22

Sample Surveys

Statistics in practice The British crime survey

22.1 Terminology used in sample surveys

22.2 Types of surveys and sampling methods

22.3 Survey errors
- Non-sampling error
- Sampling error

22.4 Simple random sampling
- Population mean
- Population total
- Population proportion
- Determining the sample size

22.5 Stratified random sampling
- Population mean
- Population total
- Population proportion
- Determining the sample size

22.6 Cluster sampling
- Population mean
- Population total
- Population proportion
- Determining the sample size

22.7 Systematic sampling

Learning Objectives

After studying this chapter and doing the exercises, you should be able to:

1. Explain what is meant by the following terms: probabilistic sampling, non-probabilistic sampling, convenience sampling, judgement sampling, simple random sampling, stratified random sampling and cluster sampling.
2. Outline the advantages and disadvantages of these different methods of sampling.
3. Calculate point estimates of the population mean, the population total and the population proportion for simple random samples, stratified random samples and single-stage cluster samples, and construct interval estimates for these parameters.
4. Make estimates of the sample size needed to achieve a given precision in estimating these population parameters, for simple random sampling and stratified random sampling.

22.1 Terminology used in sample surveys

In Chapter 1 we presented the following definitions of an element, a population, and a sample.

- An **element** is the entity on which data are collected.
- A **population** is the collection of all the elements of interest.
- A **sample** is a subset of the population.

To illustrate these concepts, consider the following situation. The PC Shop (PCS), a manufacturer of personal computers and peripherals, would like to collect data about the characteristics of individuals who purchased a PCS personal computer. To obtain such data, a sample survey of PCS personal computer owners could be conducted. The *elements* in this sample survey would be individuals who purchased a PCS personal computer. The *population* would be the collection of all people who purchased a PCS personal computer, and the *sample* would be the subset of PCS personal computer owners who are surveyed.

In sample surveys it is necessary to distinguish between the target population and the sampled population. The **target population** is the population we want to make inferences about, while the **sampled population** is the population from which the sample is actually selected. It is important to understand that these two populations are not always the same. In the PCS example, the target population consists of all people who purchased a PCS personal computer. The sampled population, however, might be all owners who sent warranty registration cards back to PCS. Not every person who buys a PCS personal computer sends in the warranty card, so the sampled population would differ from the target population. Conclusions drawn from a sample survey apply only to the sampled population. Whether these conclusions can be extended to the target population depends on the judgement of the analyst. The key issue is whether the correspondence between the sampled population and the target population on the characteristics of interest is close enough to allow this extension.

Before sampling, the population must be divided into **sampling units**. In some cases, the sampling units are simply the elements. In other cases, the sampling units are groups of the elements. For example, suppose we want to survey chartered engineers

STATISTICS IN PRACTICE

The British crime survey *Home Office, UK Government*

The British Crime Survey (BCS) is an annual survey, the results of which receive great media attention, because crime has become such an important political and economic issue in the UK. For example, on 21 July 2005, after publication of the results from the 2004/2005 British Crime Survey, the *Guardian* newspaper headlined an article 'Overall crime down by 44 per cent since 1995', followed by the sub-headline 'Continued fall marred by apparent increase in violent incidents'.

The BCS aims to measure the amount of crime being experienced by private individuals in England and Wales, including crime that is not reported to the police. It therefore complements the information from official police statistics. The survey also assesses people's attitudes to crime and to the criminal justice system. The Home Office, who sponsor the survey, describe it on their website as 'a very important source of information about levels of crime and other Home Office issues'.

Metropolitan Police, United Kingdom

© Tomasz Resiak

The survey involves interviews with over 40 000 individuals each year. It is designed to be representative of adults aged over 16 living in private residential accommodation in England and Wales. The design for the core sample of interviewees uses a probabilistic sampling method known as stratified, two-stage cluster sampling. Addresses of households are selected randomly from the Postcode Address File (PAF), a database maintained primarily for postal delivery purposes by the UK Post Office. The PAF is stratified by Police Force Area and by population density (i.e. divided into sub-groups using these criteria) before random selection of addresses takes place. One aim in the sample design is to achieve a minimum of 600–700 interviews in the core sample in each Police Force Area. Consequently, smaller Police Force Areas are sampled disproportionately to achieve this objective.

Random sampling from the PAF is initially of postcode sectors (i.e. geographical groupings of postal delivery addresses), then 32 addresses are randomly selected from each sampled postcode sector. This is known as two-stage cluster sampling. It produces geographically clustered groups of sampled addresses. The geographical clustering tends to reduce the precision of estimates that are made from the BCS sample, but the efficiency gains outweigh this disadvantage, because the geographical clustering reduces survey costs.

In this chapter you will learn about the issues that statisticians consider in the design and execution of a complex sample survey such as the British Crime Survey.

who are involved in the design of heating and air conditioning systems for commercial buildings. If a list of all chartered engineers involved in such work were available, the sampling units would be the professional engineers we want to survey. If such a list is not available, we must find an alternative approach. A business telephone directory might provide a list of all engineering firms involved in the design of heating and air conditioning systems. Given this list, we could select a sample of the engineering firms to survey. Then, for each firm surveyed, we might interview all the professional engineers. In this case, the engineering firms would be the sampling units and the engineers interviewed would be the elements. A list of the sampling units for a particular study is called a **frame**. In the sample survey of chartered engineers, the frame is defined as

all engineering firms listed in the business telephone directory. The frame is not a list of all chartered engineers because no such list is available. The choice of a particular frame and hence the definition of the sampling units is often determined by the availability and reliability of a list. In practice, the development of the frame can be one of the most difficult and important steps in conducting a sample survey.

22.2 Types of surveys and sampling methods

The three most common types of surveys historically have been mail surveys, telephone surveys, and personal interview surveys. In recent years, online surveys have become much more widespread. Each of these survey types involves the design and administration of a questionnaire. Survey costs tend to be lower for mail, online and telephone surveys. But with well-trained interviewers, higher response rates and longer questionnaires are possible with personal interviews. Other types of surveys used to collect data do not necessarily involve questionnaires. For example, accounting firms are often hired to sample a company's inventory of goods to estimate the value of inventory on the company's balance sheet. In such surveys, someone simply counts the items and records the results.

In surveys that use questionnaires, the design of the questionnaire is critical. The designer must resist the temptation to include questions that *might* be of interest, because every question adds to the length of the questionnaire. Long questionnaires can lead to respondent fatigue, especially in mail, online and telephone surveys. However, if personal interviews are to be used, a longer and more complex questionnaire is feasible. A large body of knowledge exists concerning the phrasing, sequencing, and grouping of questions for a questionnaire. These issues are discussed in more comprehensive books on survey sampling. Several sources for this type of information are listed in the bibliography.

Sample surveys can also be classified in terms of the sampling method used. With **probabilistic sampling**, the probability of obtaining each possible sample can be computed. With a **non-probabilistic sampling** method, this probability is unknown. Non-probabilistic sampling methods should not be used if the researcher wants to make statements about the precision of the estimates. In contrast, probabilistic sampling methods can be used to develop confidence intervals that provide bounds on the sampling error. In the following sections, four of the most popular probabilistic sampling methods are discussed: simple random sampling, stratified random sampling, cluster sampling and systematic sampling. Although statisticians prefer to use a probabilistic sampling method, non-probabilistic sampling methods often are necessary. The advantages of non-probabilistic sampling methods are their low expense and ease of implementation. The disadvantage is that statistically valid statements cannot be made about the precision of the estimates. Two of the more common non-probabilistic methods are convenience sampling and judgement sampling. With **convenience sampling**, the units included in the sample are chosen because of accessibility. For example, a professor conducting a research study at a university may ask student volunteers to participate in the study simply because they are in the professor's class. In this case, the sample of students is referred to as a convenience sample. In some situations, convenience sampling is the only practical approach. For example, to sample a shipment of oranges, an inspector might select oranges haphazardly from several crates, because labelling each orange in the entire shipment to create a frame, and using a probabilistic sampling method, would be impractical. Wildlife captures and volunteer panels for consumer research are other examples of convenience samples. Although convenience sampling

is a relatively easy approach to sample selection and data gathering, it is impossible to evaluate the 'goodness' of the sample statistics obtained in terms of their ability to estimate the population parameters of interest. A convenience sample may provide good results or it may not. There is no statistically justified procedure for making any statistical inferences from the sample results.

