CHAPTER 7

Sampling and Sampling Distributions

CONTENTS

STATISTICS in PRACTICE

MEADWESTVACO CORPORATION*

STAMFORD, CONNECTICUT

MeadWestvaco Corporation, a leading producer of packaging, coated and specialty papers, and specialty chemicals, employs more than 17,000 people. It operates worldwide in 30 countries and serves customers located in approximately 100 countries. MeadWestvaco's internal consulting group uses sampling to provide a variety of information that enables the company to obtain significant productivity benefits and remain competitive.

For example, MeadWestvaco maintains large woodland holdings, which supply the trees, or raw material, for many of the company's products. Managers need reliable and accurate information about the timberlands and forests to evaluate the company's ability to meet its future raw material needs. What is the present volume in the forests? What is the past growth of the forests? What is the projected future growth of the forests? With answers to these important questions MeadWestvaco's managers can develop plans for the future, including long-term planting and harvesting schedules for the trees.

How does MeadWestvaco obtain the information it needs about its vast forest holdings? Data collected from sample plots throughout the forests are the basis for learning about the population of trees owned by the company. To identify the sample plots, the timberland holdings are first divided into three sections based on location and types of trees. Using maps and random numbers, MeadWestvaco analysts identify random samples of 1/5- to 1/7-acre plots in each section of the forest. MeadWestvaco foresters collect data from these sample plots to learn about the forest population.

Random sampling of its forest holdings enables MeadWestvaco Corporation to meet future raw material needs. © Robert Crum/Shutterstock.com.

Foresters throughout the organization participate in the field data collection process. Periodically, two-person teams gather information on each tree in every sample plot. The sample data are entered into the company's continuous forest inventory (CFI) computer system. Reports from the CFI system include a number of frequency distribution summaries containing statistics on types of trees, present forest volume, past forest growth rates, and projected future forest growth and volume. Sampling and the associated statistical summaries of the sample data provide the reports essential for the effective management of MeadWestvaco's forests and timberlands.

In this chapter you will learn about simple random sampling and the sample selection process. In addition, you will learn how statistics such as the sample mean and sample proportion are used to estimate the population mean and population proportion. The important concept of a sampling distribution is also introduced.

*The authors are indebted to Dr. Edward P. Winkofsky for providing this Statistics in Practice.

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.

The reason we select a sample is to collect data to make an inference and answer research questions about a population.

Let us begin by citing two examples in which sampling was used to answer a research question about a population.

1. Members of a political party in Texas were considering supporting a particular candidate for election to the U.S. Senate, and party leaders wanted to estimate the proportion of registered voters in the state favoring the candidate. A sample of 400 registered voters in Texas was selected and 160 of the 400 voters indicated a preference for the candidate. Thus, an estimate of the proportion of the population of registered voters favoring the candidate is 160/400 = .40.
2. A tire manufacturer is considering producing a new tire designed to provide an increase in mileage over the firm's current line of tires. To estimate the mean useful life of the new tires, the manufacturer produced a sample of 120 tires for testing. The test results provided a sample mean of 36,500 miles. Hence, an estimate of the mean useful life for the population of new tires was 36,500 miles.

A sample mean provides an estimate of a population mean, and a sample proportion provides an estimate of a population proportion. With estimates such as these, some estimation error can be expected. This chapter provides the basis for determining how large that error might be.

It is important to realize that sample results provide only *estimates* of the values of the corresponding population characteristics. We do not expect exactly .40, or 40%, of the population of registered voters to favor the candidate, nor do we expect the sample mean of 36,500 miles to exactly equal the mean mileage for the population of all new tires produced. The reason is simply that the sample contains only a portion of the population. Some sampling error is to be expected. With proper sampling methods, the sample results will provide "good" estimates of the population parameters. But how good can we expect the sample results to be? Fortunately, statistical procedures are available for answering this question.

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

The sampled population for the tire mileage example is more difficult to define because the sample of 120 tires 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 the tires that could have been made by the production process at that particular point in time. In this sense the sampled population is considered infinite, making it impossible to construct a frame to draw the sample from. In Section 7.2, we discuss how to select a random sample in such a situation.

In this chapter, we show how simple random sampling can be used to select a sample from a finite population and describe how a random sample can be taken from an infinite population that is generated by an ongoing process. We then show how data obtained from a 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. As we will show, knowledge of the appropriate sampling distribution enables us to make statements about how close the sample estimates are to the corresponding population parameters. The last section discusses some alternatives to simple random sampling that are often employed in practice.

7.1 The Electronics Associates Sampling Problem

The director of personnel for Electronics Associates, Inc. (EAI), has been assigned the task of developing a profile of the company's 2500 managers. The characteristics to be identified include the mean annual salary for the managers and the proportion of managers having completed the company's management training program.

WEB file
EAI

Using the 2500 managers as the population for this study, we can find the annual salary and the training program status for each individual by referring to the firm's personnel records. The data set containing this information for all 2500 managers in the population is in the file named EAI.

Using the EAI data and the formulas presented in Chapter 3, we computed the population mean and the population standard deviation for the annual salary data.

$$\text{Population mean:} \quad \mu = \$51{,}800$$
$$\text{Population standard deviation:} \quad \sigma = \$4000$$

The data for the training program status show that 1500 of the 2500 managers completed the training program.

Numerical characteristics of a population are called **parameters**. Letting p denote the proportion of the population that completed the training program, we see that $p = 1500/2500 = .60$. The population mean annual salary ($\mu = \$51{,}800$), the population standard deviation of annual salary ($\sigma = \$4000$), and the population proportion that completed the training program ($p = .60$) are parameters of the population of EAI managers.

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.

Now, suppose that the necessary information on all the EAI managers was not readily available in the company's database. The question we now consider is how the firm's director of personnel can obtain estimates of the population parameters by using a sample of managers rather than all 2500 managers in the population. Suppose that a sample of 30 managers will be used. Clearly, the time and the cost of developing a profile would be substantially less for 30 managers than for the entire population. If the personnel director 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. Let us explore the possibility of using a sample for the EAI study by first considering how we can identify a sample of 30 managers.

7.2 Selecting a Sample

In this section we describe how to select a sample. We first describe how to sample from a finite population and then describe how to select a sample from an infinite population.

Sampling from a Finite Population

Other methods of probability sampling are described in Section 7.8

Statisticians recommend selecting a probability sample when sampling from a finite population because a probability sample allows them to make valid statistical inferences about the population. The simplest type of probability sample is one in which each sample of size n has the same probability of being selected. It is called a simple random sample. A simple random sample of size n from a finite population of size N is defined as follows.

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.

We describe how Excel, Minitab, and StatTools can be used to generate a simple random sample in the chapter appendices.

One procedure for selecting a simple random sample from a finite population is to use a table of random numbers 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. Sampling n elements in this way will satisfy the definition of a simple random sample from a finite population.

To select a simple random sample from the finite population of EAI managers, we first construct a frame by assigning each manager a number. For example, we can assign the

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

managers the numbers 1 to 2500 in the order that 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 having an equal chance of occurring. Because the largest number in the population list of EAI managers, 2500, has four digits, we will select random numbers from the table in sets or groups of four digits. Even though we may start the selection of random numbers anywhere in the table and move systematically in a direction of our choice, we will use the first row of Table 7.1 and move from left to right. The first 7 four-digit random numbers are

6327 1599 8671 7445 1102 1514 1807

Because the numbers in the table are random, these four-digit numbers are equally likely.

The random numbers in the table are shown in groups of five for readability.

We can now use these four-digit random numbers 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. It does not correspond to one of the numbered managers in the population, and hence is discarded. The second number, 1599, is between 1 and 2500. Thus 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 before identifying managers number 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.

In implementing this simple random sample selection process, it is possible that a random number used previously may appear again in the table before the complete sample of 30 EAI managers has been selected. Because we do not want to select a manager more than one time, any previously used random numbers are ignored because the corresponding manager is already included in the sample. 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. However, sampling without replacement is the sampling procedure used most often in practice. When we refer to simple random sampling, we will assume the sampling is without replacement.

Sampling from an Infinite Population

Sometimes we want to select a sample from a population, but the population is infinitely large or the elements of the population are being generated by an ongoing process for which there is no limit on the number of elements that can be generated. Thus, it is not possible to develop a list of all the elements in the population. This is considered the infinite population case. With an infinite population, we cannot select a simple random sample because we cannot construct a frame consisting of all the elements. In the infinite population case, statisticians recommend selecting what is called a random sample.

RANDOM SAMPLE (INFINITE POPULATION)

A **random sample** of size *n* from an infinite population is a sample selected such that the following conditions are satisfied.

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

Care and judgment must be exercised in implementing the selection process for obtaining a random sample from an infinite population. Each case may require a different selection procedure. Let us consider two examples to see what we mean by the conditions (1) each element selected comes from the same population and (2) each element is selected independently.

