associated with the multinomial logit model. The resulting scores are predicted logits associated with each latent class t. For example, for responses Y1=j, Y2=k, Y3=m, Y4=s to 4 nominal indicators, the logit associated with cluster t is:

\[ \text{Logit}(t) = a[t] + b1[j,t] + b2[k,t] + b3[m,t] + b4[s,t] \]

Thus, to obtain the posterior membership probabilities for latent class t0 given this response pattern, use the following formula:

\[
\text{Prob(cluster}= t0 | Y1=j, Y2=k, Y3=m, Y4=s) = \frac{\exp(\text{Logit}[t_0])}{\sum_t \exp(\text{Logit}[t])} = \frac{\exp(a[t_0]+b1[j,t_0]+b2[k,t_0]+b3[m,t_0]+b4[s,t_0])}{\sum_t \exp(a[t]+b1[j,t]+b2[k,t]+b3[m,t]+b4[s,t])}
\]

For further details, see the tutorial “Using XLSTAT-LG to estimate latent class cluster models”.

**Classification:** Activate this option to display a table containing the posterior membership probability and the modal assignment for each of the cases based on the current model.

**Charts tab:**

**Profile plot:** The profile plot is constructed from the conditional probabilities for the nominal variables and means for the other indicators as displayed in the columns of the Profile table. The quantities associated with the selected clusters are plotted and connected. For the scale types ordinal, continuous, count, and numeric covariate, prior to plotting the class-specific means, they are re-scaled to always lie within the 0-1 range. Scaling of these "0-1 Means" is
accomplished by subtracting the lowest observed value from the class-specific means and dividing the results by the range, which is simply the difference between the highest and the lowest observed value. The advantage of such scaling is that these numbers can be depicted on the same scale as the class-specific probabilities for nominal variables. For nominal variables containing more than 2 categories, all categories are displayed simultaneously. For dichotomous variables specified as nominal, by default only the last category is displayed.

Results

Summary Sheet

Summary (descriptive) statistics: For the dependent variables and the quantitative explanatory variables, XLSTAT displays the number of observations, the number of observations with missing data, the number of observations with no missing data, the mean, and the unbiased standard deviation. For the nominal explanatory variables, the number and frequency of cases belonging to each level are displayed.

Summary Statistics:

- Model Name: The models are named after the number of classes the model contains.
- LL: The likelihood-ratio goodness-of-fit value for the current model.
- BIC(LL), AIC(LL), AIC3(LL): BIC, AIC and AIC3 (based on LL). In addition to model fit, these statistics take into account the parsimony (df or Npar) of the model. When comparing models, the lower the BIC, AIC and AIC3 value the better the model.
- Npar: Number of parameters.
- L²: Likelihood ratio chi-squared. Not available if the model contains 1 or more continuous indicators.
- df: Degrees of freedom for L2.
- p-value: Model fit p-value for L2.
- Class.Err.: Expected classification error. The expected proportion of cases misclassified when classification of cases is based on modal assignment (i.e., assigned to the class having the highest membership probability). The closer this value is to 0 the better.

Model Output Sheet
Model Summary Statistics:

Model:

Number of cases: This is the number of cases used in model estimation. This number may be less than the original number of cases on the data file if missing cases have been excluded.

- Number of replications: Total number of observations
- Number of parameters (Npar): This is the number of distinct parameters estimated.
- Seed (random numbers): The seed required to reproduce this model.
- Best seed: The single best seed that can reproduce this model more quickly using the number of starting sets =0.

Estimation summary:

- EM iterations: number of EM iterations used.
- Log-posterior: Log-posterior value.
- $L^2$: The likelihood-ratio goodness-of-fit value for the current model.
- Final convergence value: Final convergence value.
- Newton-Raphson iteration: Number of Newton-Raphson iterations used.
- Log-posterior: Log-posterior value.
- $L^2$: The likelihood-ratio goodness-of-fit value for the current model.
- Final convergence value: Final convergence value.

Chi-Square statistics:

- Degrees of freedom (df): The degrees of freedom for the current model.
- $L^2$: The likelihood-ratio goodness-of-fit value for the current model. If the bootstrap p-value for the L2 statistic has been requested, the results will be displayed here.
- $X^2$ and Cressie-Read: These are alternatives to $L^2$ that should yield a similar p-value according to large sample theory if the model specified is valid and the data is not sparse.
- BIC, AIC, AIC3 and CAIC (based on $L^2$): In addition to model fit, these statistics take into account the parsimony (df or Npar) of the model. When comparing models, the lower the BIC, AIC, AIC3 and CAIC value the better the model.
• SABIC (based on $L^2$): Sample size adjusted BIC, an information criterion similar to BIC but with $\log(N)$ replaced by $\log((N+2)/24)$.

• Dissimilarity Index: A descriptive measure indicating how much the observed and estimated cell frequencies differ from one another. It indicates the proportion of the sample that needs to be moved to another cell to get a perfect fit.

Log-likelihood statistics:

• Log-likelihood (LL): LN(Likelihood) displayed here.

• **Log-prior:** this is the term in the function maximized in the parameter estimation that is associated with the Bayes constants. This term equals 0 if all Bayes constants are set to 0.

• **Log-posterior:** this is the term in the function that is maximized in the parameter estimation. The value of the log-posterior function is obtained as the sum of the log-likelihood and log-prior values.

• BIC, AIC, AIC3 and CAIC (based on LL): these statistics (information criteria) weight fit and parsimony by adjusting the LL to account for the number of parameters in the model. The lower the value, the better the model.

• SABIC (based on LL): Sample size adjusted BIC, an information criterion similar to BIC but with $\log(N)$ replaced by $\log((N+2)/24)$.

Classification statistics:

• Classification errors: When classification of cases is based on modal assignment (to the class having the highest membership probability), the proportion of cases that are estimated to be misclassified is reported by this statistic. The closer this value is to 0 the better.

• Reduction of errors (Lambda), Entropy R2, Standard R2: These pseudo R-squared statistics indicate how well one can predict class memberships based on the observed variables (indicators and covariates). The closer these values are to 1 the better the predictions.

• Classification Log-likelihood: Log-likelihood value under the assumption that the true class membership is known.

• EN: Entropy.

• CLC: CL*2

• AWE: Similar to BIC, but also takes classification performance into account.
• ICL-BIC: BIC-2*En

Classification table:
• **Modal table**: Cross-tabulates modal class assignments.
• **Proportional table**: Cross-tabulates probabilistic class assignments.

Profile:
• **Cluster Size**: Size of each cluster
• **Indicators**: The body of the table contains (marginal) conditional probabilities that show how the clusters are related to the Nominal or Ordinal indicator variables. These probabilities sum to 1. For indicators specified as Continuous, the body of the table contains means instead of probabilities. For indicators specified as Ordinal, means are displayed in addition to the conditional probabilities within each cluster (column).
• **s.e. (standard errors)**: standard errors for the (marginal) conditional probabilities.
• **Profile plot**: The probabilities and means that appear in the Profile Output, are displayed graphically in the Profile Plot

Frequencies / Residuals:
Table of observed vs. estimated expected frequencies (and residuals). Note: Residuals having magnitude greater than 2 are statistically significant. This output is not reported in the case of 1 or more continuous indicators.

Bivariate residuals:
• **Indicators**: a table containing the bivariate residuals (BVRs) for a model. Large BVRs suggest violation of the local independence assumption.

Scoring equation: regression coefficients associated with the multinomial logit model.

Classification: Outputs for each observation the posterior class memberships and the modal assignment based on the current model.
**Estimation Warnings**

**WARNING: negative number of degrees of freedom.**

This warning indicates that the model contains more parameters than cell counts. A necessary (but not sufficient) condition for identification of the parameters of a latent class model is that the number of degrees of freedom is nonnegative. This warning thus indicates that the model is not identified. The remedy is to use a model with fewer latent classes.

**WARNING: # boundary or non-identified parameter(s)**

This warning is derived from the rank of the information matrix (Hessian or its outer-product approximation). When there are non-identified parameters, the information matrix will not be full rank. The number reported is the rank deficiency, which gives an indication of the number of non-identified parameters.

Note that there are two problems associated with this identification check. The first is that boundary estimates also yield rank deficiencies. In other words, when there is a rank deficiency, we do not know whether it is caused by boundaries or non-identified parameters. The XLSTAT-LG Bayes Constants prevent boundaries from occurring, which solves the first problem related to this message. However, a second problem is that this identification check cannot always detect non-identification when Bayes Constants are used; that is, Bayes Constants can make an otherwise non-identified model appear to be identified.

**WARNING: maximum number of iterations reached without convergence**

This warning is provided if the maximum specified EM and Newton-Raphson iterations are reached without meeting the tolerance criterion. If the (by default very strict) tolerance is almost reached, the solution is probably be ok. Otherwise, the remedy is to reestimate the model with a sharper EM tolerance and/or more EM iterations, which makes sure that the switch from EM to Newton-Raphson occurs later. The default number of 50 Newton-Raphson iterations will generally be more than sufficient.

**WARNING: estimation procedure did not converge (# gradients larger than 1.0e-3)**

This message may be related to the previous message, in which case the same remedy may be used. If the previous message is not reported, this indicates that there is a more serious non-convergence problem. The algorithms may have gotten trapped in a very flat region of the parameters space (a saddle point). The best remedy is to re-estimate the model with a different seed, and possibly with a larger number of Start Sets and more Iterations per set.
Example

A tutorial on latent class clustering is available on the Addinsoft website:

http://www.xlstat.com/demo-lcc.htm

References


Latent class regression

This tool is part of the XLSTAT-LG module. Use this tool to classify cases into meaningful clusters (latent classes) that differ on one or more parameters from latent class (LC) Regression models. LC Regression simultaneously classifies cases and estimates separate regression coefficients based on linear, logistic, multinomial, ordinal, binomial count or Poisson regression models.

Description

The latent class regression feature of XLSTAT is part of the XLSTAT-LG module, a powerful clustering tool based on Latent GOLD® 5.0:

Latent class analysis (LCA) involves the construction of latent classes (LC) which are unobserved (latent) subgroups or segments of cases. The latent classes are constructed based on the observed (manifest) responses of the cases on a set of indicator variables. Cases within the same latent class are homogeneous with respect to their responses on these indicators, while cases in different latent classes differ in their response patterns. Formally, latent classes are represented by K distinct categories of a nominal latent variable X. Since the latent variable is categorical, LC modeling differs from more traditional latent variable approaches such as factor analysis, structural equation models, and random-effects regression models since these approaches are based on continuous latent variables.

XLSTAT-LG contains separate modules for estimating two different model structures - LC Cluster models and LC Regression models - which are useful in somewhat different application areas. To better distinguish the output across modules, latent classes are labeled 'clusters' for LC Cluster models and 'classes' for LC Regression models. In this manual we also refer to latent classes using the term 'segments'.

The LC Regression Model:

- Is used to predict a dependent variable as a function of predictor variables (Regression model).
- Includes a K-category latent variable X to cluster cases (LC model)
- Each category represents a homogeneous subpopulation (segment) having identical regression coefficients (LC Regression Model).
- Each case may contain multiple records (Regression with repeated measurements).
The appropriate model is estimated according to the scale type of the dependent variable:

- Continuous - Linear regression model (with normally distributed residuals)
- Nominal (with more than 2 levels) - Multinomial logistic regression
- Ordinal (with more than 2 ordered levels) - Adjacent-category ordinal logistic regression model
- Count: Log-linear Poisson regression
- Binomial Count: Binomial logistic regression model

Note that a dichotomous dependent variable can be analyzed using either nominal, ordinal, or a binomial count as its scale type without any difference in the model results.

For either of the two model structures:

- Diagnostic statistics are available to help determine the number of latent classes, clusters, or segments
- For models containing $K > 1$ classes, covariates can be included in the model to improve classification of each case into the most likely segment.

Copyright ©2014 Statistical Innovations Inc. All rights reserved.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
Click this button to reload the default options.

Click this button to delete the data selections.

Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

**Y / Dependent variables:** Select the dependent variable here. If the ‘Column labels’ option is activated make sure that the headers of the variable(s) have also been selected.

Note: In case multiple dependent variables are selected, multiple independent regression analysis are performed. Separate output for each dependent variable will be provided and only a single scale type can be selected for all the dependent variables.

**Response type:** Select the scale type of the dependent variable. The dependent variable may be Nominal, Ordinal, Continuous, Binomial, or Count.

- **Nominal.** This setting should be used for categorical variables where the categories have no natural ordering. If the dependent variable is set to Nominal, the multinomial logit model is used.

- **Ordinal.** This setting should be used for categorical variables where the categories are ordered (either from high to low or low to high). The adjacent category logit model, also known as the baseline category logit model is specified.

- **Continuous.** This setting should be used when the variable is continuous. If the dependent variable is set to Continuous, the normal linear Regression model is used.

- **Binomial.** This setting should be used when the variable represents binomial counts. If the dependent variable is set to Binomial Count, the binomial model is used and you can also specify a variable to be used as an exposure (see Exposure). During the scan, the program checks to make sure that the exposure, if specified, is larger than any observed count.

- **Counts.** This setting should be used when the variable represents Poisson counts. If the dependent variable is set to Count, the Poisson model is used and you can also specify an additional variable to be used as an exposure (see Exposure).

**Exposure.** The Exposure field is active only if the scale type for the dependent variable has been specified to be Binomial or Count. (For other scale types, no exposure variable is used.)
For dependent variables specified as Binomial or Count, the exposure is specified by designating a variable as the exposure variable or, if no such variable is designated, by entering a value in the exposure constant box which appears to the right of the Exposure variable box. The use of an exposure variable allows the exposure to vary over cases.

By default, the value in the Exposure constant box is 1, a value often used to represent the Poisson exposure. To change the exposure constant, highlight the value in the exposure constant box and type in the desired value. Alternatively, you can select an exposure variable.

When the scale type is specified as Binomial, the value of the dependent variable represents the number of 'successes' in $N$ trials. In this case, the exposure represents the number of trials (the values for $N$), and hence should never take on a value lower than the value of the dependent variable and hence typically should be higher than the default constant of 1. Before the actual model estimation, XLSTAT-LG checks each case and will provide a warning message if this condition is not met for one or more cases. An exposure variable should be designated if the number of trials is not the same for all cases.

**Explanatory variables.** Select any variable(s) to be used as predictors of the dependent variable. Predictors may be treated as Nominal or Numeric. If no predictors are selected, the model will contain an intercept only.

- **Numeric.** This setting should be used for an ordinal or continuous covariate or predictor.
- **Nominal.** This setting should be used for categorical variables where the categories have no natural ordering.

**Range:** Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Activate this option to display the results in a new worksheet of the active workbook.

**Workbook:** Activate this option to display the results in a new workbook.

**Column labels:** Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

**Observation labels:** Activate this option if labels are available for the N observations. Then select the corresponding data. If the ‘Column labels’ option is activated you need to include a header in the selection.

With repeated measures data (multiple records per case) the Observation labels variable serves as a case ID variable, which groups the records from each case together so that they are assigned to the same fold during cross-validation. If this option is not activated, labels for the observations are automatically generated by XLSTAT (Obs1, Obs2 …), so that each case contains a single record.
**Replication weights**: Activate this option to assign a Replication Weight. A common application of replication weights is in the estimation of certain kinds of allocation models, where respondents assign a fixed number of points to each of J alternatives. For each case, the assigned points are used as replication weights to weight each of J responses. A weighted multinomial logit model is estimated.

**Case weights**: Activate this option if you want to weight the observations. If you do not activate this option, all weights are set to 1. The weights must be non-negative values. Setting a case weight to 2 is equivalent to repeating the same observation twice. If the ‘Variable labels’ option is activated, make sure that the header (first row) has also been selected.

**Number of clusters**:

**from**: Enter a number between 1-25.

**to**: Enter a number between 1-25.

Note: To specify a fixed number of clusters \(K\): use from \(K\) to \(K\). For example, to estimate a 2 class model: from 2 to 2.

**Use separate sheets**: Activate this option if you want the program to produce separate sheets for each cluster model estimated. A separate sheet with summary statistics for all models estimated will also be produced.

**Options** tab:

Parameter estimation occurs using an iterative algorithm which begins using the Expectation-Maximization (EM) algorithm until either the maximum number of EM iterations (Iterations EM) or the EM convergence criterion (Tolerance(EM)) is reached. Then, the program switches to perform Newton Raphson (NR) iterations which continue until the maximum number of NR iterations (Iterations Newton-Raphson) or the overall converge criterion (Tolerance) is reached. The program also stops iterating when the change in the log-posterior is negligible (smaller than \(10^{-12}\)). A warning is given if one of the elements of the gradient is larger than \(10^{-3}\).