In using the non-probability sampling technique referred to as **judgement sampling**, a person knowledgeable on the subject of the study selects sampling units that he or she feels are most representative of the population. Although judgement sampling is often a relatively easy way to select samples, users of the survey results must recognize that the quality of the results is dependent on the judgement of the person selecting the sample. Consequently, caution must be exercised in using judgement samples to make statistical inferences about a population parameter.

Both probabilistic and non-probabilistic sampling methods can be used to select a sample. The advantage of non-probabilistic methods is that they are generally inexpensive and easy to use. However, if it is necessary to provide statements about the precision of the estimates, a probabilistic sampling method should be used. Nevertheless, at times some researchers apply a statistical method designed for a probability sample to the data gathered from a non-probabilistic sample. In doing so, the researcher may argue that the non-probabilistic sample can be treated as though it were a random sample in the sense that it is representative of the population. However, this argument should be questioned. One should be cautious in using a non-probabilistic sample, particularly a convenience sample, to make statistical inferences about population parameters.

22.3 Survey errors

Two types of errors can occur in conducting a survey. One type, **sampling error**, is defined as the magnitude of the difference between an unbiased point estimate calculated from the sample and the population parameter. In other words, sampling error is the error that occurs because not every element in the population is surveyed. The second type, **non-sampling error**, refers to all other types of errors that can occur when a survey is conducted, such as measurement error, interviewer error, and processing error. Although sampling error can occur only in a sample survey, some types of non-sampling errors can occur in both a census and a sample survey.

Non-sampling error

One of the most common types of non-sampling error occurs whenever we incorrectly measure the characteristic of interest. Measurement error can occur in a census or a sample survey. For either type of study, the researcher must exercise care to ensure that any measuring instruments (e.g., the questionnaire) are properly calibrated and that the people who take the measurements are properly trained. Attention to detail is the best precaution in most situations.

Errors due to non-response are a concern to both the statistician responsible for designing the survey and the manager using the results. This type of non-sampling error occurs when data cannot be obtained for some of the units surveyed or when only partial data are obtained. The problem is most serious when a bias is created. For example, if interviews were conducted to assess women's attitudes toward working outside the home, making house calls only during the daytime would create an obvious bias because women who work outside the home would be excluded from the sample.

Non-sampling errors due to lack of respondent knowledge are common in technical surveys. For example, suppose building managers were surveyed to obtain detailed information about the types of ventilation systems used in office buildings. Managers of large office buildings may be especially knowledgeable about such systems because they may attend training seminars and have support staff to help keep them current. In contrast, managers of smaller office buildings may be less knowledgeable about such systems because of the wide variety of duties they must perform. This difference in knowledge can significantly affect the survey results.

Two other types of non-sampling error are selection error and processing error. Selection errors occur when an inappropriate sampling unit is included in the survey. Suppose a sample survey was designed to develop a profile of men with beards. If some interviewers interpreted the statement 'men with beards' to include men with moustaches while other interviewers did not, the resulting data would be flawed. Processing errors occur whenever data are incorrectly recorded or incorrectly transferred from recording forms, such as from questionnaires to computer files.

Although some non-sampling errors will occur in most surveys, they can be minimized by careful planning, including ensuring that the sampled population corresponds closely to the target population, proper interviewer training, good questionnaire design and pre-testing, and careful management of the process of coding and transferring the data to the computer. The final report on a survey should include some discussion of the likely impact of non-sampling errors on the results.

Sampling error

Recall the PC Shop (PCS) sample survey mentioned previously. Suppose PCS wants to estimate the mean age of people who have purchased a PCS personal computer. If the entire population of PCS personal computer owners could be surveyed (a census) and non-sampling errors were not present, we could determine the mean age exactly. But what if less than 100 per cent of the population of PCS owners can be surveyed? In this case, there will most likely be a difference between the sample mean and the population mean. The absolute value of this difference is the sampling error. In practice, it is not possible to know what the sampling error will be for any one particular sample because the population mean is unknown. However, it is possible to provide probability statements about the size of the sampling error.

Sampling error occurs because a sample, and not the entire population, is surveyed. Even though sampling error cannot be avoided, it can be controlled. Selecting an appropriate sampling method or design is one important way to control this type of error. In the following sections we will discuss four probabilistic sampling methods: simple random sampling, stratified random sampling, cluster sampling and systematic sampling.

22.4 Simple random sampling

The definition of simple random sampling was presented in Chapter 7:

> A simple random sample of size n from a finite population of size N is a sample selected such that every possible sample of size n has the same probability of being selected.

To conduct a sample survey using **simple random sampling**, we begin by constructing a frame or list of all elements in the sampled population. Then a selection procedure, based on the use of random numbers, is used to ensure that each element in the sampled population has the same probability of being selected. In this section we show how estimates of a population mean, total and proportion are made for sample surveys that use simple random sampling.

Population mean

In Chapter 8 we showed that the sample mean $\bar{x}$ provides an estimate of the population mean μ, and the sample standard deviation s provides an estimate of the population standard deviation σ. For a sample of size n, the t distribution can be used to provide the following interval estimate of μ.

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} \tag{22.1}$$

In expression (22.1), $s/\sqrt{n}$ is the estimate of $\sigma_{\bar{x}}$, the standard error of the mean.

When a simple random sample of size n is selected from a finite population of size N, an estimate of the standard error of the mean is

$$s_{\bar{x}} = \sqrt{\frac{N-n}{N}}\left(\frac{s}{\sqrt{n}}\right) \tag{22.2}$$

Using $s_{\bar{x}}$ as an estimate of $\sigma_{\bar{x}}$, the interval estimate of the population mean becomes

$$\bar{x} \pm t_{\alpha/2}s_{\bar{x}} \tag{22.3}$$

In a sample survey it is common practice to use a value of $t = 2$ when developing interval estimates. Hence, when simple random sampling is used, an approximate 95 per cent confidence interval estimate of the population mean is given by the following expression.

Approximate 95 per cent confidence interval estimate of the population mean

$$\bar{x} \pm 2s_{\bar{x}} \tag{22.4}$$

As an example, consider the situation of the publisher of *On Board*, a specialist magazine aimed at the surf-boarding and sail-boarding communities. The magazine currently has $N = 8000$ subscribers. A simple random sample of $n = 484$ subscribers shows mean annual income to be €37 500 with a standard deviation of €9040. An unbiased estimate of the mean annual income of all subscribers is given by $\bar{x}$ = €37 500. Using the sample results and equation (22.2), we obtain the following estimate of the standard error of the mean.

$$s_{\bar{x}} = \sqrt{\frac{8000-484}{8000}}\left(\frac{9040}{\sqrt{484}}\right) = 398$$

Therefore, using formula (22.4), we find that an approximate 95 per cent confidence interval estimate of the mean annual income for the magazine subscribers is

$$37\,500 \pm 2(398) = 37\,500 \pm 796$$

or €36 704 to €38 296.

The number added to and subtracted from a point estimate to create an interval estimate is called the **bound on the sampling error**. For example, in the *On Board* sample survey, an estimate of the standard error of the point estimator is $s_{\bar{x}}$ = €398, and the bound on the sampling error is 2(€398) = €796.

The preceding procedure can be used to compute an interval estimate for other population parameters such as the population total or the population proportion. In these cases, the approximate 95 per cent confidence interval can be written as

Point estimator ± 2(Estimate of the standard error of the point estimator)

Population total

Consider the problem facing Northern Electricity and Gas (NEG). As part of an energy usage study, NEG needs to estimate the *total* floor area for the 500 schools in its service area. We will denote the total floor area for the 500 schools as X. In other words, X denotes the population total. Note that if μ, the mean floor for the 500 schools, were known, the value of X could be computed by multiplying together N and μ. However, because μ is unknown, a point estimate of X is obtained by multiplying together N and $\bar{x}$. We shall denote the point estimator of X as $\hat{X}$.

Point estimator of a population total

$$\hat{X} = N\bar{x} \tag{22.5}$$

An estimate of the standard error of this point estimator is given by

$$s_{\hat{X}} = Ns_{\bar{x}} \tag{22.6}$$

where

$$s_{\bar{x}} = \sqrt{\frac{N-n}{N}}\left(\frac{s}{\sqrt{n}}\right) \tag{22.7}$$

Note that equation (22.7) is the formula for the estimated standard error of the mean. With this standard error and equation (22.6), an approximate 95 per cent confidence interval for the population total is given by the following expression.