A common quality control application involves a production process where there is no limit on the number of elements that can be produced. The conceptual population we are sampling from is all the elements that could be produced (not just the ones that are produced) by the ongoing production process. Because we cannot develop a list of all the elements that could be produced, the population is considered infinite. To be more specific, let us consider a production line designed to fill boxes of a breakfast cereal with a mean weight of 24 ounces of breakfast cereal per box. Samples of 12 boxes filled by this process are periodically selected by a quality control inspector to determine if the process is operating properly or if, perhaps, a machine malfunction has caused the process to begin underfilling or overfilling the boxes.

With a production operation such as this, the biggest concern in selecting a random sample is to make sure that condition 1, the sampled elements are selected from the same population, is satisfied. To ensure that this condition is satisfied, the boxes must be selected at approximately the same point in time. This way the inspector avoids the possibility of selecting some boxes when the process is operating properly and other boxes when the process is not operating properly and is underfilling or overfilling the boxes. With a production process such as this, the second condition, each element is selected independently, is satisfied by designing the production process so that each box of cereal is filled independently. With this assumption, the quality control inspector only needs to worry about satisfying the same population condition.

As another example of selecting a random sample from an infinite population, consider the population of customers arriving at a fast-food restaurant. Suppose an employee is asked to select and interview a sample of customers in order to develop a profile of customers who visit the restaurant. The customer arrival process is ongoing and there is no way to obtain a list of all customers in the population. So, for practical purposes, the population for this ongoing process is considered infinite. As long as a sampling procedure is designed so that all the elements in the sample are customers of the restaurant and they are selected independently, a random sample will be obtained. In this case, the employee collecting the sample needs to select the sample from people who come into the restaurant and make a purchase to ensure that the same population condition is satisfied. If, for instance, the employee selected someone for the sample who came into the restaurant just to use the restroom, that person would not be a customer and the same population condition would be violated. So, as long as the interviewer selects the sample from people making a purchase at the restaurant, condition 1 is satisfied. Ensuring that the customers are selected independently can be more difficult.

The purpose of the second condition of the random sample selection procedure (each element is selected independently) is to prevent selection bias. In this case, selection bias would occur if the interviewer were free to select customers for the sample arbitrarily. The interviewer might feel more comfortable selecting customers in a particular age group and might avoid customers in other age groups. Selection bias would also occur if the interviewer selected a group of five customers who entered the restaurant together and asked all of them to participate in the sample. Such a group of customers would be likely to exhibit similar characteristics, which might provide misleading information about the population of customers. Selection bias such as this 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 elements (customers) are selected independently.

McDonald's, the fast-food restaurant leader, implemented a random sampling procedure for this situation. The sampling procedure was based on the fact that some customers presented discount coupons. Whenever a customer presented a discount coupon, the next customer served was asked to complete a customer profile questionnaire. Because arriving customers presented discount coupons randomly and independently of other customers, this sampling procedure ensured that customers were selected independently. As a result, the sample satisfied the requirements of a random sample from an infinite population.

Situations involving sampling from an infinite population are usually associated with a process that operates over time. Examples include parts being manufactured on a production line, repeated experimental trials in a laboratory, transactions occurring at a bank, telephone calls arriving at a technical support center, and customers entering a retail store. In each case, the situation may be viewed as a process that generates elements from an infinite population. As long as the sampled elements are selected from the same population and are selected independently, the sample is considered a random sample from an infinite population.

NOTES AND COMMENTS

1. In this section we have been careful to define two types of samples: a simple random sample from a finite population and a random sample from an infinite population. In the remainder of the text, we will generally refer to both of these as either a *random sample* or simply a *sample*. We will not make a distinction of the sample being a "simple" random sample unless it is necessary for the exercise or discussion.
2. Statisticians who specialize in sample surveys from finite populations use sampling methods that provide probability samples. With a probability sample, each possible sample has a known probability of selection and a random process is used to select the elements for the sample. Simple random sampling is one of these methods. In Section 7.8, we describe some other probability sampling methods: stratified random sampling, cluster sampling, and systematic sampling. We use the term "simple" in simple random sampling to clarify that this is the probability sampling method that assures each sample of size n has the same probability of being selected.
3. The number of different simple random samples of size n that can be selected from a finite population of size N is

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

In this formula, $N!$ and $n!$ are the factorial formulas 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 obtained.

Exercises

Methods

SELF test

1. Consider a finite population with five elements labeled A, B, C, D, and E. Ten possible simple random samples of size 2 can be selected.
   a. List the 10 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.
2. Assume a finite population has 350 elements. Using the last three digits of each of the following five-digit random numbers (e.g., 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

SELF test

3. *Fortune* publishes data on sales, profits, assets, stockholders' equity, market value, and earnings per share for the 500 largest U.S. industrial corporations (*Fortune* 500, 2012). Assume that you want to select a simple random sample of 10 corporations from the *Fortune* 500 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 10 corporations that would be selected.
4. The 10 most active stocks on the New York Stock Exchange on March 6, 2006, are shown here (*The Wall Street Journal,* March 7, 2006).

AT&T	Lucent	Nortel	Qwest	Bell South
Pfizer	Texas Instruments	Gen. Elect.	iShrMSJpn	LSI Logic

   Exchange authorities decided to investigate trading practices using a sample of three of these stocks.
   a. Beginning with the first random digit in column 6 of Table 7.1, read down the column to select a simple random sample of three stocks for the exchange authorities.
   b. Using the information in the third Note and Comment, determine how many different simple random samples of size 3 can be selected from the list of 10 stocks.
5. A student government organization is interested in estimating the proportion of students who favor a mandatory "pass-fail" grading policy for elective courses. A list of names and addresses of the 645 students enrolled during the current quarter 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 10 students who would be selected using simple random sampling. The three-digit random numbers begin with 816, 283, and 610.
6. The *County and City Data Book,* published by the Census Bureau, lists information on 3139 counties throughout the United States. Assume that a national study will collect data from 30 randomly selected counties. Use four-digit random numbers from the last column of Table 7.1 to identify the numbers corresponding to the first five counties selected for the sample. Ignore the first digits and begin with the four-digit random numbers 9945, 8364, 5702, and so on.

7. Assume that we want to identify a simple random sample of 12 of the 372 doctors practicing in a particular city. The doctors' names are available from a local medical organization. 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.

8. The following stocks make up the Dow Jones Industrial Average (*Barron's,* July 30, 2012).

1. 3M
2. AT&T
3. Alcoa
4. American Express
5. Bank of America
6. Boeing
7. Caterpillar
8. Chevron
9. Cisco Systems
10. Coca-Cola
11. Disney
12. DuPont
13. ExxonMobil
14. General Electric
15. Hewlett-Packard
16. Home Depot
17. IBM
18. Intel
19. Johnson & Johnson
20. Kraft Foods
21. McDonald's
22. Merck
23. Microsoft
24. J.P. Morgan
25. Pfizer
26. Procter & Gamble
27. Travelers
28. United Technologies
29. Verizon
30. Wal-Mart

Suppose you would like to select a sample of six of these companies to conduct an in-depth study of management practices. Use the first two digits in each row of the ninth column of Table 7.1 to select a simple random sample of six companies.

9. *The Wall Street Journal* provides the net asset value, the year-to-date percent return, and the three-year percent return for 555 mutual funds (*The Wall Street Journal,* April 25, 2003). Assume that a simple random sample of 12 of the 555 mutual funds will be selected for a follow-up study on the size and performance of mutual funds. Use the fourth column of the random numbers in Table 7.1, beginning with 51102, to select the simple random sample of 12 mutual funds. Begin with mutual fund 102 and use the *last* three digits in each row of the fourth column for your selection process. What are the numbers of the 12 mutual funds in the simple random sample?

10. Indicate which of the following situations involve sampling from a finite population and which involve sampling from an infinite population. In cases where the sampled population is finite, describe how you would construct a frame.
    a. Obtain a sample of licensed drivers in the state of New York.
    b. Obtain a sample of boxes of cereal produced by the Breakfast Choice company.
    c. Obtain a sample of cars crossing the Golden Gate Bridge on a typical weekday.
    d. Obtain a sample of students in a statistics course at Indiana University.
    e. Obtain a sample of the orders that are processed by a mail-order firm.

7.3 Point Estimation

Now that we have described how to select a simple random sample, let us return to the EAI problem. A simple random sample of 30 managers and the corresponding data on annual salary and management training program participation are as 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. Participation in the management training program is indicated by Yes in the management training program 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

TABLE 7.2 ANNUAL SALARY AND TRAINING PROGRAM STATUS FOR A SIMPLE RANDOM SAMPLE OF 30 EAI MANAGERS

Annual Salary ($)	Management Training Program	Annual Salary ($)	Management Training Program
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

sample mean and the sample standard deviation s. Using the formulas for a sample mean and a sample standard deviation presented in Chapter 3, the sample mean is

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

and the sample standard deviation is

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

To estimate p, the proportion of managers in the population who completed the management training program, we use the corresponding sample proportion $\bar{p}$. Let x denote the number of managers in the sample who completed the management training program. The data in Table 7.2 show that $x = 19$. Thus, with a sample size of $n = 30$, the sample proportion is

$$\bar{p} = \frac{x}{n} = \frac{19}{30} = .63$$

By making the preceding computations, we perform the statistical procedure called *point estimation.* We refer to the sample mean $\bar{x}$ as the **point estimator** of the population mean μ, the sample standard deviation s as the point estimator of the population standard deviation σ, and the sample proportion $\bar{p}$ as the point estimator of the population proportion p. The numerical value obtained for $\bar{x}$, s, or $\bar{p}$ is called the **point estimate**. Thus, 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 .63 is the point estimate of p. Table 7.3 summarizes the sample results and compares the point estimates to the actual values of the population parameters.

