7 Sampling and Sampling Distributions

CHAPTER CONTENTS

LEARNING OBJECTIVES

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

1 Explain the terms simple random sample, sampling with replacement and sampling without replacement.

2 Select a simple random sample from a finite population using random number tables.

3 Explain the terms parameter, statistic, point estimator and unbiasedness.

4 Identify relevant point estimators for a population mean, population standard deviation and population proportion.

5 Explain the term sampling distribution.

6 Describe the form and characteristics of the sampling distribution:

6.1 of the sample mean, when the sample size is large or when the population is normal.

6.2 of the sample proportion when the sample size is large.

In Chapter 1, we defined the terms *element*, *population* and *sample*:

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

STATISTICS IN PRACTICE
Copyright and Public Lending Right

How would you feel if the size of your income was determined each year by a random sampling procedure? This is the situation that often exists, for at least part of annual income, for musicians and other artists who receive copyright payments for the performance or broadcasting of their work. Even in this 21st–century world of large databases and sophisticated communication, it is not always possible, or it is too costly, to maintain 100 per cent checks on what is being broadcast over TV, radio and the Internet, so an alternative is to sample.

In a similar vein, many book authors receive payments through a Public Lending Right (PLR) scheme. This is particularly so for authors of fiction, or of popular non-fiction, whose books are available for loan in public libraries. A PLR scheme is intended to compensate authors for potential loss of income because their books are available in public libraries, and are therefore borrowed rather than bought by readers. The website www.plrinternational.com listed 30 countries in mid-2012 with established PLR schemes. All except Australia, Canada and New Zealand were in Europe.

The UK PLR scheme was set up in 1979. From the outset, it was decided that it would be too costly to try and collect data from all libraries in the UK. Data on lending are collected from a sample of libraries. Similar decisions have been made in many of the other countries that operate PLR schemes. The current UK sample is reckoned to cover about 15 per cent of all library authorities in the UK (there are over 200). The PLR scheme in Finland, as another example, is estimated to cover about 10 per cent of all library loans.

The examples from copyright and PLR are cases where the sampling schemes involved can influence the income of individuals – the copyright holders or authors. The website for the UK PLR scheme acknowledges, for example, that authors of books with a 'local interest' – local history, say – are likely to qualify for PLR payments only if the library sample for the year contains library authorities in the relevant geographical area.

Companies and governments often make important decisions based on sample data. This chapter examines the basis and practicalities of scientific sampling.

The reason we sample is to collect data to make an inference and answer a research question about a population. Numerical characteristics of a population (e.g. population mean, population standard deviation) are called **parameters**. Numerical characteristics of a sample (e.g. sample mean, sample standard deviation) are called **sample statistics**. Primary purposes of statistical inference are to make estimates and test hypotheses about population parameters using sample statistics.

Here are two situations in which samples provide estimates of population parameters:

1 A European car tyre manufacturer developed a new tyre designed to provide an increase in tyre lifetime. To estimate the mean lifetime (in kilometres or miles) of the new tyre, the manufacturer

selected a sample of 120 new tyres for testing. The test results provided a sample mean of 56 000 kilometres (35 000 miles). Therefore, an estimate of the mean tyre lifetime for the population of new tyres was 56 000 kilometres.

2 Members of an African government were interested in estimating the proportion of registered voters likely to support a proposal for constitutional reform to be put to the electorate in a national referendum. The time and cost associated with contacting every individual in the population of registered voters were prohibitive. A sample of 5000 registered voters was therefore selected, and 2810 of the 5000 voters indicated support for the proposal. An estimate of the proportion of the population of registered voters supporting the proposal was 2810/5000 = 0.562.

These two examples illustrate some of the reasons why samples are used. In the tyre lifetime example, collecting the data on tyre life involves wearing out each tyre tested. Clearly it is not feasible to test every tyre in the population. A sample is the only realistic way to obtain the tyre lifetime data. In the example involving the referendum, contacting every registered voter in the population is in principle possible, but the time and cost are prohibitive. Consequently, a sample of registered voters is preferred.

It is important to realize that sample results provide only *estimates* of the values of the population characteristics, because the sample contains only a portion of the population. A sample mean provides an estimate of a population mean, and a sample proportion provides an estimate of a population proportion. Some estimation error can be expected. This chapter provides the basis for determining how large the estimation error might be. With proper sampling methods, the sample results will provide 'good' estimates of the population parameters.

Let us define some of the terms used in sampling. The **sampled population** is the population from which the sample is drawn, and a **sampling frame** is a list of the elements from which the sample will be selected. In the second example above, the sampled population is all registered voters in the country, and the sampling frame is the list of all registered voters. Because the number of registered voters is finite, this is an illustration of sampling from a finite population. In Section 7.2, we consider how a simple random sample can be selected from a finite population.

The sampled population for the tyre lifetime example is more difficult to define. The sample of 120 tyres was obtained from a production process at a particular point in time. We can think of the sampled population as the conceptual population of all tyres that could be made by the production process under similar conditions to those prevailing at the time of sampling. In this context, the sampled population is considered infinite, making it impossible to construct a sampling frame. In Section 7.2, we consider how to select a random sample in such a situation.

We first show how simple random sampling can be used to select a sample from finite and from infinite populations. We then show how data from a simple random sample can be used to compute estimates of a population mean, a population standard deviation, and a population proportion. In addition, we introduce the important concept of a sampling distribution. Knowledge of the appropriate sampling distribution enables us to make statements about how close the sample estimates might be to the corresponding population parameters.

EAI

7.1 THE EAI SAMPLING PROBLEM

The head of personnel services for E-Applications & Informatics plc (EAI) has been given the task of constructing a profile of the company's 2500 managers. The characteristics to be identified include the mean annual salary and the proportion of managers who have completed the company's management training programme. The 2500 managers are the population for this study. We can find the annual salary and training programme status for each individual by referring to the firm's personnel records. The data file containing this information for all 2500 managers in the population is on the online platform, in the file 'EAI'.

Using the EAI data set and the formulae from Chapter 3, we calculate the population mean and the population standard deviation for the annual salary data.

$$\text{Population mean: } \mu = €51\,800$$

$$\text{Population standard deviation: } \sigma = €4000$$

The data set shows that 1500 of the 2500 managers completed the training programme. Let π denote the proportion of the population that completed the training programme: $\pi = 1500/2500 = 0.60$. The population mean annual salary (μ = €51 800), the population standard deviation of annual salary (σ = €4000), and the population proportion that completed the training programme ($\pi = 0.60$) are parameters of the population of EAI managers.

Now, suppose the necessary information on all the EAI managers was *not* readily available in the company's database. How can the head of personnel services obtain estimates of the population parameters by using a sample of managers, rather than all 2500 managers in the population? Suppose a sample of 30 managers will be used. Clearly, the time and the cost of constructing a profile would be substantially less for 30 managers than for the entire population. If the head of personnel could be assured that a sample of 30 managers would provide adequate information about the population of 2500 managers, working with a sample would be preferable to working with the entire population. Often the cost of collecting information from a sample is substantially less than from a population, especially when personal interviews must be conducted to collect the information.

First we consider how we can identify a sample of 30 managers.

7.2 SIMPLE RANDOM SAMPLING

Several methods can be used to select a sample from a population. One important method is **simple random sampling**. The definition of a simple random sample and the process of selecting such a sample depend on whether the population is *finite* or *infinite*. We first consider sampling from a finite population, because the EAI sampling problem involves a finite population of 2500 managers.

Sampling from a finite population

Simple random sample (finite population)

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

One procedure for selecting a simple random sample from a finite population is to choose the elements for the sample one at a time in such a way that, at each step, each of the elements remaining in the population has the same probability of being selected.

To select a simple random sample from the population of EAI managers, we first assign each manager a number. We can assign the managers the numbers 1 to 2500 in the order their names appear in the EAI personnel file. Next, we refer to the table of random numbers shown in Table 7.1. Using the first row of the table, each digit, 6, 3, 2, …, is a random digit with an equal chance of occurring. The random numbers in the table are shown in groups of five for readability. Because the largest number in the population list, 2500, has four digits, we shall select random numbers from the table in groups of four digits. We may start the selection of random numbers anywhere in the table and move systematically in a direction of our choice. We shall use the first row of Table 7.1 and move from left to right. The first seven four-digit random numbers are

6327 1599 8671 7445 1102 1514 1807

These four-digit numbers are equally likely, because the numbers in the table are random. We use them to give each manager in the population an equal chance of being included in the random sample.