Sometimes, for example in the case of models with many parameters, it is more efficient to use only the EM algorithm. This is accomplished by setting Iterations Newton-Raphson to 0. With very large models, one may also consider suppressing the computation of standard errors (and associated Wald statistics).

**Convergence**:

**Tolerance(EM)**: Expectation-Maximization (EM) Tolerance is the sum of absolute relative changes of parameter values in a single iteration as long as the EM algorithm is used. It determines when the program switches from EM to Newton-Raphson (if the NR iteration limit
has been set to \( > 0 \). Increasing the EM Tolerance will switch faster from EM to NR. To change this option, click the value to highlight it, then type in a new value. You may enter any non-negative real number. The default is 0.01. Values between 0.01 and 0.1 (1\% and 10\%) are reasonable.

**Tolerance**: Overall Tolerance (Tolerance) is the sum of absolute relative changes of parameter values in a single iteration. It determines when the program stops its iteration. The default is \( 1.0 \times 10^{-8} \) which specifies a tight convergence criterion. To change this option, click the value to highlight it, then type in a new value. You may enter any non-negative real number.

**Iterations**:

**EM**: Maximum number of EM iterations. The default is 250. If the model does not converge after 250 iterations, this value should be increased. You also may want to increase this value if you set Newton-Raphson iterations = 0. To change this option, click the value to highlight it, then type in a new value. You may enter any non-negative integer.

**Newton-Raphson**: Maximum number of NR iterations. The default is 50. If the model does not converge after 50 iterations, this value should be increased. To change this option, click the value to highlight it, then type in a new value. You may enter any non-negative integer. A value of 0 is entered to direct XLSTAT-LG to use only EM, which may produce faster convergence in models with many parameters or in models that contain continuous indicators.

**Start values**:

The best way to prevent ending up with a local solution is the use of multiple sets of starting values since different sets of starting values may yield solutions with different log-posterior values. The use of such multiple sets of random starting values is automated. This procedure increases considerably the probability of finding the global solution, but in general does not guarantee that it will be found in a single run. To reduce the likelihood of obtaining a local solution, the following options can be used to either increasing the number of start sets, the number of iterations per set, or both.

**Random sets**: The default is 16 for the number of random sets of starting values to be used to start the iterative estimation algorithm. Increasing the number of sets of random starting values for the model parameters reduces the likelihood of converging to a local (rather than global) solution. To change this option, click the value to highlight it, then type in a new value. You may enter any non-negative integer. Using either the value 0 or 1 results in the use of a single set of starting values.

**Iterations**: This option allows specification of the number of iterations to be performed per set of start values. XLSTAT-LG first performs this number of iterations within each set and subsequently twice this number within the best 10\% of the start sets. For some models, many more than 50 iterations per set may need to be performed to avoid local solutions.
Seed (random numbers): The default value of 123456789 means that the Seed is obtained during estimation using a pseudo random number generator based on clock time. Specifying a non-negative integer different from 0, yields the same result each time.

To specify a particular numeric seed (such as the Best Start Seed reported in the Model Summary Output for a previously estimated model), click the value to highlight it, then type in a non-negative integer. When using the Best Start Seed, be sure to deactivate the Random Sets option (using Random Sets = 0).

Tolerance: Indicates the convergence criterion to be used when running the model of interest with the various start sets. The definition of this tolerance is the same as the one that is used for the EM and Newton-Raphson Iterations.

Bayes Constants:

The Bayes options can be used to eliminate the possibility of obtaining boundary solutions. You may enter any non-negative real value. Separate Bayes constants can be specified for three different situations:

Latent: The default is 1. Increase the value to increase the weight allocated to the Dirichlet prior which is used to prevent the occurrence of boundary zeroes in estimating the latent distribution. The number can be interpreted as a total number of added cases that is equally distributed among the classes (and the covariate patterns). To change this option, click the value to highlight it, then type in a new value.

Categorical: The default is 1. Increase the value to increase the weight allocated to the Dirichlet prior which is used in estimating multinomial models with variables specified as Ordinal or Nominal. This number can be interpreted as a total number of added cases to the cells in the models for the indicators to prevent the occurrence of boundary solutions. To change this option, click the value to highlight it, then type in a new value.

Error variance: The default is 1. Increase the value to increase the weight allocated to the inverse-Wishart prior which is used in estimating the error variance-covariance matrix in models for continuous dependent variables or indicators. The number can be interpreted as the number of pseudo-cases added to the data, each pseudo-case having a squared error equal to the total variance of the indicator concerned. Such a prior prevents variances of zero from occurring. To change this option, click the value to highlight it, then type in a new value.

For technical details, see section 7.3 of Vermunt and Magidson (2013a).

Class Independent:

Various restrictions are available for intercepts and predictor effects. In addition, for models with continuous dependent variables, restrictions are available for error variances.
• **Error variances:** This option indicates that the error covariances are restricted to be equal across classes (class independent).

• **Predictors (1 or more).** This option indicates that the predictors are restricted to be equal across classes (class independent).

• **Intercept.** This option indicates that the intercept is restricted to be equal across classes (class independent).

**Missing data** tab:

**Do not accept missing data:** Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

**Remove observations:** Activate this option to remove the observations with missing data.

**Outputs** tab:

**Descriptive statistics:** Activate this option to display descriptive statistics for the variables selected.

**Statistics:** Activate this option to display the following statistics about the model(s).

**Chi-square:** Activate this option to display various chi-square based statistics related to model fit.

**Log-likelihood:** Activate this option to display log-likelihood statistics.

**Classification:** Activate this option to display the Classification Table, which cross-tabulates modal and probabilistic class assignment.

**Parameters:**

**Standard errors:** Activate this option to display the standard errors of the parameters. The standard (Hessian) computation method makes use of the second-order derivatives of the log-likelihood function called the Hessian matrix.

**Wald tests:** Activate this option to display the Wald statistics.

**Frequencies / Residuals:** Activate this option to display the observed and expected frequencies along with the standardized residuals for a model. This output is not available if at least one indicators is continuous. This output is not reported in the case 1 or more continuous indicators.
**Iteration details:** Activate this option to display technical information associated with the performance of the estimation algorithm, such as log-posterior and log-likelihood values at convergence:

- EM algorithm,
- Newton algorithm.

When applicable, this file also contains warning messages concerning non-convergence, unidentified parameters and boundary solutions.

**Estimated values:** Activate this option to display the predicted values information (the probability of responding to each category) to the data. The following variables (and variable names) will be shown:

- `pred_1` - the predicted prob of responding in the first category
- `pred_2` - the predicted prob of responding in the second category
- `pred_dep` - the predicted value (weighted average of the category scores, with the predicted probs as the weights)

**Classification:** Activate this option to display a table containing the posterior membership probability and the modal assignment for each of the cases based on the current model.

**Nominal coding:**

**Effect (default).** By default, the Parameter Output contains effect coding for nominal indicators, dependent variable, active covariates and the latent classes (clusters).

Use either of these options to change to dummy coding.

- **a1=0 (Dummy First).** Selection of this option causes dummy coding to be used with the first category serving as the reference category.
- **an=0 (Dummy Last).** Selection of this option causes dummy coding to be used with the last category serving as the reference category.

**Charts tab:**

**Profile plot:** Activate this option to display the profile plot.
Results

Summary Sheet

Summary (descriptive) statistics: For the dependent variables and the quantitative explanatory variables, XLSTAT displays the number of observations, the number of observations with missing data, the number of observations with no missing data, the mean, and the unbiased standard deviation. For the nominal explanatory variables, the number and frequency of cases belonging to each level are displayed.

Summary Statistics:

- **Model Name**: The models are named after the number of classes the model contains.
- **LL**: The likelihood-ratio goodness-of-fit value for the current model.
- **BIC(LL), AIC(LL), AIC3(LL)**: BIC, AIC and AIC3 (based on LL). In addition to model fit, these statistics take into account the parsimony (df or Npar) of the model. When comparing models, the lower the BIC, AIC and AIC3 value the better the model.
- **Npar**: Number of parameters.
- **L²**: Likelihood ratio chi-squared. Not available if the model contains 1 or more continuous indicators.
- **df**: Degrees of freedom for \( L^2 \).
- **p-value**: Model fit p-value for \( L^2 \).
- **Class.Err.**: Expected classification error. The expected proportion of cases misclassified when classification of cases is based on modal assignment (i.e., assigned to the class having the highest membership probability). The closer this value is to 0 the better.

Model Output Sheet

Model Summary Statistics:

Model:

- **Number of cases**: This is the number of cases used in model estimation. This number may be less than the original number of cases on the data file if missing cases have been excluded.
### Number of replications:
Total number of observations.

### Number of parameters (Npar):
This is the number of distinct parameters estimated.

### Seed (random numbers):
The seed required to reproduce this model.

### Best seed:
The single best seed that can reproduce this model more quickly using the number of starting sets = 0.

### Estimation summary:
- EM iterations: number of EM iterations used.
- Log-posterior: Log-posterior value.
- \( L^2 \): The likelihood-ratio goodness-of-fit value for the current model.
- Final convergence value: Final convergence value.
- Newton-Raphson iteration: number of Newton-Raphson iterations used.
- Log-posterior: Log-posterior value.
- \( L^2 \): The likelihood-ratio goodness-of-fit value for the current model.
- Final convergence value: Final convergence value.

### Chi-Square statistics:
- Degrees of freedom (df): The degrees of freedom for the current model.
- \( L^2 \): The likelihood-ratio goodness-of-fit value for the current model. If the bootstrap p-value for the \( L^2 \) statistic has been requested, the results will be displayed here.
- \( X^2 \) and Cressie-Read: These are alternatives to \( L^2 \) that should yield a similar p-value according to large sample theory if the model specified is valid and the data is not sparse.
- BIC, AIC, AIC3 and CAIC (based on \( L^2 \)): In addition to model fit, these statistics take into account the parsimony (df or Npar) of the model. When comparing models, the lower the BIC, AIC, AIC3 and CAIC value the better the model.
- SABIC (based on \( L^2 \)): Sample size adjusted BIC, an information criterion similar to BIC but with \( \log(N) \) replaced by \( \log((N+2)/24) \).
- Dissimilarity Index: A descriptive measure indicating how much the observed and estimated cell frequencies differ from one another. It indicates the proportion of the sample that needs to be moved to another cell to get a perfect fit.
Log-likelihood statistics:

Log-likelihood (LL): displayed here.

- Log-prior: this is the term in the function maximized in the parameter estimation that is associated with the Bayes constants. This term equals 0 if all Bayes constants are set to 0.

- Log-posterior: this is the function that is maximized in the parameter estimation. The value of the log-posterior function is obtained as the sum of the log-likelihood and log-prior values.

- BIC, AIC, AIC3 and CAIC (based on LL): these statistics (information criteria) weight fit and parsimony by adjusting the LL to account for the number of parameters in the model. The lower the value, the better the model.

- SABIC (based on LL): Sample size adjusted BIC, an information criterion similar to BIC but with log(N) replaced by log((N+2)/24).

Classification statistics:

- Classification errors: When classification of cases is based on modal assignment (to the class having the highest membership probability), the proportion of cases that are estimated to be misclassified is reported by this statistic. The closer this value is to 0 the better.

- Reduction of errors (Lambda), Entropy R2, Standard R2: These pseudo R-squared statistics indicate how well one can predict class memberships based on the observed variables (indicators and covariates). The closer these values are to 1 the better the predictions.

- Classification Log-likelihood: Log-likelihood value under the assumption that the true class membership is known.

- EN: Entropy.

- CLC: CL*2

- AWE: Similar to BIC, but also takes classification performance into account.

- ICL-BIC: BIC-2*En

Classification table:

- **Modal table**: Cross-tabulates modal class assignments.
• **Proportional table**: Cross-tabulates probabilistic class assignments.

**Prediction statistics:**

The columns in this table correspond to:

- Baseline: prediction error of the baseline model (also referred to as null-model)
- Model: the prediction error of the estimated model.
- R²: the proportional reduction of errors in the estimated model compared to the baseline model

The rows in this table correspond to:

- Squared Error: Average prediction error based on squared error.
- Minus Log-likelihood: Average prediction error based on minus the log-likelihood.
- Absolute Error: Average prediction error based on absolute error.
- Prediction error: Average prediction error based on proportion of prediction errors (for categorical variables only).

For technical information, see section 8.1.5 of Vermunt and Magidson (2013a).

**Prediction table**: For nominal and ordinal dependent variables, a prediction table that cross-classifies observed and against estimated values is also provided.

**Parameters**:

- R²: class-specific and overall R² values. The overall R² indicates how well the dependent variable is overall predicted by the model (same measure as appearing in Prediction Statistics). For ordinal, continuous, and (binomial) counts, these are standard R² measures. For nominal dependent variables, these can be seen as weighted averages of separate R² measures for each category, where each category is represented by a dummy variable = 1 for that category and 0 for all other categories.

- **Intercept**: intercept of the linear regression equation.

- **s.e.**: standard errors of the parameters.

- **z-value**: z-test statistics corresponding to the parameter tests.

- **Wald**: Wald statistics are provided in the output to assess the statistical significance of the set of parameter estimates associated with a given variable. Specifically, for each
variable, the Wald statistic tests the restriction that each of the parameter estimates in that set equals zero (for variables specified as Nominal, the set includes parameters for each category of the variable). For Regression models, by default, two Wald statistics (Wald, Wald(=)) are provided in the table when more than 1 class has been estimated. For each set of parameter estimates, the Wald(=) statistic considers the subset associated with each class and tests the restriction that each parameter in that subset equals the corresponding parameter in the subsets associated with each of the other classes. That is, the Wald(=) statistic tests the equality of each set of regression effects across classes.

- **p-value**: measures of significance for the estimates.
- **Mean**: means for the regression coefficients.
- **Std.Dev**: standard deviations for the regression coefficients.

**Classification**: Outputs for each observation the posterior class memberships and the modal assignment based on the current model.

**Estimation Warnings**

**WARNING**: negative number of degrees of freedom.

This warning indicates that the model contains more parameters than cell counts. A necessary (but not sufficient) condition for identification of the parameters of a latent class model is that the number of degrees of freedom is nonnegative. This warning thus indicates that the model is not identified. The remedy is to use a model with fewer latent classes.

**WARNING**: # boundary or non-identified parameter(s)

This warning is derived from the rank of the information matrix (Hessian or its outer-product approximation). When there are non-identified parameters, the information matrix will not be full rank. The number reported is the rank deficiency, which gives an indication of the number of non-identified parameters.

Note that there are two problems associated with this identification check. The first is that boundary estimates also yield rank deficiencies. In other words, when there is a rank deficiency, we do not know whether it is caused by boundaries or non-identified parameters. The XLSTAT-LG Bayes Constants prevent boundaries from occurring, which solves the first problem related to this message. However, a second problem is that this identification check cannot always detect non-identification when Bayes Constants are used; that is, Bayes Constants can make an otherwise non-identified model appear to be identified.

**WARNING**: maximum number of iterations reached without convergence
This warning is provided if the maximum specified EM and Newton-Raphson iterations are reached without meeting the tolerance criterion. If the (by default very strict) tolerance is almost reached, the solution is probably be ok. Otherwise, the remedy is to reestimate the model with a sharper EM tolerance and/or more EM iterations, which makes sure that the switch from EM to Newton-Raphson occurs later. The default number of 50 Newton-Raphson iterations will generally be more than sufficient.

WARNING: estimation procedure did not converge (# gradients larger than 1.0e-3)

This message may be related to the previous message, in which case the same remedy may be used. If the previous message is not reported, this indicates that there is a more serious non-convergence problem. The algorithms may have gotten trapped in a very flat region of the parameters space (a saddle point). The best remedy is to re-estimate the model with a different seed, and possibly with a larger number of Start Sets and more Iterations per set.

Example

A tutorial on latent class regression is available on the Addinsoft website:

http://www.xlstat.com/demo-lcr.htm

References

Vermunt J.K. (2010). Latent class modeling with covariates: Two improved three-step approaches. Political Analysis, 18, 450-469. Link:
http://members.home.nl/jeroenvermunt/lca_three_step.pdf


http://www.statisticalinnovations.com/technicalsupport/Lgetchnical.pdf

Dose effect analysis

Use this function to model the effects of a dose on a response variable, if necessary taking into account an effect of natural mortality.

Description

This tool uses logistic regression (Logit, Probit, complementary Log-log, Gompertz models) to model the impact of doses of chemical components (for example a medicine or phytosanitary product) on a binary phenomenon (healing, death).

More information on logistic regression is available in the help section on this subject.