Approximate 95 per cent confidence interval estimate of the population total

$$N\bar{x} \pm 2s_{\hat{X}} \tag{22.8}$$

Suppose that in the NEG study a simple random sample of $n = 50$ schools is selected from the population of $N = 500$ schools. The sample mean is $\bar{x} = 2000$ square metres and the sample standard deviation is $s = 400$ square metres. Using equation (22.5), we find that the point estimator of the population total is

$$\hat{X} = (500)(2000) = 1\,000\,000$$

Equation (22.7) can be used to compute an estimate of the standard error of the mean.

$$s_{\hat{X}} = \sqrt{\frac{500 - 50}{500}}\left(\frac{400}{\sqrt{50}}\right) = 53.67$$

Then, using equation (22.6), we can obtain an estimate of the standard error of $\hat{X}$.

$$s_{\hat{X}} = (500)(53.67) = 26\,833$$

Therefore, using expression (22.8), we find that an approximate 95 per cent confidence interval estimate of the total floor area for the 500 schools in NEG's service area is

$$1\,000\,000 \pm 2(26\,833) = 1\,000\,000 \pm 53\,666$$

or 946 334 to 1 053 666 square metres.

Population proportion

The population proportion π is the fraction of the elements in the population with some characteristic of interest. In a market research study, for example, one might be interested in the proportion of consumers preferring a certain brand of product. The sample proportion p is an unbiased point estimator of the population proportion. An estimate of the standard error of the proportion is given by

$$s_p = \sqrt{\left(\frac{N - n}{N}\right)\left(\frac{p(1 - p)}{n}\right)} \tag{22.9}$$

An approximate 95 per cent confidence interval estimate of the population proportion is given by the following expression.

Approximate 95 per cent confidence interval estimate of the population proportion

$$p \pm 2s_p \tag{22.10}$$

As an illustration, suppose that in the Northern Electricity and Gas sampling problem, NEG would also like to estimate the proportion of the 500 schools in its service area that use gas as fuel for heating. If 35 of the 50 sampled schools indicate the use of gas, the point estimate of the proportion of the 500 schools in the population that use natural gas is $p = 35/50 = 0.70$. Using equation (22.9), we can compute an estimate of the standard error of the proportion.

$$s_p = \sqrt{\left(\frac{500 - 50}{500}\right)\left(\frac{0.7(1 - 0.7)}{50}\right)} = 0.0621$$

Therefore, using expression (22.10), we find that an approximate 95 per cent confidence interval for the population proportion is

$$0.7 \pm 2(0.0621) = 0.7 \pm 0.1242$$

or 0.5758 to 0.8242.

As this example shows, the width of the confidence interval can be rather large when a population proportion is being estimated. In general, large sample sizes are needed to obtain precise estimates of population proportions. A report of a sample survey of 529 mutual fund investors conducted by Louis Harris & Associates stated, 'Results should be accurate to within 4.3 percentage points'. This statement implies an approximate 95 per cent confidence interval with a width of 0.086. For large populations, samples of $n = 1000$ or more are often used.

Determining the sample size

An important consideration in sample design is the choice of sample size. The best choice usually involves a trade-off between cost and precision. Larger samples provide greater precision (tighter bounds on the sampling error), but are more costly. Often the budget for a study will dictate how large the sample can be. In other cases, the size of the sample must be large enough to provide a specified level of precision.

A common approach to choosing the sample size is to first specify the precision desired and then determine the smallest sample size providing that precision. In this context, the term *precision* refers to the size of the approximate confidence interval. Smaller confidence intervals provide more precision. Because the size of the approximate confidence interval depends on the bound B on the sampling error, choosing a level of precision amounts to choosing a value for B. Let us see how this approach works in choosing the sample size necessary to estimate the population mean.

Recall that the bound on the sampling error is '2 times the estimate of the standard error of the point estimator'. Using the expression in equation (22.2) for the standard error for the mean, we have

$$B = 2\sqrt{\frac{N - n}{N}}\left(\frac{s}{\sqrt{n}}\right) \tag{22.11}$$

Solving equation (22.11) for n will provide a bound on the sampling error equal to B. Doing so yields

$$n = \frac{Ns^2}{N\left(\frac{B^2}{4}\right) + s^2} \tag{22.12}$$

Once a desired level of precision has been selected (by choosing a value for B), equation (22.12) can then be used to find the value of n that will provide the desired level of precision. Using equation (22.12) to choose n for a practical study presents problems, however. In addition to specifying the desired bound on the sampling error B, one must know the sample variance s^2, but s^2 will not be known until the sample is actually taken.

Cochran* suggests several ways to develop an estimate of s^2 in practice. Three of them are stated as the following:

1. Take the sample in two stages. Use the value of s^2 found in stage 1 in equation (22.12); the resulting value of n is what the size of the total sample must be. Then, select the number of additional units needed at stage 2 to provide the total sample size determined in stage 1.
2. Use the results of a pilot survey or pretest to estimate s^2.
3. Use information from a previous sample.

Let us now consider an example involving the estimate of the population mean for starting salaries of graduates of a particular university. Suppose, with $N = 5000$ graduates, we want to construct an approximate 95 per cent confidence interval with a width of at most €500. To provide such a confidence interval, $B = 250$. Before using equation (22.12) to determine the sample size, we need an estimate of s^2. Suppose a study of starting salaries conducted last year found that $s = €1500$. We can use the data from this previous sample to estimate s^2. Using $B = 250$, $s = 1500$, and $N = 5000$, we can now use equation (22.12) to determine the sample size.

$$n = \frac{5000(1500)^2}{5000\left(\frac{(250)^2}{4}\right) + (1500)^2} = 139.97$$

Rounding up, we see that a sample size of 140 will provide an approximate 95 per cent confidence interval of width €500. Keep in mind, however, that this calculation is based on the initial estimate of $s = €1500$. If s turns out to be larger in this year's sample survey, the resulting approximate confidence interval will have a width greater than €500. Consequently, if cost considerations permit, the survey designer might choose a sample size of, say, 150 to provide added assurance that the final approximate 95 per cent confidence interval will have a width less than €500.

The formula for determining the sample size necessary for estimating a population total with a bound B on the sampling error is as follows.

$$n = \frac{Ns^2}{\left(\frac{B^2}{4N}\right) + s^2} \tag{22.13}$$

In the previous example, we wanted to estimate the mean starting salary with a bound on the sampling error of $B = 250$. Suppose we are also interested in estimating the total salary of all 5000 graduates with a bound of €1 million. We can use equation (22.13) with $B = 1\,000\,000$ to find the sample size needed to provide such a bound on the population total.

$$n = \frac{5000(1500)^2}{\left(\frac{(1\,000\,000)^2}{4(5000)}\right) + (1500)^2} = 215.31$$

Rounding up, we see that a sample size of 216 is necessary to provide an approximate 95 per cent confidence interval with a bound of €1 million. We note here that if the

* William G. Cochran, *Sampling Techniques*, 3rd ed., Wiley, 1977.

same survey is expected to provide a bound of €250 on the population mean and a bound of €1 million on the population total, a sample size of at least 216 must be used. This size will provide a tighter bound than necessary on the population mean, while providing the minimum desired precision for the population total.

To choose the sample size for estimating a population proportion, we use a formula similar to the one for the population mean. We simply substitute $p(1 - p)$ for s^2 in equation (22.12) to obtain

$$n = \frac{Np(1 - p)}{N\left(\frac{B^2}{4}\right) + p(1 - p)} \tag{22.14}$$

To use equation (22.14), we must specify the desired bound B and provide an estimate of p. If a good estimate of p is not available, we can use $p = 0.5$. This will ensure that the resulting approximate confidence interval will have a bound on the sampling error at least as small as desired.

EXERCISES

Methods

1 Simple random sampling was used to obtain a sample of $n = 50$ elements from a population of $N = 800$. The sample mean was $\bar{x} = 215$, and the sample standard deviation was found to be $s = 20$.

a. Estimate the population mean.
b. Estimate the standard error of the mean.
c. Construct an approximate 95 per cent confidence interval for the population mean.

2 Simple random sampling was used to obtain a sample of $n = 80$ elements from a population of $N = 400$. The sample mean was $\bar{x} = 75$, and the sample standard deviation was found to be $s = 8$.

a. Estimate the population total.
b. Estimate the standard error of the population total.
c. Develop an approximate 95 per cent confidence interval for the population total.