The first number, 6327, is greater than 2500. We discard it because it does not correspond to one of the numbered managers in the population. The second number, 1599, is between 1 and 2500.

TABLE 7.1 Random numbers

63271	59986	71744	51102	15141	80714	58683	93108	13554	79945
88547	09896	95436	79115	08303	01041	20030	63754	08459	28364
55957	57243	83865	09911	19761	66535	40102	26646	60147	15702
46276	87453	44790	67122	45573	84358	21625	16999	13385	22782
55363	07449	34835	15290	76616	67191	12777	21861	68689	03263
69393	92785	49902	58447	42048	30378	87618	26933	40640	16281
13186	29431	88190	04588	38733	81290	89541	70290	40113	08243
17726	28652	56836	78351	47327	18518	92222	55201	27340	10493
36520	64465	05550	30157	82242	29520	69753	72602	23756	54935
81628	36100	39254	56835	37636	02421	98063	89641	64953	99337
84649	48968	75215	75498	49539	74240	03466	49292	36401	45525
63291	11618	12613	75055	43915	26488	41116	64531	56827	30825
70502	53225	03655	05915	37140	57051	48393	91322	25653	06543
06426	24771	59935	49801	11082	66762	94477	02494	88215	27191
20711	55609	29430	70165	45406	78484	31639	52009	18873	96927
41990	70538	77191	25860	55204	73417	83920	69468	74972	38712
72452	36618	76298	26678	89334	33938	95567	29380	75906	91807
37042	40318	57099	10528	09925	89773	41335	96244	29002	46453
53766	52875	15987	46962	67342	77592	57651	95508	80033	69828
90585	58955	53122	16025	84299	53310	67380	84249	25348	04332
32001	96293	37203	64516	51530	37069	40261	61374	05815	06714
62606	64324	46354	72157	67248	20135	49804	09226	64419	29457
10078	28073	85389	50324	14500	15562	64165	06125	71353	77669
91561	46145	24177	15294	10061	98124	75732	00815	83452	97355
13091	98112	53959	79607	52244	63303	10413	63839	74762	50289

So the first manager selected for the random sample is number 1599 on the list of EAI managers. Continuing this process, we ignore the numbers 8671 and 7445 (greater than 2500) before identifying managers numbered 1102, 1514 and 1807 to be included in the random sample. This process continues until the simple random sample of 30 EAI managers has been obtained.

It is possible that a random number already used may appear again in the table before the sample of 30 EAI managers has been fully selected. Because we do not want to select a manager more than once, any previously used random numbers are ignored. Selecting a sample in this manner is referred to as **sampling without replacement**. If we selected a sample such that previously used random numbers are acceptable, and specific managers could be included in the sample two or more times, we would be **sampling with replacement**. Sampling with replacement is a valid way of identifying a simple random sample, but sampling without replacement is used more often. When we refer to simple random sampling, we shall assume that the sampling is without replacement.

Computer-generated random numbers can also be used to implement the random sample selection process. EXCEL, MINITAB and IBM SPSS all provide functions for generating random numbers.

The number of different simple random samples of size n that can be selected from a finite population of size N is:

$$\frac{N!}{n!(N-n)!}$$

$N!$, $(N - n)!$ and $n!$ are the factorial computations discussed in Chapter 4. For the EAI problem with $N = 2500$ and $n = 30$, this expression can be used to show that approximately 2.75×10^{69} different simple random samples of 30 EAI managers can be selected.

Sampling from an infinite population

In some situations, the population is either infinite, or so large that for practical purposes it must be treated as infinite. For example, suppose that a fast-food restaurant would like to obtain a profile of its customers by selecting a simple random sample of customers and asking each customer to complete a short questionnaire. The ongoing process of customer visits to the restaurant can be viewed as coming from an infinite population. In practice, a population is usually considered infinite if it involves an ongoing process that makes listing or counting every element in the population impossible. The definition of a simple random sample from an infinite population follows.

Simple random sample (infinite population)

A simple random sample from an infinite population is a sample selected such that the following conditions are satisfied:

1. Each element selected comes from the population.
2. Each element is selected independently.

For the example of a simple random sample of customers at a fast-food restaurant, any customer who comes into the restaurant will satisfy the first requirement. The second requirement will be satisfied if a sample selection procedure is devised to select the items independently and thereby avoid any selection bias that gives higher selection probabilities to certain types of customers. Selection bias would occur if, for instance, five consecutive customers selected were all friends who arrived together. We might expect these customers to exhibit similar profiles. Selection bias can be avoided by ensuring that the selection of a particular customer does not influence the selection of any other customer. In other words, the customers must be selected independently.

Infinite populations are often associated with an ongoing process that operates continuously over time. For example, parts being manufactured on a production line, transactions occurring at a bank, telephone calls arriving at a technical support centre and customers entering stores may all be viewed as coming from an infinite population. In such cases, an effective sampling procedure will ensure that no selection bias occurs and that the sample elements are selected independently.

EXERCISES

Methods

1. Consider a finite population with five elements labelled A, B, C, D and E. Ten possible simple random samples of size 2 can be selected.
   a. List the ten samples beginning with AB, AC and so on.
   b. Using simple random sampling, what is the probability that each sample of size 2 is selected?
   c. Assume random number 1 corresponds to A, random number 2 corresponds to B, and so on. List the simple random sample of size 2 that will be selected by using the random digits 8 0 5 7 5 3 2.

COMPLETE SOLUTIONS

2. Assume a finite population has 350 elements. Using the last three digits of each of the following five-digit random numbers (601, 022, 448, ...), determine the first four elements that will be selected for the simple random sample.

98601 73022 83448 02147 34229 27553 84147 93289 14209

Applications

3. The EURO STOXX 50 share index is calculated using data for 50 blue-chip companies from 12 Eurozone countries. Assume you want to select a simple random sample of five companies from the EURO STOXX 50 list. Use the last three digits in column 9 of Table 7.1, beginning with 554. Read down the column and identify the numbers of the five companies that would be selected.

COMPLETE SOLUTIONS

4. A student union is interested in estimating the proportion of students who favour a mandatory 'pass–fail' grading policy for optional courses. A list of names and addresses of the 645 students enrolled during the current semester is available from the registrar's office. Using three-digit random numbers in row 10 of Table 7.1 and moving across the row from left to right, identify the first ten students who would be selected using simple random sampling. The three-digit random numbers begin with 816, 283 and 610.

5. Assume that we want to identify a simple random sample of 12 of the 372 doctors practising in a particular city. The doctors' names are available from the local health authority. Use the eighth column of five-digit random numbers in Table 7.1 to identify the 12 doctors for the sample. Ignore the first two random digits in each five-digit grouping of the random numbers. This process begins with random number 108 and proceeds down the column of random numbers.

6. Indicate whether the following populations should be considered finite or infinite.
   - **a.** All registered voters in Ireland.
   - **b.** All television sets that could be produced by the Johannesburg factory of the TV-M Company.
   - **c.** All orders that could be processed by a mail-order firm.
   - **d.** All emergency telephone calls that could come into a local police station.
   - **e.** All components that Fibercon plc produced on the second shift on 17 February 2013.

7.3 POINT ESTIMATION

We return to the EAI problem. A simple random sample of 30 managers and the corresponding data on annual salary and management training programme participation are shown in Table 7.2. The notation x_1, x_2 and so on is used to denote the annual salary of the first manager in the sample, the annual salary of the second manager in the sample and so on. Completion of the management training programme is indicated by Yes in the relevant column.