As is evident from Table 7.3, the point estimates differ somewhat from the corresponding population parameters. This difference is to be expected because a sample, and not a census of the entire population, is being used to develop the point estimates. In the next chapter, we will show how to construct an interval estimate in order to provide information about how close the point estimate is to the population parameter.

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
μ = Population mean annual salary	\$51,800	$\bar{x}$ = Sample mean annual salary	\$51,814
σ = Population standard deviation for annual salary	\$4000	s = Sample standard deviation for annual salary	\$3348
p = Population proportion having completed the management training program	.60	$\bar{p}$ = Sample proportion having completed the management training program	.63

Practical Advice

The subject matter of most of the rest of the book is concerned with 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 an amusement 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 single 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 we might encounter a significant difference between the sampled population and the target population. In such a case, we would question the validity of the point estimates being made. 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 judgment is a necessary ingredient of sound statistical practice.

Exercises

Methods

SELF test

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

5 8 10 7 10 14

a. What is the point estimate of the population mean?
b. What is the point estimate of the population standard deviation?

12. A survey question for a sample of 150 individuals yielded 75 Yes responses, 55 No responses, and 20 No Opinions.
a. What is the point estimate of the proportion in the population who respond Yes?
b. What is the point estimate of the proportion in the population who respond No?

Applications

SELF test

13. A sample of 5 months of sales data provided the following information:

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

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

WEB file

MutualFund

14. *BusinessWeek* published information on 283 equity mutual funds (*BusinessWeek,* January 26, 2004). A sample of 40 of those funds is contained in the data set MutualFund. Use the data set to answer the following questions.
   a. Develop a point estimate of the proportion of the *BusinessWeek* equity funds that are load funds.
   b. Develop a point estimate of the proportion of funds that are classified as high risk.
   c. Develop a point estimate of the proportion of funds that have a below-average risk rating.

15. Many drugs used to treat cancer are expensive. *BusinessWeek* reported on the cost per treatment of Herceptin, a drug used to treat breast cancer (*BusinessWeek,* January 30, 2006). Typical treatment costs (in dollars) for Herceptin are provided by a simple random sample of 10 patients.

4376	5578	2717	4920	4495
4798	6446	4119	4237	3814

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

16. A sample of 426 U.S. adults age 50 and older were asked how important a variety of issues were in choosing whom to vote for in the 2012 presidential election (*AARP Bulletin,* March 2012).
   a. What is the sampled population for this study?
   b. Social Security and Medicare was cited as "very important" by 350 respondents. Estimate the proportion of the population of U.S. adults age 50 and over who believe this issue is very important.
   c. Education was cited as "very important" by 74% of the respondents. Estimate the number of respondents who believe this issue is very important.
   d. Job Growth was cited as "very important" by 354 respondents. Estimate the proportion of U.S. adults age 50 and over who believe job growth is very important.
   e. What is the target population for the inferences being made in parts (b) and (d)? Is it the same as the sampled population you identified in part (a)? Suppose you later learn that the sample was restricted to members of the American Association of Retired People (AARP). Would you still feel the inferences being made in parts (b) and (d) are valid? Why or why not?

17. The American Association of Individual Investors (AAII) polls its subscribers on a weekly basis to determine the number who are bullish, bearish, or neutral on the short-term prospects for the stock market. Their findings for the week ending March 2, 2006, are consistent with the following sample results (AAII website, March 7, 2006).

Bullish	409	Neutral	299	Bearish	291

   Develop a point estimate of the following population parameters.
   a. The proportion of all AAII subscribers who are bullish on the stock market.
   b. The proportion of all AAII subscribers who are neutral on the stock market.
   c. The proportion of all AAII subscribers who are bearish on the stock market.
   d. What is the sampled population? What is the target population for parts (a), (b), and (c)? Would you be comfortable extending these results to the target population of all investors?

7.4 Introduction to Sampling Distributions

In the preceding section we said that the sample mean $\bar{x}$ is the point estimator of the population mean μ, and the sample proportion $\bar{p}$ is the point estimator of the population proportion p. For the simple random sample of 30 EAI managers shown in Table 7.2, the point estimate of μ is $\bar{x} = \$51{,}814$ and the point estimate of p is $\bar{p} = .63$. Suppose we select another simple random sample of 30 EAI managers and obtain the following point estimates:

$$\text{Sample mean: } \bar{x} = \$52{,}670$$
$$\text{Sample proportion: } \bar{p} = .70$$

Note that different values of $\bar{x}$ and $\bar{p}$ were obtained. Indeed, a second simple random sample of 30 EAI managers cannot be expected to provide the same point estimates as the first sample.

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.

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 $\bar{x}$ and $\bar{p}$. Table 7.4 contains a portion of the results obtained for 500 simple random samples, and Table 7.5 shows the frequency and relative frequency distributions for the 500 $\bar{x}$ values. Figure 7.1 shows the relative frequency histogram for the $\bar{x}$ values.

In Chapter 5 we defined a random variable as a numerical description of the outcome of an experiment. If we consider the process of selecting a simple random sample as an

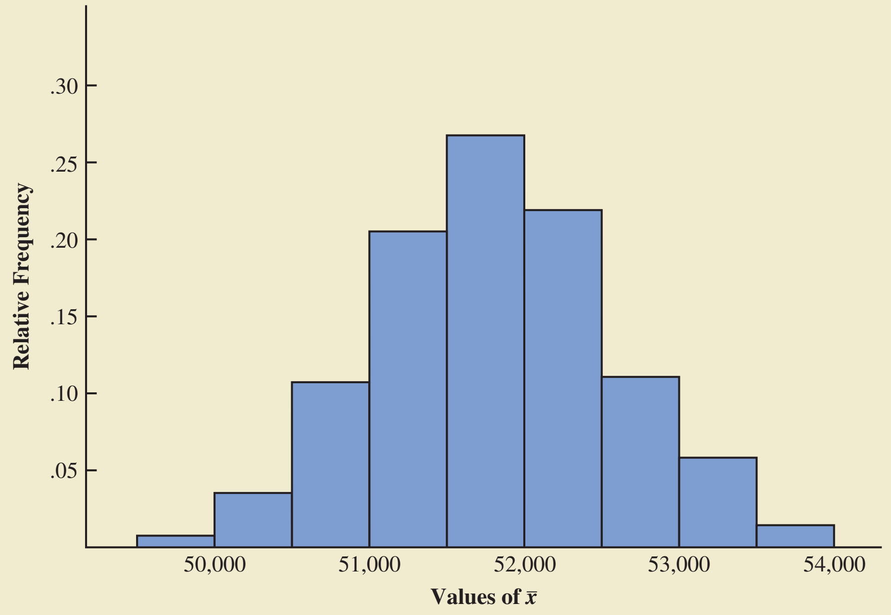
TABLE 7.4 VALUES OF $\bar{x}$ AND $\bar{p}$ FROM 500 SIMPLE RANDOM SAMPLES OF 30 EAI MANAGERS

Sample Number	Sample Mean ($\bar{x}$)	Sample Proportion ($\bar{p}$)
1	51,814	.63
2	52,670	.70
3	51,780	.67
4	51,588	.53
.	.	.
.	.	.
.	.	.
500	51,752	.50

TABLE 7.5 FREQUENCY AND RELATIVE FREQUENCY DISTRIBUTIONS OF $\bar{x}$ FROM 500 SIMPLE RANDOM SAMPLES OF 30 EAI MANAGERS

Mean Annual Salary ($)	Frequency	Relative Frequency
49,500.00–49,999.99	2	.004
50,000.00–50,499.99	16	.032
50,500.00–50,999.99	52	.104
51,000.00–51,499.99	101	.202
51,500.00–51,999.99	133	.266
52,000.00–52,499.99	110	.220
52,500.00–52,999.99	54	.108
53,000.00–53,499.99	26	.052
53,500.00–53,999.99	6	.012
Totals	500	1.000