3 Simple random sampling was used to obtain a sample of $n = 100$ elements from a population of $N = 1000$. The sample proportion was $p = 0.30$.

a. Estimate the population proportion.
b. Estimate the standard error of the proportion.
c. Develop an approximate 95 per cent confidence interval for the population proportion.

4 A sample is to be taken to develop an approximate 95 per cent confidence interval estimate of the population mean. The population consists of 450 elements, and a pilot study resulted in $s = 70$. How large must the sample be if we want to construct an approximate 95 per cent confidence interval with a width of 30?

Applications

5 There are 376 district and unitary local authorities in England and Wales. Suppose you take a simple random sample of 50 of them and find that the mean number of people living in these localities is 135 210, with a sample standard deviation of 93 030. You find that 29 of the sampled local authorities have populations of more than 100 000.

a. Construct an approximate 95 per cent confidence interval for the mean number of residents in all 376 localities.
b. Construct an approximate 95 per cent confidence interval for the total population of all 376 localities.
c. Construct an approximate 95 per cent confidence interval for the proportion of all local authorities that have populations of more than 100 000.

6 *The Wall Street Journal* conducted a survey of subscribers to its interactive edition. One question asked the 504 respondents whether they used a laptop computer when travelling; 55 per cent said they did. Another question asked respondents whether they used an express or package service when travelling; 31 per cent said they did. (*The Wall Street Journal Interactive Edition Subscriber Survey*, 2000).

a. Calculate an estimate of the standard error of the proportion for the proportion that uses a laptop computer.
b. Calculate an estimate of the standard error of the proportion for the proportion that uses an express or package service.
c. Are the estimates of the standard error the same in parts (a) and (b)? If they differ, explain why.
d. Construct an approximate 95 per cent confidence interval for the proportion that uses a laptop computer.
e. Construct an approximate 95 per cent confidence interval for the proportion that uses an express or package service.

7 A quality of life survey was conducted with employees of a manufacturing firm. Of the firm's 3000 employees, a sample of 300 were sent questionnaires. Two hundred usable questionnaires were obtained, giving a response rate of 67 per cent.

a. The mean annual salary for the sample was £23 200 with s = £3000. Construct an approximate 95 per cent confidence interval for the mean annual salary of the population.
b. Use the information in part (a) to construct an approximate 95 per cent confidence interval for the total salary of all 3000 employees.
c. There were 73 per cent of the respondents who reported that they were 'generally satisfied' with their job. Construct an approximate 95 per cent confidence interval for the population proportion.
d. Comment on whether you think the results in part (c) might be biased. Would your opinion change if you knew the respondents were guaranteed anonymity?

22.5 Stratified random sampling

In **stratified random sampling**, the population is first divided into H groups, called strata. Then for stratum h a simple random sample of size n_h is selected. The data

from the H simple random samples are combined to develop an estimate of a population parameter such as the population mean, total or proportion.

If the variability within each stratum is smaller than the variability across the strata, a stratified random sample can lead to greater precision (narrower interval estimates of the population parameters). The basis for forming the various strata depends on the judgement of the designer of the sample. Depending on the application, a population might be stratified by department, location, age, product type, industry type, sales levels, and so on.

As an example, suppose the School of Business at South Downs University wants to conduct a survey of this year's graduates to learn about their starting salaries. There are five degree programmes in the university: accounting, finance, information systems, marketing, and operations management. Of the $N = 1500$ students who graduated this year, there were $N_1 = 500$ accounting students, $N_2 = 350$ finance students, $N_3 = 200$ information systems students, $N_4 = 300$ marketing students, and $N_5 = 150$ operations management students. Analysis of previous salary data suggests more variability in starting salaries across degree programmes than within each degree programme. As a result, a stratified random sample of $n = 180$ students is selected: 45 of the 500 students who graduated in accounting ($n_1 = 45$), 40 who graduated in finance ($n_2 = 40$), 30 who graduated in information systems ($n_3 = 30$), 35 who graduated in marketing ($n_4 = 35$), and 30 who graduated in operations management ($n_5 = 30$).

Population mean

In stratified sampling an unbiased estimate of the population mean is obtained by computing a weighted average of the sample means for each stratum. The weights used are the fraction of the population in each stratum. The resulting point estimator, denoted $\bar{x}_{st}$, is defined as follows.

Point estimator of the population mean

$$\bar{x}_{st} = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right) \bar{x}_h \tag{22.15}$$

H = number of strata

$\bar{x}_h$ = sample mean for stratum h

N_h = number of elements in the population in stratum h

N = total number of elements in the population; $N = N_1 + N_2 + \cdots + N_H$

For stratified random sampling, the formula used for computing an estimate of the standard error of the mean is a function of s_h, the sample standard deviation for stratum h.

$$s_{\bar{x}_{st}} = \sqrt{\frac{1}{N^2} \sum_{h=1}^{H} N_h (N_h - n_h) \frac{s_h^2}{n_h}} \tag{22.16}$$

Using these results, we see that an approximate 95 per cent confidence interval estimate of the population mean is given by the following expression.

Approximate 95 per cent confidence interval estimate of the population mean

$$\bar{x}_{st} \pm 2s_{\bar{x}_{st}} \tag{22.17}$$

Suppose the survey of 180 graduates of the School of Business at South Downs University provided the sample results shown in Table 22.1. The sample means for each degree programme, or stratum, are €25 500 for accounting, €24 750 for finance, €28 750 for information systems, €24 000 for marketing, and €26 000 for operations management. Using these results and equation (22.15), we can compute a point estimate of the population mean.

$$\bar{x}_{st} = \left(\frac{500}{1500}\right)(25\,500) + \left(\frac{350}{1500}\right)(24\,750) + \left(\frac{200}{1500}\right)(28\,750) + \left(\frac{300}{1500}\right)(24\,000) + \left(\frac{150}{1500}\right)(26\,000) = 25\,508$$

Table 22.1 South Downs University sample survey of starting salaries of graduates

Degree programme (h)	$\bar{x}_h$	s_h	N_h	n_h
Accounting	25 500	1000	500	45
Finance	24 750	850	350	40
Information systems	28 750	1150	200	30
Marketing	24 000	800	300	35
Operations management	26 000	1125	150	30

In Table 22.2 we show a portion of the calculations needed to estimate the standard error; note that

$$\sum_{h=1}^{H} N_h(N_h - n_h)\frac{s_h^2}{n_h} = 10\,727\,259\,423$$

Hence,

$$s_{\bar{x}_{st}} = \sqrt{\left(\frac{1}{1500^2}\right)(10\,727\,259\,423)} = \sqrt{4767.67} = 69.04$$

Using equation (22.17), we find that an approximate 95 per cent confidence interval estimate of the population mean is 25 508 ± 2(69) = 25 508 ± 138, or €25 370 to €25 646.

Table 22.2 Partial calculations for the estimate of the standard error of the mean for the South Downs University sample survey

Degree programme	h	$N_h(N_h - n_h)\dfrac{s_h^2}{n_h}$
Accounting	1	$500(500 - 45)\dfrac{(1000)^2}{45} = 5\,055\,555\,556$
Finance	2	$350(350 - 40)\dfrac{(850)^2}{40} = 1\,959\,781\,250$
Information systems	3	$200(200 - 30)\dfrac{(1150)^2}{30} = 1\,498\,833\,333$
Marketing	4	$300(300 - 35)\dfrac{(800)^2}{35} = 1\,453\,714\,286$
Operations management	5	$150(150 - 30)\dfrac{(1125)^2}{30} = 759\,375\,000$
		$10\,727\,259\,423$

Population total

The point estimator of the population total (X) is obtained by multiplying together N and $\bar{x}_{st}$.

Point estimator of the population total

$$\hat{X} = N\bar{x}_{st} \tag{22.18}$$

An estimate of the standard error of this point estimator is

$$s_{\hat{X}} = Ns_{\bar{x}_{st}} \tag{22.19}$$

Hence, an approximate 95 per cent confidence interval for the population total is given by the following expression.

Approximate 95 per cent confidence interval estimate of the population total

$$N\bar{x}_{st} \pm 2s_{\hat{X}} \tag{22.20}$$

Now suppose the School of Business at South Downs University would also like to estimate the total earnings of the 1500 business graduates in order to estimate their

impact on the economy. Using equation (22.18), we obtain an unbiased estimate of the total earnings.

$$\hat{X} = (1500)25\,508 = 38\,262\,000$$

Using equation (22.19), we obtain an estimate of the standard error of the population total.

$$s_{\hat{X}} = 1500(69) = 103\,500$$

Hence, using expression (22.20), we find that an approximate 95 per cent confidence interval estimate of the total earnings of the 1500 graduates is $38\,262\,000 \pm 2(103\,500) = 38\,262\,000 \pm 207\,000$ or €38 055 000 to €386 469 000.