To estimate the value of a population parameter, we compute a corresponding characteristic of the sample, referred to as a sample statistic. For example, to estimate the population mean μ and the population standard deviation σ for the annual salary of EAI managers, we use the data in Table 7.2 to calculate the corresponding sample statistics: the sample mean and the sample standard deviation. Using the formulae from Chapter 3, the sample mean is:

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{1\,554\,420}{30} = 51\,814 \quad (€)$$

TABLE 7.2 Annual salary and training programme status for a simple random sample of 30 EAI managers

Annual salary (€)	Management training programme	Annual salary (€)	Management training programme
x_1 = 49 094.30	Yes	x_{16} = 51 766.00	Yes
x_2 = 53 263.90	Yes	x_{17} = 52 541.30	No
x_3 = 49 643.50	Yes	x_{18} = 44 980.00	Yes
x_4 = 49 894.90	Yes	x_{19} = 51 932.60	Yes
x_5 = 47 621.60	No	x_{20} = 52 973.00	Yes
x_6 = 55 924.00	Yes	x_{21} = 45 120.90	Yes
x_7 = 49 092.30	Yes	x_{22} = 51 753.00	Yes
x_8 = 51 404.40	Yes	x_{23} = 54 391.80	No
x_9 = 50 957.70	Yes	x_{24} = 50 164.20	No
x_{10} = 55 109.70	Yes	x_{25} = 52 973.60	No
x_{11} = 45 922.60	Yes	x_{26} = 50 241.30	No
x_{12} = 57 268.40	No	x_{27} = 52 793.90	No
x_{13} = 55 688.80	Yes	x_{28} = 50 979.40	Yes
x_{14} = 51 564.70	No	x_{29} = 55 860.90	Yes
x_{15} = 56 188.20	No	x_{30} = 57 309.10	No

and the sample standard deviation is:

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{325\,009\,260}{29}} = 3348 \quad (€)$$

To estimate π, the proportion of managers in the population who completed the management training programme, we use the corresponding sample proportion. Let m denote the number of managers in the sample who completed the management training programme. The data in Table 7.2 show that $m = 19$. So, with a sample size of $n = 30$, the sample proportion is:

$$p = \frac{m}{n} = \frac{19}{30} = 0.63$$

These computations are an example of the statistical procedure called *point estimation*. We refer to the sample mean as the **point estimator** of the population mean μ, the sample standard deviation as the point estimator of the population standard deviation σ, and the sample proportion as the point estimator of the population proportion π. The numerical value obtained for the sample mean, sample standard deviation or sample proportion is called a **point estimate**. For the simple random sample of 30 EAI managers shown in Table 7.2, €51 814 is the point estimate of μ, €3348 is the point estimate of σ and 0.63 is the point estimate of π.

TABLE 7.3 Summary of point estimates obtained from a simple random sample of 30 EAI managers

Population parameter	Parameter value	Point estimator	Point estimate
Population mean annual salary	μ = €51 800	Sample mean annual salary	$\bar{x}$ = €51 814
Population standard deviation for annual salary	σ = €4 000	Sample standard deviation for annual salary	s = €3 348
Population proportion who have completed the management training programme	π = 0.60	Sample proportion who have completed the management training programme	p = 0.63

Table 7.3 summarizes the sample results and compares the point estimates to the actual values of the population parameters.

The point estimates in Table 7.3 differ somewhat from the corresponding population parameters. This difference is to be expected because a sample, rather than a census of the entire population, is being used to obtain the point estimates. In the next chapter, we shall show how to construct an interval estimate in order to provide information about how close the point estimate is to the population parameter.

Practical advice

The subject matter of most of the rest of the book is statistical inference. Point estimation is a form of statistical inference. We use a sample statistic to make an inference about a population parameter. When making inferences about a population based on a sample, it is important to have a close correspondence between the sampled population and the target 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 taken. In this section, we have described the process of drawing a simple random sample from the population of EAI managers and making point estimates of characteristics of that same population. So the sampled population and the target population are identical, which is the desired situation. But in other cases, it is not as easy to obtain a close correspondence between the sampled and target populations.

Consider the case of a theme park selecting a sample of its customers to learn about characteristics such as age and time spent at the park. Suppose all the sample elements were selected on a day when park attendance was restricted to employees of a large company. Then the sampled population would be composed of employees of that company and members of their families. If the target population we wanted to make inferences about were typical park customers over a typical summer, then there might be a substantial difference between the sampled population and the target population. In such a case, we would question the validity of the point estimates being made. The park management would be in the best position to know whether a sample taken on a particular day was likely to be representative of the target population.

In summary, whenever a sample is used to make inferences about a population, we should make sure that the study is designed so that the sampled population and the target population are in close agreement. Good judgement is a necessary ingredient of sound statistical practice.

EXERCISES

Methods

7. The following data are from a simple random sample.

5 8 10 7 10 14

a. Calculate a point estimate of the population mean.
b. Calculate a point estimate of the population standard deviation.

COMPLETE SOLUTIONS

8. A survey question for a sample of 150 individuals yielded 75 Yes responses, 55 No responses and 20 No Opinion responses.

a. Calculate a point estimate of the proportion in the population who respond Yes.
b. Calculate a point estimate of the proportion in the population who respond No.

Applications

9. A simple random sample of five months of sales data provided the following information:

Month:	1	2	3	4	5
Units sold:	94	100	85	94	92

a. Calculate a point estimate of the population mean number of units sold per month.
b. Calculate a point estimate of the population standard deviation.

10. The data set Mutual Fund contains data on a sample of 40 mutual funds. These were randomly selected from 283 funds featured in *Business Week*. Use the data set to answer the following questions.

MUTUAL FUND

a. Compute a point estimate of the proportion of the *Business Week* mutual funds that are load funds.
b. Compute a point estimate of the proportion of the funds that are classified as high risk.
c. Compute a point estimate of the proportion of the funds that have a below-average risk rating.

11. In a YouGov opinion poll for the *Financial Times* in late June 2012, during the 'Euro crisis', a sample of 1033 German adults was asked 'If there were a referendum tomorrow on Germany's membership of the single currency, the euro, how would you vote?' The responses were:

To stay in the single currency	444
To bring back the Deutschmark	424
Would not vote	72
Don't know	93

Calculate point estimates of the following population parameters:

a. The proportion of all adults who would vote to stay in the single currency.
b. The proportion of all adults who vote to bring back the Deutschmark.
c. The proportion of all adults who would not vote or don't know.

12. Many drugs used to treat cancer are expensive. *Business Week* reported on the cost per treatment of Herceptin, a drug used to treat breast cancer. Typical treatment costs (in dollars) for Herceptin are provided by a simple random sample of ten patients.

COMPLETE SOLUTIONS

4376	5578	2717	4920	4495
4798	6446	4119	4237	3814

a. Calculate a point estimate of the mean cost per treatment with Herceptin.
b. Calculate a point estimate of the standard deviation of the cost per treatment with Herceptin.

7.4 INTRODUCTION TO SAMPLING DISTRIBUTIONS

For the simple random sample of 30 EAI managers in Table 7.2, the point estimate of μ is $\bar{x} = €51\,814$ and the point estimate of π is $p = 0.63$. Suppose we select another simple random sample of 30 EAI managers and obtain the following point estimates:

$$\text{Sample mean: } \bar{x} = €52\,670$$
$$\text{Sample proportion: } p = 0.70$$

Note that different values of the sample mean and sample proportion were obtained. A second simple random sample of 30 EAI managers cannot be expected to provide exactly the same point estimates as the first sample.

Now, suppose we repeat the process of selecting a simple random sample of 30 EAI managers over and over again, each time computing the values of the sample mean and sample proportion. Table 7.4 contains a portion of the results obtained for 500 simple random samples, and Table 7.5 shows the

TABLE 7.4 Values $\bar{x}$ and p from 500 simple random samples of 30 EAI managers

Sample number	Sample mean ($\bar{x}$)	Sample proportion (p)
1	51 814	0.63
2	52 670	0.70
3	51 780	0.67
4	51 588	0.53
.	.	.
.	.	
.	.	.
500	51 752	0.50

TABLE 7.5 Frequency distribution of $\overline{X}$ values from 500 simple random samples of 30 EAI managers

Mean annual salary (€)	Frequency	Relative frequency
49 500.00–49 999.99	2	0.004
50 000.00–50 499.99	16	0.032
50 500.00–50 999.99	52	0.104
51 000.00–51 499.99	101	0.202
51 500.00–51 999.99	133	0.266
52 000.00–52 499.99	110	0.220
52 500.00–52 999.99	54	0.108
53 000.00–53 499.99	26	0.052
53 500.00–53 999.99	6	0.012
	Totals 500	1.000

frequency and relative frequency distributions for the 500 values. Figure 7.1 shows the relative frequency histogram for the values.