FIGURE 7.1 RELATIVE FREQUENCY HISTOGRAM OF $\bar{x}$ VALUES FROM 500 SIMPLE RANDOM SAMPLES OF SIZE 30 EACH



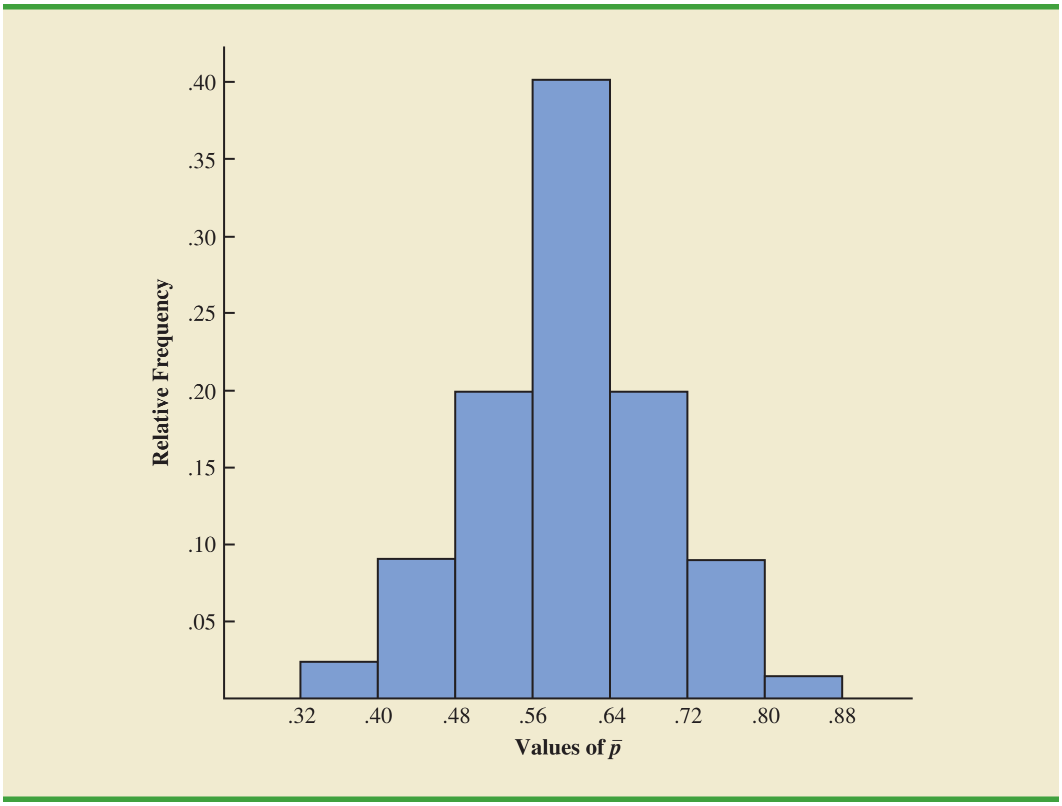
experiment, the sample mean $\bar{x}$ is the numerical description of the outcome of the experiment. Thus, the sample mean $\bar{x}$ is a random variable. As a result, just like other random variables, $\bar{x}$ has a mean or expected value, a standard deviation, and a probability distribution. Because the various possible values of $\bar{x}$ are the result of different simple random samples, the probability distribution of $\bar{x}$ is called the **sampling distribution** of $\bar{x}$. Knowledge of this sampling distribution and its properties will enable us to make probability statements about how close the sample mean $\bar{x}$ 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 $\bar{x}$. 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 is near the population mean $\mu = \$51{,}800$. We will describe the properties of the sampling distribution of $\bar{x}$ more fully in the next section.

The 500 values of the sample proportion $\bar{p}$ are summarized by the relative frequency histogram in Figure 7.2. As in the case of $\bar{x}$, $\bar{p}$ is a random variable. If every possible sample of size 30 were selected from the population and if a value of $\bar{p}$ were computed for each sample, the resulting probability distribution would be the sampling distribution of $\bar{p}$. 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 $\bar{p}$.

In practice, we select only one simple random sample from the population. 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 for the sample statistics $\bar{x}$ and $\bar{p}$. 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 $\bar{x}$. In Section 7.6 we show the characteristics of the sampling distribution of $\bar{p}$.

FIGURE 7.2 RELATIVE FREQUENCY HISTOGRAM OF $\bar{p}$ VALUES FROM 500 SIMPLE RANDOM SAMPLES OF SIZE 30 EACH



7.5 Sampling Distribution of $\bar{x}$

In the previous section we said that the sample mean $\bar{x}$ is a random variable and its probability distribution is called the sampling distribution of $\bar{x}$.

SAMPLING DISTRIBUTION OF $\bar{x}$

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

This section describes the properties of the sampling distribution of $\bar{x}$. Just as with other probability distributions we studied, the sampling distribution of $\bar{x}$ has an expected value or mean, a standard deviation, and a characteristic shape or form. Let us begin by considering the mean of all possible $\bar{x}$ values, which is referred to as the expected value of $\bar{x}$.

Expected Value of $\bar{x}$

In the EAI sampling problem we saw that different simple random samples result in a variety of values for the sample mean $\bar{x}$. Because many different values of the random variable $\bar{x}$ are possible, we are often interested in the mean of all possible values of $\bar{x}$ that can be generated by the various simple random samples. The mean of the $\bar{x}$ random variable is the expected value of $\bar{x}$. Let $E(\bar{x})$ represent the expected value of $\bar{x}$ 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, $E(\bar{x})$ and μ are equal.

The expected value of $\bar{x}$ equals the mean of the population from which the sample is selected.

EXPECTED VALUE OF $\bar{x}$

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

where

$$\begin{aligned} E(\bar{x}) &= \text{the expected value of } \bar{x} \\ \mu &= \text{the population mean} \end{aligned}$$

This result shows that with simple random sampling, the expected value or mean of the sampling distribution of $\bar{x}$ is equal to the mean of the population. In Section 7.1 we saw that the mean annual salary for the population of EAI managers is $\mu = \$51,800$. Thus, 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 **unbiased**. Thus, equation (7.1) shows that $\bar{x}$ is an unbiased estimator of the population mean μ.

Standard Deviation of $\bar{x}$

Let us define the standard deviation of the sampling distribution of $\bar{x}$. We will use the following notation.

$$\begin{aligned} \sigma_{\bar{x}} &= \text{the standard deviation of } \bar{x} \\ \sigma &= \text{the standard deviation of the population} \\ n &= \text{the sample size} \\ N &= \text{the population size} \end{aligned}$$

It can be shown that the formula for the standard deviation of $\bar{x}$ depends on whether the population is finite or infinite. The two formulas for the standard deviation of $\bar{x}$ follow.

STANDARD DEVIATION OF $\bar{x}$

Finite Population	*Infinite Population*	
$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}\left(\frac{\sigma}{\sqrt{n}}\right)$	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$	**(7.2)**

In comparing the two formulas 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 $\sqrt{(N-n)/(N-1)}$ is close to 1. As a result, the difference between the values of the standard deviation of $\bar{x}$ for the finite and infinite population cases becomes negligible. Then, $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ becomes a good approximation to the standard deviation of $\bar{x}$ even though the population is finite. This observation leads to the following general guideline, or rule of thumb, for computing the standard deviation of $\bar{x}$.

USE THE FOLLOWING EXPRESSION TO COMPUTE THE STANDARD DEVIATION OF $\bar{x}$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{7.3}$$

whenever

1. The population is infinite; or
2. The population is finite *and* the sample size is less than or equal to 5% of the population size; that is, $n/N \le .05$.

Problem 21 shows that when $n/N \le .05$, the finite population correction factor has little effect on the value of $\sigma_{\bar{x}}$.

The term standard error is used throughout statistical inference to refer to the standard deviation of a point estimator.

In cases where $n/N > .05$, the finite population version of formula (7.2) should be used in the computation of $\sigma_{\bar{x}}$. Unless otherwise noted, throughout the text we will assume that the population size is "large," $n/N \le .05$, and expression (7.3) can be used to compute $\sigma_{\bar{x}}$.

To compute $\sigma_{\bar{x}}$, we need to know σ, the standard deviation of the population. To further emphasize the difference between $\sigma_{\bar{x}}$ and σ, we refer to the standard deviation of $\bar{x}$, $\sigma_{\bar{x}}$, as the **standard error** of the mean. In general, the term *standard error* refers to the standard deviation of a point estimator. Later we will 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. Let us now 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 $n/N = 30/2500 = .012$. Because the sample size is less than 5% of the population size, we can ignore the finite population correction factor and use equation (7.3) to compute the standard error.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4000}{\sqrt{30}} = 730.3$$

Form of the Sampling Distribution of $\bar{x}$

The preceding results concerning the expected value and standard deviation for the sampling distribution of $\bar{x}$ are applicable for any population. The final step in identifying the characteristics of the sampling distribution of $\bar{x}$ is to determine the form or shape of the sampling distribution. We will 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 selecting a random sample has a normal, or nearly normal, distribution. When the population has a normal distribution, the sampling distribution of $\bar{x}$ is normally distributed for any sample size.