Population proportion

An unbiased estimate of the population proportion, π, for stratified random sampling is a weighted average of the proportions for each stratum. The weights used are the fraction of the population in each stratum. The resulting point estimator, denoted p_{st} is defined as follows.

Point estimator of the population proportion

$$p_{st} = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right) p_h \qquad (22.21)$$

where
H = the number of strata
p_h = the sample proportion for stratum h
N_h = the number of elements in the population in stratum h
N = the total number of elements in the population: $N = N_1 + N_2 + \cdots + N_H$

An estimate of the standard error of p_{st} is given by

$$s_{p_{st}} = \sqrt{\frac{1}{N^2} \sum_{h=1}^{H} N_h(N_h - n_h)\left[\frac{p_h(1-p_h)}{n_h}\right]} \qquad (22.22)$$

Hence, an approximate 95 per cent confidence interval estimate of the population proportion is given by the following expression.

Approximate 95 per cent confidence interval estimate of the population proportion

$$p_{st} \pm 2s_{p_{st}} \qquad (22.23)$$

In the South Downs University survey, the university wants to know the proportion of graduates receiving a starting salary of €26 000 or more. The results of the sample survey of 180 graduates show that 63 received starting salaries of €26 000 or more and that 16 of the 63 graduated in accounting, three graduated in finance, 29 graduated in information systems, 0 graduated in marketing, and 15 graduated in operations management.

Using equation (22.21), we can compute the point estimate of the proportion receiving starting salaries of 18 000 or more.

$$p_{st} = \left(\frac{500}{1500}\right)\left(\frac{16}{45}\right) + \left(\frac{350}{1500}\right)\left(\frac{3}{40}\right) + \left(\frac{200}{1500}\right)\left(\frac{29}{30}\right) + \left(\frac{300}{1500}\right)\left(\frac{0}{35}\right) + \left(\frac{150}{1500}\right)\left(\frac{15}{30}\right) = 0.3149$$

In Table 22.3 we show a portion of the calculations needed to estimate the standard error; note that

$$\sum_{h=1}^{H} N_h(N_h - n_h)\left[\frac{p_h(1 - p_h)}{n_h - 1}\right] = 1533.11$$

Hence,

$$s_{p_{st}} = \sqrt{\frac{1}{1500^2}(1533.11)} = 0.0261$$

Using expression (22.23), we find that an approximate 95 per cent confidence interval for the proportion of graduates receiving starting salaries of €26 000 or more is $0.3149 \pm 2(0.0261) = 0.3149 \pm 0.0522$, or 0.2627 to 0.3671.

Table 22.3 Partial calculations for the estimate of the standard error of p_{st} for the Lakeland College sample survey

Degree programme	h	$N_h(N_h - n_h)\left[\frac{p_h(1 - p_h)}{n_h}\right]$
Accounting	1	$500(500 - 45)\left[\frac{(16/45)(29/45)}{45}\right] = 1158.41$
Finance	2	$350(350 - 40)\left[\frac{(3/40)(37/40)}{40}\right] = 188.18$
Information systems	3	$200(200 - 30)\left[\frac{(29/30)(1/30)}{30}\right] = 36.52$
Marketing	4	$300(300 - 35)\left[\frac{(0/35)(35/35)}{55}\right] = 0.0000$
Operations management	5	$150(150 - 30)\left[\frac{(15/30)(15/30)}{30}\right] = 150.00$
		1533.11

Determining the sample size

With stratified random sampling we can think of choosing a sample size as a two-step process. First, a total sample size n must be chosen. Second, we must decide how to assign the sampled units to the various strata. Alternatively, we could first decide how large a sample to take in each stratum and then sum the stratum sample sizes to obtain the total sample size. It is often of interest to calculate estimates of the mean, total and proportion for the individual strata, therefore a combination of these two approaches is often employed. An overall sample size n and an allocation that will provide the necessary precision for the overall population parameter of interest are found. Then, if the sample sizes in some of the strata are not large enough to provide the precision necessary for the estimates within the strata, the sample sizes for those strata are adjusted upward as necessary. In this subsection we discuss some of the issues pertinent to allocating the total sample to the various strata and present a method for choosing the total sample size and making the allocation.

The allocation task is to decide what fraction of the total sample should be assigned to each stratum. This fraction determines how large the simple random sample will be in each stratum. The factors considered most important in making the allocation follow:

1. The number of elements in each stratum
2. The variance of the elements within each stratum
3. The cost of selecting elements within each stratum

Generally, larger samples should be assigned to the larger strata and to the strata with larger variances. Conversely, to get the most information for a given cost, smaller samples should be allocated to the strata where the cost per unit of sampling is greatest.

The individual stratum variances often differ greatly. For example, suppose that in a particular study we are interested in determining the mean number of employees per building. Because variability is greater in a stratum with larger buildings than in one with smaller buildings, a proportionately larger sample should be taken in such a stratum. The cost of selection can be an important consideration when significant interviewer travel between sampled units is necessary in some of the strata but not in others. This issue frequently arises when some of the strata involve rural areas and others involve cities.

In many surveys the cost per unit of sampling is approximately the same for each stratum (e.g. mail and telephone surveys). In such cases, the cost of sampling can be ignored in making the allocation. We present here the appropriate formulae for choosing the sample size and making the allocation in such cases. More advanced texts on sampling provide formulae for the case when sampling costs vary significantly across strata. The formulae we present in this section will minimize the total sampling cost for a given level of precision. This method, known as *Neyman allocation,* allocates the total sample n to the various strata as follows.

$$n_h = n\left(\frac{N_h s_h}{\sum_{h=1}^{H} N_h s_h}\right) \tag{22.24}$$

Equation (22.24) shows that the number of units allocated to a stratum increases with the stratum size and standard deviation. Note that to make this allocation, we need to first determine the total sample size n. Given a specified level of precision B,

we can use the following formulae to choose the total sample size when estimating the population mean and the population total.

Sample size when estimating the population mean

$$n = \frac{\left(\sum_{h=1}^{H} N_h s_h\right)^2}{N^2\left(\frac{B^2}{4}\right) + \sum_{h=1}^{H} N_h s_h^2} \tag{22.25}$$

Sample size when estimating the population total

$$n = \frac{\left(\sum_{h=1}^{H} N_h s_h\right)^2}{\frac{B^2}{4} + \sum_{h=1}^{H} N_h s_h^2} \tag{22.26}$$

As an example, suppose a Ford dealer wants to survey the customers who purchased a Mondeo, Focus or Fiesta to obtain information the dealer feels will be helpful in determining future advertizing. In particular, suppose the dealer wants to estimate the mean monthly income for these customers, with a bound on the sampling error of €100. The dealer's 600 Mondeo, Focus and Fiesta customers are divided into three strata: 100 Mondeo owners, 200 Focus owners, and 300 Fiesta owners. A pilot survey was used to estimate the standard deviation in each stratum; the results are s_1 = €1300, s_2 = €900, and s_3 = €500 for the Mondeo, Focus and Fiesta owners, respectively.

The first step in choosing a sample size for this survey is to use equation (22.25) to determine the total sample size necessary to provide a bound of B = €100 on the estimate of the population mean. First, we compute

$$\sum_{h=1}^{3} N_h s_h = 100(1300) + 200(900) + 300(500) = 460\,000$$

Next, we compute

$$\sum_{h=1}^{3} N_h s_h^2 = 100(1300)^2 + 200(900)^2 + 300(500)^2 = 406\,000\,000$$

Substituting these values into equation (22.25), we can determine the total sample size needed to provide a bound on the sampling error of B = €100.

$$n = \frac{(460\,000)^2}{\frac{(600)^2(100)^2}{4} + 406\,000\,000} = 162$$

A total sample size of 162 will provide the precision desired. To allocate the total sample to the three strata, we use equation (22.24).

$$n_1 = 162\left(\frac{100(1300)}{460\,000}\right) = 46$$

$$n_2 = 162\left(\frac{200(900)}{460\,000}\right) = 63$$

$$n_1 = 162\left(\frac{300(500)}{460\,000}\right) = 53$$

We would therefore recommend sampling 46 Mondeo owners, 63 Focus owners, and 53 Fiesta owners for a total sample size of 162 customers.