In Chapter 5 we defined a random variable as a numerical description of the outcome of an experiment. If we consider selecting a simple random sample as an experiment, the sample mean is a numerical description of the outcome of the experiment. So, the sample mean is a random variable. In accordance with the naming conventions for random variables described in Chapters 5 and Chapter 6 (i.e. use of capital letters for names of random variables), we denote this random variable $\overline{X}$. Just like other random variables, $\overline{X}$ has a mean or expected value, a standard deviation and a probability distribution. Because the various possible values of $\overline{X}$ are the result of different simple random samples, the probability distribution of $\overline{X}$ is called the **sampling distribution** of $\overline{X}$. Knowledge of this sampling distribution will enable us to make probability statements about how close the sample mean is to the population mean μ.

Let us return to Figure 7.1. We would need to enumerate every possible sample of 30 managers and compute each sample mean to completely determine the sampling distribution of $\overline{X}$. However, the histogram of 500 $\bar{x}$ values gives an approximation of this sampling distribution. From the approximation we observe the bell-shaped appearance of the distribution. We note that the largest concentration of the $\bar{x}$ values and the mean of the 500 $\bar{x}$ values are near the population mean μ = €51 800. We shall describe the properties of the sampling distribution of $\overline{X}$ more fully in the next section.

The 500 values of the sample proportion are summarized by the relative frequency histogram in Figure 7.2. As in the case of the sample mean, the sample proportion is a random variable, which we denote P. If every possible sample of size 30 were selected from the population and if a value p were computed for each sample, the resulting distribution would be the sampling distribution of P. The relative frequency histogram of the 500 sample values in Figure 7.2 provides a general idea of the appearance of the sampling distribution of P.

FIGURE 7.1
Relative frequency histogram of sample mean values from 500 simple random samples of size 30 each

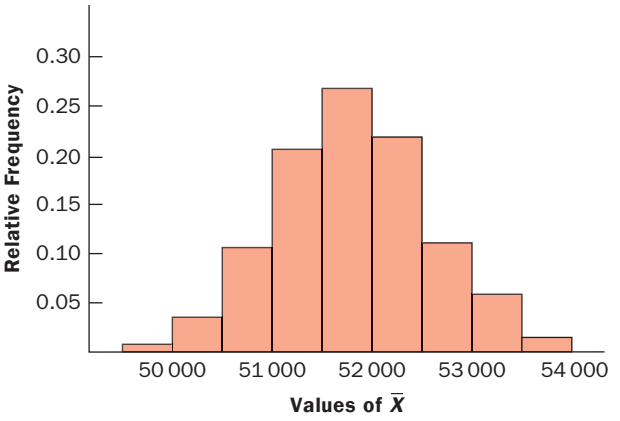
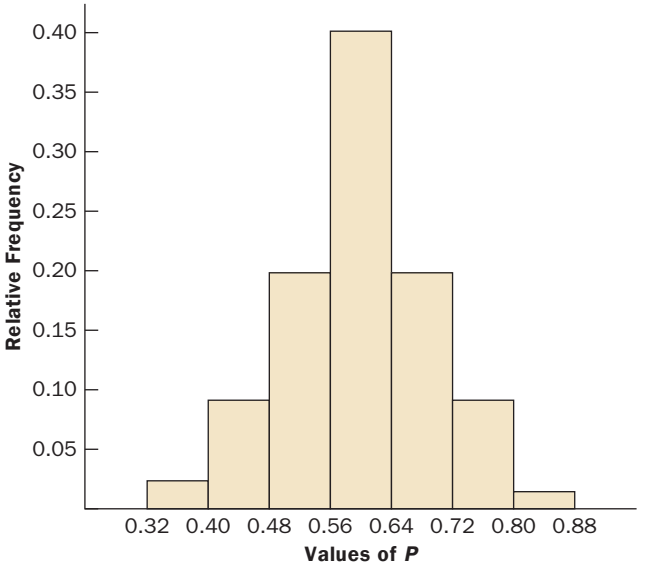


FIGURE 7.2
Relative frequency histogram of sample proportion values from 500 simple random samples of size 30 each



In practice, we select only one simple random sample from the population for estimating population characteristics. We repeated the sampling process 500 times in this section simply to illustrate that many different samples are possible and that the different samples generate a variety of values $\overline{x}$ and p for the sample statistics $\overline{X}$ and P. The probability distribution of any particular sample statistic is called the sampling distribution of the statistic. In Section 7.5 we show the characteristics of the sampling distribution of $\overline{X}$. In Section 7.6 we show the characteristics of the sampling distribution of P. The ability to understand the material in subsequent chapters depends heavily on the ability to understand and use the sampling distributions presented in this chapter.

7.5 SAMPLING DISTRIBUTION OF $\overline{X}$

This section describes the properties of the sampling distribution of $\overline{X}$. Just as with other probability distributions we have studied, the sampling distribution of $\overline{X}$ has an expected value or mean, a standard deviation and a characteristic shape or form. We begin by considering the expected value of $\overline{X}$.

Sampling distribution of $\overline{X}$

The sampling distribution of $\overline{X}$ is the probability distribution of all possible values of the sample mean.

Expected value of $\overline{X}$

Consider the $\overline{X}$ values generated by the various possible simple random samples. The mean of all these values is known as the expected value of $E(\overline{X})$. Let $\overline{X}$ represent the expected value of $\overline{X}$, and μ represent the mean of the population from which we are selecting a simple random sample. It can be shown that with simple random sampling, $E(\overline{X})$ and μ are equal.

Expected value of $\overline{X}$

where

$$E(\overline{X}) = \mu \tag{7.1}$$

$E(\overline{X}) =$ the expected value of $\overline{X}$

$\mu =$ the mean of the population from which the sample is selected

In Section 7.1 we saw that the mean annual salary for the population of EAI managers is μ = 51 800. So according to equation (7.1), the mean of all possible sample means for the EAI study is also €51 800.

When the expected value of a point estimator equals the population parameter, we say the point estimator is an **unbiased** estimator of the population parameter.

Unbiasedness

The sample statistic Q is an unbiased estimator of the population parameter θ if

$$E(Q) = \theta$$

where $E(Q)$ is the expected value of the sample statistic Q.

Figure 7.3 shows the cases of unbiased and biased point estimators. In the illustration showing the unbiased estimator, the mean of the sampling distribution is equal to the value of the population parameter. The estimation errors balance out in this case, because sometimes the value of the point estimator may be less than θ and other times it may be greater than θ.

In the case of a biased estimator, the mean of the sampling distribution is less than or greater than the value of the population parameter. In the illustration in Panel B of Figure 7.3, $E(Q)$ is greater than θ; the sample statistic has a high probability of overestimating the value of the population parameter. The amount of the bias is shown in the figure.

Equation (7.1) shows that $\overline{X}$ is an unbiased estimator of the population mean μ.

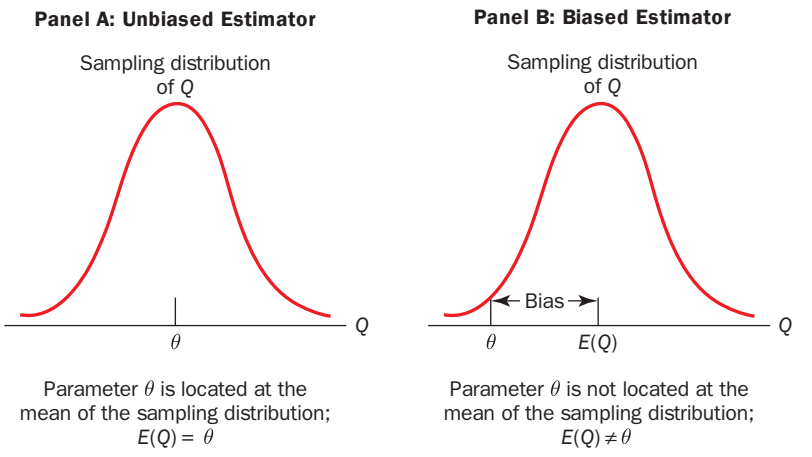


FIGURE 7.3 Examples of unbiased and biased point estimators

Standard deviation of $\overline{X}$

It can be shown that with simple random sampling, the standard deviation of $\overline{X}$ depends on whether the population is finite or infinite. We use the following notation.