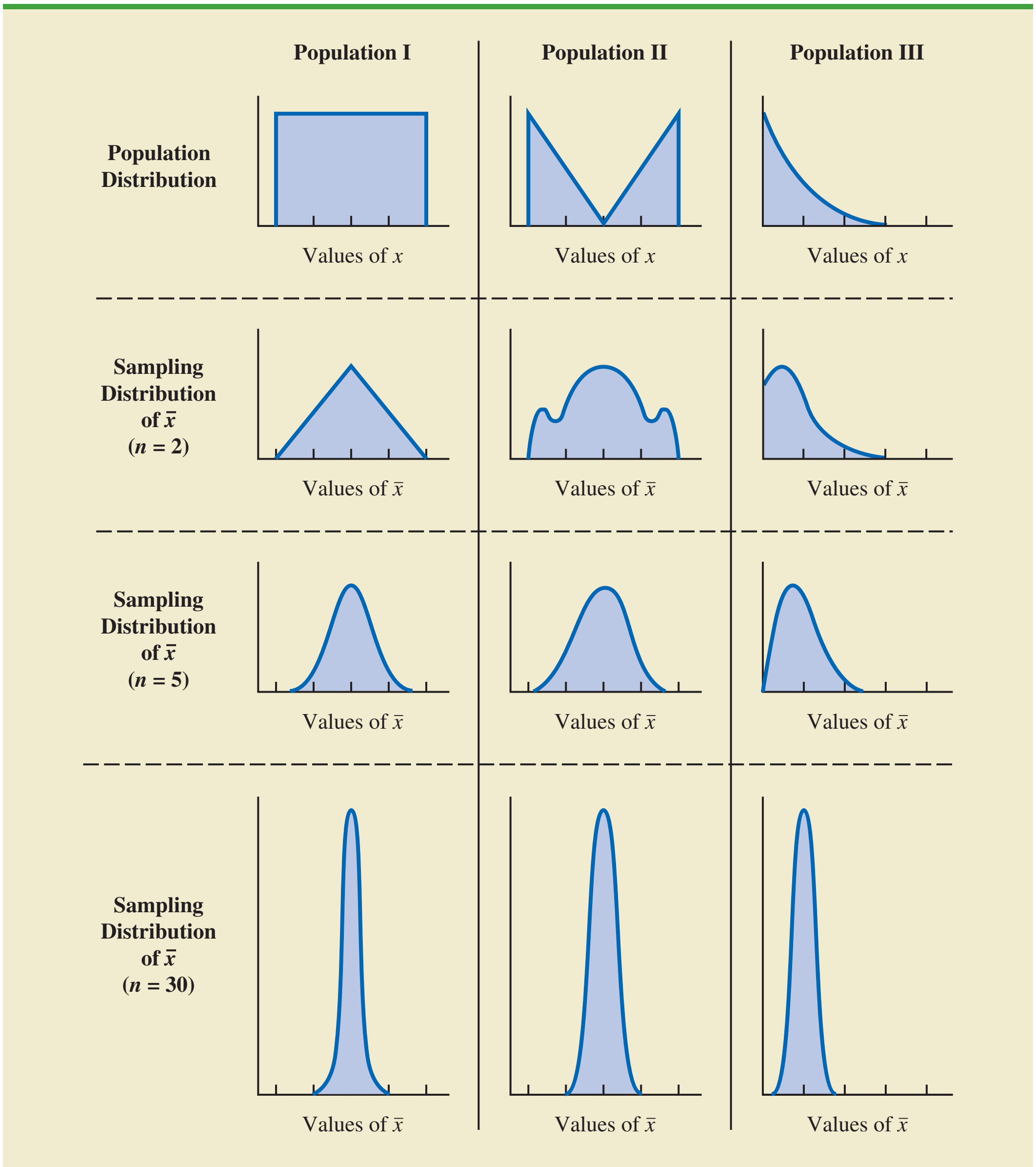
Population does not have a normal distribution. When the population from which we are selecting a random sample does not have a normal distribution, the **central limit theorem** is helpful in identifying the shape of the sampling distribution of $\bar{x}$. A statement of the central limit theorem as it applies to the sampling distribution of $\bar{x}$ follows.

CENTRAL LIMIT THEOREM

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

Figure 7.3 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 are normally distributed. Population I follows a uniform distribution. Population II is often called the rabbit-eared distribution. It is symmetric, but the more likely

FIGURE 7.3 ILLUSTRATION OF THE CENTRAL LIMIT THEOREM FOR THREE POPULATIONS



values fall in the tails of the distribution. Population III is shaped like the exponential distribution; it is skewed to the right.

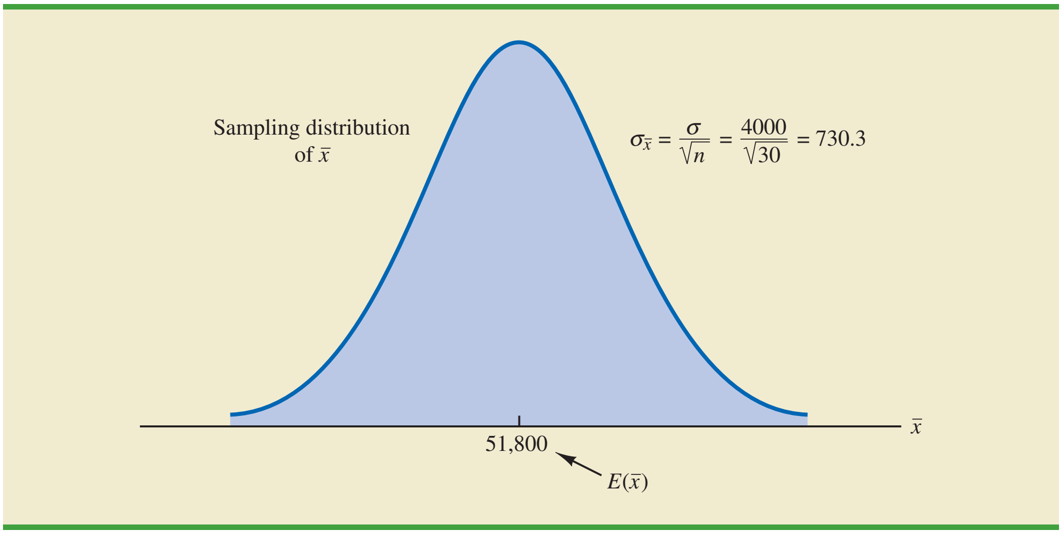
The bottom three panels of Figure 7.3 show the shape of the sampling distribution for samples of size $n = 2$, $n = 5$, and $n = 30$. When the sample size is 2, we see that the shape of each sampling distribution is different from the shape of the corresponding population distribution. For samples of size 5, we see that the shapes of the sampling distributions for populations I and II begin to look similar to the shape of a normal distribution. Even though the shape of the sampling distribution for population III begins to look similar to the shape of a normal distribution, some skewness to the right is still present. Finally, for samples of size 30, the shapes of each of the three sampling distributions are approximately normal.

From a practitioner standpoint, we often want to know how large the sample size needs to be before the central limit theorem applies and we can assume that the shape of the sampling distribution is approximately normal. Statistical researchers have investigated this question by studying the sampling distribution of $\bar{x}$ for a variety of populations and a variety of sample sizes. General statistical practice is to assume that, for most applications, the sampling distribution of $\bar{x}$ can be approximated by a normal distribution whenever the sample is size 30 or more. In cases where the population is highly skewed or outliers are present, samples of size 50 may be needed. Finally, if the population is discrete, the sample size needed for a normal approximation often depends on the population proportion. We say more about this issue when we discuss the sampling distribution of $\bar{p}$ in Section 7.6.

Sampling Distribution of $\bar{x}$ for the EAI Problem

Let us return to the EAI problem where we previously showed that $E(\bar{x}) = \$51{,}800$ and $\sigma_{\bar{x}} = 730.3$. 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 $\bar{x}$ is normally distributed. 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 of $\bar{x}$ can be approximated by a normal distribution. In either case, we are comfortable proceeding with the conclusion that the sampling distribution of $\bar{x}$ can be described by the normal distribution shown in Figure 7.4.

FIGURE 7.4 SAMPLING DISTRIBUTION OF $\bar{x}$ FOR THE MEAN ANNUAL SALARY OF A SIMPLE RANDOM SAMPLE OF 30 EAI MANAGERS



Practical Value of the Sampling Distribution of $\bar{x}$

Whenever a simple random sample is selected and the value of the sample mean is used to estimate the value of the population mean μ, we cannot expect the sample mean to exactly equal the population mean. The practical reason we are interested in the sampling distribution of $\bar{x}$ is that it can be used to provide probability information about the difference between the sample mean and the population mean. To demonstrate this use, let us return to the EAI problem.

Suppose the personnel director believes the sample mean will be an acceptable estimate of the population mean if the sample mean is within \$500 of the population mean. However, 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 personnel director's request in probability terms. That is, the personnel director is concerned with the following question: 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?