To determine the sample size when estimating a population proportion, we simply substitute $\sqrt{p_h(1-p_h)}$ for s_h in equation (22.25); the result is

$$n = \frac{\left(\sum_{h=1}^{H} N_h\sqrt{p_h(1-p_h)}\right)^2}{N^2\left(\frac{B^2}{4}\right) + \sum_{h=1}^{H} N_h p_h(1-p_h)} \tag{22.27}$$

Once the total sample size for the population proportion estimate has been determined, allocation to the various strata is again made by using equation (22.24) with $\sqrt{p_h(1-p_h)}$ substituted for s_h.

Another type of allocation that is sometimes used with stratified simple random sampling is called *proportional allocation*. In this approach, the sample size allocated to each stratum is given by the following formula.

$$n_h = n\left(\frac{N_h}{N}\right) \tag{22.28}$$

Proportional allocation is appropriate when the stratum variances are approximately equal and the cost per unit of sampling is about the same across strata. In the case where the stratum variances are equal, proportional allocation and the Neyman procedure result in the same allocation.

An advantage of stratified random sampling is that estimates of population parameters for each stratum are automatically available as a by-product of the sampling procedure. For example, besides obtaining an estimate of the average starting salary for all graduates in the South Downs University sampling problem, we obtained an estimate of the average starting salary for each degree programme. Because each of the starting salary estimates was based on a simple random sample from each stratum, the procedure for constructing an approximate confidence interval estimate when a simple random sample is selected (see equation (22.4)) can be used to compute an approximate 95 per cent confidence interval estimate for the mean in each stratum. In a similar manner, interval estimates for the population total and the population proportion for each stratum can be constructed by using equations (22.8) and (22.10) respectively.

EXERCISES

Methods

8 A stratified random sample was taken with the following results.

Stratum (h)	$\bar{x}_h$	s_h	p_h	N_h	n_h
1	138	30	0.50	200	20
2	103	25	0.78	250	30
3	210	50	0.21	100	25

a. Construct an estimate of the population mean for each stratum.
b. Construct an approximate 95 per cent confidence interval for the population mean in each stratum.
c. Construct an approximate 95 per cent confidence interval for the overall population mean.

9 Reconsider the sample results in exercise 8.

a. Construct an estimate of the population total for each stratum.
b. Construct a point estimate of the total for all 550 elements in the population.
c. Construct an approximate 95 per cent confidence interval for the population total.

10 Reconsider the sample results in exercise 8.

a. Construct an approximate 95 per cent confidence interval for the proportion in each stratum.
b. Construct a point estimate of the population proportion for the 550 elements in the population.
c. Estimate the standard error of the population proportion.
d. Construct an approximate 95 per cent confidence interval for the population proportion.

11 A population was divided into three strata with $N_1 = 300$, $N_2 = 600$, and $N_3 = 500$. From a past survey, the following estimates for the standard deviations in the three strata are available: $s_1 = 150$, $s_2 = 75$, $s_3 = 100$.

a. Suppose an estimate of the population mean with a bound on the error of estimate of $B = 20$ is required. How large must the sample be? How many elements should be allocated to each stratum?
b. Suppose a bound of $B = 10$ is desired. How large must the sample be? How many elements should be allocated to each stratum?
c. Suppose an estimate of the population total with a bound of $B = 15\,000$ is requested. How large must the sample be? How many elements should be allocated to each stratum?

Applications

12 An accounting firm works for a number of clients in the banking, insurance, and share-dealing sectors: $N_1 = 50$ banks, $N_2 = 38$ insurance companies, and $N_3 = 35$ share-dealing firms. A marketing research firm has been hired to survey

the accounting firm's clients in these three sectors. The survey will ask a variety of questions about both the clients' businesses and their satisfaction with services provided by the accounting firm. Suppose an approximate 95 per cent confidence interval is requested for the mean number of employees for the 123 clients, with a bound on the error of estimation of $B = 30$.

a. Suppose a pilot study finds $s_1 = 80$, $s_2 = 150$, and $s_3 = 45$. Choose a total sample size, and explain how the sample size should be allocated to the three strata.
b. Suppose the pilot test is called into question and a decision is made to assume the stratum standard deviations are all equal to 100 in choosing the sample size. Choose a total sample size, and determine how many elements should be sampled in each stratum.

13 A stratified simple random sample is to be taken of a bank's customers to learn about a variety of attitudinal and demographic issues. The stratification is to be based on savings account balances as of June 30, 2001. A frequency distribution follows showing the number of accounts in each stratum, together with the standard deviation of account balances by stratum.

Stratum (£)	Accounts	Standard Deviation of Account Balances (£)
0.00–1000.00	3000	80
1000.01–2000.00	600	150
2000.01–5000.00	250	220
5000.00–10 000.00	100	700
Over 10 000	50	3000

a. Assuming the cost per unit sampled is approximately equal across strata, determine the total number of persons to include in the sample. Assume we want a bound on the error of estimate of the population mean for savings account balances of $B = £20$.
b. Use the Neyman allocation procedure to determine the number to be sampled for each stratum.

22.6 Cluster sampling

Cluster sampling requires that the population be divided into N groups of elements called clusters, such that each element in the population belongs to one and only one cluster. For example, suppose we want to survey registered voters in the UK. One approach would be to develop a frame consisting of all registered voters in the UK and then select a simple random sample of voters from this frame. Alternatively, in cluster sampling, we might choose to define the frame as the list of the $N = 646$ parliamentary constituencies in the UK. In this approach, each constituency or cluster would consist of a group of registered voters, and each registered voter in the UK would belong to one and only one cluster.

Suppose we select a simple random sample of $n = 50$ of the 646 parliamentary constituencies. At this point, we could collect data for *all* registered voters in each of the 50 sampled clusters, an approach referred to as *single-stage cluster sampling,* or we could select a simple random sample of registered voters from each of the 50 sampled clusters, an approach referred to as *two-stage cluster sampling.* In either case, formulae are available for using the sample results to construct point and interval estimates of population parameters such as the population mean, total or proportion. In this chapter, however, we consider only single-stage cluster sampling; more advanced texts on sampling present results for two-stage cluster sampling.

Both stratified and cluster sampling share the characteristic that they divide the population into groups of elements. The reasons for choosing cluster sampling, however, differ from the reasons for choosing stratified sampling. Cluster sampling tends to provide better results when the elements within the clusters are heterogeneous (not alike), whereas stratified sampling works well when the elements within each stratum are homogeneous (alike). In the ideal case, each cluster would be a small-scale version of the entire population. In this case, sampling a small number of clusters would provide good information about the characteristics of the entire population.

One of the primary applications of cluster sampling involves area sampling, where the clusters are counties, parliamentary constituencies, cities, or other well-defined geographic sections of the population. Because data are collected from only a sample of the total geographic areas or clusters available, and the elements within the clusters are typically close to one another, significant savings in time and cost can be realized when a data collector or interviewer is sent to a sampled unit. As a result, even if a larger total sample size is required, cluster sampling may be less costly than either simple random sampling or stratified random sampling. In addition, cluster sampling can minimize the time and the cost associated with developing the frame or list of elements to be sampled because cluster sampling does not require that a list of every element in the population be developed. One needs only a list of the elements in the clusters sampled.

To illustrate cluster sampling, let us consider a survey conducted by a professional accounting association (PAA) of the 12 000 practising qualified accountants in a particular country. As part of the survey, the PAA collected information on income, gender, and factors related to the accountant's lifestyle. Because personal interviews were needed to obtain all the desired information, the PAA used a cluster sample to minimize the total travel and interviewing cost. The frame consisted of all firms that were registered to practise accounting in the country. Suppose there are $N = 1000$ clusters, or firms, registered to practise accounting in the country, and that a simple random sample of $n = 10$ firms is to be selected.

In presenting the formulae for cluster sampling that are needed to construct approximate 95 per cent confidence interval estimates of the population mean, total and proportion, we will use the following notation.

N = number of clusters in the population

n = number of clusters selected in the sample

M_i = number of elements in cluster i

M = number of elements in the population; $M = M_1 + M_2 + \cdots + M_N$

$\bar{M} = M/N$ = average number of elements in a cluster

x_i = total of all observations in cluster i

a_i = number of observations in cluster i with a certain characteristic

For the PAA's sample survey we have the following information.

$$N = 1000$$

$$n = 10$$

$$M = 12\,000$$

$$\bar{M} = 12\,000/1000 = 12$$

Table 22.4 shows the values of M_i and x_i for each of the sampled clusters as well as the number of qualified accountants who were female in the sampled firms (a_i).

