Chapter 10 - Processing

10.0 Introduction

Processing transforms survey responses obtained during collection into a form that is suitable for tabulation and data analysis. It includes all data handling activities – automated and manual – after collection and prior to estimation. It is time-consuming, resource-intensive and has an impact on the quality and cost of the final data. It is therefore important that it be well planned, the quality of its implementation monitored and corrective actions taken when required.

The extent and order of the processing activities depend, among others, on the nature of the data that are collected, the collection method, the survey budget and objectives in terms of data quality. Coding, for instance, can occur before or after data capture while editing typically occurs throughout the survey. The following is an example of processing activities for a paper questionnaire:

i. After collection, check the data on the questionnaire. This step ensures that all necessary information has been received and legibly recorded, that interviewer notes have been reviewed and some preliminary edits performed to check for gross errors and inconsistencies.

ii. Code any data on the questionnaire that requires coding (e.g., written answers to open questions).

iii. Capture the data electronically into a computer. Data capture may be followed by more coding.

iv. Perform detailed editing, then imputation. Questionnaires that fail one or more checks are put aside for further examination, either for follow-up with the respondent, or for imputation.

v. Perform outlier detection to identify suspicious or extreme values.

vi. Store the data on a database to facilitate data manipulation during the post-processing activities.

In order to streamline processing, several of the above activities – namely, capture, editing and coding – may be combined through automation using computer-assisted collection methods.

Since errors are likely to occur at each stage of processing – particularly for those manual and repetitive activities such as coding, capture and editing – processing should be monitored and corrective actions taken when necessary in order to maintain or improve quality. This is done by implementing quality control and quality assurance procedures.

The purpose of this chapter is to cover the main processing activities: coding, data capture, editing, imputation, outlier detection and treatment and creating a database. For details on quality assurance and quality control, see **Appendix B - Quality Assurance and Quality Control**.

10.1 Coding

Coding is the process of assigning a numerical value to responses to facilitate data capture and processing in general. As stated in **Chapter 3 - Introduction to Survey Design**, coding entails either assigning a code to a given response or comparing the response to a set of codes and selecting the one that best describes the response.

Difficulties in determining a set of response categories to a question was discussed in **Chapter 5 - Questionnaire Design**. In the case of closed questions, the response categories are determined before collection, with the numerical code usually appearing on the questionnaire beside each response category. For open questions, coding occurs after collection and may be either manual or automated. Manual coding requires interpretation and judgement on the part of the coder, and may vary between coders.

When determining the coding scheme, the goal should be to classify responses into a meaningful set of exhaustive and mutually exclusive categories that bring out the essential pattern of responses. For some questions, coding may be straightforward (e.g., marital status). In other cases, such as geography, industry and occupation, a standard coding system may exist. But for many questions no standard coding system exists and determining a good coding scheme is a nontrivial task. The coding scheme should be consistent and logical and take into account how detailed the codes should be in light of the purpose of the survey and tabulations or data analysis to be performed. It is best to start with a broad list, since too few categories can be misleading and a large *other* category can be uninformative. Categories can always be collapsed but it is difficult to split categories after the fact.

As mentioned in **Chapter 5 - Questionnaire Design**, the type of interview should be considered when determining the response categories for a closed question. For a self-enumeration survey, supplying a list of 50 categories on paper to respondents is feasible (though not ideal); for a telephone interview, listing 50 response categories over the telephone is impractical. To simplify coding, ideally all questions on a questionnaire would be closed with a short list of response categories. In practice, this is not always possible and sometimes open questions are necessary.

10.1.1 Pre-coding Closed Questions

The response categories for closed questions can be pre-coded on the questionnaire. For paper questionnaires, boxes for codes can be placed either next to the item to be coded or down the margin of the questionnaire. This greatly improves the efficiency of data capture after collection: instead of typing in the selected response category, a numeric code is captured (numerical codes are also easier to analyse than strings of words). With computer-assisted collection methods, the codes are automatically captured when the interviewer or respondent selects an answer.

For example, the following coding system was used for the 2002 Public Service Employee Survey:

In the past three years how many promotions have you had?
○ *none*
○ *one*
○ *more than one*

The benefits of closed questions were discussed in **Chapter 5 - Questionnaire Design**. The main benefits are that they are less burdensome to respondents, and data collection, capture and analysis are cheaper, faster and easier than for open questions. One disadvantage of closed questions is that the respondent's natural wording is unknown. This can make it difficult to verify the quality of the coding. For example, if an open question were used to determine a person's occupation, the respondent's description of his or her work could lead to a different occupation code than if the respondent or interviewer were to select from amongst a choice of occupational codes provided by a closed question.

10.1.2 Manual Coding of Open Questions

When manually coding open questions, the coder (typically after collection) must read, interpret and manually convert a written response to an open question into a numeric code. This numeric code is then either recorded on the questionnaire or entered into a computer. In order to assign a code, the coder may simply be required to note whether the answer contains a key word or reference to a particular item. Sometimes, coding is based on the answer to only one question, sometimes it is based on the answer to several related questions. In this case, the quality of the coding depends very much on the clarity and completeness of the written response and how well the initial editing was done, the soundness of the coding scheme and the skill of the coder.

Coders must be trained carefully since applying the coding scheme requires taking into consideration the following:

- the number of possible answers;
- complexity (judgement);
- possible ambiguity in the response (i.e., the quality of the response).

Variability between individual coders is inevitable. A detailed check of a coder's first batch of questionnaires should be performed to detect mistakes and identify whether or not further training is required. After that, periodic checks should be done on the quality of the coding and corrective action taken if required. This is often done using quality control methods (see **Appendix B - Quality Assurance and Quality Control**).

10.1.3 Automated Coding of Open Questions

Traditionally, coding open questions has been a manual operation. However, due to advances in technology, resource constraints and, most importantly, concerns about timeliness and quality, coding is becoming more and more an automated process.

In general, two files are input to an automated coding system. One file contains the survey responses that are to be coded, referred to as the write-in file. The second file is called the reference file and contains typical written responses (or phrases) and their corresponding numeric codes.

Most often, the first step of automated coding is parsing. Parsing is the process of standardising a phrase in order to allow the computer to recognise equivalent phrases. Parsing usually consists of the removal of extraneous characters, such as punctuation, and the removal of double words, trivial words, certain suffixes and prefixes, etc. Both the survey write-in files and the reference file are parsed before continuing.

The next step involves, for every write-in in the survey file, a search of the reference file for an exact match. If found, the code in the reference file is copied to the survey file and the record[1] is considered coded. However, if an exact match cannot be found, an attempt can be made to find the closest match possible among the reference file records. Every record on the reference file is assigned a score, which is a measure of how similar the reference file phrase is to the survey response. Scores are evaluated based on predetermined parameters (which are specified to reduce the risk of making an error) and if a close-enough match is found, the code is transferred to the survey response and the record is considered coded.

[1] In this chapter, *questionnaire* generally refers to the hard-copy document whereas *record* refers to the electronic version of the completed questionnaire.

Sometimes several reference file records are found with similar scores and sometimes no reference file record is found to be close to the survey response. In these situations, the records are generally sent to a small manual coding operation staffed with expert coders who are responsible for coding any records left uncoded at the end of the automated phase and for quality control of the output of the automated system (see **Appendix B - Quality Assurance and Quality Control**).

10.2 Data Capture

Data capture is the transformation of responses into a machine-readable format. With paper-based collection methods, capture occurs after collection (usually after the questionnaire has been groomed and some preliminary edits have been performed). For example, data capture might consist of a clerk (referred to as a keyer) manually typing into a computer the reported values on the questionnaire. With computer-based collection methods, capture occurs at the time of collection.

There are several ways to improve the efficiency of data capture. One is to use computer-assisted data collection methods. The main benefit of computer-assisted methods is that collection and capture are combined, resulting in an integrated, faster and more efficient data collection and capture process than paper-based methods. One disadvantage of computer-assisted methods is that the software programs require extensive development and testing. (For more on the advantages and disadvantages of computer-assisted data collection, see **Chapter 4 - Data Collection Methods**.)

