Chapter 6 - Sample Designs

6.0 Introduction

Chapter 3 - Introduction to Survey Design stated that during the planning phase the statistical agency must decide whether to conduct a census or sample survey. If the decision is a sample survey, then the agency needs to plan how to select the sample. ***Sampling is a means of selecting a subset of units from a population for the purpose of collecting information for those units to draw inferences about the population as a whole***.

There are two types of sampling: non-probability and probability sampling. The one chosen depends primarily on whether reliable inferences are to be made about the population. Non-probability sampling, discussed in section 6.1, uses a subjective method of selecting units from a population. It provides a fast, easy and inexpensive way of selecting a sample. However, in order to make inferences about the population from the sample, the data analyst must assume that the sample is representative of the population. This is often a risky assumption to make in the case of non-probability sampling.

Probability sampling, discussed in section 6.2, involves the selection of units from a population based on the principle of randomisation or chance. Probability sampling is more complex, time consuming and usually more costly than non-probability sampling. However, because units from the population are randomly selected and each unit's inclusion probability can be calculated, reliable estimates can be produced along with estimates of the sampling error, and inferences can be made about the population.

There are several different ways in which a probability sample can be selected. The design chosen depends on a number of factors such as: the available survey frame, how different the population units are from each other (i.e., their variability) and how costly it is to survey members of the population. For a given population, a balance of sampling error with cost and timeliness is achieved through the choice of design and sample size.

The purpose of this chapter is to present different probability sample designs and factors to consider when determining which one is appropriate for a specific survey. For details on factors affecting sample size, see **Chapter 8 - Sample Size Determination and Allocation**.

6.1 Non-Probability Sampling

Non-probability sampling is a method of selecting units from a population using a subjective (i.e., non-random) method. Since non-probability sampling does not require a complete survey frame, it is a fast, easy and inexpensive way of obtaining data. The problem with non-probability sampling is that it is unclear whether or not it is possible to generalise the results from the sample to the population. The reason for this is that the selection of units from the population for a non-probability sample can result in large biases.

For example, a common design is for the interviewer to subjectively decide who should be sampled. Since the interviewer is most likely to select the most accessible or friendly members of the population, a large portion of the population has no chance of ever being selected, and this portion of the population is likely to differ in a systematic manner from those selected members. Not only can this bias the results of the

survey, it can falsely reduce the apparent variability of the population due to a tendency to select 'typical' units and eliminate extreme values. By contrast, probability sampling avoids such bias by randomly selecting units (see section 6.2).

Due to selection bias and (usually) the absence of a frame, an individual's inclusion probability cannot be calculated for non-probability samples, so there is no way of producing reliable estimates or estimates of their sampling error. In order to make inferences about the population, it is necessary to assume that the sample is representative of the population. This usually requires assuming that the characteristics of the population follow some model or are evenly or randomly distributed over the population. This is often dangerous due to the difficulty of assessing whether or not these assumptions hold.

Non-probability sampling is often used by market researchers as an inexpensive and quick alternative to probability sampling, but it is not a valid substitute for probability sampling for the reasons delineated above. So, why bother with non-probability sampling? Non-probability sampling can be applied to studies that are used as:

- an idea generating tool;
- a preliminary step towards the development of a probability sample survey;
- a follow-up step to help understand the results of a probability sample survey.

For example, non-probability sampling can provide valuable information in the early stages of an investigation. It can be used for exploratory or diagnostic studies to gain insights into people's attitudes, beliefs, motivations and behaviours. Sometimes non-probability sampling is the only viable option – for example, sampling volunteers may be the only way of obtaining data for medical experiments.

Non-probability sampling is often used to select individuals for focus groups and in-depth interviews. For example, at Statistics Canada, non-probability sampling is used to test Census of Population questions, to ensure that the questions asked and concepts used are clear to respondents. In addition, if the content of a question is deemed to be controversial, subpopulations may be selected and tested. If, through the use of focus groups, these questions can be made acceptable to these people, they may be acceptable for all members of the population. (For more on focus groups, see **Chapter 5 - Questionnaire Design**.)

Another example of the use of non-probability sampling is for preliminary studies. If a new survey is being designed to cover a field about which very little is known, pilot surveys often use non-probability designs. For example, consider the relatively new industry of web page designer. Suppose nothing is known about the number of people working in the industry, how much they earn, or other details of the profession. A pilot survey could be designed, with questionnaires sent to a few persons known to design Web pages. Feedback from the questionnaire may provide an idea about their earnings, and may indicate that many web designers work out of their homes, are only listed under their personal phone numbers and advertise exclusively on the Internet.

The **advantages** of non-probability sampling are that:

i. It is quick and convenient.

As a general rule, non-probability samples can be quickly drawn and surveyed: it is very easy to simply walk outside and ask questions of the first hundred people encountered on the street.

ii. It is relatively inexpensive.

It usually only takes a few hours of an interviewer's time to conduct such a survey. As well, non-probability samples are generally not spread out geographically, therefore travelling expenses for interviewers are low.

iii. It does not require a survey frame.

iv. It can be useful for exploratory studies and survey development.

The **disadvantages** of non-probability sampling are that:

i. In order to make inferences about the population it requires strong assumptions about the representativeness of the sample. Due to the selection bias present in all non-probability samples, these are often dangerous assumptions to make. When inferences are to be made, probability sampling should be performed instead.

ii. It is impossible to determine the probability that a unit in the population is selected for the sample, so reliable estimates and estimates of sampling error cannot be computed.

The following sections describe five different types of non-probability sampling schemes: haphazard sampling, volunteer sampling, judgement sampling, quota sampling and modified probability sampling. Network or snowball sampling, which is less commonly used, is presented in section 6.3.

6.1.1 Haphazard Sampling

Units are selected in an aimless, arbitrary manner with little or no planning involved. Haphazard sampling assumes that the population is homogeneous: if the population units are all alike, then any unit may be chosen for the sample. An example of haphazard sampling is the 'man in the street' interview where the interviewer selects any person who happens to walk by. Unfortunately, unless the population is truly homogeneous, selection is subject to the biases of the interviewer and whoever happened to walk by at the time of sampling.

6.1.2 Volunteer Sampling

With this method, the respondents are volunteers. Generally, volunteers must be screened so as to get a set of characteristics suitable for the purposes of the survey (e.g., individuals with a particular disease). This method can be subject to large selection biases, but is sometimes necessary. For example, for ethical reasons, volunteers with particular medical conditions may have to be solicited for some medical experiments.

Another example of volunteer sampling is callers to a radio or television show, when an issue is discussed and listeners are invited to call in to express their opinions. Only the people who care strongly enough about the subject one way or another tend to respond. The silent majority does not typically respond, resulting in a large selection bias. Volunteer sampling is often used to select individuals for focus groups or in-depth interviews (i.e., for qualitative testing, where no attempt is made to generalise to the whole population).

6.1.3 Judgement Sampling

With this method, sampling is done based on previous ideas of population composition and behaviour. An expert with knowledge of the population decides which units in the population should be sampled. In other words, the expert purposely selects what is considered to be a representative sample.

Judgement sampling is subject to the researcher's biases and is perhaps even more biased than haphazard sampling.

Since any preconceptions the researcher has are reflected in the sample, large biases can be introduced if these preconceptions are inaccurate. However, it can be useful in exploratory studies, for example in selecting members for focus groups or in-depth interviews to test specific aspects of a questionnaire.

6.1.4 Quota Sampling

This is one of the most common forms of non-probability sampling. Sampling is done until a specific number of units (quotas) for various subpopulations has been selected. Quota sampling is a means for satisfying sample size objectives for the subpopulations.

The quotas may be based on population proportions. For example, if there are 100 men and 100 women in the population and a sample of 20 are to be drawn, 10 men and 10 women may be interviewed. Quota sampling can be considered preferable to other forms of non-probability sampling (e.g., judgement sampling) because it forces the inclusion of members of different subpopulations.

Quota sampling is somewhat similar to stratified sampling in that similar units are grouped together (see section 6.2.6 for stratified sampling). However, it differs in how the units are selected. In probability sampling, the units are selected randomly while in quota sampling a non-random method is used – it is usually left up to the interviewer to decide who is sampled. Contacted units that are unwilling to participate are simply replaced by units that are, in effect ignoring nonresponse bias.

Market researchers often use quota sampling (particularly for telephone surveys) instead of stratified sampling to survey individuals with particular socio-economic profiles. This is because compared with stratified sampling, quota sampling is relatively inexpensive and easy to administer and has the desirable property of satisfying population proportions. However, it disguises potentially significant selection bias.

As with all other non-probability sample designs, in order to make inferences about the population, it is necessary to assume that persons selected are similar to those not selected. Such strong assumptions are rarely valid.

6.1.5 Modified Probability Sampling

Modified probability sampling is a combination of probability and non-probability sampling. The first stages are usually based on probability sampling (see the following section). The last stage is a non-probability sample, usually a quota sample. For example, geographical areas may be selected using a probability design, and then within each region, a quota sample of individuals may be drawn.

6.2 Probability Sampling

Probability sampling is a method of sampling that allows inferences to be made about the population based on observations from a sample. In order to be able to make inferences, the sample should not be subject to selection bias. Probability sampling avoids this bias by randomly selecting units from the population (using a computer or table of random numbers). It is important to note that random does not mean arbitrary. In particular, the interviewers do not arbitrarily choose respondents since then sampling would be subject to their personal biases. Random means that selection is unbiased – it is based on chance. With probability sampling, it is never left up to the discretion of the interviewer to subjectively decide who should be sampled.