Table 22.4 Results of the PAA'S sample survey

Firm (i)	Number of Qualified Accountants (M_i)	Total Salary (€000s) for Firm i (x_i)	Number of Female Qualified Accountants (a_i)
1	8	384	2
2	25	1350	8
3	4	148	0
4	17	857	6
5	7	296	1
6	3	131	2
7	15	761	2
8	4	176	0
9	12	577	5
10	33	1880	9
Totals	128	6560	35

Population mean

The point estimator of the population mean obtained from cluster sampling is given by the following formula.

Point estimator of the population mean

$$\bar{x}_c = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} M_i} \qquad (22.29)$$

An estimate of the standard error of this point estimator is

$$s_{\bar{x}_c} = \sqrt{\left(\frac{N-n}{Nn\bar{M}^2}\right)\frac{\sum_{i=1}^{n}(x_i - \bar{x}_c M_i)^2}{n-1}} \qquad (22.30)$$

Hence, the following expression gives an approximate 95 per cent confidence interval estimate of the population mean.

Approximate 95 per cent confidence interval estimate of the population mean

$$\bar{x}_c \pm 2s_{\bar{x}_c} \qquad (22.31)$$

Using the data in Table 22.4, we obtain an estimate of the mean salary for practising qualified accountants.

$$\bar{x}_c = \frac{6560}{128} = 51.25$$

The salary data in Table 22.4, listed in thousands of euros, indicate that an estimate of the mean salary for practising qualified accountants in the country is €51 250.

In Table 22.5, we show a portion of the calculations needed to estimate the standard error; note that

$$\sum_{i=1}^{n}(x_i - \bar{x}_c M_i)^2 = 51\,281.378$$

Hence,

$$s_{\bar{x}_c} = \sqrt{\left(\frac{1000 - 10}{(1000)(10)(12)^2}\right)\frac{51\,281.378}{10 - 1}} = 1.979$$

The standard error is €1979. Using expression (22.31), we find that an approximate 95 per cent confidence interval estimate for the mean annual salary is 51 250 ± 2(1979) = 51 250 ± 3958 or €47 292 to €55 208.

Table 22.5 Partial calculations for the estimate of the standard error of the mean for the PAA's sample survey where $\bar{x}_c = 51.250$

Firm (i)	M_i	x_i	$(x_i - 51.250M_i)^2$		
1	8	384	$[384 - 51.250(8)]^2$	=	676.000
2	25	1350	$[1350 - 51.250(25)]^2$	=	4 726.563
3	4	148	$[148 - 51.250(4)]^2$	=	3 249.000
4	17	857	$[857 - 51.250(17)]^2$	=	203.063
5	7	296	$[296 - 51.250(7)]^2$	=	3 937.563
6	3	131	$[131 - 51.250(3)]^2$	=	517.563
7	15	761	$[761 - 51.250(15)]^2$	=	60.063
8	4	176	$[176 - 51.250(4)]^2$	=	841.000
9	12	577	$[577 - 51.250(12)]^2$	=	1 444.000
10	33	1880	$[1880 - 51.250(33)]^2$	=	35 626.563
Totals	128	6560			51 281.378

$$\sum_{i=1}^{n}(x_i - \bar{x}_c M_i)^2$$

Population total

The point estimator of the population total X is obtained by multiplying together M and $\bar{x}_c$.

Point estimator of the population total

$$\hat{X} = M\bar{x}_c \tag{22.32}$$

An estimate of the standard error of this point estimator is

$$s_{\hat{X}} = Ms_{\bar{x}_c} \tag{22.33}$$

Hence, an approximate 95 per cent confidence interval estimate for the population total is given by the following expression.

Approximate 95 per cent confidence interval estimate of the population total

$$M\bar{x}_c \pm 2s_{\hat{X}} \tag{22.34}$$

For the PAA's sample survey,

$$\begin{aligned} \hat{X} &= M\bar{x}_c = 12\,000(51\,250) = €615\,000\,000 \\ s_{\hat{X}} &= Ms_{\bar{x}_c} = 12\,000(1979) = €23\,748\,000 \end{aligned}$$

Hence, using expression (22.34), we find that an approximate 95 per cent confidence interval is €615 000 000 ± 2(€23 748 000) = €615 000 000 ± €47 496 000, or €567 504 000 to €662 496 000.

Population proportion

The point estimator of the population proportion obtained from cluster sampling follows.

Point estimator of the population proportion

$$p_c = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} M_i} \tag{22.35}$$

where
a_i = number of elements in cluster i with the characteristic of interest

An estimate of the standard error of this point estimator is

$$s_{p_c} = \sqrt{\left(\frac{N-n}{Nn\bar{M}^2}\right)\frac{\sum_{i=1}^{n}(a_i - p_c M_i)^2}{n-1}} \tag{22.36}$$

Hence, an approximate 95 per cent confidence interval estimate for the population proportion is given by the following expression.

Approximate 95 per cent confidence interval estimate of the population proportion

$$p_c \pm 2s_{p_c} \tag{22.37}$$

For the PAA's sample survey, we can use equation (22.35) and the data in Table 22.4 to construct an estimate of the proportion of practising qualified accountants who are women.

$$p_c = \frac{2 + 8 + \cdots + 9}{8 + 25 + \cdots + 33} = \frac{35}{128} = 0.2734$$

In Table 22.6 we show a portion of the calculations needed to estimate the standard error; note that

$$\sum_{i=1}^{n}(a_i - p_c M_i)^2 = 15.2098$$

Table 22.6 Partial calculations for the estimation of the standard error for the PAA's sample survey where $p_c = 0.2734$

Firm (i)	M_i	a_i	$(a_i - 0.2734\,M_i)^2$
1	8	2	$[2 - 0.2734(8)]^2 = 0.0350$
2	25	8	$[8 - 0.2734(25)]^2 = 1.3572$
3	4	0	$[0 - 0.2734(4)]^2 = 1.1960$
4	17	6	$[6 - 0.2734(17)]^2 = 1.8284$
5	7	1	$[1 - 0.2734(7)]^2 = 0.8350$
6	3	2	$[2 - 0.2734(3)]^2 = 1.3919$
7	15	2	$[2 - 0.2734(15)]^2 = 4.4142$
8	4	0	$[0 - 0.2734(4)]^2 = 1.1960$
9	12	5	$[5 - 0.2734(12)]^2 = 2.9556$
10	33	9	$[9 - 0.2734(33)]^2 = 0.0005$
	Totals 128	35	15.2098

$$\sum_{i=1}^{n}(a_i - p_c M_i)^2$$

Hence,

$$s_{p_c} = \sqrt{\left(\frac{1000 - 10}{(1000)(10)(12)^2}\right)\frac{15.2098}{10 - 1}} = 0.0341$$

Using expression (22.37), we find that an approximate 95 per cent confidence interval for the proportion of practising qualified accountants who are women is $0.2734 \pm 2(0.0341) = 0.2734 \pm 0.0682$ or 0.2052 to 0.3416.

Determining the sample size

Once the clusters are formed, the primary issue in choosing a sample size is selecting the number of clusters n. The procedure for cluster sampling is similar to that for other methods of sampling. An acceptable level of precision is specified by choosing a value for B, the bound on the sampling error. Then a formula is developed for finding the value of n that will provide the desired precision.

The average cluster size and the variance between clusters are key factors in deciding how many clusters to include in the sample. If the clusters are similar, the variance between them will be small and the number of clusters sampled can be smaller. Also, if the average number of elements per cluster is larger, the number of clusters sampled can be smaller. The formulae for making an exact determination of sample size are included in more advanced texts on sampling.

EXERCISES

Methods

14 A sample of four clusters is to be taken from a population with $N = 25$ clusters and $M = 300$ elements. The values of M_i, x_i and a_i for each cluster in the sample follow.

Cluster (i)	M_i	x_i	a_i
1	7	95	1
2	18	325	6
3	15	190	6
4	10	140	2
Totals	50	750	15

a. Construct point estimates of the population mean, total, and proportion.
b. Estimate the standard errors for the estimates in part (a).
c. Construct an approximate 95 per cent confidence interval for the population mean.
d. Construct an approximate 95 per cent confidence interval for the population total.
e. Construct an approximate 95 per cent confidence interval for the population proportion.