For paper-based collection methods, pre-coding closed questions can greatly improve the efficiency of manual data capture. Another option is to optically scan completed questionnaires. Optical scanning works best for closed questions and is less reliable at capturing answers to open questions. Scanning can reduce data capture errors compared to manual capture, but scanning errors do occur and should be evaluated and minimised. In the case of long questionnaires, the logistics of optical scanning are more cumbersome: staples have to be removed, questionnaire identifiers added to every page, scanners reset to read in different pages, etc. Another option is to have all answers coded on a single sheet of paper. This simplifies scanning but it can be burdensome for the interviewer to read a question from one sheet of paper and record it on another. It also limits questions to closed questions and on a large sheet full of answer boxes, it is easy for the interviewer to code the wrong answer or to code the answer in the wrong answer box. Finally, it is difficult for the interviewer to refer back to an answer given by the respondent since the questions and answers are in different places.

With paper-based collection methods it is particularly important that quality assurance and quality control procedures be implemented to minimise and correct errors introduced during data capture (see **Appendix B - Quality Assurance and Quality Control**).

10.3 Editing

In an ideal world, every questionnaire would be completed without any errors. Unfortunately, responses to some questions may be missing, incomplete or incorrect. ***Editing is the application of checks to identify missing, invalid or inconsistent entries that point to data records that are potentially in error***. Editing usually identifies nonsampling errors arising from measurement (response) errors, nonresponse or processing. The purpose of editing is to:

- better understand the survey processes and the survey data;
- detect erroneous or missing data;
- follow-up with the respondent;

- send a record to imputation;
- delete a record.

In order to identify erroneous records, edit rules are applied. Some examples of edit rules are:

- each question must have one and only one response;
- the valid responses for Question X are *1* or *2*;
- the sum of the parts for Question X cannot be less than the response to Question Y.

Editing can occur at several points throughout the survey process and ranges from simple preliminary checks performed by interviewers in the field to more complex automated verifications performed by a computer program after the data have been captured. In general, edit rules are based upon what is logically or validly possible, based upon:

- expert knowledge of the subject matter;
- other related surveys or data;
- the structure of the questionnaire and its questions;
- statistical theory.

Subject matter experts should be knowledgeable of how variables relate to one another and what answers are reasonable. Such individuals are instrumental in specifying what types of rules are appropriate. Typically, they are analysts who have experience with the types of data being edited. For example, a transportation analyst may be aware of the acceptable range of values for fuel consumption rates for various makes and models of vehicles. Analysis of other surveys or datasets relating to the same sorts of variables as the ones being edited can be useful in establishing some of the edit rules.

Equally important, the layout and structure of the questionnaire have an impact on the edit rules. Edits should verify that responses respect the logical flow of the questionnaire. This is often manifested through the use of *go to* or *skip* instructions which imply that certain questionnaire items do not apply to certain categories of respondents and therefore the respondent is to skip to another question.

There are three main categories of edits: *validity*, *consistency* and *distribution* edits. Validity and consistency edits are applied one questionnaire at a time. Validity edits verify the syntax of responses and include such things as checking for non-numeric characters reported in numeric fields and checking for missing values. The first two examples of edit rules above are validity edits. Validity edits can also check that the coded data lie within an allowed range of values. For example, a range edit might be put on the reported age of a respondent to ensure that it lies between 0 and 125 years.

Consistency edits verify that relationships between questions are respected. The third example of an edit rule above is a consistency edit. Consistency edits can be based on logical, legal, accounting or structural relationships between questions or parts of a question. The relationship between date of birth and marital status is one example where an edit might be: ‘a person less than 15 years of age cannot have any marital status other than *never married*.’ Consistency edits may also be based on the logical flow of the questionnaire, for example: ‘if Question X is answered *no* then Question Y cannot be answered’. Consistency edits may also involve the use of historical data (e.g., year-to-year ratios). In the case of household surveys, there may be edits between household members.

Distribution edits are performed by looking at data across questionnaires. These attempt to identify records that are outliers with respect to the distribution of the data. Distribution edits are sometimes referred to as statistical edits (Hidiroglou and Berthelot, 1986) or outlier detection. For more information, see Section 10.5. For a discussion of nonsampling errors, see **Chapter 3 - Introduction to Survey Design**.

10.3.1 Edits During Data Collection

Edits during data collection are often referred to as field edits and generally consist of validity edits and some simple consistency edits. The purpose of editing during data collection is to:

- identify the need for improvement to the collection vehicle;
- identify the need for more training;
- detect obvious errors and perform immediate follow-up with the respondent;
- clean-up entries.

Editing during data collection may be performed by:

- the respondent (in the case of self-enumeration);
- the interviewer during the interview;
- the interviewer immediately after the interview;
- the interviewer's supervisor;
- clerical staff.

Field edits are used to identify problems with data collection procedures and the design of the questionnaire as well as the need for more interviewer training. They are also used to detect mistakes made during the interview by the respondent or the interviewer and to identify missing information during collection in order to reduce the need for follow-up later on. Editing during collection is considerably easier to implement when it is automated through a computer-assisted collection method.

For self-enumeration questionnaires, respondents may edit their own answers. In almost all interviewer-assisted surveys, some editing is performing during the interview and interviewers are instructed and trained to review the answers they record on a questionnaire immediately after the interview is concluded – either after leaving the dwelling or after hanging up the telephone. This way they still have an opportunity to detect and treat records that failed edit rules, either because the correct information may still be fresh in their mind or because they can easily and inexpensively follow-up with the respondent to ascertain the correct values. Any edit failures left unresolved are usually dealt with later by imputation.

Another purpose of field edits is to clean-up responses. Often during an interview, the interviewer writes short notes in the margin of the questionnaire or in the notes section of the CATI application. This may be because the interviewer does not know the coding scheme for an open question, or he or she may want to refer to the interviewer's manual for the interpretation of an answer. In these cases, interviewers edit their questionnaires after the interview in order to clean-up these notes.

One of the tasks assigned to supervisors is to check the work of their interviewers to detect errors and then feed this information back to the interviewer. Usually the kinds of failures detected are similar to those that could be detected by the interviewer immediately after the interview and usually there is still some opportunity to follow-up with the respondent to determine the correct values. The supervisors should also be looking for patterns of errors that occur. Lessons learned from one interviewer should be passed on to the whole team.

In many surveys, completed questionnaires are transmitted by respondents or by interviewers to a Regional Office for log-in and clerical *grooming*. This grooming often consists of the same or additional edits to those carried out by the interviewers or supervisors. Grooming includes deciphering handwritten answers, interpreting interviewer remarks, standardising measurement scales (e.g., changing a value reported in feet to metres), etc. It may also involve making sure interviewers have completed all administrative fields on the questionnaire such as response status codes (e.g., indicating a fully or partially completed questionnaire). This process provides for a systematic, independent review or edit of the questionnaire data before they are sent on to data capture. Checking the questionnaire identification codes can be an important element of this exercise since, without complete identification, questionnaires cannot be logged in or data captured. The degree of editing depends upon the available budget and the extent to which the clerical staff doing the editing can be expected to identify and resolve difficulties encountered. Where possible, this kind of editing is combined with any coding, tallying or batching of questionnaire items that might be required before data capture is started. In some cases, Regional Office staff may follow-up with the respondent to resolve important edit failures.

10.3.2 Edits After Data Collection

The most comprehensive and complicated edits are generally carried out as a separate edit and imputation stage after data collection. During data capture, edits can be carried out by keyers or automatically by computer programs, or by the computer application in the case of computer-assisted collection methods. For paper questionnaires with manual data capture, it is economical to use data capture as an opportunity to apply rules to clean the data sufficiently to make the subsequent processing stages more efficient. Generally, editing during data capture is minimised since responding to an edit failure slows down data capture. Edits during this stage of processing are mainly validity edits and simple consistency edits.

More complex edit rules are generally reserved for the separate edit stage after data capture – along with validity edits, more complex consistency edits are often performed along with selective editing and outlier detection (see section 10.5).