Natural mortality

This tool takes natural mortality into account in order to model the phenomenon studied more accurately. Indeed, if we consider an experiment carried out on insects, certain will die because of the dose injected, and others from other phenomenon. None of these associated phenomena are relevant to the experiment concerning the effects of the dose but may be taken into account. If \( p \) is the probability from a logistic regression model corresponding only to the effect of the dose, and if \( m \) is natural mortality, then the observed probability that the insect will succumb is:

\[
P(\text{obs}) = m + (1 - m) \cdot p
\]

Abbot's formula (Finney, 1971) is written as:

\[
p = \frac{P(\text{obs}) - m}{(1 - m)}
\]

The natural mortality \( m \) may be entered by the user as it is known from previous experiments, or is determined by XLSTAT.

ED 50, ED 90, ED 99

XLSTAT calculates ED 50 (or median dose), ED 90 and ED 99 doses which correspond to doses leading to an effect respectively on 50%, 90% and 99% of the population.
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

Reload default options: Click this button to reload the default options.

Delete selections: Click this button to delete the data selections.

Change data handling: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Dependent variables:

Response variable(s): Select the response variable(s) you want to model. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

Response type: Choose the type of response variable you have selected:

- **Binary variable**: If you select this option, you must select a variable containing exactly two distinct values. If the variable has value 0 and 1, XLSTAT will see to it that the high probabilities of the model correspond to category 1 and that the low probabilities correspond to category 0. If the variable has two values other than 0 or 1 (for example Yes/No), the lower probabilities correspond to the first category and the higher probabilities to the second.

- **Sum of binary variables**: If your response variable is a sum of binary variables, it must be of type numeric and contain the number of positive events (event 1) amongst those observed. The variable corresponding to the total number of events observed for this
observation (events 1 and 0 combined) must then be selected in the "Observation weights" field. This case corresponds, for example, to an experiment where a dose D (D is the explanatory variable) of a medicament is administered to 50 patients (50 is the value of the observation weights) and where it is observed that 40 get better under the effects of the dose (40 is the response variable).

Explanatory variables:

**Quantitative:** Activate this option if you want to include one or more quantitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The data selected may be of the numerical type. If the variable header has been selected, check that the "Variable labels" option has been activated.

**Qualitative:** Activate this option if you want to include one or more qualitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Variable labels" option has been activated.

**Model:** Choose the type of function to use (see description).

**Range:** Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Activate this option to display the results in a new worksheet of the active workbook.

**Workbook:** Activate this option to display the results in a new workbook.

**Variable labels:** Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

**Observation labels:** Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 ...).
**Observation weights**: This field must be entered if the "sum of binary variables" option has been chosen. Otherwise, this field is not active. If a column header has been selected, check that the "Variable labels" option has been activated.

**Options** tab:

**Firth's method**: Activate this option to use Firth's penalized likelihood (see description).

**Confidence interval (%)**: Enter the percentage range of the confidence interval to use for the various tests and for calculating the confidence intervals around the parameters and predictions. Default value: 95.

**Stop conditions**:

- **Iterations**: Enter the maximum number of iterations for the Newton-Raphson algorithm. The calculations are stopped when the maximum number if iterations has been exceeded. Default value: 100.

- **Convergence**: Enter the maximum value of the evolution of the log of the likelihood from one iteration to another which, when reached, means that the algorithm is considered to have converged. Default value: 0.000001.

**Take the log**: Activate this option so that XLSTAT uses the logarithm of the input variables in the model.

**Natural mortality parameter**:

- **Optimized**: Choose this option so that XLSTAT optimizes the value of the natural mortality parameter.

- **User defined**: Choose this option to set the value of the natural mortality parameter.

**Validation** tab:

**Validation**: Activate this option if you want to use a sub-sample of the data to validate the model.

**Validation set**: Choose one of the following options to define how to obtain the observations used for the validation:

- **Random**: The observations are randomly selected. The "Number of observations" N must then be specified.

- **N last rows**: The N last observations are selected for the validation. The "Number of observations" N must then be specified.
- **N first rows**: The N first observations are selected for the validation. The “Number of observations” N must then be specified.

- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to use for the validation.

**Prediction** tab:

**Prediction**: activate this option if you want to select data to use them in prediction mode. If activate this option, you need to make sure that the prediction dataset is structured as the estimation dataset: same variables with the same order in the selections. On the other hand, variable labels must not be selected: the first row of the selections listed below must correspond to data.

**Quantitative**: activate this option to select the quantitative explanatory variables. The first row must not include variable labels.

**Qualitative**: activate this option to select the qualitative explanatory variables. The first row must not include variable labels.

**Observations labels**: activate this option if observations labels are available. Then select the corresponding data. If this option is not activated, the observations labels are automatically generated by XLSTAT (PredObs1, PredObs2 …).

**Missing data** tab:

**Remove observations**: Activate this option to remove the observations with missing data.

**Estimate missing data**: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

- **Nearest neighbour**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.

**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the variables selected.

**Correlations**: Activate this option to display the explanatory variables correlation matrix.
**Goodness of fit statistics**: Activate this option to display the table of goodness of fit statistics for the model.

**Type III analysis**: Activate this option to display the type III analysis of variance table.

**Model coefficients**: Activate this option to display the table of coefficients for the model. Optionally, **confidence intervals** of type "profile likelihood" can be calculated (see description).

**Standardized coefficients**: Activate this option if you want the standardized coefficients (beta coefficients) for the model to be displayed.

**Equation**: Activate this option to display the equation for the model explicitly.

**Predictions and residuals**: Activate this option to display the predictions and residuals for all the observations.

**Probability analysis**: If only one explanatory variable has been selected, activate this option so that XLSTAT calculates the value of the explanatory variable corresponding to various probability levels.

**Charts** tab:

**Regression charts**: Activate this option to display regression chart:

- **Standardized coefficients**: Activate this option to display the standardized parameters for the model with their confidence interval on a chart.

- **Predictions**: Activate this option to display the regression curve.
  - **Confidence intervals**: Activate this option to have confidence intervals displayed on charts (1) and (4).

**Results**

XLSTAT displays a large number tables and charts to help in analyzing and interpreting the results.

**Summary statistics**: This table displays descriptive statistics for all the variables selected. For the quantitative variables, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed. For qualitative variables,
including the dependent variable, the categories with their respective frequencies and percentages are displayed.

**Correlation matrix**: This table displays the correlations between the explanatory variables.

**Correspondence between the categories of the response variable and the probabilities**: This table shows which categories of the dependent variable have been assigned probabilities 0 and 1.

**Goodness of fit coefficients**: This table displays a series of statistics for the independent model (corresponding to the case where the linear combination of explanatory variables reduces to a constant) and for the adjusted model.

- **Observations**: The total number of observations taken into account (sum of the weights of the observations);
- **Sum of weights**: The total number of observations taken into account (sum of the weights of the observations multiplied by the weights in the regression);
- **DF**: Degrees of freedom;
- **-2 Log(Like.)**: The logarithm of the likelihood function associated with the model;
- **$R^2$ (McFadden)**: Coefficient, like the $R^2$, between 0 and 1 which measures how well the model is adjusted. This coefficient is equal to 1 minus the ratio of the likelihood of the adjusted model to the likelihood of the independent model;
- **$R^2(Cox and Snell)$**: Coefficient, like the $R^2$, between 0 and 1 which measures how well the model is adjusted. This coefficient is equal to 1 minus the ratio of the likelihood of the adjusted model to the likelihood of the independent model raised to the power $2/Sw$, where $Sw$ is the sum of weights.
- **$R^2(Nagelkerke)$**: Coefficient, like the $R^2$, between 0 and 1 which measures how well the model is adjusted. This coefficient is equal to ratio of the $R^2$ of Cox and Snell, divided by 1 minus the likelihood of the independent model raised to the power $2/Sw$;
- **AIC**: Akaike’s Information Criterion;
- **SBC**: Schwarz’s Bayesian Criterion.

**Test of the null hypothesis H0: $Y=p0$**: The H0 hypothesis corresponds to the independent model which gives probability $p0$ whatever the values of the explanatory variables. We seek to check if the adjusted model is significantly more powerful than this model. Three tests are available: the likelihood ratio test (-2 Log(Like.)), the Score test and the Wald test. The three statistics follow a Chi$^2$ distribution whose degrees of freedom are shown.
**Type III analysis:** This table is only useful if there is more than one explanatory variable. Here, the adjusted model is tested against a test model where the variable in the row of the table in question has been removed. If the probability $Pr > LR$ is less than a significance threshold which has been set (typically 0.05), then the contribution of the variable to the adjustment of the model is significant. Otherwise, it can be removed from the model.

**Model parameters:** The parameter estimate, corresponding standard deviation, Wald's Chi$^2$, the corresponding p-value and the confidence interval are displayed for the constant and each variable of the model. If the corresponding option has been activated, the "profile likelihood" intervals are also displayed.

The **equation of the model** is then displayed to make it easier to read or re-use the model.

The table of **standardized coefficients** (also called beta coefficients) are used to compare the relative weights of the variables. The higher the absolute value of a coefficient, the more important the weight of the corresponding variable. When the confidence interval around standardized coefficients has value 0 (this can be easily seen on the chart of normalized coefficients), the weight of a variable in the model is not significant.

The **predictions and residuals** table shows, for each observation, its weight, the value of the qualitative explanatory variable, if there is only one, the observed value of the dependent variable, the model's prediction, the same values divided by the weights, the standardized residuals and a confidence interval.

If only one quantitative variable has been selected, the **probability analysis** table allows to see to which value of the explanatory variable corresponds a given probability of success.

**Example**

A tutorial on how to use the dose effect analysis is available on the Addinsoft website:

http://www.xlstat.com/demo-dose.htm

**References**


Four/Five-parameter parallel lines logistic regression

Use this tool to analyze the effect of a quantitative variable on a response variable using the four/five-parameter logistic model. XLSTAT enables you to take into account some standard data while fitting the model, and to automatically remove outliers.

Description

The four parameter logistic model writes:

\[ y = a + \frac{d-a}{1 + \left( \frac{x}{c} \right)^b} \]  \hspace{1cm} (1.1)

where \( a, b, c, d \) are the parameters of the model, and where \( x \) corresponds to the explanatory variable and \( y \) to the response variable. \( a \) and \( d \) are parameters that respectively represent the lower and upper asymptotes, and \( b \) is the slope parameter. \( c \) is the abscissa of the mid-height point which ordinate is \( (a+d)/2 \). When \( a \) is lower than \( d \), the curve decreases from \( d \) to \( a \), and when \( a \) is greater than \( d \), the curve increases from \( a \) to \( d \).

The five parameter logistic model writes:

\[ y = a + \frac{d-a}{\left[ 1 + \left( \frac{x}{c} \right)^b \right]^e} \]  \hspace{1cm} (1.2)

where the additional parameter \( e \) is the asymmetry factor.

The four parameter parallel lines logistic model writes:

\[ y = a + \frac{d-a}{1 + \left( s_0 \cdot \frac{x}{c_0} + s_1 \cdot \frac{x}{c_1} \right)^b} \]  \hspace{1cm} (2.1)

where \( s_0 \) is 1 if the observation comes from the standard sample, and 0 if not, and where \( s_1 \) is 1 if the observation is from the sample of interest, and 0 if not. This is a constrained model because the observations corresponding to the standard sample influence the optimization of the values of \( a, b, \) and \( d \). From the above writing of the model, one can understand that this model generates two parallel curves, which only difference is the positioning of the curve, the shift being given by \( (c_1-c_0) \). If \( c_1 \) is greater than \( c_0 \), the curve corresponding to the sample of interest is shifted to the right of the curve corresponding to the standard sample, and vice-versa.
The five parameter parallel lines logistic model writes:

\[ y = a + \frac{d - a}{1 + \left( s0 \cdot \frac{x}{c0} + s1 \cdot \frac{x}{c1} \right)^b e} \]  

(2.2)

XLSTAT allows to fit:

- model 1.1 or 1.2 to a standard sample or to the sample of interest,
- model 2.1 or 2.2 to the standard sample and and to the standard sample the same time.

XLSTAT allows to either fit models 1.1 or 1.2 to a given sample (A case), or to fit models 1.1 or 1.2 to the standard (0) sample and then fit models 2.1 or 2.2 to both the standard sample and the sample of interest (B case).

If the Dixon’s test option is activated, XLSTAT tests for each sample if some outliers influence too much the fit of the model. In the A case, a Dixon’s test is performed once the model 1.1 or 1.2 is fitted. If an outlier is detected, it is removed, and the model is fitted again, and so on, until no outlier is detected. In the B case, we first perform a Dixon’s test on the standard sample, then on the sample of interest, and then, the models 2.1 or 2.2 is fitted on the merged samples, without the outliers.

In the B case, and if the sum of the sample sizes is greater than 9, a Fisher’s F test is performed to detect if the a, b, d and e parameters obtained with models 1.1 or 1.2 are not significantly different from those obtained with model 2.1 or 2.2.

**Dialog box**

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

- **OK**: Click this button to start the computations.
- **Cancel**: Click this button to close the dialog box without doing any computation.
- **Help**: Click this button to display the help.
Click this button to reload the default options.

Click this button to delete the data selections.

Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**Y / Dependent variables:**

**Quantitative:** Select the response variable(s) you want to model. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

**X / Explanatory variables:**

**Quantitative:** Select the quantitative explanatory variables to include in the model. If the variable header has been selected, check that the "Variable labels" option has been activated.

**Model:**

- **4PL:** Activate this option to fit the four parameter model.
- **5PL:** Activate this option to fit the five parameter model.

**Range:** Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Activate this option to display the results in a new worksheet of the active workbook.

**Workbook:** Activate this option to display the results in a new workbook.

**Variable labels:** Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

**Observation labels:** Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).
**Subsamples**: Activate this option then select a column (column mode) or a row (row mode) containing the sample identifier(s). The identifiers must be 0 for the standard sample and 1, 2, 3 ... for the other samples that you want to compare with the standard sample. If a header has been selected, check that the "Variable labels" option has been activated.

**Options** tab:

**Initial values**: Activate this option to give XLSTAT a starting point. Select the cells which correspond to the initial values of the parameters. The number of rows selected must be the same as the number of parameters.

**Parameters bounds**: Activate this option to give XLSTAT a possible region for all the parameters of the model selected. You must then select a two-column range, the one on the left being the lower bounds and the one on the right the upper bounds. The number of rows selected must be the same as the number of parameters.

**Parameters labels**: Activate this option if you want to specify the names of the parameters. XLSTAT will display the results using the selected labels instead of using generic labels pr1, pr2, etc. The number of rows selected must be the same as the number of parameters.

**Stop conditions**:

- **Iterations**: Enter the maximum number of iterations for the algorithm. The calculations are stopped when the maximum number of iterations has been exceeded. Default value: 100.

- **Convergence**: Enter the maximum value of the evolution in the Sum of Squares of Errors (SSE) from one iteration to another which, when reached, means that the algorithm is considered to have converged. Default value: 0.00001.

**Dixon’s test**: Activate this option to use the Dixon’s test to remove outliers from the estimation sample.

**Confidence intervals**: Activate this option to enter the size of the confidence interval for the Dixon’s test.

**Missing data** tab:

**Remove observations**: Activate this option to remove the observations with missing data.
**Estimate missing data:** Activate this option to estimate missing data before starting the computations.

- **Mean or mode:** Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

- **Nearest neighbour:** Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.

**Outputs** tab:

**Descriptive statistics:** Activate this option to display descriptive statistics for the variables selected.

**Goodness of fit statistics:** Activate this option to display the table of goodness of fit statistics for the model.

**Model parameters:** Activate this option to display the values of the parameters for the model after fitting.

**Equation of the model:** Activate this option to display the equation of the model once fitted.

**Predictions and residuals:** Activate this option to display the predictions and residuals for all the observations.

**Charts** tab:

**Data and predictions:** Activate this option to display the chart of observations and the curve for the fitted function.

- **Logarithmic scale:** Activate this option to use a logarithmic scale.

**Residuals:** Activate this option to display the residuals as a bar chart.

**Results**

**Summary statistics:** This table displays for the selected variables, the number of observations, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased).

If no group or a single sample was selected, the results are shown for the model and for this sample. If several sub-samples were defined (see sub-samples option in the dialog), the model
is first adjusted to the standard sample, then each sub-sample is compared to the standard sample.

**Fisher's test assessing parallelism between curves:** The Fisher's F test is used to determine if one can consider that the models corresponding the standard sample and the sample of interest are significantly different or not. If the probability corresponding to the F value is lower than the significance level, then one can consider that the difference is significant.