$\sigma_{\overline{X}}$ = the standard deviation of $\overline{X}$
σ = the standard deviation of the population
n = the sample size
N = the population size

Standard deviation of $\overline{X}$

Finite population	Infinite population	
$\sigma_{\overline{X}} = \sqrt{\frac{N-n}{N-1}}\left(\frac{\sigma}{\sqrt{n}}\right)$	$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$	**(7.2)**

In comparing the two formulae in (7.2), we see that the factor $\sqrt{(N-n)/(N-1)}$ is required for the finite population case but not for the infinite population case. This factor is commonly referred to as the **finite population correction factor**. In many practical sampling situations, we find that the population involved, although finite, is 'large', whereas the sample size is relatively 'small'. In such cases the finite population correction factor is close to 1. As a result, the difference between the values of the standard deviation of $\overline{X}$ for the finite and infinite population cases becomes negligible. Then, $\sigma_{\overline{X}} = \sigma/\sqrt{n}$ becomes a good approximation to the standard deviation of $\overline{X}$ even though the population is finite. This observation leads to the following general guideline, or rule of thumb, for computing the standard deviation of $\overline{X}$.

Use the following expression to compute the standard deviation of $\overline{X}$

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \qquad \textbf{(7.3)}$$

whenever:

1. The population is infinite; or
2. The population is finite *and* the sample size is less than or equal to 5 per cent of the population size; that is, $n / N \leq 0.05$.

In cases where $n/N > 0.05$, the finite population version of formula (7.2) should be used in the computation of $\sigma_{\overline{X}}$. Unless otherwise noted, throughout the text we shall assume that the population size is 'large', $n/N \leq 0.05$, and expression (7.3) can be used to compute $\sigma_{\overline{X}}$.

To compute $\sigma_{\overline{X}}$, we need to know σ, the standard deviation of the population. To further emphasize the difference between $\sigma_{\overline{X}}$ and σ, we refer to $\sigma_{\overline{X}}$ as the **standard error** of the mean. The term standard error is used throughout statistical inference to refer to the standard deviation of a point estimator. Later we shall see that the value of the standard error of the mean is helpful in determining how far the sample mean may be from the population mean.

We return to the EAI example and compute the standard error of the mean associated with simple random samples of 30 EAI managers. In Section 7.1 we saw that the standard deviation of annual salary for the population of 2500 EAI managers is $\sigma = 4000$. In this case, the population is finite, with $N = 2500$. However, with a sample size of 30, we have $n/N = 30/2500 = 0.012$. Because the sample size is less than 5 per cent of the population size, we can ignore the finite population correction factor and use equation (7.3) to compute the standard error.

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{4000}{\sqrt{30}} = 730.3$$

Form of the sampling distribution of $\overline{X}$

The preceding results concerning the expected value and standard deviation for the sampling distribution of $\overline{X}$ are applicable for any population. The final step in identifying the characteristics of the sampling distribution of $\overline{X}$ is to determine the form or shape of the sampling distribution. We shall consider two cases: (1) the population has a normal distribution; and (2) the population does not have a normal distribution.

Population has a normal distribution

In many situations it is reasonable to assume that the population from which we are sampling has a normal, or nearly normal, distribution. When the population has a normal distribution, the sampling distribution of $\overline{X}$ is normally distributed for any sample size.

Population does not have a normal distribution

When the population from which we are selecting a simple random sample does not have a normal distribution, the **central limit theorem** is helpful in identifying the shape of the sampling distribution of $\overline{X}$.

Central limit theorem

In selecting simple random samples of size n from a population, the sampling distribution of the sample mean $\overline{X}$ can be approximated by a *normal distribution* as the sample size becomes large.

Figure 7.4 shows how the central limit theorem works for three different populations. Each column refers to one of the populations. The top panel of the figure shows that none of the populations is normally distributed. When the samples are of size 2, we see that the sampling distribution begins to take on an appearance different from that of the population distribution. For samples of size 5, we see all three sampling distributions beginning to take on a bell-shaped appearance. Finally, the samples of size 30 show all three sampling distributions to be approximately normally distributed. For sufficiently large samples, the sampling distribution of $\overline{X}$ can be approximated by a normal distribution. How large must the sample size be before we can assume that the central limit theorem applies? Studies of the sampling distribution of $\overline{X}$ for a variety of populations and a variety of sample sizes have indicated that, for most applications, the sampling distribution of $\overline{X}$ can be approximated by a normal distribution whenever the sample size is 30 or more.

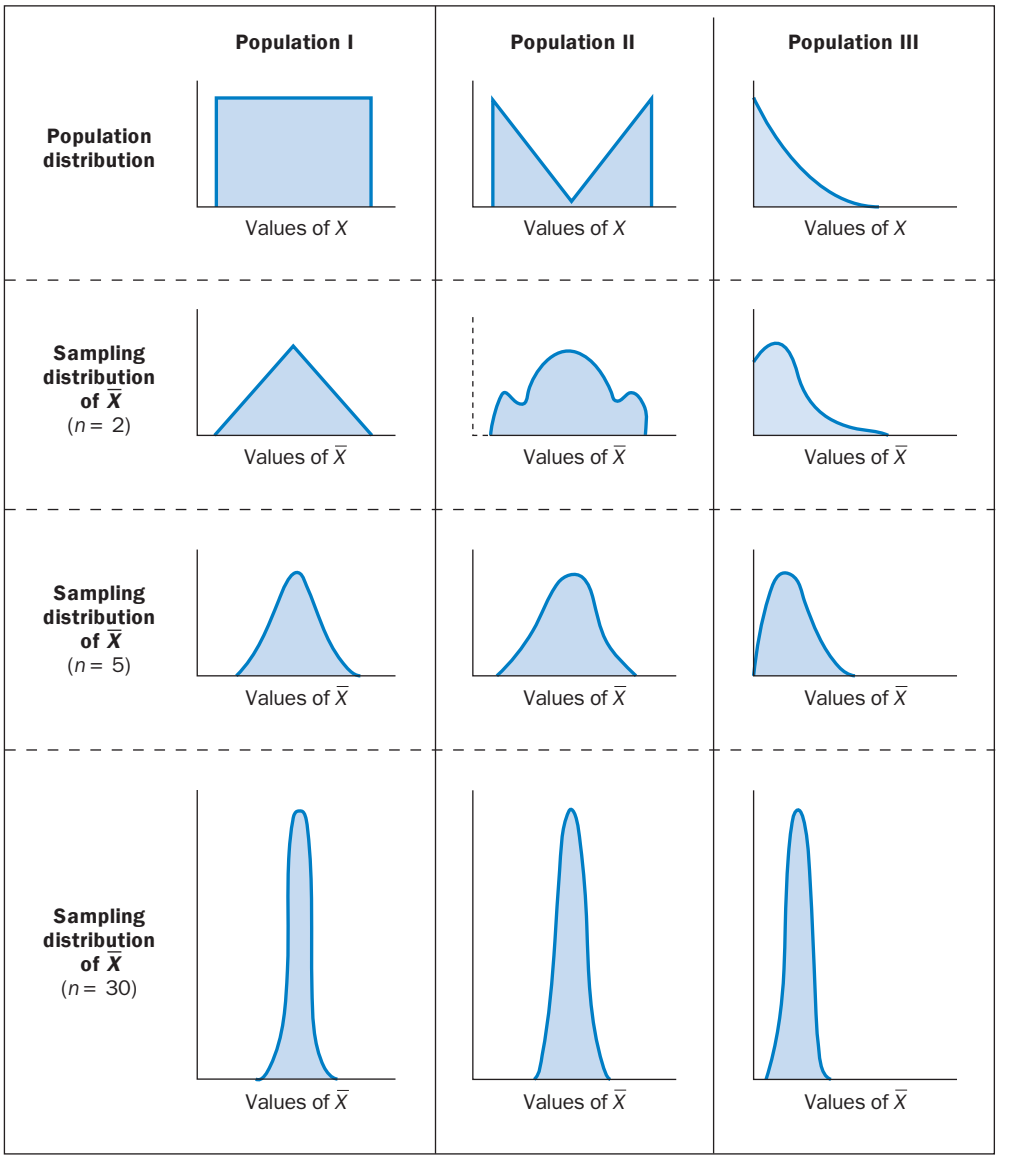


FIGURE 7.4
Illustration of the central limit theorem for three populations

The theoretical proof of the central limit theorem requires independent observations in the sample. This condition is met for infinite populations and for finite populations where sampling is done with replacement. Although the central limit theorem does not directly address sampling without replacement from finite populations, general statistical practice applies the findings of the central limit theorem when the population size is large.