For edit failures after data collection, the usual procedure is to flag the field that failed an edit and then either impute the field or exclude the record from further processing.

Most edit failures at this stage are flagged for imputation. Values that fail an edit should be flagged with a special code to indicate that an unacceptable value or invalid blank was reported. These flags are particularly useful when assessing the quality of the survey data. In some cases, the record or questionnaire may fail so many edit rules – or a small number of critical edits – that it is rendered useless for further processing. In such cases, the record is usually treated as a nonrespondent, removed from the processing stream and a nonresponse weight adjustment performed (see **Chapter 7 - Estimation** for details of weight adjustments).

10.3.3 Selective Editing

In editing, there is a trade-off between getting every record perfect and spending a reasonable amount of resources (i.e., time and money) tidying up the data. Historically, much time and effort has been spent trying to ensure that any and all survey errors are identified. Not only is over-editing the data a poor use of resources, but it can lead to biased results. Typically, the data are expected to follow a pre-defined model and when the data do not follow the model, they are said to fail an edit. If the data are changed every time an edit fails, this can severely bias the data towards the model and away from the real life situation. Also, excessive editing and follow-ups with respondents can result in high response burden and lower the respondent's co-operation on future occasions.

To avoid spending excessive time and resources editing data that have little impact on the final estimates, selective editing practices are recommended, particularly for business surveys (i.e., where the population is skewed and a few businesses dominate the estimates). The selective editing approach is based upon the idea that only critical edit failures need be treated. Selective editing generally applies to quantitative data. An example of an application of selective editing is a procedure that modifies individual records according to their potential impact on the survey estimates or through the analysis of aggregate data.

Selective edit failures may result in following-up with the respondent, excluding the record from further processing or specifying records that require imputation.

The advantages of selective editing include:
- cost savings;
- data quality can be improved by redirecting resources to high impact records or to other activities;
- timeliness can be improved by reducing processing time;
- response burden can be reduced as a result of fewer follow-ups.

The disadvantages of selective editing include:
- less attention is given to the data quality at the level of the individual unit;
- inconsistencies may be left in the data, which may give users the impression of poor data quality;
- the nonsampling error for small domains can be greater than if all questionnaires are individually edited;
- there may be resistance from data processing clerks, subject matter experts, management or data users who may have less confidence in the data.

Some selective editing approaches include:

i. Top-Down Approach

With this method, the most influential weighted data values are listed from top to bottom for a given domain of estimation and examined one by one. Data examination and verification stops when the next most influential data value does not affect the domain estimate significantly. For example, consider a sample of five businesses from a population of 100 businesses for a survey that wants to estimate the total number of employees in the population. The survey's estimate of the total number of employees is 737. The analyst feels that this estimate is too high (since he expects the average number of employees per business to be 3). In order to examine this value, the contribution of each record is examined as a proportion of the estimate. As can be seen from Table 1, the first record contributes 81.4% to the estimate of the total. Because of its influence on the estimate, this record is examined more closely. It quickly becomes evident that the number of employees reported by this company is higher than expected and the weight is much higher than for other records (perhaps due to a nonresponse adjustment). This record is consequently treated as an influential observation (see Section 10.5). Since all the other weighted values contribute only a small proportion to the overall total, they are not examined more closely.

Table 1: Example of Top-Down Editing

Record	Number of employees	Weight	Contribution to Total
1	12	50	81.4%
2	7	8	7.6%
3	3	12	4.9%
4	2	15	3.3%
5	1	15	2.0%

ii. Aggregate Method

With the aggregate method, suspicious *domain estimates* are identified. The weighted data of *all* records belonging to the domain are then examined. For example, for a survey estimating average household size, if the average household size in a given village is found to be 23, all of the weighted individual records for that village would be examined to see if there are any values that seem to be substantially higher than the others.

iii. Graphical Method

Here, the data are graphed to identify suspicious values. For example, the distribution of the data can be graphed to identify unlikely tails of the distribution.

iv. Questionnaire Score Method

Berthelot and Latouche (1992) propose the use of a score function, where each respondent is assigned a score based on some measure of size, the number of suspicious data items on the questionnaire and the relative importance of the variables. Only records with a high score are examined.

10.3.4 Manual versus Automated Edits

Editing can be automated by means of a computer program. Depending on the volume of editing to be done (in terms of the number of data items or number of questionnaires), the nature and complexity of the edit rules involved, the impact of the unit, the importance of the variables and the stage of survey processing to which the edit rules apply, manual or automated processing may be appropriate. The more complex the edit rules, the more difficult and error-prone is a manual process. Conversely, for some surveys (e.g., paper based ones) it is difficult – if not impossible – to incorporate automated edits during the data collection. Other factors that affect whether or not editing should be manual or automated include the need to monitor interviews and the need for an audit trail. Editing performed after data capture is, however, usually automated. A generally accepted principle for this editing phase – and its related imputation phase – is that it should not require reference back to the individual hard-copy questionnaire unless absolutely necessary. In other words, the electronic records resulting from data capture should contain all the necessary information required for the subsequent editing and imputation to be carried out.

10.3.5 Constraints to Editing

Some constraints to editing are:
- available resources (time, budget and people);
- available software;
- respondent burden;
- intended use of the data;
- co-ordination with imputation.

i. Resources (time, budget and people)

In a manual editing environment, the process of editing can be quite labour intensive. It is necessary to:
- develop and document the edit rules to be followed and the actions to be followed when there is an edit failure;
- train those who are going to do the editing;
- establish a mechanism for supervising and checking the work of those doing the editing (i.e., implement Quality Assurance and Quality Control procedures);
- establish a method of evaluating the impact of editing on the original data.

In an automated environment, the implication for time, cost and resources for front-end development can be enormous. Tasks include:
- developing and documenting the edit rules;
- writing a computer program or adapt a software to identify edit failures;

- testing the computer program;
- editing the survey data by running the program.

In either case, it is important to be sure that the investment in editing is worthwhile. It is a waste of resources to institute an expensive and time-consuming editing strategy to catch a handful of records whose impact on survey results is negligible. On the other hand, it is risky to have only a rudimentary editing strategy only to find there are major errors and inconsistencies in the survey responses. How many records are likely to fail the edit rules? What will be the impact of these failures on the resulting data quality? Are all the records equally important? These kinds of questions are important but not always easy to answer. The responses to these questions depend on, amongst other things, how well designed is the questionnaire, how survey-literate are the respondents and how well-trained are the interviewers.

Often, especially in the case of repeated surveys, it is desirable to analyse the raw (i.e., unedited) survey data before embarking upon an editing strategy. This allows the statistical agency to assess in advance the likely volume of edit failures and the kinds of relationships that exist between questions. Editing should, in fact, be viewed as a continuous process that does not necessarily have a start and a finish. It is a learning process aimed at constant improvement of the entire survey process over time.

ii. Software

Some specialised software packages exist for editing and imputing survey data (e.g., Statistics Canada's Generalized Edit and Imputation System, GEIS, or the Canadian Census Edit and Imputation System, CANCEIS). Such packages can allow for the use of comprehensive edit rules for a relatively small front-end investment in systems design. Alternatively, statistical agencies can program their own editing strategy.

iii. Respondent burden

One of the implications of editing questionnaires is the possibility of follow-up with the respondent to treat missing or erroneous data. In the vast majority of situations, the respondent is the most accurate source of information for questionnaire items. However, follow-up is burdensome for the respondent and costly for the statistical agency. Moreover, there may be a significant time interval between the original interview and the time of follow-up so that the respondent no longer remembers the correct answer. These considerations mean that follow-up (as a way of treating edit failures) is generally limited to edits failures identified during collection or arising from selective editing. Since follow-up after collection is generally impractical and undesirable, imputation is required.

iv. Intended use of the data

The amount of editing that is performed should depend, to a large extent, on the uses of the resulting data. Datasets or data items that will be used primarily for qualitative assessments – where decisions will not be based on precise measurements – may not need to be as rigorously edited as those which will have a strategic importance in decision making. Moreover, within a given dataset, some items may be much more important than others and it may therefore be desirable to devote more time and resources to ensuring that they are clean.