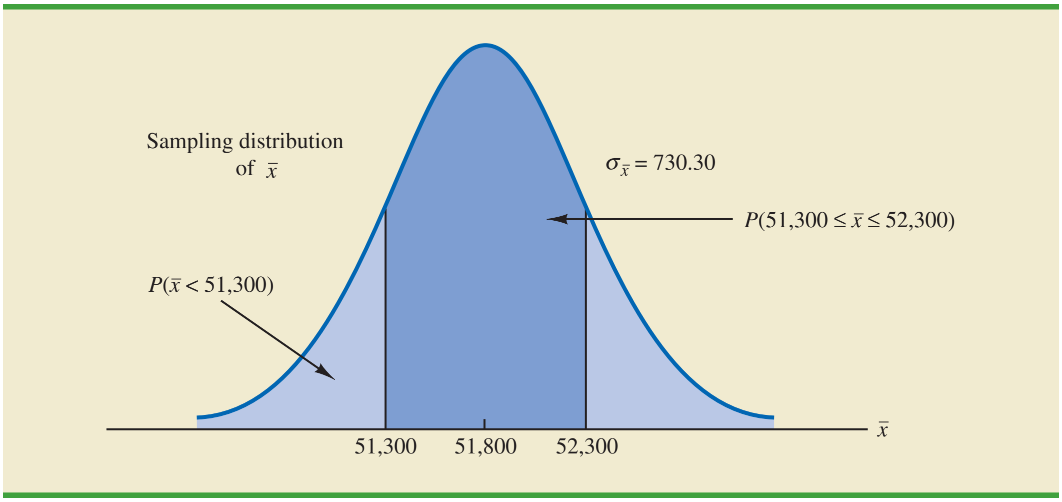
Because we have identified the properties of the sampling distribution of $\bar{x}$ (see Figure 7.4), we will use this distribution to answer the probability question. Refer to the sampling distribution of $\bar{x}$ shown again in Figure 7.5. With a population mean of \$51,800, the personnel director wants to know the probability that $\bar{x}$ is between \$51,300 and \$52,300. This probability is given by the darkly shaded area of the sampling distribution shown in Figure 7.5. Because the sampling distribution is normally distributed, with mean 51,800 and standard error of the mean 730.3, we can use the standard normal probability table to find the area or probability.

We first calculate the z value at the upper endpoint of the interval (52,300) and use the table to find the area under the curve to the left of that point (left tail area). Then we compute the z value at the lower endpoint of the interval (51,300) and use the table to find the area under the curve to the left of that point (another left tail area). Subtracting the second tail area from the first gives us the desired probability.

At $\bar{x} = 52{,}300$, we have

$$z = \frac{52{,}300 - 51{,}800}{730.30} = .68$$

FIGURE 7.5 PROBABILITY OF A SAMPLE MEAN BEING WITHIN \$500 OF THE POPULATION MEAN FOR A SIMPLE RANDOM SAMPLE OF 30 EAI MANAGERS



Referring to the standard normal probability table, we find a cumulative probability (area to the left of $z = .68$) of .7517.

At $\bar{x} = 51{,}300$, we have

$$z = \frac{51{,}300 - 51{,}800}{730.30} = -.68$$

The area under the curve to the left of $z = -.68$ is .2483. Therefore, $P(51{,}300 \leq \bar{x} \leq 52{,}300) = P(z \leq .68) - P(z < -.68) = .7517 - .2483 = .5034$.

The sampling distribution of $\bar{x}$ can be used to provide probability information about how close the sample mean $\bar{x}$ is to the population mean μ.

The preceding computations show that a simple random sample of 30 EAI managers has a .5034 probability of providing a sample mean $\bar{x}$ that is within \$500 of the population mean. Thus, there is a $1 - .5034 = .4966$ probability that the difference between $\bar{x}$ and $\mu = \$51{,}800$ will be more than \$500. In other words, a simple random sample of 30 EAI managers has roughly a 50–50 chance of providing a sample mean within the allowable \$500. Perhaps a larger sample size should be considered. Let us explore this possibility by considering the relationship between the sample size and the sampling distribution of $\bar{x}$.

Relationship Between the Sample Size and the Sampling Distribution of $\bar{x}$

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 data provided by the larger sample size, 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 $\bar{x}$.

First note that $E(\bar{x}) = \mu$ regardless of the sample size. Thus, the mean of all possible values of $\bar{x}$ is equal to the population mean μ regardless of the sample size n. However, note that the standard error of the mean, $\sigma_{\bar{x}} = \sigma/\sqrt{n}$, is related to the square root of the sample size. Whenever the sample size is increased, the standard error of the mean $\sigma_{\bar{x}}$ decreases. With $n = 30$, the standard error of the mean for the EAI problem is 730.3. However, with the increase in the sample size to $n = 100$, the standard error of the mean is decreased to

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4000}{\sqrt{100}} = 400$$

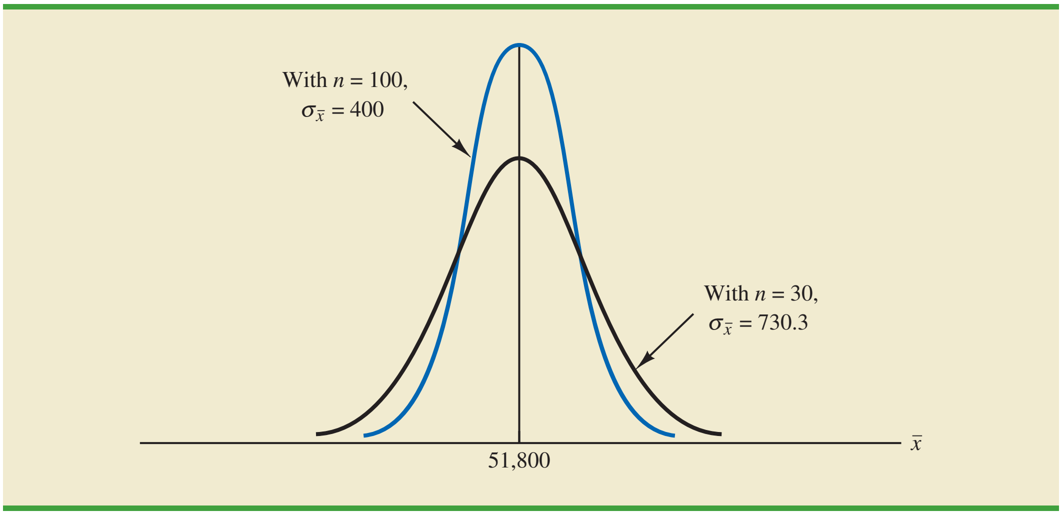
The sampling distributions of $\bar{x}$ with $n = 30$ and $n = 100$ are shown in Figure 7.6. Because the sampling distribution with $n = 100$ has a smaller standard error, the values of $\bar{x}$ have less variation and tend to be closer to the population mean than the values of $\bar{x}$ with $n = 30$.

We can use the sampling distribution of $\bar{x}$ for the case with $n = 100$ to compute the probability that a simple random sample of 100 EAI managers will provide a sample mean that is 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 probability table to find the area or probability.

At $\bar{x} = 52{,}300$ (see Figure 7.7), we have

$$z = \frac{52{,}300 - 51{,}800}{400} = 1.25$$

FIGURE 7.6 A COMPARISON OF THE SAMPLING DISTRIBUTIONS OF $\bar{x}$ FOR SIMPLE RANDOM SAMPLES OF $n = 30$ AND $n = 100$ EAI MANAGERS



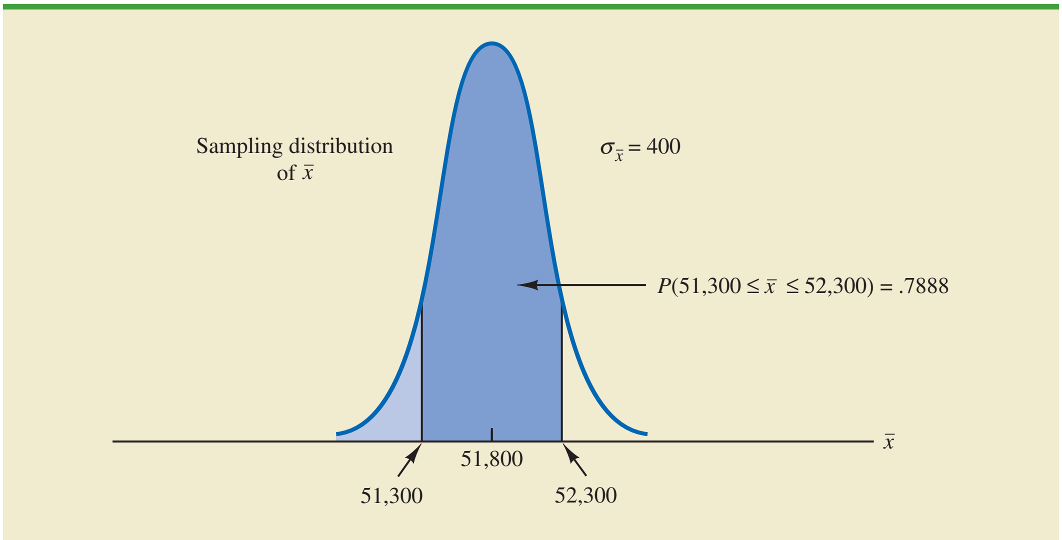
Referring to the standard normal probability table, we find a cumulative probability corresponding to $z = 1.25$ of .8944.

At $\bar{x} = 51{,}300$, we have

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

The cumulative probability corresponding to $z = -1.25$ is .1056. Therefore, $P(51{,}300 \leq \bar{x} \leq 52{,}300) = P(z \leq 1.25) - P(z \leq -1.25) = .8944 - .1056 = .7888$. Thus, by increasing the sample size from 30 to 100 EAI managers, we increase the probability of obtaining a sample mean within $500 of the population mean from .5034 to .7888.