**Goodness of fit coefficients:** This table shows the following statistics:

- The number of observations;
- The number of degrees of freedom (DF);
- The determination coefficient $R^2$;
- The sum of squares of the errors (or residuals) of the model (SSE or SSR respectively);
- The means of the squares of the errors (or residuals) of the model (MSE or MSR);
- The root mean squares of the errors (or residuals) of the model (RMSE or RMSR);

**Model parameters:** This table displays the estimator and the standard error of the estimator for each parameter of the model. It is followed by the equation of the model.

**Predictions and residuals:** This table displays giving for each observation the input data and corresponding prediction and residual. The outliers detected by the Dixon's test, if any, are displayed in bold.

**Charts:** On the first chart are displayed in blue color, the data and the curve corresponding to the standard sample, and in red color, the data and the curve corresponding to the sample of interest. A chart that allows to compare predictions and observed values as well as the bar chart of the residuals are also displayed.

**Example**

A tutorial on how to use the four parameters logistic regression tool is available on the Addinsoft website:

[http://www.xlstat.com/demo-4pl.htm](http://www.xlstat.com/demo-4pl.htm)
References


XLSTAT-PLSPM

XLSTAT-PLSPM is a module of XLSTAT that is dedicated to the component based structural equation modeling in particular with methods such as Partial Least Squares Path Modeling (PLS-PM / PLS-SEM), Generalized Structured Components Analysis (GSCA) or Regularized Generalized Canonical Correlation Analysis (RGCCA). These are innovative methods for representing complex relationships between observed variables and latent variables.

The XLSTAT-PLSPM methods can be used in different fields such as in marketing for the analysis of consumer satisfaction. 3 levels of display are proposed (Classic, Expert, Marketing) in order to adapt to different types of users.

Description

Partial Least Squares Path Modeling (PLS-PM) is a statistical approach for modeling complex multivariable relationships (structural equation models) among observed and latent variables. Since a few years, this approach has been enjoying increasing popularity in several sciences (Esposito Vinzi et al., 2007). Structural Equation Models include a number of statistical methodologies allowing the estimation of a causal theoretical network of relationships linking latent complex concepts, each measured by means of a number of observable indicators.

The first presentation of the finalized PLS approach to path models with latent variables has been published by Wold in 1979 and then the main references on the PLS algorithm are Wold (1982 and 1985).

Herman Wold opposed LISREL (Jöreskog, 1970) "hard modeling" (heavy distribution assumptions, several hundreds of cases necessary) to PLS "soft modeling" (very few distribution assumptions, few cases can suffice). These two approaches to Structural Equation Modeling have been compared in Jöreskog and Wold (1982).

From the standpoint of structural equation modeling, PLS-PM is a component-based approach where the concept of causality is formulated in terms of linear conditional expectation. PLS-PM seeks for optimal linear predictive relationships rather than for causal mechanisms thus privileging a prediction-relevance oriented discovery process to the statistical testing of causal hypotheses. Two very important review papers on PLS approach to Structural Equation Modeling are Chin (1998, more application oriented) and Tenenhaus et al. (2005, more theory oriented).

Furthermore, PLS Path Modeling can be used for analyzing multiple tables and it is directly related to more classical data analysis methods used in this field. In fact, PLS-PM may be also viewed as a very flexible approach to multi-block (or multiple table) analysis by means of both the hierarchical PLS path model and the confirmatory PLS path model (Tenenhaus and Hanafi, 2007). This approach clearly shows how the "data-driven" tradition of multiple table analysis can be somehow merged in the "theory-driven" tradition of structural equation modeling so as
to allow running the analysis of multi-block data in light of current knowledge on conceptual relationships between tables.

Other methods such as Generalized Structured Components Analysis (GSCA) or Regularized Generalized Canonical Correlation Analysis (RGCCA) have been introduced to tackle the weakness of PLS-PM.

The PLS Path Modeling algorithm

A PLS Path model is described by two models: (1) a measurement model relating the manifest variables to their own latent variable and (2) a structural model relating some endogenous latent variables to other latent variables. The measurement model is also called the outer model and the structural model the inner model.

1. Manifest variables standardization

There exist four options for the standardization of the manifest variables depending upon three conditions that eventually hold in the data:

- Condition 1: The scales of the manifest variables are comparable. For instance, in the ECSI example the item values (between 0 and 100) are comparable. On the other hand, for instance, weight in tons and speed in km/h would not be comparable.

- Condition 2: The means of the manifest variables are interpretable. For instance, if the difference between two manifest variables is not interpretable, the location parameters are meaningless.

- Condition 3: The variances of the manifest variables reflect their importance.

If condition 1 does not hold, then the manifest variables have to be standardized (mean 0 and variance 1).

If condition 1 holds, it is useful to get the results based on the raw data. But the calculation of the model parameters depends upon the validity of the other conditions:

- Condition 2 and 3 do not hold: The manifest variables are standardized (mean 0 variance 1) for the parameter estimation phase. Then the manifest variables are rescaled to their original means and variances for the final expression of the weights and loadings.

- Condition 2 holds, but not condition 3: The manifest variables are not centered, but are standardized to unitary variance for the parameter estimation phase. Then the manifest variables are rescaled to their original variances for the final expression of the weights and loadings (to be defined later).
- Conditions 2 and 3 hold: Use the original variables.

Lohmöller (1989) introduced a standardization parameter to select one of these four options:

<table>
<thead>
<tr>
<th>Variable scales are comparable</th>
<th>Means are interpretable</th>
<th>Variance is related to variable importance</th>
<th>Mean</th>
<th>Variance</th>
<th>Rescaling</th>
<th>METRIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>no</td>
<td>no</td>
<td>0</td>
<td>1</td>
<td>no</td>
<td>1</td>
</tr>
<tr>
<td>Yes</td>
<td>no</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>yes</td>
<td>2</td>
</tr>
<tr>
<td>Yes</td>
<td>yes</td>
<td>original</td>
<td>1</td>
<td>yes</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>yes</td>
<td>original</td>
<td>1</td>
<td>yes</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

With METRIC=1 being “Standardized, weights on standardized MV”, METRIC=2 being “Standardized, weights on raw MV”, METRIC=3 being “Reduced, weights on raw MV” and METRIC=4 being “Raw MV”.

2. The measurement model

A latent variable (LV) $\xi$ is an unobservable variable (or construct) indirectly described by a block of observable variables $x_h$ which are called manifest variables (MV) or indicators. There are three ways to relate the manifest variables to their latent variables, respectively called the reflective way, the formative one, and the MIMIC (Multiple effect Indicators for Multiple Causes) way.

2.1. The reflective way

2.1.1. Definition

In this model each manifest variable reflects its latent variable. Each manifest variable is related to its latent variable by a simple regression:

\[
x_h = \pi_{h0} + \pi_h \xi + \epsilon_h,
\]

where $\xi$ has mean $m$ and standard deviation 1. It is a reflective scheme: each manifest variable $x_h$ reflects its latent variable $\xi$. The only hypothesis made on model (1) is called by H. Wold the predictor specification condition:

\[
E(x_h \mid \xi) = \pi_{h0} + \pi_h \xi.
\]

This hypothesis implied that the residual $\epsilon_h$ has a zero mean and is uncorrelated with the latent variable $\xi$.  

1368
2.1.2. Check for unidimensionality

In the reflective way the block of manifest variables is unidimensional in the meaning of factor analysis. On practical data this condition has to be checked. Three main tools are available to check the unidimensionality of a block: use of principal component analysis of each block of manifest variables, Cronbach's $\alpha$ and Dillon-Goldstein's $\rho$.

a) Principal component analysis of a block

A block is essentially unidimensional if the first eigenvalue of the correlation matrix of the block MVs is larger than 1 and the second one smaller than 1, or at least very far from the first one. The first principal component can be built in such a way that it is positively correlated with all (or at least a majority of) the MVs. There is a problem with MV negatively correlated with the first principal component.

b) Cronbach's $\alpha$

Cronbach's $\alpha$ can be used to check unidimensionality of a block of $p$ variables $x_h$ when they are all positively correlated. Cronbach has proposed the following procedure for standardized variables:

\[ \alpha = \frac{\sum \text{cor}(x_h, x_{h'})}{p + \sum \text{cor}(x_h, x_{h'}) \times \frac{p}{p-1}}. \]

The Cronbach's alpha is also defined for original (raw) variables as:

\[ \alpha = \frac{\sum \text{cov}(x_h, x_{h'})}{\text{var} \left( \sum_h x_h \right) \times \frac{p}{p-1}}. \]

A block is considered as unidimensional when the Cronbach's alpha is larger than 0.7.

c) Dillon-Goldstein's $\rho$

The sign of the correlation between each MV $x_h$ and its LV $\xi$ is known by construction of the item and is supposed here to be positive. In equation (1) this hypothesis means that all the loadings $\pi_h$ are positive. A block is unidimensional if all these loadings are large.
The Goldstein-Dillon's $\rho$ is defined by:

\[
\rho = \frac{\left( \sum_{h=1}^{p} \pi_h \right)^2 \text{Var}(\xi)}{\left( \sum_{h=1}^{p} \pi_h \right)^2 \text{Var}(\xi) + \sum_{h=1}^{p} \text{Var}(\varepsilon_h)}.
\]

Let's now suppose that all the MVs $x_h$ and the latent variable $\xi$ are standardized. An approximation of the latent variable $\xi$ is obtained by standardization of the first principal component $t_1$ of the block MVs. Then $\pi_h$ is estimated by $\text{cor}(x_h, t_1)$ and, using equation (1), $\text{Var}(\varepsilon_h)$ is estimated by $1 - \text{cor}^2(x_h, t_1)$. So we get an estimate of the Dillon-Goldstein's $\hat{\rho}$:

\[
\hat{\rho} = \frac{\left[ \sum_{h=1}^{p} \text{cor}(x_h, t_1) \right]^2}{\left[ \sum_{h=1}^{p} \text{cor}(x_h, t_1) \right]^2 + \sum_{h=1}^{p} \left[ 1 - \text{cor}^2(x_h, t_1) \right]}
\]

A block is considered as unidimensional when the Dillon-Goldstein's $\hat{\rho}$ is larger than 0.7. This statistic is considered to be a better indicator of the unidimensionality of a block than the Cronbach's alpha (Chin, 1998, p.320).

PLS Path Modeling is a mixture of *a priori* knowledge and data analysis. In the reflective way, the a priori knowledge concerns the unidimensionality of the block and the signs of the loadings. The data have to fit this model. If they do not, they can be modified by removing some manifest variables that are far from the model. Another solution is to change the model and use the formative way that will now be described.

### 2.2. The formative way

In the formative way, it is supposed that the latent variable $\xi$ is generated by its own manifest variables. The latent variable $\xi$ is a linear function of its manifest variables plus a residual term:

\[
\xi = \sum_{h} \omega_h x_h + \delta.
\]

In the formative model the block of manifest variables can be multidimensional. The predictor specification condition is supposed to hold as:
This hypothesis implies that the residual vector $\delta$ has a zero mean and is uncorrelated with the MVs $x_h$.

2.3. The MIMIC way

The MIMIC way is a mixture of the reflective and formative ways.

The measurement model for a block is the following:

\[ x_h = \pi_{h0} + \pi_h \xi + \varepsilon_h, \quad \text{for } h = 1 \text{ to } p_1 \]

where the latent variable is defined by:

\[ \xi = \sum_{h=p_1+1}^{p} \omega_h x_h + \delta. \]

The $p_1$ first manifest variables follow a reflective way and the $(p - p_1)$ last ones a formative way. The predictor specification hypotheses still hold and lead to the same consequences as before on the residuals.

3. The structural model

The causality model leads to linear equations relating the latent variables between them (the structural or inner model):

\[ \xi_j = \beta_{j0} + \sum_i \beta_{ji} \xi_i + v_j. \]

The predictor specification hypothesis is still applied.

A latent variable, which never appears as a dependent variable, is called an exogenous variable. Otherwise it is called an endogenous variable.

4. The Estimation Algorithm

4.1. Latent variables Estimation

The latent variables $\xi$ are estimated according to the following procedure.
4.1.1. Outer estimate $y_j$ of the standardized latent variable ($\xi_j - m_j$)

The standardized latent variables (mean = 0 and standard deviation = 1) are estimated as linear combinations of their centered manifest variables:

$$ y_j \propto \pm \left[ \sum w_{jh} (x_{jh} - \bar{x}_{jh}) \right], $$

where the symbol "$\propto$" means that the left variable represents the standardized right variable and the "$\pm$" sign shows the sign ambiguity. This ambiguity is solved by choosing the sign making $y_j$ positively correlated to a majority of $x_{jh}$.

The standardized latent variable is finally written as:

$$ y_j = \sum \tilde{w}_{jh} (x_{jh} - \bar{x}_{jh}). $$

The coefficients $w_{jh}$ and $\tilde{w}_{jh}$ are both called the outer weights.

The mean $m_j$ is estimated by:

$$ \hat{m}_j = \sum \tilde{w}_{jh} \bar{x}_{jh}, $$

and the latent variable $\xi_j$ by

$$ \hat{\xi}_j = \sum \tilde{w}_{jh} x_{jh} = y_j + \hat{m}_j. $$

When all manifest variables are observed on the same measurement scale, it is nice to express (Fornell (1992)) latent variables estimates in the original scale as:

$$ \hat{\xi}^*_{ij} = \frac{\sum \tilde{w}_{jh} x_{jh}}{\sum \tilde{w}_{jh}}. $$

Equation (16) is feasible when all outer weights are positive. Finally, most often in real applications, latent variables estimates are required on a 0-100 scale so as to have a reference scale to compare individual scores. From the equation (16), for the i-th observed case, this is easily obtained by the following transformation:

$$ \hat{\xi}_{0-100}^{ij} = 100 \times \left( \frac{\hat{\xi}^*_{ij} - x_{\min}}{x_{\max} - x_{\min}} \right), $$
where $x_{\min}$ and $x_{\max}$ are, respectively, the minimum and the maximum value of the measurement scale common to all manifest variables.

### 4.1.2. Inner estimate $z_j$ of the standardized latent variable ($\xi_j - m_j$)

The inner estimate $z_j$ of the standardized latent variable ($\xi_j - m_j$) is defined by:

\[
(18) \\
z_j \propto \sum_{j': \xi_j \text{ is connected with } \xi_j} e_{jj'} y_{j'}
\]

where the inner weights $e_{jj'}$ are equal to the signs of the correlations between $y_j$ and the $y_{j'}$'s connected with $y_j$. Two latent variables are connected if there exists a link between the two variables: an arrow goes from one variable to the other in the arrow diagram describing the causality model. This choice of inner weights is called the **centroid scheme**.

**Centroid scheme:**

This choice shows a drawback in case the correlation is approximately zero as its sign may change for very small fluctuations. But it does not seem to be a problem in practical applications.

In the original algorithm, the inner estimate is the right term of (18) and there is no standardization. We prefer to standardize because it does not change anything for the final inner estimate of the latent variables and it simplifies the writing of some equations.

Two other schemes for choosing the inner weights exist: the factorial scheme and the path weighting (or structural) scheme. These two new schemes are defined as follows:

**Factorial scheme:**

The inner weights $e_{ji}$ are equal to the correlation between $y_i$ and $y_j$. This is an answer to the drawbacks of the centroid scheme described above.

**Path weighting scheme (structural):**

The latent variables connected to $\xi_j$ are divided into two groups: the predecessors of $\xi_j$, which are latent variables explaining $\xi_j$, and the followers, which are latent variables explained by $\xi_j$.

For a predecessor $\xi_{j'}$ of the latent variable $\xi_j$, the inner weight $e_{jj'}$ is equal to the regression coefficient of $y_j$ in the multiple regression of $y_j$ on all the $y_{j'}$'s related to the predecessors of $\xi_j$. If $\xi_{j'}$ is a successor of $\xi_j$ then the inner weight $e_{jj'}$ is equal to the correlation between $y_j$ and $y_{j'}$. 

1373
These new schemes do not significantly influence the results but are very important for theoretical reasons. In fact, they allow to relate PLS Path modeling to usual multiple table analysis methods.

*The Horst scheme:*

The internal weights $e_{ji}$ are always 1. This is one of the first scheme developed for the PLS Path Modeling.

### 4.2. The PLS algorithm for estimating the weights

#### 4.2.1. Estimation modes for the weights $w_{jh}$

There are three classical ways to estimate the weights $w_{jh}$: Mode A, Mode B and Mode C.

*Mode A:*

In mode A the weight $w_{jh}$ is the regression coefficient of $z_j$ in the simple regression of $x_{jh}$ on the inner estimate $z_i$:

$$\begin{equation}
(19) \quad w_{jh} = \text{cov}(x_{jh}, z_i),
\end{equation}$$

as $z_j$ is standardized.