15 A sample of six clusters is to be taken from a population with $N = 30$ clusters and $M = 600$ elements. The following table shows values of M_i, x_i and a_i for each cluster in the sample.

Cluster (i)	M_i	x_i	a_i
1	35	3 500	3
2	15	965	0
3	12	960	1
4	23	2 070	4
5	20	1 100	3
6	25	1 805	2
Totals	130	10 400	13

a. Construct point estimates of the population mean, total, and proportion.
b. Construct an approximate 95 per cent confidence interval for the population mean.
c. Construct an approximate 95 per cent confidence interval for the population total.
d. Construct an approximate 95 per cent confidence interval for the population proportion.

Applications

16 A public utility is conducting a survey of mechanical engineers to learn more about the factors influencing the choice of heating, ventilation, and air conditioning (HVAC) equipment for new commercial buildings. A total of 120 firms in the utility's service area are engaged in designing HVAC systems. The sampling plan is to use cluster sampling with each firm representing a cluster. For each firm in the sample, all of the mechanical engineers will be interviewed. Approximately 500 mechanical engineers are believed to be employed by the 120 firms. A sample of ten firms was taken. Among other things, the age of each respondent was recorded as well as whether the respondent had attended the local university.

Cluster (i)	M_i	*Total of Respondents' Ages*	*Number Attending Local University*
1	12	520	8
2	1	33	0
3	2	70	1
4	1	29	1
5	6	270	3
6	3	129	2
7	2	102	0
8	1	48	1
9	9	337	7
10	13	462	12
Totals	50	2000	35

a. Estimate the mean age of mechanical engineers engaged in this type of work.
b. Estimate the proportion of mechanical engineers in the utility's service area who attended the local university.
c. Construct an approximate 95 per cent confidence interval for the mean age of mechanical engineers designing HVAC systems for commercial buildings.
d. Construct an approximate 95 per cent confidence interval for the proportion of mechanical engineers in the utility's service area who attended the local university.

17 A public agency is interested in learning more about the people living in nursing homes in a particular city. A total of 100 nursing homes are caring for 4800 people in the city and a cluster sample of six homes has been taken. Each person in the six homes has been interviewed. A portion of the sample results follows.

Home	Residents	Average Age of Residents	Number of Disabled Residents
1	14	61	12
2	7	74	2
3	96	78	30
4	23	69	8
5	71	73	10
6	29	84	22

a. Calculate an estimate of the mean age of nursing home residents in this city.
b. Construct an approximate 95 per cent confidence interval for the proportion of disabled persons in the city's nursing homes.
c. Estimate the total number of disabled persons residing in nursing homes in this city.

22.7 Systematic sampling

Systematic sampling is often used as an alternative to simple random sampling. In some sampling situations, especially those with large populations, it can be time-consuming to select a simple random sample by first finding a random number and then counting or searching through the frame until the corresponding element is found. Systematic sampling offers an alternative to simple random sampling in such cases. For example, if a sample size of 50 from a population containing 5000 elements is desired, we might sample one element for every $5000/50 = 100$ elements in the population. A systematic sample for this case would involve randomly selecting one of the first 100 elements from the frame. The remaining sample elements are identified by starting with the first sampled element, and then selecting every 100th element that follows in the frame. In effect, the sample of 50 is identified by moving systematically through the population and identifying every 100th element after the first randomly selected element. The sample of 50 will often be easier to select in this manner than it would be if simple random sampling were used. A systematic sample generated with a random starting position has properties similar to a simple random sample providing the frame can be considered to be a random ordering of the elements in the population.

Summary

We provided a brief introduction to the field of survey sampling. The purpose of survey sampling is to collect data for the purpose of making estimates of population parameters such as the population mean, total or proportion. Survey sampling as a method of data collection can be contrasted with conducting experiments to generate data. When survey sampling is used, the design of the sampling plan is of critical importance in determining which existing data will be collected. When experiments are employed, experimental design issues are of critical importance in determining which data will be generated, or created.

Two types of errors can occur with sample surveys: sampling error and non-sampling error. Sampling error is the error that occurs because a sample, and not the entire population, is used to estimate a population parameter. Non-sampling error refers to all the other types of errors that can occur, such as measurement, interviewer, non-response, and processing error. Non-sampling errors are controlled by proper questionnaire design, thorough training of interviewers, careful verification of data, and so on. Sampling errors are minimized by a proper choice of sample design and by selecting an appropriate sample size.

We discussed four commonly used sample designs in this chapter: simple random sampling, stratified random sampling, cluster sampling, and systematic sampling. The objective of sample design is to get the most precise estimates for the least cost. When the population can be divided into strata so that the elements within each stratum are relatively homogeneous, stratified random sampling will provide more precision (smaller approximate confidence intervals) than simple random sampling. When the elements can be grouped in clusters so that all the elements in a cluster are close together geographically, cluster sampling often reduces interviewer cost; in these situations, cluster sampling will often provide the most precision for a given outlay. Systematic random sampling was presented as an alternative to simple random sampling.

Key terms

Bound on the sampling error
Cluster sampling
Convenience sampling
Element
Frame
Judgement sampling
Non-probabilistic sampling
Non-sampling error
Population
Probabilistic sampling
Sample
Sampled population
Sampling error
Sampling unit
Simple random sample
Stratified random sampling
Systematic sampling
Target population

Key formulae

Simple random sampling

Interval estimate of the population mean

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} \qquad (22.1)$$

Estimate of the standard error of the mean

$$s_{\bar{x}} = \sqrt{\frac{N-n}{N}}\left(\frac{s}{\sqrt{n}}\right) \tag{22.2}$$

Interval estimate of the population mean

$$\bar{x} \pm t_{\alpha/2} s_{\bar{x}} \tag{22.3}$$

Approximate 95 per cent confidence interval estimate of the population mean

$$\bar{x} \pm 2s_{\bar{x}} \tag{22.4}$$

Point estimator of a population total

$$\hat{X} = N\bar{x} \tag{22.5}$$

Approximate 95 per cent confidence interval estimate of the population total

$$N\bar{x} \pm 2s_{\hat{X}} \tag{22.8}$$

Approximate 95 per cent confidence interval estimate of the population proportion

$$p \pm 2s_p \tag{22.10}$$

Stratified random sampling

Point estimator of the population mean

$$\bar{x}_{st} = \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)\bar{x}_h \tag{22.15}$$

Approximate 95 per cent confidence interval estimate of the population mean

$$\bar{x}_{st} \pm 2s_{\bar{x}_{st}} \tag{22.17}$$

Point estimator of the population total

$$\hat{X} = N\bar{x}_{st} \tag{22.18}$$

Approximate 95 per cent confidence interval estimate of the population total

$$N\bar{x}_{st} \pm 2s_{\hat{X}} \tag{22.20}$$

Point estimator of the population proportion

$$p_{st} = \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)p_h \tag{22.21}$$

Approximate 95 per cent confidence interval estimate of the population proportion

$$p_{st} \pm 2s_{p_{st}} \tag{22.23}$$

Allocating the total sample *n* to the strata: Neyman allocation

$$n_h = n\left(\frac{N_h s_h}{\sum_{h=1}^{H} N_h s_h}\right) \tag{22.24}$$

Sample size when estimating the population mean

$$n = \frac{\left(\sum_{h=1}^{H} N_h s_h\right)^2}{N^2\left(\frac{B^2}{4}\right) + \sum_{h=1}^{H} N_h s_h^2} \tag{22.25}$$

Sample size when estimating the population total

$$n = \frac{\left(\sum_{h=1}^{H} N_h s_h\right)^2}{\frac{B^2}{4} + \sum_{h=1}^{H} N_h s_h^2} \tag{22.26}$$

Cluster sampling

Point estimator of the population mean

$$\bar{x}_c = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} M_i} \tag{22.29}$$

Approximate 95 per cent confidence interval estimate of the population mean

$$\bar{x}_c \pm 2s_{\bar{x}_c} \tag{22.31}$$

Point estimator of the population total

$$\hat{X} = M\bar{x}_c \tag{22.32}$$

Approximate 95 per cent confidence interval estimate of the population total

$$M\bar{x}_c \pm 2s_{\hat{X}} \tag{22.34}$$

Point estimator of the population proportion

$$p_c = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} M_i} \qquad \textbf{(22.35)}$$

Approximate 95 per cent confidence interval estimate of the population proportion

$$p_c \pm 2s_{p_c} \qquad \textbf{(22.37)}$$