FIGURE 7.7 PROBABILITY OF A SAMPLE MEAN BEING WITHIN $500 OF THE POPULATION MEAN FOR A SIMPLE RANDOM SAMPLE OF 100 EAI MANAGERS



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.

NOTES AND COMMENTS

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

2. 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.

Exercises

Methods

18. A population has a mean of 200 and a standard deviation of 50. A sample of size 100 will be taken and the sample mean $\bar{x}$ will be used to estimate the population mean.
    a. What is the expected value of $\bar{x}$?
    b. What is the standard deviation of $\bar{x}$?
    c. Show the sampling distribution of $\bar{x}$.
    d. What does the sampling distribution of $\bar{x}$ show?

SELF test

19. A population has a mean of 200 and a standard deviation of 50. Suppose a sample of size 100 is selected and $\bar{x}$ 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?

20. Assume the population standard deviation is $\sigma = 25$. Compute the standard error of the mean, $\sigma_{\bar{x}}$, 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?

21. Suppose a random sample of size 50 is selected from a population with $\sigma = 10$. 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

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

SELF test

23. In the EAI sampling problem (see Figure 7.5), we showed that for $n = 30$, there was .5034 probability of obtaining a sample mean within ±\$500 of the population mean.
    a. What is the probability that $\bar{x}$ 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.

24. *Barron's* reported that the average number of weeks an individual is unemployed is 17.5 weeks (*Barron's,* February 18, 2008). Assume that for the population of all unemployed individuals the population mean length of unemployment is 17.5 weeks and that the population standard deviation is 4 weeks. Suppose you would like to select a sample of 50 unemployed individuals for a follow-up study.
    a. Show the sampling distribution of $\bar{x}$, the sample mean average for a sample of 50 unemployed individuals.
    b. What is the probability that a simple random sample of 50 unemployed individuals will provide a sample mean within 1 week of the population mean?
    c. What is the probability that a simple random sample of 50 unemployed individuals will provide a sample mean within 1/2 week of the population mean?

25. The College Board reported the following mean scores for the three parts of the Scholastic Aptitude Test (SAT) (*The World Almanac,* 2009):

Critical Reading	502
Mathematics	515
Writing	494

Assume that the population standard deviation on each part of the test is $\sigma = 100$.

    a. What is the probability a sample of 90 test takers will provide a sample mean test score within 10 points of the population mean of 502 on the Critical Reading part of the test?
    b. What is the probability a sample of 90 test takers will provide a sample mean test score within 10 points of the population mean of 515 on the Mathematics part of the test? Compare this probability to the value computed in part (a).
    c. What is the probability a sample of 100 test takers will provide a sample mean test score within 10 of the population mean of 494 on the writing part of the test? Comment on the differences between this probability and the values computed in parts (a) and (b).

26. The mean annual cost of automobile insurance is \$939 (CNBC, February 23, 2006). Assume that the standard deviation is $\sigma = \$245$.
    a. What is the probability that a sample of automobile insurance policies will have a sample mean within \$25 of the population mean for each of the following sample sizes: 30, 50, 100, and 400?
    b. What is the advantage of a larger sample size when attempting to estimate the population mean?

27. The Economic Policy Institute periodically issues reports on wages of entry level workers. The institute reported that entry level wages for male college graduates were \$21.68 per hour and for female college graduates were \$18.80 per hour in 2011 (Economic Policy Institute website, March 30, 2012). Assume the standard deviation for male graduates is \$2.30, and for female graduates it is \$2.05.
    a. What is the probability that a sample of 50 male graduates will provide a sample mean within \$.50 of the population mean, \$21.68?
    b. What is the probability that a sample of 50 female graduates will provide a sample mean within \$.50 of the population mean, \$18.80?
    c. In which of the preceding two cases, part (a) or part (b), do we have a higher probability of obtaining a sample estimate within \$.50 of the population mean? Why?
    d. What is the probability that a sample of 120 female graduates will provide a sample mean more than \$.30 below the population mean?

28. The average score for male golfers is 95 and the average score for female golfers is 106 (*Golf Digest,* April 2006). Use these values as the population means for men and women and assume that the population standard deviation is $\sigma = 14$ strokes for both. A sample of 30 male golfers and another sample of 45 female golfers will be taken.
    a. Show the sampling distribution of $\bar{x}$ for male golfers.
    b. What is the probability that the sample mean is within 3 strokes of the population mean for the sample of male golfers?
    c. What is the probability that the sample mean is within 3 strokes of the population mean for the sample of female golfers?
    d. In which case, part (b) or part (c), is the probability of obtaining a sample mean within 3 strokes of the population mean higher? Why?

29. The mean preparation fee H&R Block charged retail customers last year was $183 (*The Wall Street Journal,* March 7, 2012). Use this price as the population mean and assume the population standard deviation of preparation fees is $50.
    a. What is the probability that the mean price for a sample of 30 H&R Block retail customers is within $8 of the population mean?
    b. What is the probability that the mean price for a sample of 50 H&R Block retail customers is within $8 of the population mean?
    c. What is the probability that the mean price for a sample of 100 H&R Block retail customers is within $8 of the population mean?
    d. Which, if any, of the sample sizes in parts (a), (b), and (c) would you recommend to have at least a .95 probability that the sample mean is within $8 of the population mean?

30. To estimate the mean age for a population of 4000 employees, 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 $\sigma = 8.2$ years, 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 $n/N \leq .05$?
    c. What is the probability that the sample mean age of the employees will be within ± 2 years of the population mean age?

7.6 Sampling Distribution of $\bar{p}$

The sample proportion $\bar{p}$ is the point estimator of the population proportion p. The formula for computing the sample proportion is

$$\bar{p} = \frac{x}{n}$$

where

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

As noted in Section 7.4, the sample proportion $\bar{p}$ is a random variable and its probability distribution is called the sampling distribution of $\bar{p}$.

SAMPLING DISTRIBUTION OF $\bar{p}$

The sampling distribution of $\bar{p}$ is the probability distribution of all possible values of the sample proportion $\bar{p}$.

To determine how close the sample proportion $\bar{p}$ is to the population proportion p, we need to understand the properties of the sampling distribution of $\bar{p}$: the expected value of $\bar{p}$, the standard deviation of $\bar{p}$, and the shape or form of the sampling distribution of $\bar{p}$.

Expected Value of $\bar{p}$

The expected value of $\bar{p}$, the mean of all possible values of $\bar{p}$, is equal to the population proportion p.

EXPECTED VALUE OF $\bar{p}$

$$E(\bar{p}) = p \qquad \textbf{(7.4)}$$

where

$$E(\bar{p}) = \text{the expected value of } \bar{p}$$
$$p = \text{the population proportion}$$

Because $E(\bar{p}) = p$, $\bar{p}$ is an unbiased estimator of p. Recall from Section 7.1 we noted that $p = .60$ for the EAI population, where p is the proportion of the population of managers who participated in the company's management training program. Thus, the expected value of $\bar{p}$ for the EAI sampling problem is .60.

Standard Deviation of $\bar{p}$

Just as we found for the standard deviation of $\bar{x}$, the standard deviation of $\bar{p}$ depends on whether the population is finite or infinite. The two formulas for computing the standard deviation of $\bar{p}$ follow.

STANDARD DEVIATION OF $\bar{p}$

Finite Population *Infinite Population*

$$\sigma_{\bar{p}} = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{p(1-p)}{n}} \qquad \sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} \qquad \textbf{(7.5)}$$

Comparing the two formulas 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 $\bar{x}$, 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 $n/N \leq .05$, we will use $\sigma_{\bar{p}} = \sqrt{p(1-p)/n}$. However, if the population is finite with $n/N > .05$, the finite population correction factor should be used. Again, unless specifically noted, throughout the text we will assume that the population size is large in relation to the sample size and thus 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 $\bar{x}$. We stated that in general the term standard error refers to the standard deviation of a point estimator. Thus, for proportions we use *standard error of the proportion* to refer to the standard deviation of $\bar{p}$. 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 program is $p = .60$. With $n/N = 30/2500 = .012$, 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, $\sigma_{\bar{p}}$ is

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{.60(1-.60)}{30}} = .0894$$

Form of the Sampling Distribution of $\bar{p}$

Now that we know the mean and standard deviation of the sampling distribution of $\bar{p}$, the final step is to determine the form or shape of the sampling distribution. The sample proportion is $\bar{p} = x/n$. For a simple random sample from a large population, the value of x 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 x/n is the same as the binomial probability of x, which means that the sampling distribution of $\bar{p}$ is also a discrete probability distribution and that the probability for each value of x/n is the same as the probability of x.

In Chapter 6 we also 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:

$$np \geq 5 \quad \text{and} \quad n(1-p) \geq 5$$

Assuming these two conditions are satisfied, the probability distribution of x in the sample proportion, $\bar{p} = x/n$, can be approximated by a normal distribution. And because n is a constant, the sampling distribution of $\bar{p}$ can also be approximated by a normal distribution. This approximation is stated as follows:

> The sampling distribution of $\bar{p}$ can be approximated by a normal distribution whenever $np \geq 5$ and $n(1-p) \geq 5$.