There are two main criteria for probability sampling: one is that the units be randomly selected, the second is that all units in the survey population have a non-zero inclusion probability in the sample and that these probabilities can be calculated. It is not necessary for all units to have the same inclusion probability, indeed, in most complex surveys, the inclusion probability varies from unit to unit.

There are many different types of probability sample designs. The most basic is simple random sampling and the designs increase in complexity to encompass systematic sampling, probability-proportional-to-size sampling, cluster sampling, stratified sampling, multi-stage sampling, multi-phase sampling and replicated sampling. Each of these sampling techniques is useful in different situations. If the objective of the survey is simply to provide overall population estimates and stratification would be inappropriate or impossible, simple random sampling may be the best. If the cost of survey collection is high and the resources are available, cluster sampling is often used. If subpopulation estimates are also desired (such as estimates by province, age group, or size of business), stratified sampling is usually performed.

Most of the more complex designs use auxiliary information on the survey frame to improve sampling. If the frame has been created from a previous census or from administrative data, there may be a wealth of supplementary information that can be used for sampling. For example, for a farm survey, the statistical agency may have the size of every farm in hectares from the last agricultural census. For a survey of people, information (e.g., age, sex, ethnic origin, etc.) may be available for everyone from the last population census. For a business survey, the statistical agency may have administrative information such as the industry (e.g., retail, wholesale, manufacturing), the type of business (e.g., food store), the number of employees, etc. In order for the auxiliary information to improve sampling, there must be a correlation between the auxiliary data and the survey variables.

The main **advantage** of probability sampling is that since each unit is randomly selected and each unit's inclusion probability can be calculated reliable estimates and an estimate of the sampling error of each estimate can be produced. Therefore, inferences can be made about the population. In fact, with a probability design, a relatively small sample can often be used to draw inferences about a large population.

The main **disadvantages** of probability sampling are that it is more difficult, takes longer and is usually more expensive than non-probability sampling. In general, the expense of creating and maintaining a good quality frame is substantial. And because probability samples tend to be more spread out geographically across the population than non-probability samples, sample sizes are generally much larger and data collection is often more costly and difficult to manage. However, for a statistical agency, the ability to make inferences from a probability sample usually far outweighs these disadvantages.

For the qualities of a good frame, see **Chapter 3 - Introduction to Survey Design**. For more information on the uses of administrative data, see **Appendix A - Administrative Data.**

6.2.1 Statistical Efficiency

Simple Random Sampling (SRS) is used as a benchmark for evaluating the efficiency of other sampling strategies. In order to understand the concept of efficient sampling, some definitions are presented here.

A parameter is a population characteristic that the client or data user is interested in estimating, for example the population average, proportion or total***. An estimator is a formula by which an estimate of the parameter is calculated from the sample and an estimate is the value of the estimator using the data from the realised sample***. ***The sampling strategy is the combination of the sample design and estimator used.***

For example, the parameter of interest might be the population average, $\overline{Y}$, which is calculated as follows:

$$\overline{Y} = \sum_{i \in U} \frac{y_i}{N}$$

where y_i is the value of the variable y for the i^{th} unit, U is the set of units in the population and there are N units in the population.

For an SRS with 100% response rate, the usual – but not the only – estimator for the population average is:

$$\hat{\overline{Y}} = \sum_{i \in S_r} \frac{y_i}{n}$$

where S_r is the set of respondents in the sample and there are n units in the sample. The value of $\sum_{i \in S_r} \frac{y_i}{n}$ for a particular sample is called the estimate.

Estimates calculated from different samples differ from one another. The ***sampling distribution of an estimator is the distribution of all the different values that the estimator can have for all possible samples from the same design from the population***. This distribution thus depends on the sampling strategy.

Estimators have certain desirable properties. One is that the estimator be unbiased or approximately unbiased. ***An estimator is unbiased if the average estimate over all possible samples is equal to the true value of the parameter***. Another desirable property of an estimator is that the sampling distribution be concentrated as closely as possible about the average (i.e., that the sampling error be small). The sampling error of an estimator is measured by its sampling variance, which is calculated as the average squared deviation about its mean calculated across all possible samples generated from the sample design. An estimator with small sampling variance is said to be *precise.* Precision increases as the sampling variance decreases. Note that an estimator can be precise but biased. *Accuracy* is a measure of both the bias and precision of the estimator: an accurate estimator has good precision and is nearly unbiased.

One sampling strategy is more efficient than another if the sampling variance of the estimator for the sampling strategy is smaller than that of another sampling strategy. So as not to confuse this type of efficiency with other types – for example, cost efficiency – this will be referred to as statistical efficiency. Statistical efficiency is an important consideration when comparing different possible designs since if one design can provide improved or equivalent precision using a smaller sample size, this can provide considerable cost savings. The following sample designs compare their efficiency relative to SRS. Formally, this is measured by calculating the design effect, presented in section 7.3.3 of **Chapter 7 - Estimation**.

For more details on estimation, factors affecting precision and estimating precision, see **Chapter 7 - Estimation**.

6.2.2 Simple Random Sampling (SRS)

The starting point for all probability sampling designs is simple random sampling (SRS). SRS is a one-step selection method that ensures that every possible sample of size n has an equal chance of being selected. As a consequence, each unit in the sample has the same inclusion probability. This probability, π, is equal to n/N, where N is the number of units in the population.

Sampling may be done with or without replacement. Sampling with replacement allows for a unit to be selected more than once. Sampling without replacement means that once a unit has been selected, it cannot be selected again. Simple random sampling with replacement (SRSWR) and simple random sampling without replacement (SRSWOR) are practically identical if the sample size is a very small fraction of the population size. This is because the possibility of the same unit appearing more than once in the sample is small. Generally, sampling without replacement yields more precise results and is operationally more convenient. For the purpose of this chapter, sampling is assumed to be without replacement unless otherwise specified.

Consider a population of five people and suppose that a sample of three is selected (SRSWOR). Label the people in the population 1, 2, 3, 4 and 5 and denote the population as the set {1, 2, 3, 4, 5}. There are ten possible samples of three people: {1, 2, 3}, {1, 2, 4}, {1, 2, 5}, {1, 3, 4}, {1, 3, 5}, {1, 4, 5}, {2, 3, 4}, {2, 3, 5}, {2, 4, 5} and {3, 4, 5}. Each of these samples has an equal chance of being selected and each individual is selected in 6 out of the 10 possible samples, thus each individual has an inclusion probability of $\pi = n / N = 3 / 5$.

To select a simple random sample, the statistical agency usually has constructed a complete frame (either a list or area frame) before sampling. On a list frame, the units are generally numbered 1 to N, although the method of assigning a unique number to each unit is not important. Next, n units from the list are chosen at random using a random number table or a computer-generated random number and the corresponding units make up the sample.

As a means of illustrating the technique of SRSWOR, consider a survey of students from a school. Assume that a suitable list of students is available or can be created from existing sources. This list serves as the survey or sampling frame. Now, suppose that the population list contains N=1530 students of which a sample of size n=90 is required. The next step is to decide how to select the 90 students.

Sample selection can be done using a table of random numbers (see Table 1). The first step involves selecting a four-digit number (four since this is the number of digits in 1530). Sampling begins by selecting a number anywhere in the table and then proceeding in any direction. The first 90 four-digit numbers that do not exceed 1530 are selected.

Suppose row 01 and column 85 - 89 are selected as the starting point. Proceeding down this column, the random numbers selected are 189, 256, 984, 744, 1441, 617, etc. Selection continues until 90 different numbers are obtained. The result is a sample that consists of students with the corresponding numbers in the listing of the population. (Since the method under discussion is SRSWOR, any number that appears more than once is ignored). Although a random number table was used above to illustrate the manual selection of a simple random sample, practically speaking, a computer program would randomly select units.

SRS has a number of **advantages** over other probability sampling techniques, including:

i. It is the simplest sampling technique.

ii. It requires no additional (auxiliary) information on the frame in order to draw the sample.

The only information that is required is a complete list of the survey population and contact information.

iii. It needs no technical development.

The theory behind SRS is well established, so that standard formulas exist to determine the sample size, population estimates and variance estimates and these formulas are easy to use.

The **disadvantages** of SRS are:

i. It makes no use of auxiliary information even if such information exists on the survey frame. This can result in estimates being less statistically efficient than if another sample design had been used.

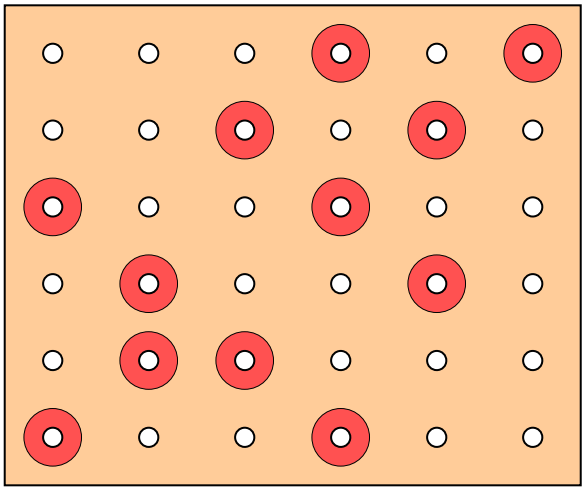
ii. It can be expensive if personal interviews are used, since the sample may be widely spread out geographically.

iii. It is possible to draw a ‘bad’ SRS sample. Since all samples of size n have an equal chance of being included in the sample, it is possible to draw a sample that is not well dispersed and that poorly represents the population.