*Mode B:*

In mode B the vector $w$ of weights $w_j$ is the regression coefficient vector in the multiple regression of $z_j$ on the manifest centered variables $(x_{jh} - \bar{x}_{jh})$ related to the same latent variable $\xi_j$:

$$\begin{equation}
(20) \quad w = (X_j^T X_j)^{-1}X_j^T z_i,
\end{equation}$$

where $X_j$ is the matrix with columns defined by the centered manifest variables $x_{jh} - \bar{x}_{jh}$ related to the $j$-th latent variable $\xi_i$.

Mode A is appropriate for a block with a reflective measurement model and Mode B for a formative one. Mode A is often used for an endogenous latent variable and mode B for an exogenous one. Modes A and B can be used simultaneously when the measurement model is the MIMIC one. Mode A is used for the reflective part of the model and Mode B for the formative part.
In practical situations, mode B is not so easy to use because there is often strong multicollinearity inside each block. When this is the case, PLS regression may be used instead of OLS multiple regression. As a matter of fact, it may be noticed that mode A consists in taking the first component from a PLS regression, while mode B takes all PLS regression components (and thus coincides with OLS multiple regression). Therefore, running a PLS regression and retaining a certain number of significant components may be meant as a new intermediate mode between mode A and mode B.

*Mode C (centroid):*

In mode C the weights are all equal in absolute value and reflect the signs of the correlations between the manifest variables and their latent variables:

\[
(21) \quad w_jh = \text{sign}(\text{cor}(x_{jh}, z_j)).
\]

These weights are then normalized so that the resulting latent variable has unitary variance. Mode C actually refers to a formative way of linking manifest variables to their latent variables and represents a specific case of Mode B whose comprehension is very intuitive to practitioners.

### 4.2.2. Estimating the weights

The starting step of the PLS algorithm consists in beginning with an arbitrary vector of weights \( w_jh \). These weights are then standardized in order to obtain latent variables with unitary variance.

A good choice for the initial weight values is to take \( w_{jh} = \text{sign}(\text{cor}(x_{jh}, \xi_h)) \) or, more simply, \( w_{jh} = \text{sign}(\text{cor}(x_{jh}, \xi_h)) \) for \( h = 1 \) and 0 otherwise or they might be the elements of the first eigenvector from a PCA of each block.

Then the steps for the outer and the inner estimates, depending on the selected mode, are iterated until convergence (guaranteed only for the two-blocks case, but practically always encountered in practice even with more than two blocks).

After the last step, final results are yielded for the inner weights \( \tilde{w}_{jh} \), the standardized latent variable \( y_j = \sum \tilde{w}_{jh} (x_{jh} - \bar{x}_{jh}) \), the estimated mean \( \hat{m}_j = \sum \tilde{w}_{jh} \bar{x}_{jh} \) of the latent variable \( \xi_j \), and the final estimate \( \hat{\xi}_j = \sum \tilde{w}_{jh} x_{jh} = y_j + \hat{m}_j \) of \( \xi_j \). The latter estimate can be rescaled according to transformations (16) and (17).
The latent variable estimates are sensitive to the scaling of the manifest variables in Mode A, but not in mode B. In the latter case, the outer LV estimate is the projection of the inner LV estimate on the space generated by its manifest variables.

### 4.3. Estimation of the structural equations

The structural equations (11) are estimated by individual OLS multiple regressions where the latent variables $\xi_j$ are replaced by their estimates $\hat{\xi}_j$. As usual, the use of OLS multiple regressions may be disturbed by the presence of strong multicollinearity between the estimated latent variables. In such a case, PLS regression may be applied instead.

### 5. Missing Data Treatment

In XLSTAT-PLSPM, there exists a specific treatment for missing data (Lohmöller, 1989):

1. When some cells are missing in the data, means and standard deviations of the manifest variables are computed on all the available data.

2. All the manifest variables are centered.

3. If a unit has missing values on a whole block $j$, the value of the latent variable estimate $y_j$ is missing for this unit.

4. If a unit $i$ has some missing values on a block $j$ (but not all), then the outer estimate $y_{ji}$ is defined by:

   $$ y_{ji} = \sum_{j: x_{jh} \text{ exists}} \hat{w}_{jh} (x_{jhi} - \bar{x}_{jh}) . $$

   That means that each missing data of variable $x_{jh}$ is replaced by the mean $\bar{x}_{jh}$.

5. If a unit $i$ has some missing values on its latent variables, then the inner estimate $z_{ji}$ is defined by:

   $$ z_{ji} = \sum_{k: \hat{z}_k \text{ is connected with } \hat{\xi}_j \text{ and } y_{ji} \text{ exists}} e_{jk} y_{ki} . $$

   That means that each missing data of variable $y_k$ is replaced by its mean 0.
6. The weights \( w_{jh} \) are computed using all the available data on the basis of the following procedures:

- For mode A: The outer weight \( w_{jh} \) is the regression coefficient of \( z_j \) in the regression of \( (x_{jh} - \bar{x}_{jh}) \) on \( z_j \) calculated on the available data.

- For mode B: When there are no missing data, the outer weight vector \( w_j \) is equal to:

  \[
  w_j = (X_j'X_j)^{-1}X_j'z_i.
  \]

  The outer weight vector \( w_j \) is also equal to

  \[
  w_j = [\text{Var}(X_j)]^{-1}\text{Cov}(X_j,z_i),
  \]

  where \( \text{Var}(X_j) \) is the covariance matrix of \( X_j \) and \( \text{Cov}(X_j,z_i) \) the column vector of the covariances between the variables \( x_{jh} \) and \( z_i \).

  When there are missing data, each element of \( \text{Var}(X_j) \) and \( \text{Cov}(X_j,z_i) \) is computed using all the pairwise available data and \( w_j \) is computed using the previous formula.

  This pairwise deletion procedure shows the drawback of possibly computing covariances on different sample sizes and/or different statistical units. However, in the case of few missing values, it seems to be very robust. This justifies why the blindfolding procedure that will be presented in the next section yields very small standard deviations for parameters.

7. The path coefficients are the regression coefficients in the multiple regressions relating some latent variables to some others. When there are some missing values, the procedure described in point 6 (Mode B) is also used to estimate path coefficients.

Nevertheless, missing data can be also treated with other classical procedures, such as mean imputation, listwise deletion, multiple imputation, the NIPALS algorithm (discussed below) and so on so forth.

6. Model Validation

A path model can be validated at three levels: (1) the quality of the measurement model, (2) the quality of the structural model, and (3) each structural regression equation.

6.1. Communality and redundancy

The communality index measures the quality of the measurement model for each block. It is defined, for block \( j \), as:
\begin{equation}
\text{Communality}_j = \frac{1}{p_j} \sum_{h=1}^{p_j} \text{cor}^2(x_{jh}, y_j).
\end{equation}

The average communality is the average of all the $\text{cor}^2(x_{jh}, y_j)$:
\begin{equation}
\overline{\text{Communality}} = \frac{1}{p} \sum_{j=1}^{p} p_j \text{Communality}_j,
\end{equation}
where $p$ is total number of manifest variables in all blocks.

The redundancy index measures the quality of the structural model for each endogenous block. It is defined, for an endogenous block $j$, as:
\begin{equation}
\text{Redundancy}_j = \text{Communality}_j \times R^2\left(y_j, \{y_j's\text{ explaining } y_j}\right).
\end{equation}

The average redundancy for all endogenous blocks can also be computed.

A global criterion of goodness-of-fit (GoF) can be proposed (Amato, Esposito Vinzi and Tenenhaus 2004) as the geometric mean of the average communality and the average $R^2$:
\begin{equation}
\text{GoF} = \sqrt{\text{Communality} \times R^2}.
\end{equation}

As a matter of fact, differently from LISREL, PLS Path Modeling does not optimize any global scalar function so that it naturally lacks of an index that can provide the user with a global validation of the model (as it is instead the case with $\chi^2$ and related measures in LISREL). The GoF represents an operational solution to this problem as it may be meant as an index for validating the PLS model globally, as looking for a compromise between the performances of the measurement and the structural model, respectively.

\subsection*{6.2. The Blindfolding approach: cross-validated communality and redundancy}

The cv-communality (cv stands for cross-validated) index measures the quality of the measurement model for each block. It is a kind of cross-validated R-square between the block MVs and their own latent variable calculated by a blindfolding procedure.

The quality of each structural equation is measured by the cv-redundancy index (i.e. Stone-Geisser’s $Q^2$). It is a kind of cross-validated R-square between the manifest variables of an endogenous latent variable and all the manifest variables associated with the latent variables explaining the endogenous latent variable, using the estimated structural model.
Following Wold (1982, p. 30), the cross-validation test of Stone and Geisser fits soft modeling like hand in glove. In PLS Path Modeling statistics on each block and on each structural regression are available.

The significance levels of the regression coefficients can be computed using the usual Student’s t statistic or using cross-validation methods like jack-knife or bootstrap.

Here is the description of the blindfolding approach proposed by Herman Wold:

1. The data matrix is divided into G groups. The value G = 7 is recommended by Herman Wold. We give in the following table an example on a dataset made by 12 statistical units and 5 variables. The first group is related to letter a, the second one to letter b, and so on.

<table>
<thead>
<tr>
<th>(x_1)</th>
<th>(x_2)</th>
<th>(x_3)</th>
<th>(x_4)</th>
<th>(x_5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>f</td>
<td>d</td>
<td>b</td>
<td>g</td>
</tr>
<tr>
<td>b</td>
<td>g</td>
<td>e</td>
<td>c</td>
<td>a</td>
</tr>
<tr>
<td>c</td>
<td>a</td>
<td>f</td>
<td>d</td>
<td>b</td>
</tr>
<tr>
<td>d</td>
<td>b</td>
<td>g</td>
<td>e</td>
<td></td>
</tr>
<tr>
<td>e</td>
<td>c</td>
<td>a</td>
<td>f</td>
<td>d</td>
</tr>
<tr>
<td>f</td>
<td>d</td>
<td>b</td>
<td>g</td>
<td>e</td>
</tr>
<tr>
<td>g</td>
<td>e</td>
<td>c</td>
<td>a</td>
<td>f</td>
</tr>
<tr>
<td>a</td>
<td>f</td>
<td>d</td>
<td>b</td>
<td>g</td>
</tr>
<tr>
<td>b</td>
<td>g</td>
<td>e</td>
<td>c</td>
<td>a</td>
</tr>
<tr>
<td>c</td>
<td>a</td>
<td>f</td>
<td>d</td>
<td>b</td>
</tr>
<tr>
<td>d</td>
<td>b</td>
<td>g</td>
<td>e</td>
<td>c</td>
</tr>
<tr>
<td>e</td>
<td>c</td>
<td>a</td>
<td>f</td>
<td>d</td>
</tr>
</tbody>
</table>

2. Each group of cells is removed at its turn from the data. So a group of cells appears to be missing (for example all cells with letter a).

3. A PLS model is run G times by excluding each time one of the groups.

4. One way to evaluate the quality of the model consists in measuring its capacity to predict manifest variables using other latent variables. Two indices are used: communality and redundancy.

5. In the communality option, we get prediction for the values of the centered manifest variables not included in the analysis, using the latent variable estimate, by the following formula:

\[
\text{Pred}(x_{jh(i)} - \bar{x}_{jh}) = \hat{\pi}_{jh(i)} y_{ji},
\]

where \(\hat{\pi}_{jh(i)}\) and \(y_{ji}\) are computed on data where the i-th value of variable \(x_{jh}\) is missing.

The following terms are computed:

- Sum of squares of observations for one MV: \(SSO_{jh} = \sum_i (x_{jh(i)} - \bar{x}_{jh})^2\).

- Sum of squared prediction errors for one MV: \(SSE_{jh} = \sum_i (x_{jh(i)} - \bar{x}_{jh} - \hat{\pi}_{jh(i)} y_{ji})^2\).
- Sum of squares of observations for Block $j$: \[ \text{SSO}_j = \sum_h \text{SSO}_{jh}. \]

- Sum of squared prediction errors for Block $j$: \[ \text{SSE}_j = \sum_h \text{SSE}_{jh}. \]

- CV-Communality measure for Block $j$: \[ H_j^2 = 1 - \frac{\text{SSE}_j}{\text{SSO}_j}. \]

The index $H_j^2$ is the cross-validated communality index. The mean of the cv-communality indices can be used to measure the global quality of the measurement model if they are positive for all blocks.

6. In the redundancy option, we get a prediction for the values of the centred manifest variables not used in the analysis by using the following formula:

\[
\text{Pred}(x_{jh(i)} - \overline{x}_{jh}) = \hat{\pi}_{jh(i)} \text{Pred}(y_{j(i)}),
\]

where $\hat{\pi}_{jh(i)}$ is the same as in the previous paragraph and $\text{Pred}(y_{j(i)})$ is the prediction for the $i$-th observation of the endogenous latent variable $y_j$ using the regression model computed on data where the $i$-th value of variable $x_{jh}$ is missing.

The following terms are also computed:

- Sum of squared prediction errors for one MV:

\[
\text{SSE}_{jh} = \sum_i (x_{jh} - \overline{x}_{jh} - \hat{\pi}_{jh(i)} \text{Pred}(y_{j(i)}))^2
\]

- Sum of squared prediction errors for block $j$:

\[
\text{SSE}_j = \sum_h \text{SSE}_{jh}
\]

- CV-Redundancy measure for an endogenous block $j$:

\[
F_j^2 = 1 - \frac{\text{SSE}_j}{\text{SSO}_j}
\]

The index $F_j^2$ is the cross-validated redundancy index. The mean of the various cv-redundancy indices related to the endogenous blocks can be used to measure the global quality of the structural model if they are positive for all endogenous blocks.

6.3. Resampling: Jackknife and Bootstrap
The significance of PLS-PM parameters, coherently with the distribution-free nature of the estimation method, is assessed by means of non-parametric procedures. As a matter of fact, besides the classical blindfolding procedure, Jackknife and Bootstrap resampling options are available.

### 6.3.1. Jackknife

The Jackknife procedure builds resamples by deleting a certain number of units from the original sample (with size N). The default option consists in deleting 1 unit at a time so that each Jackknife sub-sample is made of N-1 units. Increasing the number of deleted units leads to a potential loss in robustness of the t-statistic because of a smaller number of sub-samples. The complete statistical procedure is described in Chin (1998, p.318-320).

### 6.3.2. Bootstrap

The Bootstrap samples, instead, are built by resampling with replacement from the original sample. The procedure produces samples consisting of the same number of units as in the original sample. The number of resamples has to be specified. The default is 100 but a higher number (such as 200) may lead to more reasonable standard error estimates.

We must take into account that, in PLS-PM, latent variables are defined up to the sign. It means that $y_j = \sum \tilde{w}_{jh} (x_{jh} - \bar{x}_{jh})$ and -$y_j$ are both equivalent solutions. In order to remove this indeterminacy, Wold (1985) suggests retaining the solution where the correlations between the manifest variables $x_{jh}$ and the latent variable $y_j$ show a majority of positive signs. Referring to the signs of the elements in the first eigenvector obtained on the original sample is also a way of controlling the sign in the different bootstrap re-samples.

### GSCA (Generalized Structured Component Analysis)

This method introduced by Hwang and Takane (2011), allows to optimize a global function using an algorithm called Alternating Least Square algorithm (ALS).

GSCA lies in the tradition of component analysis. It substitutes components for factors as in PLS. Unlike PLS, however, GSCA offers a global least squares optimization criterion, which is consistently minimized to obtain the estimates of model parameters. GSCA is thus equipped with an overall measure of model fit while fully maintaining all the advantages of PLS (e.g., less restricted distributional assumptions, no improper solutions, and unique component score estimates). In addition, GSCA handles more diverse path analyses, compared to PLS.

Let $Z$ denote an $N$ by $J$ matrix of observed variables. Assume that $Z$ is columnwise centered and scaled to unit variance. Then, the model for GSCA may be expressed as

$$ZV = ZWA + E,$$
\[
P = GA + E, \quad (1)
\]

where \(P = ZV\), and \(G = ZW\). In (1), \(P\) is an \(N\) by \(T\) matrix of all endogenous observed and composite variables, \(G\) is an \(N\) by \(D\) matrix of all exogenous observed and composite variables, \(V\) is a \(J\) by \(T\) matrix of component weights associated with the endogenous variables, \(W\) is a \(J\) by \(D\) matrix of component weights for the exogenous variables, \(A\) is a \(D\) by \(T\) supermatrix consisting of a matrix of component loadings relating components to their observed variables, denoted by \(C\), in addition to a matrix of path coefficients between components, denoted by \(B\), that is, \(A = [C, B]\), and \(E\) is a matrix of residuals.