Alternatively, some records in a dataset may carry more importance than others and may contribute significantly to survey estimates. This is especially true in business surveys where 5% of the companies may contribute 95% of the total earnings in a given industry. Focussing on the most influential fields or records is one of the purposes of selective editing (section 10.3.3) and outlier detection (section 10.5).

v. Co-ordination with imputation

Editing alone is of minimal value if some action is not taken to treat items that fail the edit rules. When the respondent is not followed up, this corrective action is generally referred to as imputation. The dual actions of editing and imputation are closely related. It is therefore important to consider how the latter will be done when developing specifications for the former. In many cases, imputation is done when the edit failure is detected (before proceeding with the examination of subsequent rules). This approach is desirable in situations where it is obvious what action should be performed given the nature of the question or given the answers to related questions. Frequently, however, imputation is performed as a separate step after all the data have been processed through all the edit rules.

10.3.6 Guidelines for Editing

The following are some guidelines for editing:

i. Edits should be developed by staff who have expertise in the subject matter, questionnaire design, data analysis and with other similar surveys.

ii. Editing should be performed at several stages of the survey.

iii. Edits applied at each stage should not contradict edits at some other stage (edits applied throughout collection and processing should be consistent with each other).

iv. Editing should be used to provide information about the survey process, either in the form of quality measures for the current survey or to suggest improvements for future surveys.

v. When starting a survey, some assumptions are made about the data. During editing, it is possible to test the validity of these assumptions. For example, it may become obvious that some range edits were too strict or that some sequencing edits failed too frequently, indicating inappropriate edit rules (or some problems with the questionnaire). This information should be used to adjust the edits in the future (or to improve the design of the questionnaire).

vi. Information on the types of edits performed and the impact of editing on the survey data should be communicated to users.

vii. Quality assurance and quality control procedures should be applied to minimise and correct errors introduced during editing (see **Appendix B - Quality Assurance and Quality Control**).

10.4 Imputation

Imputation is a process used to determine and assign replacement values to resolve problems of missing, invalid or inconsistent data. This is done by changing some of the responses and all of the missing values on the record being edited to ensure that a plausible, internally consistent record is created. Some problems are rectified earlier by contact with the respondent or through manual study of the questionnaire but, as mentioned previously, it is usually impossible to resolve all problems this way so imputation is used to handle the remaining edit failures.

One alternative to imputation is to let the user treat the missing, invalid or inconsistent data. This approach is not recommended. If the user decides to ignore or delete all records with edit failures, this could result in the loss of a great deal of data since many records may be affected. If the user attempts to replace the missing data, this can lead to inconsistent estimates by different users and can undermine the reputation of the statistical agency that conducted the survey. Since the user has access to fewer variables for imputation than the statistical agency it is likely that the user cannot treat the edit failures as well.

Note that in the case of total nonresponse – when very little or no data have been collected – a common approach is to perform a nonresponse weight adjustment (see **Chapter 7 - Estimation)**.

10.4.1 Methods of Imputation

The methods of imputation can be grouped into two categories – stochastic or deterministic. *Deterministic imputation* means that, given the respondent data, there is only one possible imputed value. *Stochastic imputation* has an element of randomness: if imputation were repeated on the same dataset, deterministic methods would impute the same value each time while stochastic methods might impute a different value each time.

Methods of deterministic imputation include:

- deductive;
- mean value;
- ratio/regression;
- sequential hot-deck;
- sequential cold-deck;
- nearest-neighbour.

With the exception of deductive imputation, each deterministic method has a stochastic counterpart. When imputing quantitative data, this can be achieved by adding a random residual from an appropriate distribution or model to the imputed value. The stochastic counterpart of sequential hot-deck is random hot-deck imputation. Stochastic imputation may better preserve the frequency structure of the dataset and may restore more realistic variability in the imputed values than deterministic methods.

With the exception of donor imputation methods where one donor can be used to impute all the missing or inconsistent data for a recipient record, the following methods consider the imputation of one item at a time.

10.4.1.1 Deductive Imputation

Deductive imputation is a method whereby a missing or inconsistent value can be deduced with certainty. Often this is based upon the pattern of responses given to other items on the questionnaire. Usually deductive imputation is performed before any other method. For example, in a sum of four items, if the total is reported to be 100 with two items reported to be 60 and 40 and the other two left blank, then it can be deduced that the 2 missing values are zero.

More commonly, imputation must substitute a value that is not known for certain to be true. The following provides brief descriptions of some common imputation methods. For all of these methods, it is best to group together similar records, as is done for nonresponse weight adjustments (see **Chapter 7 - Estimation**). These groupings are referred to as imputation classes.

10.4.1.2 Mean Value Imputation

With mean value imputation, the missing or inconsistent value is replaced with the mean value for the imputation class. For example, suppose that a questionnaire for a housing survey is missing the value for the monthly rent payment for an apartment. The missing value can be imputed by the average monthly rent payment for respondents who correctly reported their monthly rent (the imputation class could consist of respondents in the same geographic area as the questionnaire requiring imputation).

For the missing data, imputing the mean value is equivalent to applying the same nonresponse weight adjustment to all respondents in the same imputation class. It assumes that nonresponse is uniform and that nonrespondents have similar characteristics to respondents.

While mean value imputation may produce reasonable point estimates (i.e., estimates of totals, means, etc.), it destroys distributions and multivariate relationships by creating an artificial spike at the class mean. This artificially lowers the estimated sampling variance of the final estimates if conventional formulas for the sampling variance are used.

To avoid disrupting the distribution of the data, mean value imputation is often used as a last resort, when no auxiliary information is available or when there are very few records to impute.

10.4.1.3 Ratio/Regression Imputation

Ratio/regression imputation uses auxiliary information or valid responses from other records to create a ratio or regression model that makes use of the relationship that exists between two or more variables. For example, ratio imputation uses the following model:

$$y_i = Rx_i + \varepsilon_i$$

where

y_i is the value of the y variable for the i^{th} unit,
x_i is the value of a related x variable for the i^{th} unit,
R is the slope of the line (i.e., the change in y_i for one unit increase in x_i),
ε_i is assumed to be a random error variable with mean 0 and variance equal to σ^2.

In other words, the model assumes that y_i is approximately linearly related to x_i and that observed values of y_i deviate above and below this line by a random amount, ε_i.

Values of y_i could then be imputed by:

$$\tilde{y}_i = \frac{\bar{y}}{\bar{x}} x_i$$

where:

$\tilde{y}_i$ is the imputed value for variable y for record i,
$\bar{x}$ is the average reported x-value for the imputation class,
$\bar{y}$ is the average reported y-value for the imputation class.

For example, suppose a questionnaire on employment, payrolls and hours contains an invalid entry for the payroll, y_i, for a two-week period, but the number of paid employees, x_i, is properly reported and the industry of the firm is known. Using other questionnaires on the current file within this industry (i.e., imputation class) where both the payroll and the number of paid employees are correctly reported, it is possible to determine the ratio between the payroll and the number of employees. This ratio (payroll to number of employees) can then be applied to the number of employees on the questionnaire requiring imputation to determine a value for the payroll.

The assumption made here is that the ratio or regression model fit to the questionnaires with valid data (i.e., which passed all edits) in the imputation class applies equally well to the questionnaires that failed edits in the imputation class. If this is not true, serious bias can be introduced.

The accuracy of the imputed values depends to a large extent on the existence of variables closely related to the variable being imputed, the degree of sophistication used in the mathematical calculations and whether or not the calculation is restricted to an imputation class or the whole dataset. One advantage of this method is that it may preserve relationships between variables. Also, ratio and regression estimators are likely to generate more stable imputed values than simple averages. However, this method of imputation can artificially induce relationships at the data analysis stage. And, like most other imputation methods (with the exception of deductive imputation), it lowers the estimated sampling variance of the final estimates if conventional variance formulas are used.

Previous value imputation, also called carry-over or carry-forward imputation, is a special case of ratio/regression imputation where the value for the current occasion is imputed by adjusting the previous occasion's value for growth. It is frequently used for quantitative variables in business survey applications.