Table 1: Excerpt of a Table of Random Numbers

	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99
00	59311	58030	52098	87024	14194	82848	04190	96574	90464	29065
01	98567	76364	77204	27062	53402	96621	43918	01896	83991	51141
02	10363	97518	51400	98342	24830	61891	27101	37855	06235	33516
03	86852	19558	64432	99612	53537	59798	32803	67708	15297	28612
04	11258	24591	36863	31721	81305	94335	34936	02566	80972	08188
05	95068	84628	35911	33020	70659	80428	39936	31855	34334	64865
06	54463	47437	73804	36239	18739	72824	83671	39892	60518	37092
07	16874	62677	57412	31389	56869	62233	80827	73917	82402	84420
08	92484	63157	76593	03205	84869	72389	96363	52887	01087	66591
09	15669	56689	35682	53256	62300	81872	35213	09840	34471	74441
10	99116	75486	84989	23476	52967	67104	39495	39100	17217	74073
11	15696	10703	65178	90637	63110	17622	53988	71087	84148	11670
12	97720	15369	51269	69620	03388	13699	33423	67453	43269	56720
13	11666	13841	71681	98000	35979	39719	81899	07449	47985	46967
14	71628	73130	78783	75691	41632	09847	61547	18707	85489	69944
15	40501	51089	99943	91843	41995	88931	73631	69361	05375	15417
16	22518	55576	98215	82068	10798	82611	36584	67466	69377	40054
17	75112	30485	62173	02132	14878	92879	22281	16783	86352	00077
18	08327	02671	98191	84342	90813	49268	95441	15496	20168	09271
19	60251	45548	02146	05597	48228	81366	34598	72856	66762	17002
20	57430	82270	10421	00540	43648	75888	66049	21511	47676	33444
21	73528	39559	34434	88596	54086	71693	43132	14414	79949	85193
22	25991	65959	70769	64721	86413	33475	42740	06175	82758	66248
23	78388	16638	09134	59980	63806	48472	39318	35434	24057	74739
24	12477	09965	96657	57994	59439	76330	24596	77515	09577	91871
…	…	…	…	…	…	…	…	…	…	…
45	12900	71775	29845	60774	94924	21810	38636	33717	67598	82521
46	75086	23537	49639	33595	31484	97588	28617	17979	78749	35234
47	99445	51434	29181	09993	38190	42553	68922	52125	91077	40197
48	26075	31671	45386	36583	93459	48599	52022	41330	60650	91321
49	13636	93596	23377	51133	95126	61496	42474	45141	46660	42338

Simple Random Sample (illustrated, n=12)



6.2.3 Systematic Sampling (SYS)

In systematic sampling (SYS), units are selected from the population at regular intervals. Systematic sampling is sometimes used when the statistical agency would like to use SRS but no list is available, or when the list is roughly random in order in which case SYS is even simpler to conduct than SRS. A Sampling interval and a random start are required. When a list frame is used and the population size, N, is a multiple of the sample size, n, every k^{th} unit is selected where the interval k is equal to N/n. The random start, r, is a single random number between 1 and k, inclusively. The units selected are then*: r, r+k, r+2k, ..., r+(n-1)k*. Like SRS, each unit has an inclusion probability, π, equal to n/N but, unlike SRS, not every combination of n units has an equal chance of being selected: SYS can only select samples in which the units are separated by k. Thus, under this method, only k possible samples can be drawn from the population.

To illustrate SYS, suppose a population contains N=54 units and a sample of size n=9 units is to be drawn. The sampling interval would be $k=N/n=54/9=6$. Next, a random number between 1 and k=6, say 2, is chosen. The population units selected for the sample are then numbered: 2, 8, 14, 20, 26, 32, 38, 44 and 50. With a sampling interval of 6 and a population of size 54, there are only 6 possible SYS samples, while for a simple random sample of size 6, there are over 25 million possible samples.

One advantage of systematic sampling is that it can be used when no list of the population units is available in advance. In this case, a conceptual frame can be constructed by sampling every k^{th} person until the end of the population is reached.

One problem with SYS is that the sample size, n, is not known until after the sample has been selected. Another problem arises when the sampling interval, k, matches some periodicity in the population. For example, suppose that a survey of traffic flow is to be conducted in an area and only one day of the week can be sampled, in other words k is every 7^{th} day. The survey's estimated traffic flow will be dramatically different if the sampled days are all Sundays as opposed to all Tuesdays. Of course, if the sampling period is every 5^{th} day, then every day of the week could be surveyed. Unfortunately, in most cases, periodicity is not known in advance.

If N cannot be evenly divided by n, the sampling interval for SYS is not a whole number. In this case, k could be set equal to the nearest whole number, but then the sample size would vary from sample to sample. For example, suppose that N=55 and n=9, then k=55/9=6.1. If k is assumed to be 6, and if r=2, the sample contains those units numbered: 2, 8, 14, 20, 26, 32, 38, 44 and 50. If the random start is $r=1$ and every sixth unit is selected, then the sample consists of units: 1, 7, 13, 19, 25, 31, 37, 43, 49 and 55. In this case, the sample is of size 10, not 9. Another approach is to set each of the values $r, r+k, r+2k, ..., r+(n-1)k$ to the nearest whole number. With this approach, the realised sample size is fixed. For example, suppose again that N=55 and n=9, so that k=55/9=6.1. If r=1, the sample consists of units 1, 7, 13, 19, 25, 31, 38, 44 and 50.

Alternatively, if N cannot be evenly divided by n then, to avoid a variable sample size, *circular systematic sampling* could be performed. With this method, the population units are thought to exist on a circle and modular counting is used. The value of k is set equal to the whole number nearest to N/n, but now the random start, r, can be between 1 and N, rather than 1 and k (i.e., the first unit can be anywhere on the list). The selected units, as before, are: r, $r+k$, $r+2k$, ..., $r+(n-1)k$. If the j^{th} unit is such that $r+(j-1)k > N$, then the selected unit is $r+(j-1)k - N$. That is, when the end of the list is reached, sampling continues at the beginning of the list. The advantage of the circular method is that each unit has an equal chance of being in the sample. For example, using the previous example, suppose that N=55 and n=9 and k=6. A random start, r, between 1 and 55 is selected, say r=42. Then the selected population units are: 42, 48, 54, 5, 11, 17, 23, 29 and 35.

SYS has a number of **advantages** depending on the circumstances and objective of the survey:

i. It is a proxy for SRS when there is no frame.

ii. It does not require auxiliary frame information, like SRS.

iii. It can result in a sample that is better dispersed than SRS (depending on the sampling interval and how the list is sorted).

iv. It has a well-established theory, just like SRS, and so estimates can be easily calculated.

v. It is simpler than SRS since only one random number is required.

The **disadvantages** of SYS are:

i. It can result in a 'bad' sample if the sampling interval matches some periodicity in the population.

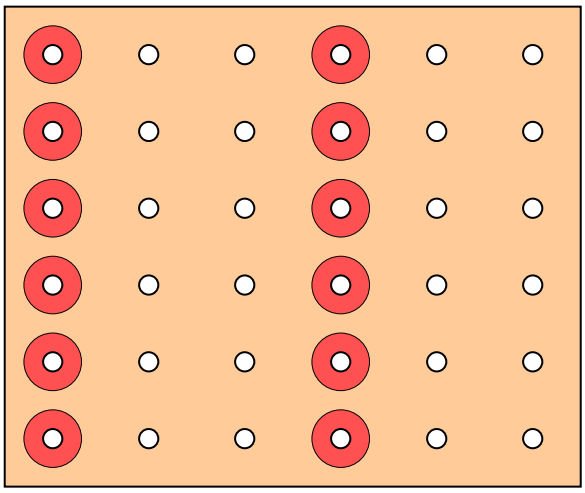
ii. Like SRS, it does not use any auxiliary information that might be available on the frame, and thus it can result in an inefficient sampling strategy.

iii. The final sample size is not known in advance when a conceptual frame is used.

iv. It does not have an unbiased estimator of the sampling variance. In order to do variance estimation, the systematic sample is often treated as if it were a simple random sample. This is only appropriate when the list is sorted randomly. (For more information on variance estimation for a systematic sample, see Cochran (1977) or Lohr (1999).)

v. It can lead to a variable sample size if the population size, N, cannot be evenly divided by the desired sample size, n (but this can be avoided using circular SYS).

Systematic Sample (illustrated, n=12, N=36, k=3)



SRS and circular SYS are both equal probability sample designs, since every possible sample has exactly the same chance of being selected. Not all sampling techniques result in equal probabilities. The sample designs described in the following sections can result in unequal probabilities. It is important to remember that in probability sampling, the criterion is not that all units have the same inclusion probability but that all units have a known non-zero inclusion probability. Often, sampling with unequal probabilities can improve the statistical efficiency of the sampling strategy.

6.2.4 Probability-Proportional-to-Size (PPS) Sampling

Probability-proportional-to-size (PPS) sampling is one technique that uses auxiliary data and yields unequal probabilities of inclusion. If population units vary in size and these sizes are known, such information can be used during sampling to increase the statistical efficiency.