We estimate the unknown parameters \(V, W, \) and \(A\) in such a way that the sum of squares of the residuals, \(E = ZV - ZWA = P - GA\), is as small as possible. This amounts to minimizing

\[
f = SS(ZV - ZWA)
= SS(P - GA), \quad (2)
\]

with respect to \(V, W, \) and \(A\), where \(SS(X) = \text{trace}(X'X)\). The components in \(P\) and/or \(G\) are subject to normalization for identification purposes.

We cannot solve (2) in an analytic way since \(V, W, \) and \(A\) can comprise zero or any fixed elements. Instead, we develop an alternating least squares (ALS) algorithm (de Leeuw, Young, & Takane, 1976) to minimize (2). In general, ALS can be viewed as a special type of the FP algorithm where the fixed point is a stationary (accumulation) point of a function to be optimized.

The proposed ALS algorithm consists of two steps: In the first step, \(A\) is updated for fixed \(V\) and \(W\). In the second step, \(V\) and \(W\) are updated for fixed \(A\). (Hwang and Takane, 2004)

**RGCCA (Regularized Generalized Canonical Correlation Analysis)**

This method introduced by Tenenhaus et al. (2005), allows to optimize a global function using an algorithm very similar to the PLSPM algorithm.

Unlike the PLS approach, the results of the RGCCA are correlations between latent variables and between manifest variables and their associated latent variables (there is no regression at the end of the algorithm).

The RGCCA is based on a simple iterative algorithm similar to that of the PLS approach which is as follows:

1 - Initialization of the outer weights in the same way as in the PLSPM algorithm.

2 - Standardization of the outer weights using the tau parameter:
3 - Computation of the internal components of each latent variable depending on the scheme used (the inner schemes are the same as in PLSPM).

\[ z_j^s = \sum_{k<j} c_{jk} e_{jk} X_k w_{k}^{s+1} + \sum_{k>j} c_{jk} e_{jk} X_k w_{k}^{s} \]

With \( e_{jk} \) being the inner weight and \( c_{jk} = 1 \) if the latent variables \( j \) and \( k \) are related.

4 - Outer weights are updated:

\[ w_j^{s+1} = \left[ (z_j^s)^T X_j \left[ \tau_j I + \left( 1 - \tau_j \right) \frac{1}{n} X_j^T X_j \right]^{-1} X_j^T w_j \right]^{-\frac{1}{2}} \left[ \tau_j I + \left( 1 - \tau_j \right) \frac{1}{n} X_j^T X_j \right]^{-1} X_j^T w_j \]

We repeat steps 3 and 4 until convergence of the algorithm.

Once the algorithm has converged, we obtain results that optimize specific functions depending on the choice of the tau parameter.

Tau is a parameter that has to be set for each latent variable. It enables you to adjust the “mode” associated to the latent variable. If \( \tau = 0 \), then we will be in the case of mode B and the results of PLSPM and RGCCA are similar. When \( \tau = 1 \), we find ourselves in the new mode A (as stated by M. Tenenhaus). This mode is close to the mode A of PLSPM while optimizing a given function. When \( \tau \) varies between 0 and 1, the latent variable mode stands in between mode A and mode B. For more details on RGCCA see Tenenhaus et al. (2011).

In the framework of RGCCA, XLSTAT-PLS-PMM allows to use the Ridge RGCCA mode. This mode search for the optimal tau parameter using the Schäfer and Stimmer (2005) formula reproduced in Tenenhaus et al. (2011).

The NIPALS algorithm

The roots of the PLS algorithm are in the NILES (Non linear Iterative LEast Squares estimation), which later became NIPALS (Non linear Iterative PArtial LEast Squares), algorithm for Principal Component Analysis (Wold, 1966). We now remind the original algorithm of H. Wold and show how it can be included in the PLS-PM framework. The interests of the NIPALS algorithm are double as it shows: how PLS handles missing data and how to extend the PLS approach to more than one dimension.
The original NIPALS algorithm is used to run a PCA in presence of missing data. This original algorithm can be slightly modified to go into the PLS framework by standardizing the principal components. Once this is done, the final step of the NIPALS algorithm is exactly the Mode A of the PLS approach when only one block of data is available. This means that PLS-PM can actually yield the first-order results of a PCA whenever it is applied to a block of reflective manifest variables.

The other dimensions are obtained by working on the residuals of $X$ on the previous standardized principal components.

The PLS approach for two sets of variables

PLS Path Modeling can be also used so as to find the main data analysis methods to relate two sets of variables. Table 1 shows the complete equivalence between PLS Path Modeling of two data tables and four classical multivariate analysis methods. In this table, the use of the deflation operation for the research of higher dimension components is mentioned.

Table 1: Equivalence between the PLS algorithm applied to two blocks of variables $X_1$ and $X_2$ and various method

<table>
<thead>
<tr>
<th></th>
<th>Canonical Correlation Analysis</th>
<th>Inter-Battery Factor Analysis</th>
<th>PLS Regression of $X_2$ on $X_1$</th>
<th>Redundancy Analysis of $X_2$ with respect to $X_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computation mode for $X_1$</td>
<td>B (deflation)</td>
<td>A (deflation)</td>
<td>A (deflation)</td>
<td>B (deflation)</td>
</tr>
<tr>
<td>Computation mode for $X_2$</td>
<td>B (deflation)</td>
<td>A (deflation)</td>
<td>A (no deflation)</td>
<td>A (no deflation)</td>
</tr>
</tbody>
</table>

The analytical demonstration of the above mentioned results can be found in Tenenhaus et al., 2005.

The PLS approach for J sets of variables

The various options of PLS Path Modeling (Modes A or B for outer estimation; centroid, factorial or path weighting schemes for inner estimation) allow to find also many methods for multiple tables analysis: Generalized Canonical Analysis (the Horst's one (1961) and the Carroll's one (1968)), Multiple Factor Analysis (Escofier & Pagès, 1994), Lohmöller's split principal component analysis (1989), Horst's maximum variance algorithm (1965).
The links between PLS and these methods have been studied on practical examples in Guinot, Latreille and Tenenhaus (2001) and in Pagès and Tenenhaus (2001).

Let us consider a situation where J blocks of variables $X_1, \ldots, X_J$ are observed on the same set of statistical units. For estimating these latent variables $\xi_i$, Wold (1982) has proposed the hierarchical model defined as follows:

- A new block $X$ is constructed by merging the J blocks $X_1, \ldots, X_J$ into a super block.
- The super block $X$ is summarized by one latent variable $\xi$.
- A path model connects each exogenous LV $\xi_i$ to the endogenous LV $\xi$.

An arrow scheme describing a hierarchical model for three blocks of variables is shown in Figure 1.

![Figure 1: A hierarchical model for a PLS analysis of J blocks of variables.](image)

Table 2 summarizes the links between Hierarchical PLS-PM and several multiple table analysis organized with respect to the choice of the outer estimation mode (A or B) and of the inner estimation scheme (Centroid, Factorial or Path Weighting).

Table 2: PLS Path modeling and Multiple Table Analysis
In the methods described in Table 2, the higher dimension components are obtained by re-running the PLS model after deflation of the $X$-block.

It is also possible to obtain higher dimension orthogonal components on some $X_j$-blocks (or on all of them). The hierarchical PLS model is re-run on the selected deflated $X_j$-blocks.

The orthogonality control for higher dimension components is a tremendous advantage of the PLS approach (see Tenenhaus (2004) for more details and an example of application).

Finally, PLS Path Modeling may be meant as a general framework for the analysis of multiple tables. It is demonstrated that this approach recovers usual data analysis methods in this context but it also allows for new methods to be developed when choosing different mixtures of estimation modes and schemes in the two steps of the algorithm (internal and external estimation of the latent variables) as well as different orthogonality constraints. Therefore, we can state that PLS Path Modeling provides a very flexible environment for the study of a multi-block structure of observed variables by means of structural relationships between latent variables. Such a general and flexible framework also enriches the data analysis methods with non-parametric validation procedures (such as bootstrap, jackknife and blindfolding) for the estimated parameters and fit indices for the different blocks that are more classical in a modeling approach than in data analysis.

**Multigroup comparison tests in PLS path modeling**

Two tests in order to compare parameters between groups are included in XLSTAT-PLSPM:

- An adapted t test based on bootstrap standard errors.

- A permutation test.

**The multigroup t test:**

Wynne Chin was the first to use this test to compare path coefficients. This test uses the estimates obtained from the bootstrap sampling in a parametric sense via t-tests. You make a parametric assumption and take the standard errors for the structural paths provided by the
bootstrap samples, and then calculate the t-test for the difference in path coefficients between groups.

\[ t = \frac{\left| \beta_{ij}^{G_1} - \beta_{ij}^{G_2} \right|}{\sqrt{\frac{(n_1-1)^2}{n_1 + n_2 - 2} \cdot SE_{G_1}^2 + \frac{(n_2-1)^2}{n_1 + n_2 - 2} \cdot SE_{G_2}^2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \]

where \( n_1 \) and \( n_2 \) are the sizes of the groups, and \( SE^2_{G_i} \) is the variance of coefficient \( \beta_{ij} \) obtained using the bootstrap sampling. This would follow a Student distribution with \( n_1 + n_2 - 2 \) degrees of freedom. This approach works reasonably well if the two samples are not far from normality and if the two variances are not too different.

**The permutation tests:**

Permutation tests offer a nonparametric alternative to t tests that fits well to PLS path modeling. They have been used together with PLS Path Modeling in Chin (2008) and Jakobowicz (2007). The principle is simple:

- Select a statistic \( S \). In the case of PLS Path Modeling, we take the absolute value of the difference on a parameter between two groups of observations. 
- Compute the value of this statistic on the two original samples associated to the groups \( \rightarrow S_{obs} \). 
- Randomly permute the elements of the two samples and compute the \( S \) statistic \( \rightarrow S_{perm} \). Repeat this step \( N_{perm} \) times (with \( N_{perm} \) very large). 
- The p-value is obtained with the following formula:

\[ p-value = \frac{1}{N_{perm}} + 1 \sum_{i=1}^{N_{perm}} I(S_{obs} < S_{perm_i}) \]

Function \( I(.) \) being 1 when \( S_{obs} < S_{perm_i} \) and 0 otherwise.

**Segmentation using the REBUS algorithm (Esposito Vinzi et al., 2008)**

When heterogeneity at the units’ level is present, it can be useful to obtain clusters of units which behave similarly on a defined model. The REBUS algorithm (Response Based procedure for detecting unit segments in PLS path modelling) offers a way to find clusters using a defined PLS model and residuals obtained using PLS Path Modeling. It is based on a simple algorithm:

1- Apply PLS Path Modeling on the entire dataset.
2- Compute residuals of the manifest and latent variables and apply an agglomerative hierarchical clustering on the residuals.

3- Apply PLS Path Modeling on each class of units using the same PLS model. You obtain as many sub-models as classes.

4- Compute residuals of the manifest and latent variables for each sub-model and compute the CM index for each unit and each class.

5- Allocate the units to the closest model (in term of CM).

6- Update the composition of the classes.

7- Repeat 3 to 6 until stable classes are obtained.

The CM index for unit i and class k is:

\[
CM_{ik} = \frac{\sum_{j=1}^{p} \sum_{h=1}^{m_k} \left[ \frac{e_{ijh}/\text{Com}(\xi_{ijh},x_{ij})}{\sum_{j=1}^{p} \sum_{h=1}^{m_k} \frac{e_{ijh}^2/\text{Com}(\xi_{ijh},x_{ij})}{\sum_{j=1}^{p} \sum_{h=1}^{m_k} e_{ijh}^2/\text{Com}(\xi_{ijh},x_{ij})}} \right]}{\left( N - m_k - 1 \right)}
\]

Where Com() is the communality index, \( e_{ijh} \) is the measurement model residual for the i\(^{th} \) unit in the k\(^{th} \) latent class, corresponding to the h\(^{th} \) manifest variable in the j\(^{th} \) block, i.e. the communality residual; \( f_{ijh}^{*} \) is the structural model residual for the i\(^{th} \) unit in the k\(^{th} \) latent class, corresponding to the j\(^{th} \) endogenous block; \( N \) is the number of units; \( m_k \) is the number of dimensions. Since all blocks are supposed to be reflective, there will always be one dimension (latent variable) per block.

**XLSTAT-PLSPM offers some options to apply the REBUS method:**

- Number of classes: The user can choose between automatic where the number of classes is set during the cluster analysis and manual where the number of classes has to be entered manually.
- Convergence: The algorithm converges when a fixed percentage of the units are staying in the same class from one iteration to another (the default value is 95%).

The REBUS method can only be applied if all blocks are reflective (mode A) and if no groups are selected.

The global quality index of the model allows to judge the quality of the segmentation. This index is equivalent to the GoF when a single class is used. When several classes are obtained, it is obtained as follows:

\[
GQI = \sqrt{\frac{1}{N} \sum_{k=1}^{K} \frac{n_k}{N} \left[ \frac{1}{P_q} \sum_{q=1}^{P_q} \sum_{p=1}^{P_q} \left( 1 - \frac{\sum_{i=1}^{n_q} e_{ipq}^2}{\sum_{i=1}^{n_q} \left( x_{ipq} - \bar{x}_{p,qk} \right)^2} \right) \right]}
\]

\[
\times \sqrt{\frac{1}{J} \sum_{j=1}^{J} \left( 1 - \frac{\sum_{i=1}^{n_k} f_{jk}^2}{\sum_{i=1}^{n_k} \left( \bar{x}_{ipq} - \bar{x}_{p,qk} \right)^2} \right) \]

with \( n_k \) being the size of class \( k \), \( P_q \) being the number of manifest variable associated to the latent variable \( q \), \( e^2 \) is the residual obtained from the manifest variables and \( f^2 \) is the residual obtained from the scores of the latent variables.

In addition to the GQI, XLSTAT-PLSPM displays the overall improvement of the GQI and the decomposition of the GQI based on the outer model and on the inner model.

**The Marketing display for the analysis of consumer satisfaction**

In addition to the Classic and Expert displays, XLSTAT-PLSPM Proposes a Marketing display that provides the user with an interface that is specially designed for marketing and satisfaction analysis.

This display sets default values to numerous parameters and proposes a simplified way to deal with variables: the scales of manifest and latent variables are selected from the beginning (see the **Fitting the Model** paragraph).

By default, the PLS-Path Modeling method, the path weighting (structural) scheme as well as OLS regressions are used. Initial weights are obtained with the values of the first principal component analysis eigenvector and one single dimension is displayed.

The segmentation option allows to carry out REBUS analysis to extract classes of homogeneous observations and multigroup tests can be defined in the options tab.

This display also proposes charts and simulation tables that helps evaluating the impact of changing variables (manifest and latent) on a target latent variable.
Model loading from the model library

XLSTAT-PLSPM lets you easily load existing structural models. Use the load model icon and choose the model to load among inside the model library.

The loaded model replaces the model that is already open but keeps your data and your results. However, the manifest and latent variables association operation should be done in addition.

Below is a list of available models in the library (.ppmxmod files):

- ECSI (European Customer Satisfaction Index)
- ACSI (American Customer Satisfaction Index)
- SCSB (Swedish Customer Satisfaction Barometer)
- Norwegian Customer Satisfaction Barometer (NCSB),
- Swiss Index of Customer satisfaction (SWICS)
- Korean Customer Satisfaction Index (KCSI)
- Malaysian Customer Satisfaction Index(MCSI)
- Simplified ECSI (excludes the complaints variable)

Projects

XLSTAT-PLSPM projects are special Excel workbook templates. When you create a new project, its default name starts with PLSPMBook. You can then save it to the name you want, but make sure you use the “Save” or “Save as” command of the XLSTAT-PLSPM toolbar to save it in the folder dedicated to the PLSPM projects using the *.ppm extension with Excel 2003 and prior, and *.ppmx with Excel 2007 and later.

A raw XLSTAT-PLSPM project contains two sheets that cannot be removed:

- D1: This sheet is empty and you need to add all the input data that you want to use into that worksheet.

- PLSPMGraph: This sheet is blank and is used to design the model. When you select this sheet, the "Path modeling" toolbar is displayed. It is made invisible when you leave that sheet.

Once a model has been designed, you can run the optimization. Results sheets are then added after the PLSPMGraph sheet.

It is possible to record a model before adding new variables and to reload it later (see the Toolbars section for additional information).
Each time you create a new project, XLSTAT asks you what display you wish to use. You may choose among 3 options depending on your objectives and your expertise.