Sampling distribution of $\overline{X}$ for the EAI problem

For the EAI problem, we previously showed that $E(\overline{X})$ = €51 800 and $\sigma_{\overline{X}}$ = €730.3 At this point, we do not have any information about the population distribution; it may or may not be normally distributed. If the population has a normal distribution, the sampling distribution of $\overline{X}$ is normally distributed.

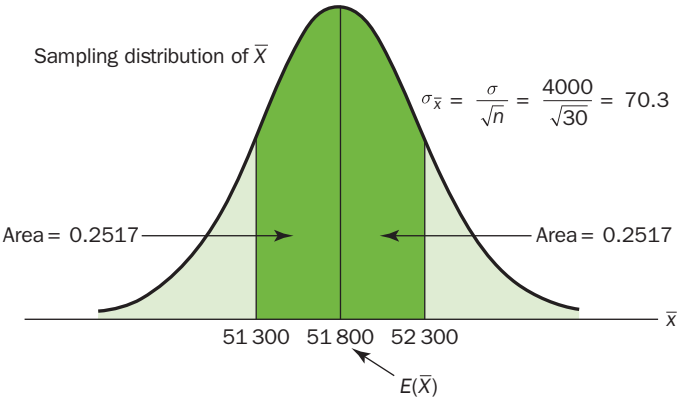


FIGURE 7.5
Sampling distribution of $\overline{X}$ for the mean annual salary of a simple random sample of 30 EAI managers, and the probability of $\overline{X}$ being within €500 of the population mean

If the population does not have a normal distribution, the simple random sample of 30 managers and the central limit theorem enable us to conclude that the sampling distribution can be approximated by a normal distribution. In either case, we can proceed with the conclusion that the sampling distribution can be described by the normal distribution shown in Figure 7.5.

Practical value of the sampling distribution of $\overline{X}$

We are interested in the sampling distribution of $\overline{X}$ because it can be used to provide probability information about the difference between the sample mean and the population mean. Suppose the head of personnel services believes the sample mean will be an acceptable estimate if it is within €500 of the population mean. It is not possible to guarantee that the sample mean will be within €500 of the population mean. Indeed, Table 7.5 and Figure 7.1 show that some of the 500 sample means differed by more than €2000 from the population mean. So we must think of the head of personnel's request in probability terms. What is the probability that the sample mean computed using a simple random sample of 30 EAI managers will be within €500 of the population mean?

We can answer this question using the sampling distribution of $\overline{X}$. Refer to Figure 7.5. With μ = €51 800, the personnel manager wants to know the probability that $\overline{X}$ is between €51 300 and €52 300. The darkly shaded area of the sampling distribution shown in Figure 7.5 gives this probability. Because the sampling distribution is normally distributed, with mean 51 800 and standard error of the mean 730.3, we can use the table of areas for the standard normal distribution to find the area or probability. At $\overline{X}$ = 51 300 we have

$$z = \frac{51300 - 51800}{730.3} = -0.68$$

Referring to the standard normal distribution table, we find the cumulative probability for $z = -0.68$ is 0.2483. Similar calculations for $\overline{X}$ = 52 300 show a cumulative probability for $z = +0.68$ of 0.7517. So the probability that the sample mean is between 51 300 and 52 300 is 0.7517 − 0.2483 = 0.5034.

These computations show that a simple random sample of 30 EAI managers has a 0.5034 probability of providing a sample mean that is within €500 of the population mean. Hence, there is a 1 − 0.5034 = 0.4966 probability that the difference between $\overline{X}$ and μ will be more than €500. In other words, a simple random sample of 30 EAI managers has a roughly 50/50 chance of providing a sample mean within the allowable €500. Perhaps a larger sample size should be considered. We explore this possibility by considering the relationship between the sample size and the sampling distribution of $\overline{X}$.

Relationship between sample size and the sampling distribution of $\overline{X}$

Suppose that in the EAI sampling problem we select a simple random sample of 100 EAI managers instead of the 30 originally considered. Intuitively, it would seem that with more sample data, the sample mean based on $n = 100$ should provide a better estimate of the population mean than the sample mean based on $n = 30$. To see how much better, let us consider the relationship between the sample size and the sampling distribution of $\overline{X}$.

First note that $E(\overline{X}) = \mu$, i.e. $\overline{X}$ is an unbiased estimator of μ, regardless of the sample size n. However, the standard error of the mean, $\sigma_{\overline{X}}$, is related to the square root of the sample size. The value of $\sigma_{\overline{X}}$ decreases when the sample size increases. With $n = 30$, the standard error of the mean for the EAI problem is 730.3. With the increase in the sample size to $n = 100$, the standard error of the mean decreases to:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{4000}{\sqrt{100}} = 400$$

The sampling distributions of $\overline{X}$ with $n = 30$ and $n = 100$ are shown in Figure 7.6. Because the sampling distribution with $n = 100$ has a smaller standard error, the values of $\overline{X}$ have less variation and tend to be closer to the population mean than the values of $\overline{X}$ with $n = 30$.

We can use the sampling distribution of $\overline{X}$ for $n = 100$ to compute the probability that a simple random sample of 100 EAI managers will provide a sample mean within €500 of the population mean. Because the sampling distribution is normal, with mean 51 800 and standard error of the mean 400, we can use the standard normal distribution table to find the area or probability. At $\overline{X} = 51\,300$ (Figure 7.7), we have:

$$z = \frac{51\,300 - 51\,800}{400} = -1.25$$

Referring to the standard normal probability distribution table, we find a cumulative probability for $z = -1.25$ of 0.1056. With a similar calculation for $\overline{X} = 52\,300$, we see that the probability of the sample mean being between 51 300 and 52 300 is $0.8944 - 0.1056 = 0.7888$. By increasing the sample size from 30 to 100 EAI managers, we have increased the probability of obtaining a sample mean within €500 of the population mean from 0.5034 to 0.7888.

The important point in this discussion is that as the sample size is increased, the standard error of the mean decreases. As a result, the larger sample size provides a higher probability that the sample mean is within a specified distance of the population mean.

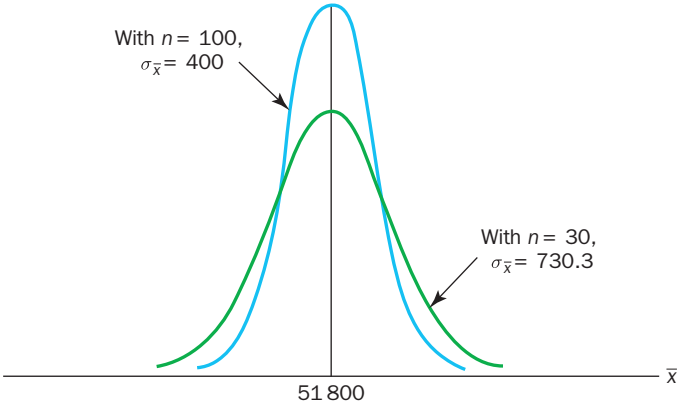
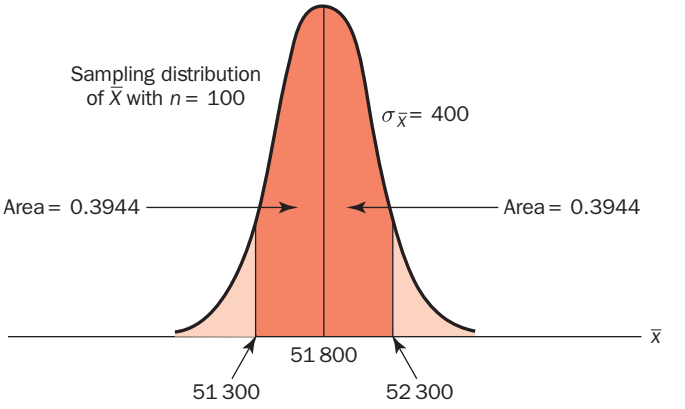


FIGURE 7.6
A comparison of the sampling distributions of $\overline{X}$ for simple random samples of $n = 30$ and $n = 100$ EAI managers

FIGURE 7.7
The probability of a sample mean being within €500 of the population mean when a simple random sample of 100 EAI managers is used



In presenting the sampling distribution of $\overline{X}$ for the EAI problem, we have taken advantage of the fact that the population mean $\mu = 51\,800$ and the population standard deviation $\sigma = 4000$ were known. However, usually the values μ and σ that are needed to determine the sampling distribution of $\overline{X}$ will be unknown. In Chapter 8 we shall show how the sample mean $\overline{X}$ and the sample standard deviation S are used when μ and σ are unknown.