PPS can yield dramatic increases in precision if the size measures are accurate and the variables of interest are correlated with the size of the unit. For less accurate size measures, it is better to create size groupings and perform stratified sampling (Section 6.2.6).

A good example of a PPS size variable is area. Farm surveys often use PPS, where the size measure is the size of the farm in hectares. Admittedly, the size of a farm can grow (or shrink) if the farmer buys or sells land, but for the most part, farm size is constant from year to year. In addition, typical questions for farm surveys, such as income, crop production, livestock holdings and expenses are often correlated with land holdings. Other size measures for business surveys include the number of employees, annual sales and the number of locations, although these variables are more likely to change from year to year.

In PPS sampling, the size of the unit determines the inclusion probability. Using farms as an example, this means that a farm with an area of 200 hectares has twice the probability of being selected as a farm with 100 hectares.

To illustrate, assume that there is a population of six farms and that the client is interested in estimating the total expenses of this farming population by sampling one farm. (A sample of size one is used for the purpose of illustration; in practice, a statistical agency rarely selects only one unit.). Suppose that there is a stable size measure for each farm (the size of the farm in hectares) and, to illustrate the efficiency gains over SRS, assume that each farm's expenses are known. (Obviously, in real life, if the expenses were known, there would be no need to conduct the survey.)

Consider the following list of farms:

Table 2: Population Values

Sampling Unit: Farm	Auxiliary Frame Information: Size of Farm in Hectares	Survey Variable of Interest: Expenses ($)
1	50	26,000
2	1,000	470,000
3	125	63,800
4	300	145,000
5	500	230,000
6	25	12,500
Total	**2,000**	**947,300**

For this population of six farms, the true total expenses are $947,300. A simple random sample could be selected, where each sample contains one unit and each unit has an inclusion probability of 1/6. Six different SRS samples of size n=1 are possible. Consider the results from SRS (see table 3). To do so, some estimation concepts (explained in detail in **Chapter 7 - Estimation**) must be introduced. For a sample of size one, the total expenses for the population is estimated by multiplying the sampled unit's expenses by the unit's weight. This weight is the average number of units in the survey population that the sampled unit represents and is the inverse of the inclusion probability.

For the PPS sample, the sampling variability is much lower. The estimates from the six possible samples now only range from a low of $920,000 to a high of $1.04 million – much better than SRS (see table 4). (The PPS inclusion probability is calculated as the size of the farm divided by the total size of all farms).

In this example, it was assumed that there is a relationship between expenses and the size of the farm, an assumption that obviously is valid here, or PPS would not have been as successful as it was. Indeed, if the variables of interest and the size variable were not correlated, PPS might not have been any better than SRS, and could have been worse.

The main **advantage** of PPS sampling is that it can improve the statistical efficiency of the sampling strategy by using auxiliary information. This can result in a dramatic reduction in the sampling variance compared with SRS or even stratified sampling (Section 6.2.6).

Table 3: Possible SRS Samples of Size *n*=1

Sample (Farm Selected)	Inclusion Probability (π)	Design Weight ($1/\pi$)	Expenses ($)	Population Estimate of Total Expenses ($)
Sample 1 (Farm 1)	1/6	6	26,000	156,000
Sample 2 (Farm 2)	1/6	6	470,000	2,820,000
Sample 3 (Farm 3)	1/6	6	63,800	382,800
Sample 4 (Farm 4)	1/6	6	145,000	870,000
Sample 5 (Farm 5)	1/6	6	230,000	1,380,000
Sample 6 (Farm 6)	1/6	6	12,500	75,000
		Average Sample Estimate		**947,300**

Notice the large sampling variability in the SRS estimates, ranging from $75,000 to $2.8 million. PPS can give estimates with much smaller sampling variability.

Table 4: Possible PPS Samples of Size *n*=1

Sample (Farm Selected)	Size of Farm	Inclusion Probability (π)	Design Weight ($1/\pi$)	Expenses ($)	Population Estimate of Total Expenses ($)
Sample 1 (Farm 1)	50	50/2,000	2,000/50	26,000	1,040,000
Sample 2 (Farm 2)	1,000	1,000/2,000	2,000/1000	470,000	940,000
Sample 3 (Farm 3)	125	125/2,000	2,000/125	63,800	1,020,800
Sample 4 (Farm 4)	300	300/2,000	2,000/300	145,000	966,667
Sample 5 (Farm 5)	500	500/2,000	2,000/500	230,000	920,000
Sample 6 (Farm 6)	25	25/2,000	2,000/25	12,500	1,000,000
			Average Sample Estimate		**947,300**

The **disadvantages** of PPS sampling are:

i. It requires a survey frame that contains good quality, up-to-date auxiliary information for all units on the frame that can be used as size measures.

ii. It is inappropriate if the size measures are not accurate or stable. In such circumstances, it is better to create size groupings and perform stratified sampling.

iii. It is not always applicable, since not every population has a stable size measure that is correlated with the main survey variables.

iv. It can result in a sampling strategy that is less statistically efficient than SRS for survey variables that are not correlated with the size variables.

v. Estimation of the sampling variance of an estimate is more complex.

vi. Frame creation is more costly and complex than SRS or SYS, since the size of each unit in the population needs to be measured and stored.

6.2.4.1 Methods of PPS Sampling

How is a PPS sample drawn? There are many PPS sampling schemes, however, three commonly used techniques are the random method, the systematic method and the randomised systematic method. (The following assumes that the size measures are integer values.)

i. The random method for PPS sampling

- for each unit in the population, cumulate the size measures for units up to and including itself;
- determine the range corresponding to each unit in the population, that is, from (but not including) the cumulative sum for the previous unit to the cumulative sum for the current unit;
- select a random number between 0 (if dealing with non-integer size measures) or 1 (for integer size measures) and the total cumulative size and select the unit whose range contains the random number;
- repeat previous step until *n* units have been selected.

To illustrate using the farm example:

Table 5: PPS Sampling using the Random Method

Farm	Size	Cumulative Size	Range
1	50	50	1-50
2	1000	1050	51-1050
3	125	1175	1051-1175
4	300	1475	1176-1475
5	500	1975	1476-1975
6	25	2000	1976-2000

For a sample containing three units, three random numbers between 1 and 2000 are selected. Suppose these numbers are: 1697, 624 and 1109. Then the farms selected are: farm 5, farm 2 and farm 3.

In the case of the random method for PPS sampling without replacement, if more than one unit is selected, complications arise both in attempting to keep probabilities directly proportional to size and in estimating the sampling variances of survey estimates. This becomes even more complicated when more than two or three units are selected with PPS without replacement, and in fact, is the subject of considerable research. Much of this research is contained in the writings of Horvitz and Thompson (1952), Yates and Grundy (1953), Rao, Hartley and Cochran (1962), Fellegi (1963), and Brewer and Hanif (1983).

ii. The systematic method

- for each unit in the population, cumulate the size measures for units up to and including itself;
- determine the range corresponding to each unit in the population, that is, from (but not including) the cumulative sum for the previous unit to the cumulative sum for the current unit;
- determine the sampling interval, k=(total cumulative size)/n;

- determine a random start, *r*, between 0 (if dealing with non-integer size measures) or 1 (for integer size measures) and *k;*

- select those units whose range contains the random numbers *r, r+k, r+2k, ... r+(n-1)k*.

iii. The randomised systematic method

In this scheme, the list is randomised prior to the application of systematic sampling. Just as with systematic sampling, if the list is used in its original order, some possible samples may be eliminated. By randomising the list, the number of potential samples that can be drawn is increased.

Note that these methods do pose certain problems. For example, for the systematic and randomised systematic methods, if the size of any unit is greater than the interval, it may be selected more than once. This problem can only be overcome by placing such large units into separate strata and sampling them independently (Section 6.2.6). A second problem is the difficulty of estimating sampling variances.

6.2.5 Cluster Sampling

Cluster sampling is the process of randomly selecting complete groups (clusters) of population units from the survey frame. It is usually a less statistically efficient sampling strategy than SRS and is performed for several reasons. The first reason is that sampling clusters can greatly reduce the cost of collection, particularly if the population is spread out and personal interviews are conducted. The second reason is that it is not always practical to sample individual units from the population. Sometimes, sampling groups of the population units is much easier (e.g., entire households). Finally, it allows the production of estimates for the clusters themselves (e.g., average revenue per household).

Cluster sampling is a two-step process. First, the population is grouped into clusters (this may consist of natural clustering, e.g., households, schools). The second step is to select a sample of clusters and interview all units within the selected clusters.

The survey frame may dictate the method of sampling. Until now, the focus has been on sampling individual units of the population from a list frame. If the units of the population are naturally grouped together, it is often easier to create a frame of these groups and sample them than try to create a list frame of all individual units in the population. For example, the client may be interested in sampling teachers but only have available a list of schools. In the case of household or farm surveys, many countries do not have complete and up-to-date lists of the people, households or farms for any large geographic area, but they do have maps of the areas. In this case an area frame could be created, with the geographical areas divided into regions (clusters), the regions sampled and everyone within the region interviewed. Different sample designs can be used to select the clusters, such as SRS, SYS or PPS. A common design uses PPS where sampling is proportional to the size of the cluster.

There are a number of considerations to bear in mind for cluster sampling. In order for estimates to be statistically efficient, the units within a cluster should be as different as possible. Otherwise, if the units within a cluster are similar, they all provide similar information and interviewing one unit would be sufficient.