Ratio and regression estimation are explained in more detail in **Chapter 11 - Data Analysis**.

10.4.1.4 Hot-Deck Imputation

Hot-deck imputation uses information from a donor record that has usually passed all edits to replace missing or inconsistent values for a recipient record. In order to find a donor record that is similar to the recipient record, variables that are related to those requiring imputation are identified to create imputation classes. The set of records in the imputation class which have passed all the edits is the donor pool for records in the imputation class requiring imputation. Hot-deck imputation can be used to impute quantitative or qualitative data, but generally uses qualitative variables to create the imputation classes. The two main types of hot-deck imputation are *sequential* and *random* hot-deck imputation.

With sequential hot-deck, the data are processed sequentially within the imputation class, one record at a time (i.e., sorted in some order). Imputation is performed by replacing the missing item on a questionnaire with the corresponding value from the previous clean responding donor in the imputation class on the data file. Sequential hot-deck is a deterministic imputation method if the same method of sorting is used each time. With random hot-deck imputation, donors are selected at random within the imputation class. Random hot-deck is a stochastic method of imputation.

To illustrate hot-deck imputation, consider the example of imputing the smoking status of a respondent. Suppose that there are two possible responses: smoker and non-smoker. To find a donor record, imputation classes are created based on age group and sex since these variables are related to a person's smoking status. Suppose that the record requiring imputation is for a female in the 15-24 age category. The set of donors is all respondent females aged 15-24 who reported their smoking status. A donor could be selected either randomly (i.e., random hot-deck) or sequentially, by sorting the list of donors and selecting one (i.e., sequential hot-deck).

The advantage of donor imputation methods (hot deck imputation and nearest neighbour, see section

10.4.1.6) is that since similar donors (i.e., companies, households, etc.) should have similar characteristics, the imputed value should be fairly close to the actual value. And with donor imputation, the multivariate distribution of the data can usually be preserved.

There are, however, some disadvantages. One disadvantage of sequential hot-deck is that it often leads to multiple use of the same donor. If one donor is used repeatedly, this can distort the distribution of the data and artificially lower the estimated sampling variance. Another disadvantage is that good auxiliary information and at least a partial response (for example, household income, age, sex, etc.) is needed to create the imputation classes, and these are not always available for the records requiring imputation. Also, care must be taken if the imputation class is small or the nonresponse rate in the imputation class is high as this may lead to no donor being found. (This is true for all methods of imputation that use imputation classes.)

In order to ensure that it is always possible to find a donor record, *hierarchical* hot-deck imputation can be used. Hierarchical imputation uses more than one level of imputation class. When a donor cannot be found for the initial most detailed imputation class, imputation classes are collapsed in a hierarchical fashion until a level is reached where a donor can be found.

For more details on donor imputation, see section 10.4.3.

10.4.1.5 Cold-Deck Imputation

Cold-deck imputation is similar to hot-deck imputation, the difference is that hot-deck imputation uses donors from the current survey, while cold-deck imputation uses donors from another source. Often cold-deck imputation uses historical data from an earlier occasion of the same survey or from a census. If the donors are selected in a random manner, then imputation is stochastic, otherwise it is deterministic.

10.4.1.6 Nearest-Neighbour Imputation

For surveys with largely quantitative data (e.g., business surveys with reported sales and inventory), it may be necessary or preferable to find a donor record by matching on quantitative data. Nearest-neighbour imputation selects a donor record based on matching variables. With this method of imputation, the goal is not necessarily to find a donor record that matches the recipient exactly on the matching variables. Instead, the goal is to find the donor that is closest to the recipient in terms of the matching variables within the imputation class – i.e., to find the nearest neighbour. This closeness is defined by a distance measure between two observations calculated using matching variables (e.g., to impute missing inventory, find the nearest-neighbour with respect to reported sales in the imputation class).

Caution must be exercised when implementing nearest neighbour methods in cases where the scale of the matching variables is quite different (e.g., currency and land areas). In most instances some form of transformation of the variables should be done in order to standardise the scale.

10.4.1.7 Deterministic Imputation with Random Residuals

Deterministic methods for quantitative data can be made stochastic by adding random residuals, for example, by imputing the mean value and adding a random residual:

$$\tilde{y}_i = \bar{y} + e_i *$$

where

$\tilde{y}_i$ is the imputed value for variable y for record i,

$\bar{y}$ is the mean for the imputation class,

e_i * is a random model residual selected from the respondents or drawn from a distribution.

One way to select e_i * is as follows. For the set of respondents in an imputation class, calculate residuals as follows:

$$e_{i(r)} = y_{i(r)} - \bar{y}_r$$

where

$y_{i(r)}$ is the reported y-value for the i^{th} respondent,

$\bar{y}_r$ is the average reported y-value for the imputation class.

Then, one can set e_i * by randomly selecting from all values of $e_{i(r)}$ in the imputation class.

See Kalton and Kasprzyk (1986) for a discussion of approaches to stochastic imputation.

10.4.2 Determining which Values to Impute

Fields that fail an edit rule due to nonresponse or invalid data that are not resolved through respondent follow-up should be imputed. For all other edit failures, since it is best to preserve as much of the respondent data as possible, imputing all edit failures is not recommended. Instead, it is best to impute a minimum set of fields for a record. The Fellegi/Holt framework (Fellegi and Holt, 1976) is one such method of identifying the fields that require imputation. Three criteria are used to determine which fields should be imputed:

- the data in each record should be made to satisfy all edits by changing the fewest possible items of data (fields);
- as much as possible, the frequency structure of the data file should be maintained;
- imputation rules should be derived from the corresponding edit rules without explicit specification.

A key feature of the Fellegi/Holt editing approach is that the edit rules are not specific to a particular imputation method. For each failed edit record, it first proceeds through a step of error localisation in which it determines the minimal set of variables (fields) to impute, as well as the acceptable ranges(s) of values to impute. In most implementations, a single donor is selected from passed edit records by matching on the basis of other variables involved in the edits but not requiring imputation. The method searches for a single exact match and can be extended to take account of other variables not explicitly involved in the edits. Occasionally, no suitable donor can be found and a default imputation method must be available.

For example, consider that a survey has an age/marital status edit to identify people who are married and under the age of 16; and an age/level of education edit to identify people who have a university education and are under the age of 18. Suppose that the survey data have a record that fails both of these edits: there is a ten-year old married woman with a university education. In order for this record to pass both edits, both the individual’s marital status and level of education could be changed, or simply the age could be changed. The Fellegi/Holt framework recommends the latter.

10.4.3 Donor Imputation Issues

The following issues require consideration when developing a donor imputation system (i.e., hot-deck, cold-deck or nearest neighbour imputation):

i. How will a donor record be found for the recipient?

The goal is to find a donor record for each recipient that is similar to the recipient. Serious thought needs to be given to the imputation classes or matching variables used – it is important that there be a strong relationship between the variables requiring imputation and those used to select donors. For methods that require the formation of imputation classes, it is important that imputation classes be large enough that a sufficient number of potential donors are available, but not so large that the records within a donor pool are dissimilar.

ii. Should all fields on a recipient record be imputed from a single donor?

This is desirable because taking all fields from one record preserves joint distributions between variables. For example, in a labour force survey, if both occupation and personal income are flagged for imputation, there is an obvious advantage to imputing both of these variables using the same donor record, since this will preserve the interrelationship between income and occupation. Another advantage of single donor imputation is that since the donor must have passed all edits, it can be used to impute all missing values (i.e., facilitating imputation).

One problem with donor imputation is that if too many matching variables are used (i.e., variables used to create imputation classes in the case of hot-deck and cold-deck), there is the risk that no suitable donor may be found. Another problem is that the matching variables used to impute one field may not be suitable for another, particularly if the variables requiring imputation are not related. Consider a multi-purpose health survey in which a person’s height and the number of cigarettes smoked daily are flagged for imputation. In this case, a different set of matching fields might be appropriate for each field requiring imputation.