It is also possible to load saved models (see the toolbars paragraph for more details).

**Options**

To display the options dialog box, click the button of the “XLSTAT-PLSPM” toolbar. Use this dialog box to define the general options of the XLSTAT-PLSPM module.

**General tab:**

**Display:** Because PLS Path Modeling and XLSTAT-PLSPM are complex, three display styles are available. The classical one including all commonly used options and the expert one lets you use more advanced features. Finally, the Marketing display lets you obtain a configuration adapted to marketing and satisfaction analysis.

**Path for the XLSTAT-PLSPM projects:** This path can be modified if and only if you have administrator rights on the machine. You can then modify the folder where the user’s files are saved by clicking the [...] button that will display a box where you can select the appropriate folder. The folder must be accessible for reading and writing to all types of users.

**Format tab:**

Use these options to set the format of the various objects that are displayed on the PLSPMGraph sheet:

- **Latent variables:** You can define the color and the size of the border line of the ellipses that represent the latent variables, as well as the color of the background, and the color and the size of the font.

- **Manifest variables:** You can define the color and the size of the border line of the rectangles that represent the manifest variables, as well as the color of the background, and the color and the size of the font.

- **Arrows (MV-LV):** You can define the color and the size of the arrows between the manifest and the latent variables.

- **Arrows (LV-LV):** You can define the color and the size of the arrows between two latent variables.
Note 1: So that changes are taken into account once you click the OK button, you need to optimize the display by clicking on the button.

Note 2: these options do not prevent you from changing the format of one or more objects on the PLSPMGraph sheet. Using the drawing toolbar of Excel, you can change the fill, the borders of the objects.

**Toolbars**

XLSTAT-PLSPM has two toolbars, “XLSTAT-PLSPM” and “Path modeling”.

The “XLSTAT-PLSPM” toolbar can be displayed by clicking the XLSTAT-PLSPM icon in the XLSTAT toolbar.

Click this icon to open a new PLSPM project (see Projects for more details).

Click this icon to open an existing PLSPM project.

Click this icon to save the current PLSPM project. This icon is only active if changes have been made in the project.

Click this icon to save the project in a new folder or under a new name.

Click this icon to display the XLSTAT-PLSPM options dialog box.

Click this icon if you want to continue using XLSTAT but not XLSTAT-PLSPM. This allows to free some memory space.

**Excel 2003 and lower:**

The second toolbar, “Path modeling” is only visible when you are on the PLSPMGraph sheet of a PLSPM project.

Click this icon to add latent variables. If you double click this icon, you can add several latent variables in a row without having to click this button each time.
Click this icon to add an arrow between two latent variables. If you double click this icon, you can add several arrows in a row without having to click this button each time. When adding an arrow, select first the latent variable that will be at the beginning of the arrow by clicking on it, then leave the mouse button down, and drag the cursor till the latent variable that will be at the end of the arrow.

Click this icon to hide the manifest variables. If a latent variable is selected when you click this icon, it will only hide the corresponding manifest variables.

Click this icon to display the manifest variables. If a latent variable is selected when you click this icon, it will only show the corresponding manifest variables.

Click this icon to optimize the display.

Click this icon to define groups. Once groups are defined, a list with the group names is displayed on the PLSPMGraph sheet, and the icon becomes ; click this icon to remove the groups definition.

Click this icon to archive the current model in the project.

Click this icon to load a previously saved model.

Click this icon to delete one or more previously saved models.

Unprotected/Protected(1)/Protected(2): The first option allows the user to modify the model and the position of the objects. The second option allows the user to modify the position of the objects. The third option does not allow the user to move the objects or to delete them.

Click this icon to completely remove all the objects on the PLSPMGraph sheet.

Click this icon to display the results of the model, if the latter has already been fitted. If the results are already displayed, the following icon is displayed ; click it to hide the results.

Click this icon to display a dialog box that allows you to choose which results should be displayed or not.

Click this icon to start the optimization of the model, and then display the results on both the PLSPMGraph sheet, and on the results sheet.

Excel 2007 and later:

The second toolbar, “Path modeling” is included in the PLSPMGraph sheet of a PLSPM project.
Click this icon to add latent variables. If you double click this icon, you can add several latent variables in a row without having to click this button each time.

Click this icon to define the manifest variables, after selecting the latent variable on the diagram. You can also use the keyboard shortcut Ctrl+M.

Click this icon to hide the manifest variables. If a latent variable is selected when you click this icon, it will only hide the corresponding manifest variables.

Click this icon to display the manifest variables. If a latent variable is selected when you click this icon, it will only show the corresponding manifest variables.

Click this icon to optimize the display.

Click this icon to define groups. Once groups are defined, a list with the group names is displayed on the PLSPMGraph sheet, and the icon becomes 

Click this icon to remove the groups definition.

Click this icon to archive the current model in the project.

Click this icon to load a previously saved model or to load a model from a model library (see the description of PLS Path modeling).

Click this icon to delete one or more previously saved models.

Unprotected/Protected(1)/Protected(2): The first option allows the user to modify the model and the position of the objects. The second option allows the user to modify the position of the objects. The third option does not allow the user to move the objects or to delete them.

To add a link between two latent variables: When adding a link, select first the latent variable that will be at the beginning of the arrow by clicking on it, then select the latent variable that will be at the end of the arrow. Click this icon when both variables are selected.
You can also use the keyboard shortcut Ctrl+L. Use the keyboard shortcut Ctrl+R to reverse the direction of the link.

1. Click this icon to transform all the links into double direction links. This can also be obtained with the keyboard shortcut Ctrl+D.

2. Click this icon to change the position of the manifest variables. You can also use the keyboard shortcut Ctrl+O.

3. Click this icon to rename a latent variable.

4. Click this icon to completely remove all the objects on the PLSPMGraph sheet.

5. Click this icon to display the results of the model, if the latter has already been fitted. If the results are already displayed, the following icon is displayed; click it to hide the results.

6. Click this icon to display a dialog box that allows you to choose which results should be displayed or not.

7. Click this icon to start the optimization of the model, and then display the results on both the PLSPMGraph sheet, and on the results sheet.

**Adding manifest variables**

Once one or more latent variables have been added on the PLSPMGraph document using the appropriate icon of the "Path modeling" toolbar, you can define the manifest variables that correspond to these variables (with Excel 2003 and lower, make a double click on the latent variable, with Excel 2007 and later, click the Manifest variables icon). A latent variable must have manifest variables, even if it is a superblock variable (a variable that is not directly related to manifest variables but to latent variables with arrows going from the latent variables to the superblock variable – the superblock variable inherits the manifest variables of the constitutive latent variables).
For a superblock, you need to add all the manifest variables of the parent latent variables. This is made easy with the XLSTAT interface.

To add manifest variables, you can

- Double-click the latent variable;
- Click the right button of the mouse, and select “Add manifest variables”.

These actions lead to the display of a dialog box which options are:

**General** tab:

**Name of the latent variable**: Enter the name of the latent variable.

**Manifest variables**: Select on the D1 sheet the data that correspond to the manifest variables. The input variables can be either quantitative or qualitative.

- **Quantitative**: Activate this option if you want to use quantitative variables and then select these variables.
- **Qualitative**: Activate this option if you want to use qualitative variables and then select these variables.

**Variable labels**: Activate this option if the first row of the data selections includes a header.

**Position**: Select the position where the manifest variables should be positioned relatively to the latent variable.

**Mode**: Select the mode option that determines how the latent variable is constructed from the manifest variables. The available options are “Mode A” (reflective way, arrows go from the latent variable to the manifest variables), “Mode B” (formative way, arrows go from the manifest variables to the latent variable), “Centroid”, “PCA”, “PLS”, and mode ”MIMIC” (a mixture of Mode A and Mode B). If Mode MIMIC is selected, you need to select a column with one row per manifest variable (and a header if the “Variable labels” option is checked), with As
for the variables with Mode A, and Bs for the variables with mode B. See the description section for more information on the modes. The "Automatic" mode is only available for superblocks. It allows to make that the mode for each manifest variable corresponds to its mode in the latent variable that is used to create the superblock. “Centroid”, “PCA”, “PLS”, and “Automatic” modes are only available in the expert display. The RGCCA mode allows to enter the value of the tau parameter and the Ridge RGCCA mode automatically optimize the value of the tau parameter. These two modes can only be applied with the RGCCA method (see the description section).

Deflation (expert display): Select the deflation mode. The deflation is used when computing the model on the second and above dimensions of the model.

- **No deflation**: Whatever the dimension, the scores of the latent variable remain constant.

- **External**: For the successive dimensions, the residuals are computed from the outer model.

Dimension (expert display): You can choose the number of dimensions to be studied.

Invert sign: Activate this option if you want to reverse the sign of the latent variable. This option is useful if you notice that the influence of a latent variable is the opposite of what it should be.

Supperblock (expert display): You can activate this option only if latent variables have already been created, and if manifest variables were selected for the latter. The list displays the latent variables for which manifest variables have already been defined. The superblock tab appears. You can then select the latent variables that are used to build the superblock variable.

Interaction (expert display): You can activate this option only if latent variables have already been created. An interaction latent variable is the product of two latent variables that have the same successor. The interaction variable will have the same successor as the two variables that were used to create it. The interaction tab appears.

Superblock tab:

The list of all latent variables is displayed. Select latent variables to be included in the superblock.

Interaction tab:

Generating latent variable: Select two of the latent variables explaining the latent variable to which the interaction variable is connected.
Treatment of the manifest variables: Select which transformation of the manifest variable prior to the product you wish to apply. Three options are available: raw manifest variables, standardized manifest variables and mean centered manifest variables.

Options (PLS) tab (expert display):

Options in the structural model (PLS):

Stop conditions:

- **Automatic**: Activate this option so that XLSTAT automatically determines the number of components to keep.

- **Max components**: Activate this option to set the maximum number of components to take into account in the model.

Options for PLS regression in the measurement model (only active, if the “PLS” mode is selected):

Stop conditions:

- **Automatic**: Activate this option so that XLSTAT automatically determines the number of components to keep.

- **Max components**: Activate this option to set the maximum number of components to take into account in the model.

Defining groups

If a qualitative variable is available and if you believe that the model could be different for the various categories of that variable, you may use it to define groups.

To define groups, go to the “PLSPMGraph” sheet, then click the appropriate icon. This displays a “Groups” dialog box, which entries are:

**Groups**: Select on the D1 sheet the data that correspond to the qualitative variable that indicates to which group each observation belongs.

**Column label**: Activate this option if the first row of the selection corresponds to a header.

**Sort alphabetically**: Activate this option if you want that XLSTAT sorts alphabetically the names of the groups (the categories of the selected qualitative variable). If this option is not activated, the categories are listed in their order or appearance.
Once you click **OK**, a list is added at the top right corner of the PLSPMGraph sheet. Once the model has been computed, you can use this list to display the results of the group you want on the PLSPMGraph sheet. The results of the model that corresponds to each group are also displayed on distinct sheets.

Note: if you want to remove the group information, click the appropriate button of the “Path modeling” toolbar.

Once groups are defined, multigroup tests can be performed.

**Fitting the model**

Once you have designed the model on the PLSPMGraph sheet, and once the manifest variables have been defined for each latent variable, you can click the appropriate icon of the “Path modeling” toolbar to display the “Run” dialog box that lets you define additional options before fitting the model.

**General** tab:

**Treatment of the manifest variables (classic and expert displays only):** Choose if and how the manifest variables should be transformed.

- **Standardized, weights on standardized MV:** The manifest variables are standardized before fitting the model, and the corresponding outer weights are estimated.

- **Standardized, weights on raw MV:** The manifest variables are standardized before fitting the model, and the outer weights are estimated for the raw variables.

- **Reduced, weights on raw MV:** The manifest variables are reduced (divided by the standard deviation) before fitting the model, and the corresponding outer weights are estimated.

- **Raw MV:** The manifest variables are not transformed.

**Initial weights (classic and expert displays only):** Choose which initial value should be used for outer weight initialization.

**Values of the first eigenvector.** The initial values are the values associated to the first eigenvector.

**Signs of the coordinates of the first eigenvector.** Instead of taking the values of the first eigenvector only take the sign.
-1 for max of first eigenvector, else +1.
-1 for min of first eigenvector, else +1.
+1 for variable with max variance, else 0.
-1 for the last manifest variable, else +1.

**Weights:** Activate this option if the observations are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the “Variable labels” option is activated.

**Method (classic and expert displays only):** Choose the method to be used. You can choose between PLSPM, GSCA and RGCCA.

**REBUS (expert display only):** Activate this option if you want to apply the REBUS algorithm on your model. When you activate this option the REBUS tab is reachable. When using REBUS, all blocks of manifest variables have to be reflective (mode A), and manifest and latent variables have to be standardized. The NIPALS and Lohmöller options are not available for the treatment of missing data.

**RGCCA (expert display only):** Activate this option if you want to apply the RGCCA. All the blocks have to be defined with either mode A, mode B or the RGCCA mode.

**Range:** Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Activate this option to display the results in a new worksheet of the active workbook.

**Workbook:** Activate this option to display the results in a new workbook.

**Variable labels:** Activate this option if the first row of the data selections includes a header.

**Observation labels:** Activate this option if observations labels are available. Then select the corresponding data. If the “Variable labels” option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

**Scale of the manifest variables (Marketing display):** allows to choose the scale on which the manifest variables are presented. Two options are available.

- Automatic: manifest variables are standardized.
- Select: enter the minimum and maximum values of the manifest variables (uniform scale) in order to work on the original variables.
Scale of the latent variables (Marketing display):

- 0-100 scale: latent variables scores are given on a 0-100 scale.
- MV scale: latent variables scores are given on the same scale as the manifest variables.

Options tab:

Internal estimation (classic and expert displays only): Select the internal estimation method (see the description section for additional details).

- Structural: The inner weights are equal to the correlation between the latent variables when estimating an explanatory (predecessor) latent variable. Otherwise they are equal to the OLS regression coefficients.
- Factorial: The inner weights are equal to the correlation between the latent variables.
- Centroid: The inner weights are equal to the sign of the correlation between the latent variables.
- PLS: The inner weights are equal to the correlation between the latent variables when estimating an explanatory (predecessor) latent variable. Otherwise they are equal to the PLS regression coefficients.

Regression (expert display only): Select the regression method that is used to estimate path coefficients

- OLS: Ordinary Least Squares regression.
- PLS: Partial Least Squares regression.

Dimensions (expert display only): Enter the number of dimensions up to which the model should be computed.

Stop conditions:

- Iterations: Enter the maximum number of iterations for the algorithm. The calculations are stopped when the maximum number of iterations has been exceeded. Default value: 100.
- Convergence: Enter the value of the difference of criterion between two steps which, when reached, means that the algorithm is considered to have converged. Default value: 0.0001.
**Confidence intervals**: Activate this option to compute the confidence intervals. Then choose the method used to compute the intervals:

- **Bootstrap**: Activate this option to use a bootstrap method. Then enter the number of “Re-samples” generated to compute the bootstrap confidence intervals.

- **Jackknife**: Activate this option to use a jackknife method. Then enter the “Group size” that is used to generate the samples to compute the jackknife confidence intervals.

**Confidence interval (%)**: Enter the size in % of the confidence intervals.

**Resampled estimates (expert display only)**: Activate this option to display estimates of standardized loading and path coefficients for each sample being generated. Standard deviation and bounds of the confidence intervals for indirect effects are also displayed.

**Model quality (classic and expert displays only)**:

- **Blindfolding**: Activate this option to check the model quality using the blindfolding approach (see the description section for additional details). Cross-validated values for redundancy and communality will be computed.

**Segmentation (Marketing display)**: Activate this option to look for classes of homogeneous observations with regards to the specified model. The REBUS algorithm is used. A truncation can also be applied.

**Comparisons (Marketing display)**: if groups have been selected, permutation t-tests can be activated here.

**Missing data tab**:

- **Do not accept missing data**: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

- **Remove observations**: Activate this option to remove the observations that contain missing data.

- **Use NIPALS**: Activate this option to use the NIPALS algorithm to handle missing data (see the description section for additional details).

- **Lohmöller**: Activate this option to use the Lohmöller’s procedure to handle missing data: pairwise deletion to compute sample means and standard deviation, and mean imputation to compute the scores. If PLS regression is used instead of OLS regression, then model is applied on the available data.

- **Use the ipsative mean**: Activate this option to use the mean of the latent variables to estimate missing data in the manifest variables.
• **Renormalize**: Activate this option to renormalize external weights for each observation when missing data have been found.