Unfortunately, units within a cluster frequently have similar characteristics and therefore are more homogeneous than units randomly selected from the general population. This results in a sampling procedure that is less efficient than SRS. For example, suppose that for a city of 100,000, two samples are

drawn. For the first sample, cluster sampling is used and one city block, containing 400 residents, is selected at random. For the second sample, SRS is used to select 400 people from the list of 100,000 residents. The 400 residents in the SRS sample are likely to be far more diverse in terms of income, age, occupation, and educational background (to name only a few variables) than the 400 people in the cluster sample who all live on the same city block.

The statistical efficiency of cluster sampling depends on how homogeneous the units within the clusters are, how many population units are in each cluster and the number of clusters sampled. When neighbouring units are similar, it is more statistically efficient to select many small clusters rather than a few, larger clusters. However, when personal interviews are conducted, the more dispersed the sample is, the more expensive the survey. The statistical agency must strike a balance between the optimal number and size of clusters, and the cost.

There can be logistical difficulties with cluster sampling. If the survey frame is an area frame based on a map and the sampling unit is a cluster of dwellings, it can be difficult to determine if a dwelling is in a cluster or not. Some basic rules should be created to determine which units are in a cluster. For example, having a rule saying *dwellings belong to the cluster in which their main entrance (front door) lies* would eliminate most problems (usually, the entire dwelling is either in or out of the boundary of the cluster). If a dwelling seems to be evenly divided between clusters, toss a coin to avoid bias. In the Canadian Labour Force Survey, clusters are determined by drawing a line down the middle of the street. This makes it easy to determine whether a dwelling is in the sample or not. (For more information on these practical considerations, see **Chapter 9 - Data Collection Operations**).

The **advantages** of cluster sampling are:

i. It can greatly reduce the cost of collection by having a less dispersed sample than SRS. This is particularly important when the population is spread out and personal interviews are conducted, since savings can be achieved by reducing the travel time of interviewers, especially for rural populations.

ii. It is easier to apply than SRS or SYS to populations that are naturally clustered (e.g., households, schools) and for certain conceptual populations, such as people crossing a border during a specific time interval. For such populations, it may be difficult, expensive or impossible to construct a list of all individual units of the population, required by SRS.

iii. It allows the production of estimates for the clusters themselves. For example, estimates of the average number of teachers per school (where schools are clusters).

iv. It can be more statistically efficient than SRS if the units within the clusters are heterogeneous (different) with respect to the study variables and the clusters are homogeneous (similar). But in practice this is usually not the case.

The **disadvantages** of cluster sampling are:

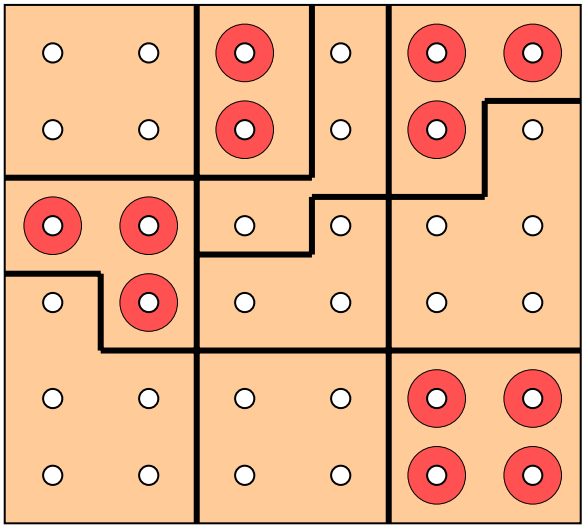
i. It can be less statistically efficient than SRS if the units within the clusters are homogeneous with respect to the study variables. This is frequently the case, since units within a cluster tend to have similar characteristics. However, to offset this loss in statistical efficiency, the number of clusters selected can be increased.

ii. Its final sample size is not usually known in advance, since it is not usually known how many units are within a cluster until after the survey has been conducted.

iii. Its survey organisation can be more complex than for other methods.

iv. Its variance estimation will be more complex than for SRS if clusters are sampled without replacement.

Cluster Sample (illustrated, 4 clusters are sampled)



6.2.6 Stratified Sampling (STR)

With stratified sampling, the population is divided into homogeneous, mutually exclusive groups called strata, and then independent samples are selected from each stratum. Any of the sample designs mentioned in this chapter can be used to sample within strata, from the simpler methods such as SRS or SYS to the more complex methods such as PPS, cluster, multi-stage or multi-phase sampling (discussed later in this chapter). For example, with cluster sampling, it is very common to first stratify, then draw the cluster sample. This is called stratified cluster sampling.

A population can be stratified by any variables that are available for all units on the frame prior to the survey being conducted. For instance, this information may simply be the address of the unit, allowing stratification by province, or there may be income data on the frame, allowing stratification by income group. Commonly used stratification variables include: age, sex, geography (e.g., province), income, revenues, household size, size of business, type of business, number of employees, etc.

There are three main reasons for stratification. The first is to make the sampling strategy more efficient than SRS or SYS. The second is to ensure adequate sample sizes for specific domains of interest for which analysis is to be performed. The third is to protect against drawing a ‘bad’ sample.

First, for a given sample size and estimator, stratification may lead to lower sampling error or, conversely, for a given sampling error, to a smaller sample size. Note that, while both cluster sampling and stratification group units in the population, with stratified sampling, samples are drawn within each stratum, while for cluster sampling, samples of clusters are drawn and everyone in the cluster surveyed. And while stratification generally increases the precision of estimation with respect to SRS, clustering generally decreases it (since neighbouring units are usually similar).

In order to improve the statistical efficiency of a sampling strategy with respect to SRS, there must be strong homogeneity within a stratum (i.e., units within a stratum should be similar with respect to the variable of interest) and the strata themselves must be as different as possible (with respect to the same variable of interest). Generally, this is achieved if the stratification variables are correlated with the survey variable of interest. The reason why stratification can increase the precision of the estimates relative to SRS is explained by Cochran (1977):

> If each stratum is homogeneous, in that the measurements vary little from one unit to another, a precise estimate of any stratum mean can be obtained from a small sample in that stratum. These estimates can be combined in a precise estimate for the whole population.

Stratification is particularly important in the case of *skewed populations* (i.e., when the distribution of values of a variable is not symmetric, but leans to the right or the left). For example, business and farm surveys often have highly skewed populations – the few large business establishments and farms often have large values for variables of interest (e.g., revenues, expenditures, number of employees). In such cases, a few population units can exert a large influence on estimates – if they happen to be selected in the sample, they can greatly increase the estimate, and if they are not selected, the estimate will be much lower. In other words, these units can increase the sampling variability of the estimate. Therefore, such units should be placed in a stratum by themselves to ensure that they do not represent other, potentially much smaller, units in the population.

To stratify businesses, a size variable based on the number of employees, for example, is often used. If the size variable has three values – small, medium and large – the statistical efficiency is improved if the large businesses have similar sales, the medium businesses have similar sales, and the small businesses have similar sales and if large and medium businesses and the medium and small business have quite different sales. Similarly, for a sample design using area frames, the proper representation of large cities can be ensured by placing them in a separate stratum, and sampling each stratum separately.

In the previous example, it was reasonable to stratify by the number of employees, since this is a measure of the size of the company and is likely to be highly related to sales. However, if a survey is interested in the age of its employees, it makes no sense to stratify by the number of employees since there is no correlation. Also, stratification that is statistically efficient for one survey variable may not work well for others. Usually the stratification variables are chosen based on their correlation with the most important survey variables. This means that for those, less important, survey variables that are uncorrelated to the stratification variables, estimates for a stratified sample can be less efficient than SRS.

The second reason for stratification is to ensure adequate sample sizes for known ***domains of interest***. When designing a survey, often the overall goal is to estimate a total. How many people were unemployed last month? What were the total retail sales last month?

In addition to overall totals, the client often requires estimates for subgroups of the population, called domains. For example, the client may wish to know how many men were unemployed and compare this with the number of women who were unemployed. Similarly, the client may want to know the sales last month for clothing stores, or for all retail stores in a certain province. Creating estimates for subgroups is called domain estimation. If domain estimates are required, the ability to calculate them with a large enough sample in each domain should be incorporated into the sample design. If the information is available on the frame, the easiest way to do this is to ensure that strata exactly correspond to the domains of interest.

The third reason for stratifying is to protect against drawing a 'bad' sample. In the case of SRS, the selection of the sample is left entirely to chance. Stratified sampling attempts to restrict the possible samples to those that are less extreme by ensuring that at least certain parts of the population are represented in the sample. For example, to ensure that both men and women are included in the sample, the survey frame should be stratified by sex (assuming this auxiliary variable is available on the frame).

In addition to these reasons, stratification is often used for operational or administrative convenience. It can enable the statistical agency to control the distribution of fieldwork among its regional offices. For example, if data collection is conducted by province, then stratification by province is appropriate, in which case the provincial regional office can be given their portion of the sample.

Once the population has been divided into strata, the statistical agency needs to determine how many units should be sampled from each stratum. This step is referred to as allocation of the sample and is covered in **Chapter 8 - Sample Size Determination and Allocation.**

Inclusion probabilities usually vary from stratum to stratum; it depends on how the sample is allocated to each stratum. To calculate the inclusion probabilities for most sample designs, the size of the sample and the size of the population in each stratum must be considered. To illustrate, consider a population with N=1000 units stratified into two groups: one stratum has N_1=250 units and the other has N_2=750 units. Suppose that SRS is used to select n_1=50 units from the first stratum and n_2=50 units from the second the probability, π_2, that a unit in the second stratum is selected is $\pi_2 = 50/750 = 1/15$. Units thus have different probabilities of inclusion – a unit in the first stratum is more likely to be selected than one in the second.