Often with donor procedures, the imputation is broken down into several stages with certain sets of fields being imputed at each stage. As a result, several donors may be involved in completing a single deficient record. If this is of concern, certain key imputed fields can be used to create imputation classes in succeeding stages to preserve internal integrity.

iii. Can a donor record be used to impute more than one recipient?

If several recipient records are imputed by the same donor, the impact on the final survey estimates can be significant. Limiting the number of times a record is used as a donor has the effect of spreading the donor usage around and avoiding over-use of a particular donor. If the response rate in a particular imputation class is very low, limiting the number of times a donor is used may result in some matches not being very good (i.e., the donor may not be very similar to the recipient record) and may result in some recipients not finding a donor. At the same time, over-use of a donor (especially if the donor has unique characteristics making it quite different from others in the population) can have a substantial effect on survey estimates. If there is no limit placed on the number of times that a record can be used as a donor, there should be a method of identifying records which are used often as donors. If any of these records have suspicious or outlying fields, some investigation may be warranted to see if the final survey results were distorted by the imputation process.

iv. What is done for recipients for whom a suitable donor cannot be found?

There are some recipients for whom a donor record cannot be found. Usually a back-up procedure is used for such recipients (e.g., hierarchical hot-deck or cold-deck imputation or mean value imputation).

v. Does the survey deal with quantitative or qualitative data?

Some imputation methods are more appropriate for qualitative variables while others are more appropriate for quantitative variables. Hot-deck methods were developed in order to treat qualitative data while nearest neighbour imputation was developed for quantitative data. Nowadays, both methods have found their use in both situations, including mixed problems.

10.4.4 Variance Estimation for Imputed Data

All of the imputation methods presented produce a single imputed value for each missing or inconsistent value. All distort, to some extent, the original distribution of values for a variable and can lead to inappropriate variance estimates when standard variance estimators are used. This can lead to confidence intervals that are too narrow and spurious declarations of significance. The extent of the distortion varies considerably depending on the amount of imputation done and the method used.

When imputation is carried out – assuming there are no other nonsampling errors – the variance of an estimate has two components: one due to sampling (the sampling variance) and one due to imputation (the imputation variance). The sampling variance component is usually underestimated in the presence of imputed data since traditional formulas assume a 100% response rate. One benefit of stochastic imputation methods is that they add some noise to the completed dataset. As a result, when stochastic imputation is used, the sampling variance of an estimate can be, most of the time, correctly estimated using traditional methods. However, to determine the total variance of the estimate, the imputation variance must still be estimated.

It is important to estimate both the sampling and the imputation components of the total variance not only to draw correct inferences, but also to know the relative importance of the sampling variance and imputation variance. This can help inform users of data quality and help allocate survey resources between sample size and edit/imputation processes.

Multiple imputation is a method, proposed by Rubin (1987), which addresses this problem by properly imputing several, say *m*, times, each value requiring imputation (for a definition of proper imputation, see Rubin, 1987 or Binder and Weiman, 1996). From the completed dataset, *m* estimates can be produced for the item. From these, a single combined estimate is produced along with a pooled variance estimate that expresses the uncertainty about which value to impute. However, multiple imputation requires more work for data processing, storage and computation of estimates.

In the case of single imputation, the important variance estimation methods have been extended to the case of data files containing imputed data. The approaches are described in Särndal (1992), Rao and Shao (1992), Rao and Sitter (1995) and Gagnon et al. (1996). A comparison of the methods is presented in Lee, Rancourt and Särndal (1994, 2001).

10.4.5 Guidelines for Imputation

Although imputation can improve the quality of the final data, care should be taken when choosing an appropriate imputation methodology. One risk with imputation is that it can destroy reported data to create records that fit preconceived models that may later turn out to be incorrect. The suitability of the imputation method depends upon the survey, its objectives, available auxiliary information and the nature of the error.

The following are some guidelines for imputation:

i. Imputed records should closely resemble the failed edit record. This is usually achieved by imputing the minimum number of variables, thereby preserving as much respondent data as possible. The underlying assumption (which is not always true in practice) is that a respondent is more likely to make only one or two errors than several.

ii. Good imputation has an audit trail for evaluation purposes. Imputed values should be flagged and the methods and sources of imputation clearly identified. The unimputed and imputed values of the record's fields should be retained so that the degree and effects of imputation can be evaluated.

iii. Imputed records should satisfy all edits.

iv. The imputation methods should be chosen carefully, considering the type of data to be imputed.

v. The imputation method should aim to reduce the nonresponse bias and preserve relationships between items as much as possible (i.e., the fit of the model underlying the imputation method should be assessed).

vi. The imputation system should be thought out, specified, programmed and tested in advance.

vii. The process should be automated, objective, reproducible and efficient.

viii. The imputation system should be able to handle any pattern of missing or inconsistent fields.

ix. For donor imputation methods, the imputed record should closely resemble the donors selected. This will tend to ensure that the combination of imputed and unimputed responses for the imputed record not only satisfy the edits but are also plausible.

10.4.6 Evaluation of Imputation Procedures

The size of the survey and its budget influence how much work can be carried out to measure imputation effects. However, users of survey data should always have at least some basic information about the degree to which survey data were modelled or estimated by imputation. In evaluating the imputation procedure, the most relevant concerns are bias and the imputation variance of the survey estimates.

If the survey budget is large, one option is to do a complete study of the effect of imputation, looking at the survey estimates with and without imputation. Cases in which discrepancies are large can be investigated and an attempt can be made to discover any bias that may exist due to imputation.

If this is not possible, at the very least, imputation should be monitored so users can be told how much

imputation was done and where. At the end of imputation, it may be useful to produce the following (some are specific to a particular method):
- the number of records which were imputed (i.e., the number of recipient records);
- the number of times each field was imputed and by what method;
- the number of records eligible to be used as donors;
- the number of records actually used as donors and the number of recipients each of these donor records imputed;
- a list (or file) indicating which donors were used for each recipient (to trace the sources of unusual imputed records);
- a list of all records for which imputation failed (e.g., because no donor was found).

It should be noted that the above information is useful for the redesign of a survey or the implementation of a similar survey. This information could be instrumental in improving the edit and imputation system, the survey questionnaire and the collection procedures. For example, if the answer to a question has a high imputation rate, this could indicate a poorly worded question (and poor data quality).

10.5 Identification and Treatment of Outliers

The identification of outliers can be viewed as a type of editing whereby suspicious records are identified. In **Chapter 7 - Estimation**, ***an outlier was defined as an observation or subset of observations that appears to be inconsistent with the remainder of the dataset***. The distinction should also be made between extreme and influential observations. An observation is influential if the combination of the reported value and the final survey weight have a large influence on the estimate. However, an extreme value need not be influential, and vice versa.

It is possible to make the distinction between univariate outliers and multivariate outliers. An observation is a univariate outlier if it is an outlier with respect to one variable. An observation is a multivariate outlier if it is an outlier with respect to two or more variables. For example, it might not be unusual to find a person with a height of 2 metres *or* a person weighing 45 kg. But someone who is *both* 2 metres high and only weighs 45 kg is an example of a multivariate outlier.

Outliers are found in every survey for almost every variable of interest. There are many reasons why outliers exist:

i. There are errors in the data (e.g., data capture errors).

ii. The outliers can be thought of as arising from another model or distribution. For example, most of the data might be considered to arise from a normal distribution, but the outliers might be thought of as arising from an exponential distribution.

iii. The outlier may be due to inherent variability of the data. What appears to be a suspicious value may simply arise from the inherent variability of the dataset – in other words, it may be a legitimate but extreme observation from the distribution. This can occur when the population is skewed, which is common for business surveys. For example, the distribution of the sales by the size of the company is typically skewed – a few very large companies often contribute to a large portion of the overall sales.

10.5.1 Identification of Outliers

The most popular outlier detection methods are univariate methods because they are simpler than multivariate methods. Traditionally, outliers are detected by measuring their relative distances from the centre of the data. For example, if $y_1, y_2, ..., y_n$ are the observed sample data and m and s are measures of the central tendency and spread of the data, respectively, then the relative distance, d_i, of y_i from the centre of the data can be defined by:

$$d_i = \frac{| y_i - m |}{s}$$

If d_i exceeds a predetermined cut-off value, then the observation is considered to be an outlier.