Note: In the case of standardized weights, the two options above lead to pairwise deletion to compute sample means and standard deviation, and mean imputation to compute the scores.

**Estimate missing data**: Activate this option to estimate missing data before starting the computations.

• **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

• **Nearest neighbour**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.

**Multigroup tests** tab (expert display):

If groups have been selected, this tab appears.

**Multigroup t test**: Activate this option to test equality between path coefficients from one group to another with a t test (the number of bootstrap sample is defined in the option tab).

• **Significance level (%)**: Enter the significance level for the t tests.

**Permutation tests**: Activate this option to test equality of parameters between two groups with a permutation test.

• **Number of permutations**: Enter the number of permutations.

• **Significance level (%)**: Enter the significance level for the t tests.

• **Path coefficients**: Activate this option to test the equality of the path coefficients.

• **Standardized loadings**: Activate this option to test the equality of the standardized loadings.

**Model quality**: Activate this option to test the equality of the quality indexes (communalities, redundancies, and GoF).

**REBUS** tab (expert display):

If the REBUS option is activated, this tab appears.

**Truncation**: Activate this option is you want XLSTAT to automatically define the truncation level, and therefore the number of classes to retain, or if you want to define the number of classes to create, or the level at which the dendrogram shall be truncated.
Stop conditions:

- **Iterations**: Enter the maximum number of iterations for the REBUS algorithm. The calculations are stopped when the maximum number of iterations has been exceeded. Default value: 100.

- **Threshold (%)**: Enter the percentage of stable units necessary to stop the algorithm. Default value: 95.

Dendrogram:

- **Horizontal**: Choose this option to display a horizontal dendrogram.

- **Vertical**: Choose this option to display a vertical dendrogram.

- **Labels**: Activate this option to display object labels (full dendrogram) or classes (truncated dendrogram) on the dendrogram.

- **Colors**: Activate this option to use colors to represent the different groups on the full dendrogram.

**Outputs** tab:

**Descriptive statistics**: Activate this option to display descriptive statistics for the selected variables.

**Model**: Activate this option to display the model specifications.

**Correlations (classic and expert designs only)**: Activate this option to display the correlation matrix.

- **Test significance**: Activate this option to test the significance of the correlations.

- **Significance level (%)**: Enter the significance level for the above tests.

**Outliers’ analysis (expert design only)**: Activate this option to display the DmodX and DModY table when PLS regression is selected.

**MV after deflation (expert design only)**: Activate this option to display the manifest variables after deflation when more than one dimension has been selected.

**Variables/Factors correlations**: Activate this option to display correlations between factors and variables.

**Inner model**: Activate this option to display the results that correspond to the inner model.
**Outer model:** Activate this option to display the results that correspond to the outer model.

**R² and communalities:** Activate this option to display the R² of the latent variables from the structural model and the communalities of the manifest variables.

**Model quality:** Activate this option to display the results of the blindfolding procedure.

**Latent variable scores (classic and expert designs only):**

- **Standardized:** Activate this option to compute and display standardized factor scores.

- **Using normalized weights:** Activate this option to display factor scores computed with normalized weights.

- **Standardized > 0-100:** Activate this option to compute standardized scores, and then transform and display the latter on a 0-100 scale.

- **Using normalized weights > 0-100:** Activate this option to compute factor scores using normalized weights, and then transform and display the factor scores on a 0-100 scale.

**Simulation table (Marketing display):** activate this option to display simulation tables that let you visualize the effects of the modification of a manifest or a latent variable on a target latent variable.

- **LV to explain:** select the target latent variable to explain. You should select an endogeneous latent variable.

- **Scale of changes:** select the scale of changes (percent or number of points). Once this option is configured, you will be able to enter the minimum, the maximum as well as the change step to obtain the range of values to test.

**IPMA (Marketing display):** Activate this option if you wish to display the tables based on IPMA (Importance Perform Analysis).

**Charts tab:**

**Coefficients plot:** activate this option to display normalized coefficients of the internal model.

**IPMA chart:** activate this option to display the IPMA charts.

**Simulation plot (Manifest variables) (Marketing display):** activate this option to display simulation plots to investigate the effects of modifying manifest variables on the score of the target latent variable.

**Simulation plot (Latent variables) (Marketing display):** activate this option to display simulation plots to investigate the effects of modifying latent variables on the score of the target latent variable.
Results options

Many results can be displayed on the PLSPMGraph sheet, once the model has been fitted. It is recommended to select only a few items in order to keep the results easy to read. To display the options dialog box, click the results icon of the “Path modeling” toolbar.

Latent variables tab:

These options allow defining which results are displayed below the latent variables.

- **Mean**: Activate this option to display the mean of the latent variable.
- **Mean (Bootstrap)**: Activate this option to display the mean of the latent variable computed using a bootstrap procedure.
- **Confidence interval**: Activate this option to display the confidence interval for the mean.
- **R²**: Activate this option to display the R-square between the latent variable and its manifest variables.
- **Adjusted R²**: Activate this option to display the adjusted R-square between the latent variable and its manifest variables.
- **R² (Boot/Jack)**: Activate this option to display the R-square between the latent variable and its manifest variables, computed using a bootstrap or jackknife procedure.
- **R² (conf. int.)**: Activate this option to display the confidence interval on the R-square between the latent variable and its manifest variables, computed using a bootstrap or jackknife procedure.
- **Communality**: Activate this option to display the communality between the latent variable and its manifest variables.
- **Redundancy**: Activate this option to display the redundancy between the latent variable and its manifest variables.
- **Communality (Blindfolding)**: Activate this option to display the communality between the latent variable and its manifest variables, computed using the blindfolding procedure.
- **Redundancy (Blindfolding)**: Activate this option to display the redundancy between the latent variable and its manifest variables, computed using the blindfolding procedure.
- **D.G. rho**: Activate this option to display the Dillon-Goldstein's rho coefficient.
- **Cronbach's alpha**: Activate this option to display the Cronbach's alpha.
• **Std. deviation (Scores):** Activate this option to display the standard deviation of the estimated latent variables’ scores

**Arrows (Latent variables) tab:**

These options allow to define which results are displayed on the arrows that relate the latent variables.

• **Correlation:** Activate this option to display the correlation coefficient between the two latent variables.

• **Contribution:** Activate this option to display the contribution of the latent variables to the R2.

• **Path coefficient:** Activate this option to display the regression coefficient that corresponds to the regression of the latent variable that is at the end of the arrow (dependent) by the latent variable at the beginning of the arrow (predecessor or explanatory).

• **Path coefficient (B/J):** Activate this option to display the regression coefficient that corresponds to the regression of the latent variable that is at the end of the arrow (dependent) by the latent variable at the beginning of the arrow (predecessor or explanatory), computed using a bootstrap or jackknife procedure.

• **Standard deviation:** Activate this option to display the standard deviation of the path coefficient.

• **Confidence interval:** Activate this option to display the confidence interval for the path coefficient.

• **Std. coeff.:** Activate this option to display standardized coefficients.

• **Student's t:** Activate this option to display the value of the Student’s t.

• **Partial correlations:** Activate this option to display partial correlations between latent variables.

• **Pr > |t|:** Activate this option to display the p-value that corresponds to the Student’s t.

• **Arrows thickness depends on:** The thickness of the arrows can be related to:
  
  o The p-value of the Student’s t (the lower the value, the thicker the arrow).
  
  o The correlation (the higher the absolute value, the thicker the arrow; blue arrows correspond to negative values, red arrows to positive values).
  
  o The contribution (the higher the value, the thicker the arrow).
**Arrows (Manifest variables) tab:**

These options allow to define which results are displayed on the arrows that relate the latent variables.

- **Weight:** Activate this option to display the weight.
- **Weight (Bootstrap):** Activate this option to display the weight computed using a bootstrap procedure.
- **Normalized weight:** Activate this option to display the normalized weight.
- **Standard deviation:** Activate this option to display the standard deviation of the weight.
- **Confidence interval:** Activate this option to display the confidence interval for the weight.
- **Correlation:** Activate this option to display the correlation coefficient between the manifest variable and the latent variable.
- **Correlation (Boot/Jack):** Activate this option to display the correlation coefficient between the manifest variable and the latent variable, computed using a bootstrap of jackknife procedure.
- **Correlation (std. deviation):** Activate this option to display the standard deviation of the correlation coefficient between the manifest variable and the latent variable, computed using a bootstrap of jackknife procedure.
- **Correlation (conf. interval):** Activate this option to display the confidence interval of the correlation coefficient between the manifest variable and the latent variable, computed using a bootstrap of jackknife procedure.
- **Communalities:** Activate this option to display the communality between the latent variable and the manifest variables.
- **Redundancy:** Activate this option to display the redundancy between the latent variable and the manifest variables.
- **Communality (Blindfolding):** Activate this option to display the communality between the latent variable and its manifest variables, computed using the blindfolding procedure.
- **Redundancy (Blindfolding):** Activate this option to display the redundancy between the latent variable and its manifest variables, computed using the blindfolding procedure.
- **Arrows thickness depends on:** The thickness of the arrows can be related to:
  - The correlation (the higher the absolute value, the thicker the arrow; blue arrows correspond to negative values, red arrows to positive values).
  - Normalized weights.
Results

The first results are general results which computation is done prior to fitting the path modeling model:

**Summary statistics**: This table displays for all the manifest variables, the number of observations, the number of missing values, the number of non-missing values, the minimum, the maximum, the mean and the standard deviation.

**Model specification (measurement model)**: This table displays for each latent variable, the number of manifest variables, the mode, the type (a latent variable which never appears as a dependent variable is called exogenous), if its sign has been inverted, the number of computed dimension and the list of all associated manifest variables.

**Model specification (structural model)**: This square matrix shows on its lower triangular part if there is an arrow that goes from the column variable to the row variable.

**Composite reliability**: This table allows to check the dimensionality of the blocks. For each latent variable, a PCA is run on the covariance or correlation matrix of the manifest variables in order to determine the dimensionality. The Cronbach’s alpha, the Dillon-Goldstein’s rho, the critical eigenvalue (that can be compared to the eigenvalues obtained from the PCA) and the condition number are displayed to facilitate the determining of the dimensionality.

**Variables/Factors correlations (Latent variable X / Dimension Y)**: These tables display for each latent variable the correlations between the manifest variables and the factors extracted from the PCA. When a block is not unidimensional, these correlations allow to identify how the corresponding manifest variables can be split into unidimensional blocks.

The results that follow are obtained once the path modeling model has been fitted:

**Goodness of fit index (Dimension Y)**: This table displays the goodness of fit index (GoF) computed using bootstrap or not and its confidence interval for

- Absolute: Value of the GoF index.
- Relative: Value of the relative GoF index obtained by dividing the absolute value by its maximum value achievable for the analyzed dataset.
- Outer model: Component of the GoF index based on the communalities (performance of the measurement model).
- Inner model: Component of the GoF index based on the R2 of the endogenous latent variables (performance of the structural model).
**Cross-loadings (Monofactorial manifest variables / Dimension Y):** This table allows to check whether a given manifest variable is really monofactorial, i.e. mostly related to its latent variable or if it is also related to other variables. Ideally, if the model has been well specified, it should appear as being mostly related to its latent variable.

**Outer model (Dimension Y):**

- **Weights (Dimension Y):** Coefficients of each manifest variable in the linear combination used to estimate the latent variable scores.

- **Standardized loadings (Dimension Y):** Correlations (standardized loadings) between each manifest variable and the corresponding latent variable. Loadings and location parameters are also displayed.

**Inner model (Dimension Y):**

- **R² (Latent variable X / Dimension Y):** Value of the R2 index for the endogenous variables in the structural equations.

- **Path coefficients (Latent variable X / 1):** Value of the regression coefficients in the structural model estimated on the standardized factor scores. The size effect ($f^2$) is also displayed.

- **Impact and contribution of the variables to Latent variable X (Dimension Y):** Value of the path coefficients and the contributions (in percent) of the predecessor latent variables to the R2 index of the endogenous latent variables.

**Bootstrap:** Values of the standardized loadings and path coefficients for each generated sample.

**Model assessment (Dimension Y):** This table summarizes important results associated to the latent variables scores.

**Correlations (Latent variables) / Dimension Y (Expert display):** Correlation matrix obtained on the latent variable scores.

**Partial correlations (Latent variables) / Dimension Y (Expert display):** Partial correlation matrix obtained on the latent variable scores.

**Direct effects (latent variable) / Dimension Y (Expert display):** This table shows direct effect between connected latent variables.

**Indirect effect (latent variable) / Dimension Y (Expert display):** This table shows the indirect effects between not directly connected latent variables. If the resampled estimates option has been selected, the standard deviations and the bounds of the confidence intervals are also displayed.
Total effect (latent variable) / Dimension Y (Expert display): This table shows the total effect between latent variables. Total effect = direct effect + indirect effect.

Discriminant validity (Squared correlations < AVE) (Dimension Y): This table allows to check whether each latent variable is really representing a concept different from the other or if some latent variables are actually representing the same concept. In this table, the R2 index for any pair of latent variables shall be smaller than the mean communalities for both variables which indicates that more variance is shared between each latent variable and its block of manifest variables than with another latent variable representing a different block of manifest variables.

IPMA (Importance Performance Matrix Analysis) tables and charts (Expert and Market displays): for each endogeneous latent variable, those tables gather importance and performance values of the latent variables. Importance is the total effect on the studied endogeneous latent variable. Performance is the score of the latent variable scaled between 0 and 100. Those indices are represented on charts.

Simulation tables and plots (Marketing display): those results can be used to understand the impact of the modification of a variable in the model on a target latent variable to explain.

- The first table gathers the most important latent variables for the prediction of the target latent variable to explain.
- The second table displays the most important manifest variables for the prediction of the target latent variable to explain.
- The following table and chart allow to visualize the impact of modifying a manifest variable on the target latent variable to explain.
- The following table and chart allow to visualize the impact of modifying a manifest variable on the score mean of the target latent variable to explain (the mean is displayed and not the change).
- The following table and chart allow to visualize the impact of modifying a latent variable on the target latent variable to explain.
- The following table and chart allow to visualize the impact of modifying a latent variable on the score mean of the target latent variable to explain (the mean is displayed and not the change).

Latent variable scores (Dimension Y):

- Mean / Latent variable scores (Dimension Y): Mean values of the individual factor scores.
- Summary statistics / Latent variable scores (Dimension Y): Descriptive statistics of the latent variable scores computed from the measurement model.
• Latent variable scores (Dimension Y): Individual latent variable scores estimated as a linear combination of the corresponding manifest variables.

• Summary statistics / Scores predicted using the structural model (Dimension Y) (expert display): Descriptive statistics of the latent variable scores computed from the structural model.

• Scores predicted using the structural model (Dimension Y) (expert display): Latent variable scores computed as the predicted values from the structural model equations.

**Model assessment / Outer model (Blindfolding):** Cross-validated values of the communalities obtained by means of the blindfolding procedure.

**Model assessment / Inner model (Blindfolding):** Cross-validated values of the redundancies obtained by means of the blindfolding procedure.

If groups are defined, some other outputs are available:

**Worksheet PLSPM (Group):** For each group, complete results are displayed in separated worksheets.

Worksheet PLSPM (Multigroup t test): For each path coefficient, results of the t test are summarized in a table. Each line represents a pair of groups.

• Difference: Absolute value of the parameter’s difference between the groups.

• t (Observed value): Observed value of the t statistic.

• T (critical value): Critical value of the t statistic.

• DF: Number of degrees of freedom.

• p-value: p-value associated to the t test.

• Alpha: Significance level

• Significant: If yes, the difference between the parameters is significant. If not, the difference is not significant.

Worksheet PLSPM (Permutation test): For each type of parameter, results of the permutation test are summarized in a table.

• Difference: Absolute value of the parameter’s difference between the groups.

• p-value: p-value associated with the permutation test.

• Alpha: Significance level.
• Significant: If yes, the difference between the parameters is significant. If not, the difference is not significant.

If the REBUS option is activated, some other outputs are available:

**Worksheet REBUS:** The dendogram obtained with the cluster analysis is displayed. For each observation, the class and the CM index is also displayed.

**Worksheet PLSPM (Class):** For each class, complete results are displayed in separated worksheets.

**Example**

A tutorial on how to use the XLSTAT-PLSPM module with Excel 2007 is available on the Addinsoft website:


A tutorial on how to use the XLSTAT-PLSPM module with Excel 2003 is available on the Addinsoft website:


A tutorial on how to compare groups with XLSTAT-PLSPM is available on the Addinsoft website:


A tutorial on how to use the REBUS method with XLSTAT-PLSPM is available on the Addinsoft Website:


**References**