EXERCISES

Methods

13. A population has a mean of 200 and a standard deviation of 50. A simple random sample of size 100 will be taken and the sample mean will be used to estimate the population mean.

a. What is the expected value of $\overline{X}$?
b. What is the standard deviation of $\overline{X}$?
c. Sketch the sampling distribution of $\overline{X}$.
d. What does the sampling distribution of $\overline{X}$ show?


COMPLETE SOLUTIONS

14. A population has a mean of 200 and a standard deviation of 50. Suppose a simple random sample of size 100 is selected and is used to estimate μ.

a. What is the probability that the sample mean will be within ±5 of the population mean?
b. What is the probability that the sample mean will be within ±10 of the population mean?

15. Assume the population standard deviation is $\sigma = 25$. Compute the standard error of the mean, $\sigma_{\overline{X}}$, for sample sizes of 50, 100, 150 and 200. What can you say about the size of the standard error of the mean as the sample size is increased?

16. Suppose a simple random sample of size 50 is selected from a population with $\sigma_{\overline{X}} = 25$. Find the value of the standard error of the mean in each of the following cases (use the finite population correction factor if appropriate).

a. The population size is infinite.
b. The population size is $N = 50\,000$.
c. The population size is $N = 5000$.
d. The population size is $N = 500$.

Applications

17. Refer to the EAI sampling problem. Suppose a simple random sample of 60 managers is used.

a. Sketch the sampling distribution of $\overline{X}$ when simple random samples of size 60 are used.
b. What happens to the sampling distribution of $\overline{X}$ if simple random samples of size 120 are used?
c. What general statement can you make about what happens to the sampling distribution of $\overline{X}$ as the sample size is increased? Does this generalization seem logical? Explain.

18. In the EAI sampling problem (see Figure 7.5), we showed that for $n = 30$, there was a 0.5034 probability of obtaining a sample mean within ± €500 of the population mean.

a. What is the probability that $\overline{X}$ is within €500 of the population mean if a sample of size 60 is used?
b. Answer part (a) for a sample of size 120.

COMPLETE SOLUTIONS

19. The Automobile Association gave the average price of unleaded petrol in Sweden as 14.63 Swedish krona (SK) per litre in June 2012. Assume this price is the population mean, and that the population standard deviation is $\sigma = 1$ SK.

a. What is the probability that the mean price for a sample of 30 petrol stations is within 0.25 SK of the population mean?
b. What is the probability that the mean price for a sample of 50 petrol stations is within 0.25 SK of the population mean?
c. What is the probability that the mean price for a sample 100 petrol stations is within 0.25 SK of the population mean?
d. Would you recommend a sample size of 30, 50 or 100 to have at least a 0.95 probability that the sample mean is within 0.25 SK of the population mean?

20. According to *Golf Digest*, the average score for male golfers is 95 and the average score for female golfers is 106. Use these values as population means. Assume that the population standard deviation is $\sigma = 14$ strokes for both men and women. A simple random sample of 30 male golfers and another simple random sample of 45 female golfers are taken.

a. Sketch the sampling distribution of $\overline{X}$ for male golfers.
b. What is the probability that the sample mean is within three strokes of the population mean for the sample of male golfers?
c. What is the probability that the sample mean is within three strokes of the population mean for the sample of female golfers?
d. In which case is the probability higher (b or c)? Why?

21. A researcher reports survey results by stating that the standard error of the mean is 20. The population standard deviation is 500.

a. How large was the sample?
b. What is the probability that the point estimate was within ± 25 of the population mean?

22. To estimate the mean age for a population of 4000 employees in a large company in Kuwait City, a simple random sample of 40 employees is selected.

a. Would you use the finite population correction factor in calculating the standard error of the mean? Explain.
b. If the population standard deviation is $\sigma = 8.2$, compute the standard error both with and without the finite population correction factor. What is the rationale for ignoring the finite population correction factor whenever $n/N \leq 0.05$?
c. What is the probability that the sample mean age of the employees will be within ± two years of the population mean age?

7.6 SAMPLING DISTRIBUTION OF *P*

The sample proportion P is a point estimator of the population proportion π. The formula for computing the sample proportion is:

$$p = \frac{m}{n}$$

where:

m = the number of elements in the sample that possess the characteristic of interest
n = sample size.

The sample proportion P is a random variable and its probability distribution is called the sampling distribution of P.

Sampling distribution of *P*

The sampling distribution of P is the probability distribution of all possible values of the sample proportion P.

To determine how close the sample proportion is to the population proportion π, we need to understand the properties of the sampling distribution of P: the expected value of P, the standard deviation of P and the shape of the sampling distribution of P.

Expected value of *P*

The expected value of P, the mean of all possible values of P, is equal to the population proportion π. P is an unbiased estimator of π.

Expected value of *P*

where:

$$E(P) = \pi \qquad \textbf{(7.4)}$$

$E(P)$ = the expected value of P
π = the population proportion

In Section 7.1 we noted that $\pi = 0.60$ for the EAI population, where π is the proportion of the population of managers who participated in the company's management training programme. The expected value of P for the EAI sampling problem is therefore 0.60.

Standard deviation of *P*

Just as we found for the standard deviation of $\overline{X}$, the standard deviation of P depends on whether the population is finite or infinite.

Standard deviation of *P*

Finite population

$$\sigma_P = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{\pi(1-\pi)}{n}}$$

Infinite population

$$\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}} \qquad (7.5)$$

Comparing the two formulae in (7.5), we see that the only difference is the use of the finite population correction factor $\sqrt{(N-n)/(N-1)}$.

As was the case with the sample mean, the difference between the expressions for the finite population and the infinite population becomes negligible if the size of the finite population is large in comparison to the sample size. We follow the same rule of thumb that we recommended for the sample mean. That is, if the population is finite with $n/N \leq 0.05$, we shall use $\sigma_P = \sqrt{\pi(1-\pi)/n}$.

However, if the population is finite with $n/N > 0.05$, the finite population correction factor should be used. Again, unless specifically noted, throughout the text we shall assume that the population size is large in relation to the sample size and so the finite population correction factor is unnecessary.

In Section 7.5 we used the term standard error of the mean to refer to the standard deviation of $\overline{X}$. We stated that in general the term standard error refers to the standard deviation of a point estimator. Accordingly, for proportions we use *standard error of the proportion* to refer to the standard deviation of *P*.

Let us now return to the EAI example and compute the standard error of the proportion associated with simple random samples of 30 EAI managers. For the EAI study we know that the population proportion of managers who participated in the management training programme is $\pi = 0.60$. With $n/N = 30/2500 = 0.012$, we can ignore the finite population correction factor when we compute the standard error of the proportion. For the simple random sample of 30 managers, σ_P is:

$$\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.60(1-0.60)}{30}} = 0.0894$$

Form of the sampling distribution of *P*

The sample proportion is $p = m/n$. For a simple random sample from a large population, the value of m is a binomial random variable indicating the number of elements in the sample with the characteristic of interest. Because n is a constant, the probability of each value of m/n is the same as the binomial probability of m, which means that the sampling distribution of P is also a discrete probability distribution.

In Chapter 6 we showed that a binomial distribution can be approximated by a normal distribution whenever the sample size is large enough to satisfy the following two conditions: $n\pi \geq 5$ and $n(1-\pi) \geq 5$. Assuming these two conditions are satisfied, the probability of m in the sample proportion, $p = m/n$, can be approximated by a normal distribution. And because n is a constant, the sampling distribution of P can also be approximated by a normal distribution. This approximation is stated as follows:

The sampling distribution of P can be approximated by a normal distribution whenever $n\pi \geq 5$ and $n(1-\pi) \geq 5$.

In practical applications, when an estimate of a population proportion is needed, we find that sample sizes are almost always large enough to permit the use of a normal approximation for the sampling distribution of *P*.