Alternatively, a tolerance interval can be given by:

$$(m - c_L s, m + c_U s)$$

where c_L and c_U are predetermined lower and upper bound values. If the population is skewed, unequal values of c_L and c_U are used. Observations falling outside of this interval are declared to be outliers.

The sample mean and variance are the statistics most frequently used to estimate the centre of the data and the spread of the data. However, since they are sensitive to outliers, they are a poor choice for the purpose of outlier detection. For example, the sample mean is shifted towards outliers if they are clustered on one side and the sample variance is greatly inflated by outliers. Therefore, the relative distance values of some outliers may appear rather small and the procedure may fail to detect them. This problem is referred to as the *masking effect*.

For this reason, one of the most popular methods of outlier detection is the *quartile method* which uses the median to estimate the centre and quartile ranges to estimate the spread of the weighted data, since these statistics are more robust (i.e., insensitive) to outliers. Quartiles divide the data into four parts: 25% of the data points are less than the first quartile, $q_{.25}$, 50% of the data points are less than the second quartile (or the median), $q_{.5}$, and 75% of the data points are less than the third quartile, $q_{.75}$. (The median and quartile ranges are discussed further in **Chapter 11 - Data Analysis**).

The lower and upper quartile ranges, h_L and h_U, are defined as:

$$h_L = q_{.5} - q_{.25}$$
$$h_U = q_{.75} - q_{.5}$$

The tolerance interval then becomes:

$$(q_{.5} - c_L h_L, q_{.5} + c_U h_U)$$

with some predetermined values for c_L and c_U which can be chosen by examining past data or based on past experience. Any observation falling outside of this interval is considered to be an outlier.

For more information on outlier detection methods, see Barnett and Lewis (1995).

10.5.2 Treatment of Outliers

Outliers detected at the editing stage of the survey process can be treated in various ways. In a manual editing system, the potential outliers are examined or followed-up and changed if they are in fact errors. In an automated editing system, outliers are often imputed. In some cases, no special treatment of outliers is performed if it is believed that they are not influential.

Outliers that are not treated in editing can be dealt with at estimation. Simply ignoring untreated outliers can result in poor estimates and an increase in the sampling variance of the estimates. Assigning the outlier a weight of one (to reduce its effect on estimates) can bias the results. The goal of outlier treatment is to decrease the impact that the outlier has on the sampling variance of the estimate without introducing too much bias.

The following approaches can be used to treat outliers during estimation:
- change the value;
- change the weight;
- use robust estimation.

i. Change the value

One example of treatment of an extreme value is Winsorization. Winsorization consists of recoding the top *k* values.

Recall that in simple random sampling (assuming a 100% response), the usual unbiased estimator of the population total *Y* is given by:

$$\hat{Y} = \frac{N}{n}\sum_{i=1}^{n} y_i$$

where *i* is the i^{th} unit in a sample of size *n*.

Assuming that y_i , *i=1,2, ..., n* are ordered values of y_i in a sample of size *n* from a population of size *N* and the *k* largest values are considered to be outliers, the one-sided *k*-times-Winsorized estimator is defined by replacing these outlier values by the $n\text{-}k^{th}$ largest value, y_{n-k}. That is:

$$\hat{Y}_W = \frac{N}{n}(\sum_{i=1}^{n-k} y_i + ky_{n-k})$$

Note that Winsorization tends to be used for one-variable situations, so it is rarely used in the multivariate sample survey situation.

ii. Change the weight

Another option is to reduce the weights of outliers so that they have a smaller impact. An example is to set the weight of the outlier to one or zero. This is seldom done because of its dramatic effect on estimates, especially for skewed populations. It can lead to serious bias – usually underestimation. For example, if two large companies contribute to the majority of retail sales, and one is identified as an outlier, eliminating this company from the estimates will seriously underestimate the total retail sales. Several estimators with reduced weights for outliers have been proposed, see Rao (1970), Hidiroglou and Srinath (1981).

iii. Robust Estimators

In classical estimation theory, the estimator of a population parameter is based on an assumption of some distribution. Typically, it is assumed that the estimator has a normal sampling distribution (see **Chapter 7 - Estimation** for the definition of a sampling distribution). The usual sample mean and variance estimators are optimal under normality. However, these estimators are extremely sensitive to outliers. Robust estimators are estimators that are less sensitive to distributional assumptions. For example, the median is more robust than the mean; interquartile ranges are more robust than the usual variance estimator. Many complex robust estimators have been proposed over the years, including Huber's M-estimators, Huber (1964).

For more information on robust estimators and outlier detection in general, see Kish (1965), Barnett and Lewis (1995), Rousseeuw and Leroy (1987), Lee et al. (1992), or Lee (1995). For more information on the mean versus the median, see **Chapter 11 - Data Analysis**.

10.6 Generating Results/ Creating a Database

After coding, data capture, edit and imputation and the detection of outliers, the data are almost ready for estimation, analysis, and publication. Before proceeding however, the format for storing the data must be established. The two main choices are a database or a flat file. A flat file is a computerised 2-dimensional arrangement of records and their corresponding values. It is easily transferable between platforms and can be read using spreadsheet software or statistical software. The main drawback to a flat file is that most statistical software must have data stored in a special format to facilitate speedy processing. When using a flat file, this special format is continually being recreated, an unnecessary inefficiency. If the data are stored in an appropriate database format, it is possible to use certain statistical and database software without needing to recreate the file. Queries can be run directly on the database. However the choice of database format may restrict the choice of statistical and database software that can be used for tabulation and analysis. It may be best to create a flat file as well as several different database files with the survey results.

Once the format for storing the data has been selected, the final (estimation) weights are calculated and the planned tabulations are produced (see **Chapter 7 - Estimation** for a description of how to calculate final weights). Usually, computer programs are written to calculate the weights and produce the tabulations. More sophisticated data analysis may also be performed. Before releasing the data, they must be examined to ensure that the confidentiality of respondents is not violated. This process, called disclosure control, may result in the suppression of some survey data. For more information on data analysis and disclosure control, see **Chapter 12 - Data Dissemination**.

10.7 Automated versus Manual Processing

In the past, almost all aspects of survey processing were done manually. However, computers now make it possible to process the data in an automated manner.

The benefits of automating coding and data capture, optical scanners, computer-assisted data collection methods and pre-coding the questionnaire have already been discussed. The arguments for using computers at collection apply to using computers for processing. Experience has shown that, in general, computers are much better at processing large volumes of information than are people. Automation can improve all aspects of data quality, in particular timeliness – it can produce results faster with fewer resources. It also ensures that procedures (for example, editing and imputation) are implemented in a consistent manner, thereby reducing nonsampling errors. Also, it is possible to use more complex methods (for example for editing, imputation, coding, quality control, etc.) and it is possible to track processing – produce reports on each processing step (e.g., number of edits and imputations performed). Automation also makes it easier to monitor and control the quality of the processing.

However, there are some drawbacks to automation, including:

- specifications must be written for each system that is to be automated. Developing a computer program for each procedure (e.g., imputation) can be time-consuming;
- operators must be trained on the software;
- automated coding, editing and imputation does not take into account any additional knowledge held by the operator.

Despite these drawbacks, it is wise to automate procedures as much as possible. The additional time required at the outset is more than offset by the time saved later in the survey process (particularly if the survey is to be repeated). As a minimum data should always be captured and then weighted and estimation performed by computer. The consistency that results from automation is important to attaining accurate and measurable results. It is also wise to take advantage of existing systems and processes, automated systems for coding, etc.

10.8 Summary

Processing is an important survey activity that converts questionnaire responses into a format suitable for tabulation and data analysis. Processing is costly, time-consuming, resource-intensive and has an impact on the final quality of the data. Automation can make it more efficient and improve the final quality of the data.