The **advantages** of stratified sampling are:

i. It can increase the precision of overall population estimates, resulting in a more efficient sampling strategy. A smaller sample can save a considerable amount on the survey, particularly data collection.

ii. It can guarantee that important subgroups, when defined as strata, are well represented in the sample, resulting in statistically efficient domain estimators.

iii. It can be operationally or administratively convenient.

iv. It can protect against selecting a 'bad' sample.

v. It allows different sampling frames and procedures to be applied to different strata (e.g., SRS in one stratum, PPS in another).

The **disadvantages** of stratified sampling are:

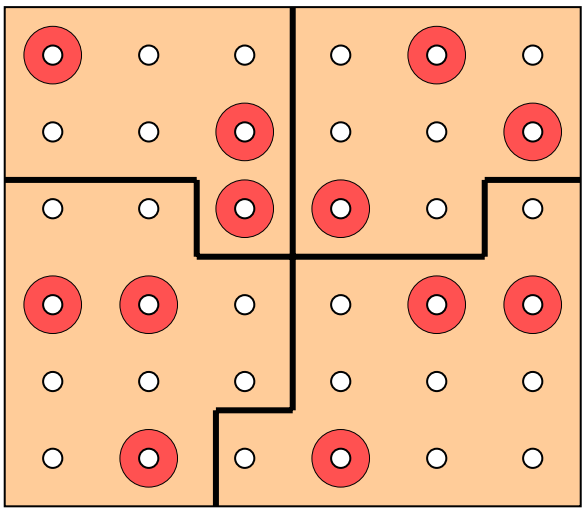
i. It requires that the sampling frame contain high quality auxiliary information for all units on the frame, not just those in the sample, that can be used for stratification.

ii. Frame creation is more costly and complex than for SRS or SYS, since the frame requires good auxiliary information.

iii. It can result in a sampling strategy that is less statistically efficient than SRS for survey variables that are not correlated to the stratification variables.

iv. Estimation is slightly more complex than for SRS or SYS.

Stratified Sample (illustrated, 4 strata, 3 units selected per stratum)



6.2.7 Multi-Stage Sampling

Thus far, the discussion has centred around one stage sample designs. Multi-stage sampling is the process of selecting a sample in two or more successive stages. The units selected at the first stage are called primary sampling units (PSU's), units selected at the second stage are called second stage units (SSU's), etc. The units at each stage are different in structure and are hierarchical (for example, people live in dwellings, dwellings make up a city block, city blocks make up a city, etc.). In two-stage sampling, the SSU's are often the individual units of the population.

A common multi-stage sample design involves two-stage cluster sampling using an area frame at the first stage to select regions (the PSU's) and then a systematic sample of dwellings (the SSU's) within a region at the second stage. With the one-stage cluster sampling presented earlier, every unit within a sampled cluster is included in the sample. In two-stage sampling, only some of the units within each selected PSU are subsampled.

Multi-stage sampling is commonly used with area frames to overcome the inefficiencies of one-stage cluster sampling, which in fact is rarely used. If the neighbouring units within a cluster are similar, then it is more statistically efficient to sample a few SSU's from many PSU's than to sample many SSU's from fewer PSU's.

Multi-stage samples can have any number of stages, but since the complexity of the design (and estimation) increases with the number of stages, designs are often restricted to two or three stages. It should be noted that the frame for the first stage is generally quite stable. For example, an area frame covering large geographical areas does not change rapidly over time. Second (and subsequent) stage frames required to sample units at subsequent stages are usually less stable. Often, these frames are list frames created in the field during collection. For example, for the geographic areas sampled at stage one, a list frame could be created for all those dwellings within the sampled areas. Note that, listing only sampled areas requires much less effort than trying to list the whole population. (See **Chapter 9 - Data Collection Operations** for details on listing).

Each stage of a multi-stage sample can be conducted using any sampling technique. Consequently, one of the chief advantages of a multi-stage sample is its flexibility. For example, within one PSU drawn at the first stage, an SRS sample may be drawn. For another PSU, there may be a measure of size that is correlated with the key survey variables, so PPS may be used within this PSU.

The Canadian Labour Force Survey (LFS) sample is an example of a multi-stage stratified sample. The country is divided into over 1,100 strata. Each stratum consists of a group of enumeration areas (EA's). EA's are geographic areas defined by the Census of Population so that the area that they cover can be canvassed by one census representative (they are created by keeping in mind the size of territory and the density of the population). The first stage of sampling is a stratified sample of clusters (EA's or groups of EA's) from within these strata. At the second stage, the clusters are mapped, all dwellings in them are listed, and the census representative selects a systematic sample of dwellings from each list. All persons within a selected dwelling are then interviewed for the survey.

Finally, note that although the examples provided thus far use an area frame at the first stage this is by no means a requirement for multi-stage sampling. An example of a multi-stage sample using a different kind of frame is a travel survey conducted at an airport. The primary sampling unit could be time – days in a month, while the second stage unit could be actual travellers. For a more complex travel survey, the second stage unit could be arriving passenger planes, while the third stage unit could be actual seats on the plane.

The **advantages** of multi-stage sampling are:

i. It can result in a more statistically efficient sampling strategy than a one-stage cluster design when clusters are homogeneous with respect to the variables of interest (i.e., a sample size reduction).

ii. It can greatly reduce the travel time and cost of personal interviews as a result of the sample being less dispersed than for other forms of sampling, such as SRS.

iii. It is not necessary to have a list frame for the entire population. All that is needed is a good frame at each stage of sample selection.

The **disadvantages** of multi-stage sampling are:

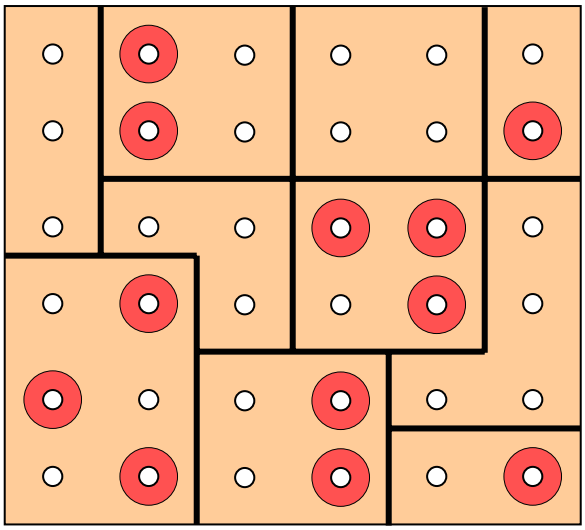
i. It is usually not as statistically efficient as SRS (although it can be more efficient than a one-stage cluster strategy).

ii. The final sample size is not always known in advance, since it is not usually known how many units are within a cluster until after the survey has been conducted. (The sample size can be controlled, however, if a fixed number of units are selected per cluster.)

iii. Its survey organisation is more complex than for one-stage cluster sampling.

iv. Its formulas for calculating estimates and sampling variance can be complex.

Multi-Stage Sample (illustrated, 2 stage cluster design, 6 PSU's selected with 1 to 3 SSU's selected within each PSU)



6.2.8 Multi-Phase Sampling

Despite the similarities in name, multi-phase sampling is quite different from multi-stage sampling. Although multi-phase sampling also involves taking two or more samples, all samples are drawn from the same frame and the units have the same structure at each phase. A multi-phase sample collects basic information from a large sample of units and then, for a subsample of these units, collects more detailed information. The most common form of multi-phase sampling is two-phase sampling (or double sampling), but three or more phases are also possible. However, as with multi-stage sampling, the more phases, the more complex the sample design and estimation.

Multi-phase sampling is useful when the frame lacks auxiliary information that could be used to stratify the population or to screen out part of the population. For example, suppose information is needed about cattle farmers, but the survey frame only lists farms, with no auxiliary information. A simple survey could be conducted whose only question is: 'Is part or all of your farm devoted to cattle farming?' With only one question, this survey should have a low cost per interview (especially if done by telephone) and consequently the agency should be able to draw a large sample.

Once the first sample has been drawn, a second, smaller sample can be drawn from amongst the cattle farmers and more detailed questions asked of these farmers. Using this method, the statistical agency avoids the expense of surveying units that are not in scope (i.e., who are not cattle farmers).

Multi-phase sampling can also be used to collect more detailed information from a subsample when there is insufficient budget to collect information from the whole sample, or when doing so would create excessive response burden. The Canadian Quarterly Retail Commodity Survey (QRCS) is one example. The first phase of the survey is the Monthly Wholesale Retail Trade Survey (MWRTS). Each month, MWRTS asks wholesale and retail establishments for two variables – their monthly sales and inventories. QRCS subsamples the retail establishments and asks them to report their sales by retail commodity, for example, clothing, electronics, foodstuffs, etc.

Similarly, multi-phase sampling can be used when there are very different costs of collection for different questions on a survey. Consider a health survey that asks some basic questions about diet, smoking, exercise and alcohol consumption. In addition, suppose the survey requires that respondents be subject to some direct measurements, such as running on a treadmill and having their blood pressure and cholesterol levels measured. It is relatively inexpensive to ask a few questions, but the medical tests require the time of a trained health practitioner and the use of an equipped laboratory, so are relatively expensive. This survey could be done as a two-phase sample, with the basic questions being asked at the first phase and only the smaller, second phase sample receiving the direct measurements.