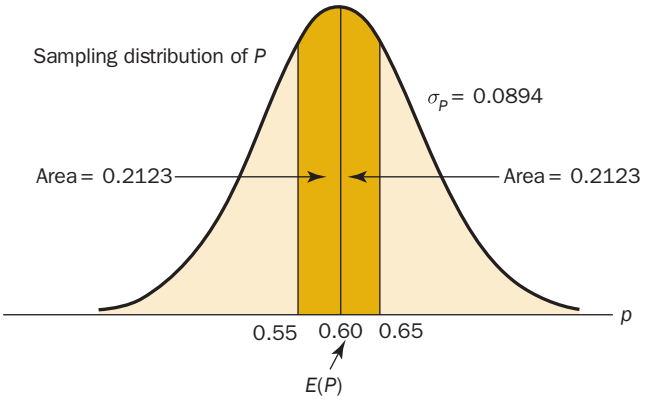


FIGURE 7.8
Sampling distribution of *P* for the proportion of EAI managers who participated in the management training programme

Recall that for the EAI sampling problem the population proportion of managers who participated in the training programme is $\pi = 0.60$. With a simple random sample of size 30, we have $n\pi = 30(0.60) = 18$ and $n(1 - \pi) = 30(0.40) = 12$. Consequently, the sampling distribution of P can be approximated by the normal distribution shown in Figure 7.8.

Practical value of the sampling distribution of *P*

The practical value of the sampling distribution of P is that it can be used to provide probability information about the difference between the sample proportion and the population proportion. For instance, suppose that in the EAI problem the head of personnel services wants to know the probability of obtaining a value of P that is within 0.05 of the population proportion of EAI managers who participated in the training programme. That is, what is the probability of obtaining a sample with a sample proportion P between 0.55 and 0.65? The darkly shaded area in Figure 7.8 shows this probability. Using the fact that the sampling distribution of P can be approximated by a normal distribution with a mean of 0.60 and a standard error of the proportion of $\sigma_P = 0.0894$, we find that the standard normal random variable corresponding to $p = 0.55$ has a value of $z = (0.55 - 0.60)/0.0894 = -0.56$. Referring to the standard normal distribution table, we see that the cumulative probability for $z = -0.56$ is 0.2877. Similarly, for $p = 0.56$ we find a cumulative probability of 0.7123. Hence, the probability of selecting a sample that provides a sample proportion P within 0.05 of the population proportion π is $0.7123 - 0.2877 = 0.4246$.

If we consider increasing the sample size to $n = 100$, the standard error of the proportion becomes:

$$\sigma_P = \sqrt{\frac{0.60(1 - 0.60)}{100}} = 0.049$$

The probability of the sample proportion being within 0.05 of the population proportion can now be calculated, again using the standard normal distribution table to find the area or probability. At $p = 0.55$, we have $z = (0.55 - 0.60)/0.049 = -1.02$. Referring to the standard normal distribution table, we see that the cumulative probability for $z = -1.02$ is 0.1539. Similarly, at $p = 0.65$ the cumulative probability is 0.8461. Hence, if the sample size is increased from 30 to 100, the probability that the sample proportion is within 0.05 of the population proportion π will increase to $0.8461 - 0.1539 = 0.6922$.

EXERCISES

Methods

23. A simple random sample of size 100 is selected from a population with $\pi = 0.40$.

a. What is the expected value of P?
b. What is the standard error of P?
c. Sketch the sampling distribution of P.

24. Assume that the population proportion is 0.55. Compute the standard error of the sample proportion, σ_P, for sample sizes of 100, 200, 500 and 1000. What can you say about the size of the standard error of the proportion as the sample size is increased?

COMPLETE SOLUTIONS

25. The population proportion is 0.30. What is the probability that a sample proportion will be within ±0.04 of the population proportion for each of the following sample sizes?

a. $n = 100$.
b. $n = 200$.
c. $n = 500$.
d. $n = 1000$.
e. What is the advantage of a larger sample size?

Applications

26. The Chief Executive Officer of Dunkley Distributors plc believes that 30 per cent of the firm's orders come from first-time customers. A simple random sample of 100 orders will be used to estimate the proportion of first-time customers.

a. Assume that the CEO is correct and $\pi = 0.30$. Describe the sampling distribution of the sample proportion P for this study?
b. What is the probability that the sample proportion P will be between 0.20 and 0.40?
c. What is the probability that the sample proportion P will be between 0.25 and 0.35?

27. Eurostat reported that, in 2011, 64 per cent of households in Spain had Internet access. Use a population proportion $\pi = 0.64$ and assume that a sample of 300 households will be selected.

a. Sketch the sampling distribution of P, the sample proportion of households that have Internet access.
b. What is the probability that the sample proportion P will be within ±0.03 of the population proportion?
c. Answer part (b) for sample sizes of 600 and 1000.

28. Advertisers contract with Internet service providers and search engines to place ads on websites. They pay a fee based on the number of potential customers who click on their ads. Unfortunately, click fraud – i.e. someone clicking on an ad solely for the purpose of driving up advertising revenue – has become a problem. Forty per cent of advertisers claim they have been a victim of click fraud. Suppose a simple random sample of 380 advertisers is taken to learn about how they are affected by this practice. Assume the population proportion $\pi = 0.40$.

a. What is the probability the sample proportion will be within ±0.04 of the population proportion experiencing click fraud?
b. What is the probability the sample proportion will be greater than 0.45?

29. In April 2012, a Gallup poll amongst a sample of 1074 Egyptian adults reported that 58 per cent thought it would be a bad thing if the military remained involved in politics after the presidential

election. Assume that the population proportion was $\pi = 0.58$, and that P is the sample proportion in a sample of $n = 1074$.

a. Sketch the sampling distribution of P.
b. What is the probability that P will be within plus or minus 0.02 of ϕ.
c. Answer part (b) for sample of 2000 adults.

COMPLETE SOLUTIONS

30. A market research firm conducts telephone surveys with a 40 per cent historical response rate. What is the probability that in a new sample of 400 telephone numbers, at least 150 individuals will cooperate and respond to the questions? In other words, what is the probability that the sample proportion will be at least 150/400 = 0.375?

31. Lura Jafari is a successful sales representative for a major publisher of university textbooks. Historically, Lura secures a book adoption on 25 per cent of her sales calls. Assume that her sales calls for one month are taken as a sample of all possible sales calls, and that a statistical analysis of the data estimates the standard error of the sample proportion to be 0.0625.

a. How large was the sample used in this analysis? That is, how many sales calls did Lura make during the month?
b. Let P indicate the sample proportion of book adoptions obtained during the month. Sketch the sampling distribution P.
c. Using the sampling distribution of P, compute the probability that Lura will obtain book adoptions on 30 per cent or more of her sales calls during a one-month period.

ONLINE RESOURCES

For the data files, online summary, additional questions and answers, and software section for Chapter 7, go to the online platform.

SUMMARY

In this chapter we presented the concepts of simple random sampling and sampling distributions.

Simple random sampling was defined for sampling without replacement and sampling with replacement. We demonstrated how a simple random sample can be selected and how the sample data can be used to calculate point estimates of population parameters.

Point estimators such as $\overline{X}$ and P are random variables. The probability distribution of such a random variable is called a sampling distribution. In particular, we described the sampling distributions of the sample mean $\overline{X}$ and the sample proportion P. We stated that $E(\overline{X}) = \mu$ and $E(P) = \pi$, i.e. they are unbiased estimators of the respective parameters. After giving the standard deviation or standard error formulae for these estimators, we described the conditions necessary for the sampling distributions of and P to follow normal distributions. Finally, we gave examples of how these normal sampling distributions can be used to calculate the probability of $\overline{X}$ or P being within any given distance of μ or π respectively.

KEY TERMS

Central limit theorem
Finite population correction factor
Parameter
Point estimate
Point estimator
Sample statistic
Sampled population
Sampling distribution
Sampling frame
Sampling with replacement
Sampling without replacement
Simple random sampling
Standard error
Target population
Unbiasedness

KEY FORMULAE

Expected value of $\bar{X}$

$$E(\bar{X}) = \mu \tag{7.1}$$

Standard deviation of $\bar{X}$ (standard error)

Finite population

$$\sigma_{\bar{X}} = \sqrt{\frac{N-n}{N-1}}\left(\frac{\sigma}{\sqrt{n}}\right)$$

Infinite population

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \tag{7.2}$$

Expected value of *P*

$$E(P) = \pi \tag{7.4}$$

Standard deviation of *P* (standard error)

Finite population

$$\sigma_P = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{\pi(1-\pi)}{n}}$$

Infinite population

$$\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}} \tag{7.5}$$