Processing usually begins with a preliminary clean-up of the questionnaire, followed by coding and data capture. This is usually followed by more detailed editing to identify missing or inconsistent data, and imputation to provide plausible substitutes for these values. Outlier detection may also be performed to identify suspicious values. Once the data are complete, consistent and valid, they are typically stored in a database.

Bibliography

Bankier, M., M. Lachance, and P. Poirier. 1999. A Generic Implementation of the Nearest neighbour imputation method. *Proceedings of the Survey Research Methods Section*. American Statistical Association. 548-553.

Barnett, V., and T. Lewis. 1995. *Outliers in Statistical Data*. John Wiley and Sons, Chichester.

Latouche, M. and J.-M. Berthelot. 1992. Use of a Score Function to Prioritize and Limit Recontacts in Editing Business Surveys. *Journal of Official Statistics,* 8(3): 389-400.

Binder, D., and S. Weimin. 1996. Frequency Valid Multiple Imputation for Surveys with a Complex Design. *Proceedings for the Section on Survey Research Methods of the American Statistical Association,* 1: 281-286.

Brick, J.M. and G. Kalton. 1996. Handling Missing Data in Survey Research. *Statistical Mathematics in Medical Research,* 5: 215-238.

Boucher, L, J.-P. S. Simard and J.-F. Gosselin. 1993. Macro-Editing, a Case Study: Selective Editing for the Annual Survey of Manufacturers Conducted by Statistics Canada, *Proceedings of the International Conference on Establishment Surveys.* American Statistical Association. Virginia.

Chambers, R.L. 1986. Outlier Robust Finite Population Estimation. *Journal of the American Statistical Association,* 81: 1063-1069.

Cox, B.G., D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge and P.S. Kott, eds. 1995. *Business Survey Methods.* John Wiley and Sons, New York.

Dielman, L. and M.P. Couper. 1995. Data Quality in a CAPI Survey: Keying Errors. *Journal of Official Statistics,* 11(2): 141-146.

Dolson, D. 1999. *Imputation Methods*. Statistics Canada.

Fay, R.E. 1996. Alternative Paradigms for the Analysis of Imputed Survey Data. *Journal of the American Statistical Association,* 91: 490-498.

Fellegi, I.P. and D. Holt. 1976. A Systematic Approach to Automatic Edit and Imputation. *Journal of the American Statistical Association,* 71: 17-35.

Gagnon, F., H. Lee, E. Rancourt and C.E. Särndal. 1996. Estimating the Variance of the Generalized Regression Estimation in the Presence of Imputation for the Generalized Estimation System. *Proceedings of the Survey Methods Section*. Statistical Society of Canada. 151-156.

Granquist, L. 1984. On the Role of Editing. *Statistisk tidskrift,* 2: 105-118.

Granquist, L. and J. Kovar. 1997. Editing of Survey Data: How Much is Enough? In Lyberg, L., P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz and D. Trewin, eds. 1997. *Survey Measurement and Process Quality*. John Wiley and Sons, New York.

Hidiroglou, M.A. 1999. Course notes for *Methods for Designing Business Survey*. Sponsored by the IASS: 52nd session of the ISI. University of Jyvaskylä, Finland.

Hidiroglou, M.A. and J.-M. Berthelot. 1986. Statistical Edit and Imputation for Periodic Surveys. *Survey Methodology,* 12(1): 73-84.

Hidiroglou, M.A. and K.P. Srinath. 1981. Some Estimators of a Population Total Containing Large Units. *Journal of the American Statistical Association,* 78: 690-695.

Huber, P.J. 1964. Robust Estimation of a Location Parameter. *Annals of Mathematical Statistics,* 35: 73-101.

Kalton, G. and D. Kasprzyk. 1982. Imputation for Missing Survey Responses. *Proceedings of the Section on Survey Research Methods*. American Statistical Association. 23-31.

Kalton, G. and D. Kasprzyk, D. 1986. The Treatment of Missing Survey Data. *Survey Methodology,* 12(1): 1-16.

Kish, L. 1965. *Survey Sampling*. John Wiley and Sons, New York.

Kovar, J.G., J. MacMillan and P. Whitridge. 1988. *Overview and Strategy for the Generalized Edit and Imputation System. (Updated February 1991)*. Statistics Canada. BSMD-88-007E/F.

Latouche, M. et J.-M. Berthelot. 1992. Use of a Score Function to Prioritize and Limit Recontacts in Editing Business Surveys. *Journal of Official Statistics*, 8(3): 389-400.

Lee, H., E. Rancourt and C.E. Särndal. 1994. Experiments with Variance Estimation from Survey Data with Imputed Values. *Journal of Official Statistics,* 10(3): 231-243.

Lee, H., E. Rancourt and C.E. Särndal. 2001. Variance Estimation from Survey Data under Single Value Imputation. *Survey Nonresponse*. John Wiley and Sons, New York.

Lee, H. 1995. Outliers in Business Surveys. In *Business Survey Methods.* Cox, B.G., D. A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge and P.S. Kott, eds. John Wiley and Sons. New York. 503-526.

Lyberg, L. and P. Dean. 1992. *Automated Coding of Survey Responses: An International Review*. Presented at the Conference of European Statisticians. Washington, D.C.

Moser, C.A. and G. Kalton. 1971. *Survey Methods in Social Investigation*. Heinemann Educational Books Limited, London.

Raj, D. 1972. *The Design of Sample Surveys*. McGraw-Hill Series in Probability and Statistics, New York.

Rancourt, E., H. Lee and C.E. Särndal. 1993. *Variance Estimation Under More than One Imputation Method*. Proceedings of the International Conference on Establishment Surveys, American Statistical Association, 374-379.

Rao, C.R. 1970. Estimation of Heteroscedastic Variances in Linear Models. *Journal of the American Statistical Association,* 65: 161-172.

Rao, J.N.K. and J. Shao. 1992. Jackknife Variance Estimation with Survey Data under Hot-deck Imputation. *Biometrika,* 79: 811-822.

Rao, J.N.K. and R.R. Sitter. 1995. Variance Estimation under Two-Phase Sampling with Application to Imputation for Missing Data. *Biometrika,* 82: 453-460.

Rao, J.N.K. 1996. On Variance Estimation with Imputed Survey Data. *Journal of the American Statistical Association,* 91: 499-506.

Rousseeuw, P.J. and A.M. Leroy. 1987. *Robust Regression and Outlier Detection*. John Wiley and Sons, New York.

Rubin, D.B. 1987. *Multiple Imputation for Nonresponse in Surveys*. John Wiley and Sons, New York.

Rubin, D.B. 1996. Multiple Imputation after 18+ Years. *Journal of the American Statistical Association,* 91: 473-489.

Sande, I.G. 1979. A Personal View of Hot-deck Imputation Procedures. *Survey Methodology*, 5(2): 238-258.

Sande, I.G. 1982. Imputation in Surveys: Coping with Reality. *The American Statistician,* 36(3). Part 1: 145-152.

Särndal, C.E. 1992. Methods for Estimating the Precision of Survey Estimates when Imputation has Been Used. *Survey Methodology,* 18(2): 242-253.

Särndal, C.E., B. Swensson and J. Wretman. 1992. *Model Assisted Survey Sampling*. Springer Verlag, New York.

Shao, J. and R.R. Sitter. 1996. Bootstrap for Imputed Survey Data. *Journal of the American Statistical Association,* 94: 254-265.

Statistics Canada. 1987. *Quality Guidelines*. Second Edition.

Statistics Canada. 1990. Course notes for *Survey Skills Development Course*.

Statistics Canada. 1998. *Statistics Canada Quality Guidelines*. Third Edition.12-539-X1E.

Statistics Canada. 1998. Course notes for *Surveys from Start to Finish*. Course code 416.

Wenzowski, M.J. 1988. Advances in Automated Coding and Computer-Assisted Coding Software at Statistics Canada. Proceedings of the 1996 Annual Research of the U.S. Census Bureau.

Yung, W. and J.N.K. Rao. 2000. Jackknife Variance Estimation under Imputation for Estimators using Poststratification Information. *Journal of the American Statistical Association,* 95: 903-915.