In addition to stratification or screening information, data collected at the first phase can be used to improve the efficiency of estimation (e.g., for regression estimation). For more on estimation, see **Chapter 7 - Estimation**.

The **advantages** of multi-phase sampling are:

i. It can greatly increase the precision of estimates (compared with SRS).

ii. It can be used to obtain auxiliary information that is not on the sampling frame (in particular, stratification information for second phase sampling).

iii. It can be used when the cost of collection for some of the survey variables is particularly expensive or burdensome for the respondent.

The **disadvantages** of multi-phase sampling are:

i. It takes longer to get results than from a one-phase survey, if results from the first phase are required to conduct the second phase.

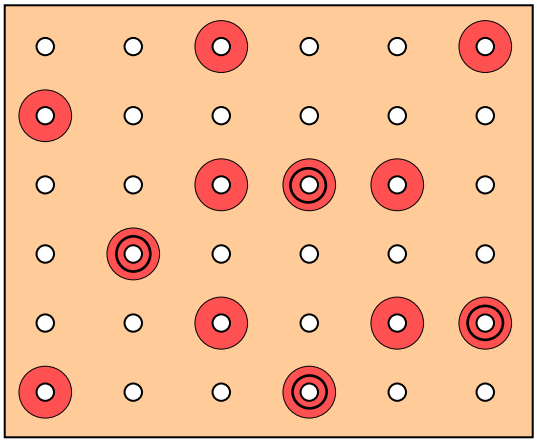
ii. It can be more expensive than a one-phase survey since it requires interviewing a sampled unit more than once.

iii. If the population is mobile or if the characteristics of interest change frequently, time delays between phases may pose problems.

iv. Its survey organisation can be complex.

v. Its formulas for the calculation of estimates and sampling variance can be quite complex.

Multi-phase sample (illustrated, 12 units selected at the first phase, 4 at the second)



6.2.9 Replicated Sampling

Replicated sampling involves the selection of a number of independent samples from a population rather than a single sample. Instead of one overall sample, a number of smaller samples, of roughly equal size, called replicates, are independently selected, each based upon the same sample design. Replicated sampling might be used in situations where preliminary results are needed quickly. Such preliminary results might be based upon the processing and analysis of a single replicate.

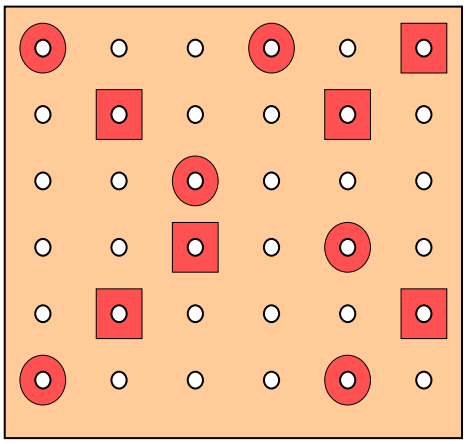
The main reason for replicated sampling is to facilitate the calculation of the sampling variance of survey estimates (sampling variance is a measure of sampling error). While it is generally possible to calculate the sampling variance based on probability samples, such calculations can be exceedingly difficult depending on the complexity of the sample design. The problem is that some mathematical expressions for sampling variance are difficult to derive and tedious and costly to program. In particular, in the case of systematic sampling, variance estimates cannot be calculated directly unless assumptions are made about the arrangement of units in the list.

Measures of sampling error are determined by examining the extent to which sample estimates, based upon all possible samples of the same size and design, differ from one another. Replicated sampling simulates this concept. Instead of drawing all possible samples (which is not practical), a reasonable number of smaller samples are selected using identical methods. For example, instead of selecting one sample of size 10,000, ten independent samples of size 1,000 could be drawn. The estimates from each of these ten samples can be compared and estimates of sampling variance derived. The reliability of the sampling variance estimates increases with the number of replicates selected. (See section 7.3.4 of **Chapter 7 - Estimation,** for an example of replicated sampling for variance estimation.)

There are a number of other procedures that use replication to estimate the sampling variance for complex sample designs. These include balanced repeated replication, jackknife and bootstrap. These techniques, which all extend the basic idea of replicated sampling, differ from one another in terms of the accuracy with which they measure the sampling variance of different types of survey estimates and their operational complexity and the situations to which they best apply.

There are drawbacks to this approach. One disadvantage of this scheme is that estimates of sampling variance, in general, tend to be less precise than if they were based directly on the statistical expressions that incorporate sample design features such as multi-stage, stratification, etc.

Replicated Sampling (illustrated, 2 samples drawn of size 6)



6.3 Special Topics in Sample Design

Sometimes, sample designs are modified to meet the special needs of a particular survey. This may be necessary because the target population is particularly difficult to locate, because the characteristic of interest is very rare in the population, because of the analytical needs of the survey or because of the method of data collection. **Chapter 4 - Data Collection Methods** presented telephone sampling designs, including Random Digit Dialling (RDD). The following sections describe other special applications of sample designs to fit particular survey needs.

6.3.1 Repeated Surveys

Surveys that are conducted once differ in several ways from repeated surveys. The aim of a repeated survey is often to study trends or changes in the characteristics of interest over a period of time.

Decisions made in the sample design of repeated surveys should take into account the possibility of deterioration in the statistical efficiency of the sampling strategy over time. A statistical agency may elect, for example, to use stratification variables that are more stable, avoiding those that may be more statistically efficient in the short run, but which change rapidly over time.

Another feature of a repeated survey is that, in general, a great deal of information is available which is useful for future design purposes. The adequacy of various features of the sample design such as the appropriateness of stratification variables and boundaries, the method of sample allocation and the size of units at various stages of a multi-stage design may be studied over time with a view to increasing statistical efficiency. Often, information required to efficiently design a one-time survey is very limited.

In the design of a repeated survey, provisions must be made to accommodate such events as births, deaths and changes in size measure. The sampling and estimation methods used in repeated surveys should incorporate these changes in a statistically efficient way with as little disruption as possible to the ongoing survey operations.

One particular type of repeated survey is a panel or *longitudinal survey*, where data are collected from the same sample units on several occasions. Such surveys usually measure changes in the characteristics of a given population with greater precision than do a series of independent samples of comparable size.
If a survey is to be repeated there are **advantages** to using a longitudinal sample, rather than doing a series of ad hoc independent samples. Some of the advantages are:

i. It reduces the sampling variance for estimates of change (i.e., $\hat{Y}_2 - \hat{Y}_1$, where $\hat{Y}_1$ is an estimate of the total at time 1 and $\hat{Y}_2$ is an estimate of the total at time 2). For example, this might be a measure of the change in the number of unemployed persons from one month to the next.

ii. It can be used to obtain information on the behaviour of respondents over time.

iii. It may reduce response errors (since respondents acquire a better understanding of the questionnaire).

iv. It may result in a cost reduction over time (development of survey, programming of computer systems, staff training, etc., are done over a long period of time).

Some of the **disadvantages** of using a longitudinal sample instead of several independent samples are:

i. Its estimation, treatment of nonresponse, etc., is more complex.

ii. It requires that the budget for the survey be guaranteed for the life of the panel. This entails a cost commitment over a long period of time.

iii. It is harder to maintain representativeness across time periods because of changes that occur in the population over time such as the addition of new units and the withdrawal of others.

iv. It may increase response error (for example, respondents' knowledge of the questionnaire may lead some to answer questions incorrectly in order to speed up the interview).

v. It can lead to higher nonresponse over time (due to respondent fatigue – the same person is surveyed repeatedly over time; difficulty tracing, etc.)

vi. Its organisation is more complex than for a one-time survey.

vii. It can result in survey-induced behaviour. For example, a respondent who is repeatedly asked about visits to the doctor may start visiting a doctor as a result of the survey.

viii. It can be difficult to define some concepts (e.g., household composition can change over time, so how is a longitudinal household defined?).

ix. If the initial sample drawn is a 'bad' sample, the statistical agency must continue with that sample.

One design that is intermediate between independent samples on successive occasions and a longitudinal sample takes the form of a rotating sample design in which part of the sample is replaced at each survey occasion.

For example, the Canadian Labour Force Survey (LFS) employs a rotation design in which households are included in the sample for six consecutive months, and every month, one sixth of the sample is replaced by a new group of households. The LFS sample is divided into six panels (or groups). Each panel in the survey is surveyed once a month for six months. At the end of its six months, the panel is removed from the survey (rotated out) and a new one is rotated in. This puts a limit on respondent burden (the average LFS interview is under 10 minutes) and gives a good sample overlap each month. An additional advantage is that the sample is refreshed each month. If the sample were never updated, then members of the sample would age over time, and families in new dwellings would never have a chance to enter the sample. As a result, the sample would no longer reflect the current population and would become biased over time.

This design offers the advantage of measuring monthly changes with greater precision, less cost and with less disruption to the field operations than would otherwise be the case if independent samples were used. It also reduces the problem of respondent burden associated with panel studies. (Nonetheless, to reflect changes in the size and structure of the population and data requirements, the LFS undergoes periodic redesigns, usually in the wake of the decennial census.)

In addition to the LFS, such designs are often used in business surveys. Note that rotating sample designs, in addition to the basic sample design, also require a methodology for how to rotate the sample. This can be complex and goes beyond the scope of this manual. For more details on rotating samples and longitudinal surveys in general, see Kalton, 1992.

6.3.2 Entry/Exit Surveys

Entry/exit surveys apply to populations crossing a border, for example people entering (or leaving) a country or users of a toll road. The problem with such populations is creating an up-to-date list frame with contact information so that the units can be interviewed or sent questionnaires. For example, suppose the client wishes to interview foreign visitors to Canada and that it is possible to obtain a list from customs of all visitors who entered the country on a particular date. One problem is how to find these people to interview them? By the time the frame is created, the travellers may have returned home, making an interview impractical. If they are still in Canada, it is unlikely that there is an address for them.

It is for these reasons that a conceptual frame and systematic sampling, or two-stage cluster sampling with systematic sampling within sampled clusters, is often used to survey such populations. The conceptual frame might be a list of the population units enumerated within a certain time interval at particular locations. For the frame to have complete coverage, these locations must be areas where the target population is concentrated. Often, entrance and exit areas are used. Exit areas are more popular since most surveys are interested in the activities the unit pursued before leaving the area.

An important consideration in the sample design – as with all sample designs – is field procedures. The operational and design challenge is to make optimal use of fieldworkers while maintaining a probability sample. An uneven flow of visitors creates a highly variable workload, making efficient staff allocation difficult. The most effective use of an interviewer's time is to interview the k^{th} visitor after *completing* the current interview, but this would be a non-probability design. Systematic sampling where one person counts people and a small team of interviewers hand out questionnaires or conduct interviews is preferable. The team size will depend on the flow density and the length of the interview, if interviews are conducted.

Data collection can be done via self-enumeration, interviews or direct observation, when appropriate. For self-enumeration questionnaires, the response rate is better if the respondent completes the questionnaire on site, rather than mailing it back to the statistical agency. Interviews obviously require more field staff, but result in higher response rates. Direct observation is very accurate and desirable, but not always applicable.

The main **advantage** of an entry/exit survey is that the frame for the final stage can be created while in the field.

The **disadvantages** of an entry/exit survey are:

i. It can be difficult to relate the survey population to a commonly understood population. This is because entry/exit surveys measure visitors, rather than people. For example, if a survey is conducted at a store, someone who visits the store more than once during the time period will be counted more than once.

ii. It can be difficult to manage field operations due to variable flows in the population. For this reason, it is recommended that interviews be kept short.

iii. It typically yields low response rates.

6.3.3 Snowball Sampling

Suppose the client wishes to find rare individuals in the population and already knows of the existence of some of these individuals and can contact them. One approach is to contact those individuals and simply ask them if they know anyone like themselves, contact those people, etc. The sample grows like a snowball rolling down a hill to hopefully include virtually everybody with that characteristic. Snowball sampling is useful for small or specialised populations such as blind, deaf, or other persons who may not belong to an organised group or such as musicians, painters, or poets, not readily identified on a survey list frame. However, snowball sampling is a method of nonprobability sampling: some individuals or sub-groups may have no chance of being sampled. In order to make inferences, strong modelling assumptions (which are usually not met) are required.

Network sampling and adaptive cluster sampling are similar sample designs that are used to target rare or specialised populations.

6.4 Summary

This chapter covered the basics of sampling. The two main types of sampling are probability sampling and non-probability sampling. Non-probability sampling is of limited use for surveys conducted by statistical agencies, since the biased selection of units does not readily permit inferences to be made about the survey population. However, it is fast and easy and can be useful for exploratory studies or during the development phase of a survey (e.g., to test the questionnaire).

Probability sampling should be used when inferences about the population are to be made based on the survey results. In a probability sample, every unit on the frame has a non-zero probability of being selected and the units are selected randomly. As a result, selection is unbiased and it is possible to calculate the probabilities of inclusion, calculate the sampling variance of estimates and make inferences

about the population. The main disadvantages of probability sampling is that it requires more time, is more costly than non-probability sampling and requires a high quality sampling frame.

The simplest probability sample designs are simple random sampling and systematic sampling, which result in equal probabilities of inclusion. More complex designs that can result in unequal probabilities of inclusion and most of which require auxiliary information include: stratified, probability-proportional-to-size, cluster, multi-stage and multi-phase sampling. Unequal probability designs are typically used to improve the statistical efficiency of the sampling strategy or to reduce the cost of sampling. Sometimes, their use is dictated by the sampling frame.

When deciding between the various possible designs, the first thing to determine is what designs are feasible given the survey frame, units on the survey frame, domains of interest, response burden, the method of data collection, budget, etc.?

Some things to consider are:
- Does the survey frame have auxiliary data that could be used to improve the efficiency of sampling? (Should the survey frame be stratified and/or should PPS be performed?)
- Does the survey frame lack auxiliary information that could be used to screen out units or that would be useful for stratification? Is data collection very expensive or burdensome? (Should two-phase sampling be performed?)
- Is the population naturally clustered or are the units on the survey frame clusters? Is the population spread out geographically and personal interviews to be conducted? (Should single-stage or multi-stage cluster sampling be performed?)

Finally, there are several special applications of sample designs that can be made depending on the specific needs of the survey.

For how to determine the size of sample required to satisfy a given level of precision, and how to compare the efficiency of different sample designs by comparing design effects, see **Chapter 8 - Sample Size Determination and Allocation**.

Bibliography

Bebbington, A.C. 1975. A Simple Method of Drawing a Sample without Replacement. *Applied Statistics,* 24(1).

Binder, D.A. 1998. Longitudinal Surveys: Why are These Surveys Different from all Other Surveys? *Survey Methodology,* 24(2): 101-108.

Brewer K.R.W. and M. Hanif. 1983. Sampling with Unequal Probabilities. Springer-Verlag, New York.

Cochran, W.G. 1977. *Sampling Techniques*. John Wiley and Sons, New York.

Conner, W.S. 1966. An Exact Formula for the Probability that Two Specified Sample Units Will Occur in a Sample Drawn with Unequal Probabilities and Without Replacement. *Journal of the American Statistical Association,* 61: 385-390.

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.

Droesbeke, J.-J., B. Fichet and P. Tassi, 1987. *Les Sondages*. Economica, Paris.

Fellegi, I.P. 1963. Sampling with Varying Probabilities Without Replacement Rotating and Non-Rotating Samples. *Journal of the American Statistical Association,* 58: 183-201.

Fink, A. 1995. *The Survey Kit*. Sage Publications, California.

Fowler, F.J. 1984. *Survey Research Methods*. 1. Sage Publications, California.

Gambino, J.G., M.P. Singh, J. Dufour, B. Kennedy and J. Lindeyer. 1998. *Methodology of the Canadian Labour Force Survey*. Statistics Canada. 71-526.

Gray, G.B. 1971. Joint Probabilities of Selection of Units in Systematic Samples. *Proceedings for the American Statistical Association*. 271-276.

Hidiroglou, M.A. 1994. Sampling and Estimation for Establishment Surveys: Stumbling Blocks and Progress. *Proceedings of the Section on Survey Research Methods*. American Statistical Association. 153-162.

Hidiroglou, M.A. and G.B. Gray. 1980. Construction of Joint Probabilities of Selection for Systematic P.P.S. Sampling. *Applied Statistics,* 29(1): 663-685.

Hidiroglou, M.A. and K.P. Srinath. 1993. Problems Associated with Designing Sub-Annual Business Surveys. *Journal of Economic Statistics,* 11: 397-405.

Horvitz, D.G. and D.J. Thompson. 1952. A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association,* 47: 663-685.

Kalton, G. 1992. *Panel Surveys: Adding the Fourth Dimension*. Proceedings of Statistics Canada Symposium 1992: Design and Analysis of Longitudinal Surveys. 7-18.

Kalton, G., J. Kordos and R. Platek, eds. 1992. *Small Area Statistics and Survey Designs*. Central Statistical Office, Warsaw. 31-75.

Kasprzyk, D., G.J. Duncan, G. Kalton and M.P. Singh, eds. 1989. *Panel Surveys*. John Wiley and Sons, New York.

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

Lavallée, P. 1998. Course notes for *Theory and Application of Longitudinal Surveys*, Statistics Canada.

Levy, P. and S. Lemeshow. 1991. *Sampling of Populations*. John Wiley and Sons, New York.

Lohr, Sharon. 1999. *Sampling: Design and Analysis*. Duxbury Press, U.S.A.

McLeod, A.I. and D.R. Bellhouse. 1983. A Convenient Algorithm for Drawing a SRS. *Applied Statistics,* 32(2).

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

Rao, J.N.K, H.O. Hartley and W.G. Cochran. 1962. On a Simple Procedure of Unequal Probability Sampling Without Replacement. *Journal of the Royal Statistical Society,* B, 27: 482-490.

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

Satin, A. and W. Shastry. 1993. *Survey Sampling: A Non-Mathematical Guide – Second Edition*. Statistics Canada. 12-602E.

Stuart, A. 1968. *Basic Ideas of Scientific Sampling*. Charles Griffin and Company Limited, London.

Thompson, M. 1997. *Theory of Sample Surveys*. Chapman and Hill, United Kingdom.

Thompson, S.K. 1992. *Sampling*. John Wiley and Sons, New York.

Yates, F. and P.M. Grundy. 1953. Selection Without Replacement from Within Strata with Probability-proportional-to-size. *Journal of the Royal Statistical Society,* B, 15: 235-261.