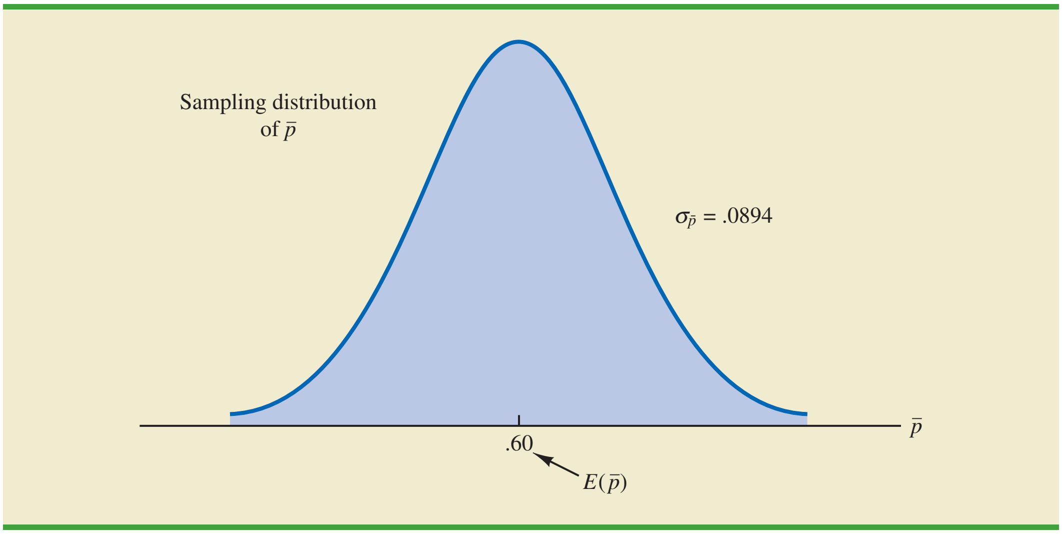
In practical applications, when an estimate of a population proportion is desired, we find that sample sizes are almost always large enough to permit the use of a normal approximation for the sampling distribution of $\bar{p}$.

Recall that for the EAI sampling problem we know that the population proportion of managers who participated in the training program is $p = .60$. With a simple random sample of size 30, we have $np = 30(.60) = 18$ and $n(1-p) = 30(.40) = 12$. Thus, the sampling distribution of $\bar{p}$ can be approximated by a normal distribution shown in Figure 7.8.

Practical Value of the Sampling Distribution of $\bar{p}$

The practical value of the sampling distribution of $\bar{p}$ 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 personnel director wants to know the probability of obtaining a value of $\bar{p}$ that is within .05 of the population proportion of EAI managers who participated in the training program. That is, what is the probability of obtaining a sample with a sample proportion $\bar{p}$ between .55 and .65? The darkly shaded

FIGURE 7.8 SAMPLING DISTRIBUTION OF $\bar{p}$ FOR THE PROPORTION OF EAI MANAGERS WHO PARTICIPATED IN THE MANAGEMENT TRAINING PROGRAM

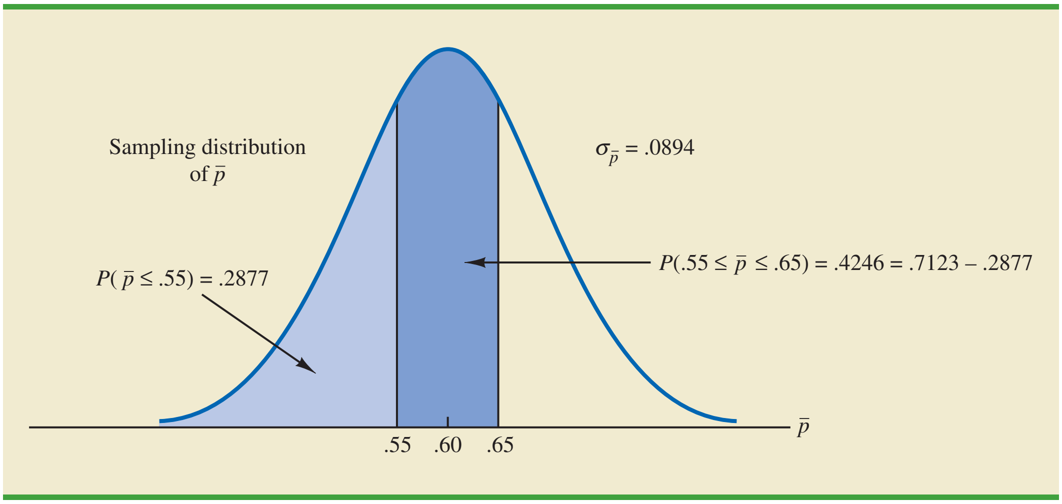


area in Figure 7.9 shows this probability. Using the fact that the sampling distribution of $\bar{p}$ can be approximated by a normal distribution with a mean of .60 and a standard error of the proportion of $\sigma_{\bar{p}} = .0894$, we find that the standard normal random variable corresponding to $\bar{p} = .65$ has a value of $z = (.65 - .60)/.0894 = .56$. Referring to the standard normal probability table, we see that the cumulative probability corresponding to $z = .56$ is .7123. Similarly, at $\bar{p} = .55$, we find $z = (.55 - .60)/.0894 = -.56$. From the standard normal probability table, we find the cumulative probability corresponding to $z = -.56$ is .2877. Thus, the probability of selecting a sample that provides a sample proportion $\bar{p}$ within .05 of the population proportion p is given by $.7123 - .2877 = .4246$.

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

$$\sigma_{\bar{p}} = \sqrt{\frac{.60(1 - .60)}{100}} = .049$$

FIGURE 7.9 PROBABILITY OF OBTAINING $\bar{p}$ BETWEEN .55 AND .65



With a sample size of 100 EAI managers, the probability of the sample proportion having a value within .05 of the population proportion can now be computed. Because the sampling distribution is approximately normal, with mean .60 and standard deviation .049, we can use the standard normal probability table to find the area or probability. At $\bar{p} = .65$, we have $z = (.65 - .60)/.049 = 1.02$. Referring to the standard normal probability table, we see that the cumulative probability corresponding to $z = 1.02$ is .8461. Similarly, at $\bar{p} = .55$, we have $z = (.55 - .60)/.049 = -1.02$. We find the cumulative probability corresponding to $z = -1.02$ is .1539. Thus, if the sample size is increased from 30 to 100, the probability that the sample proportion $\bar{p}$ is within .05 of the population proportion p will increase to $.8461 - .1539 = .6922$.

Exercises

Methods

31. A sample of size 100 is selected from a population with $p = .40$.
   a. What is the expected value of $\bar{p}$?
   b. What is the standard error of $\bar{p}$?
   c. Show the sampling distribution of $\bar{p}$.
   d. What does the sampling distribution of $\bar{p}$ show?

SELF test

32. A population proportion is .40. A sample of size 200 will be taken and the sample proportion $\bar{p}$ will be used to estimate the population proportion.
   a. What is the probability that the sample proportion will be within ±.03 of the population proportion?
   b. What is the probability that the sample proportion will be within ±.05 of the population proportion?

33. Assume that the population proportion is .55. Compute the standard error of the proportion, $\sigma_{\bar{p}}$, 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?

34. The population proportion is .30. What is the probability that a sample proportion will be within ±.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

SELF test

35. The president of Doerman Distributors, Inc., believes that 30% of the firm's orders come from first-time customers. A random sample of 100 orders will be used to estimate the proportion of first-time customers.
   a. Assume that the president is correct and $p = .30$. What is the sampling distribution of $\bar{p}$ for this study?
   b. What is the probability that the sample proportion $\bar{p}$ will be between .20 and .40?
   c. What is the probability that the sample proportion will be between .25 and .35?

36. *The Wall Street Journal* reported that the age at first startup for 55% of entrepreneurs was 29 years of age or less and the age at first startup for 45% of entrepreneurs was 30 years of age or more (*The Wall Street Journal,* March 19, 2012).
   a. Suppose a sample of 200 entrepreneurs will be taken to learn about the most important qualities of entrepreneurs. Show the sampling distribution of $\bar{p}$ where $\bar{p}$ is the sample proportion of entrepreneurs whose first startup was at 29 years of age or less.

b. What is the probability that the sample proportion in part (a) will be within ±.05 of its population proportion?

c. Suppose a sample of 200 entrepreneurs will be taken to learn about the most important qualities of entrepreneurs. Show the sampling distribution of $\bar{p}$ where $\bar{p}$ is now the sample proportion of entrepreneurs whose first startup was at 30 years of age or more.

d. What is the probability that the sample proportion in part (c) will be within ±.05 of its population proportion?

e Is the probability different in parts (b) and (d)? Why?

f. Answer part (b) for a sample of size 400. Is the probability smaller? Why?

37. People end up tossing 12% of what they buy at the grocery store (*Reader's Digest,* March 2009). Assume this is the true population proportion and that you plan to take a sample survey of 540 grocery shoppers to further investigate their behavior.

a. Show the sampling distribution of $\bar{p}$, the proportion of groceries thrown out by your sample respondents.

b. What is the probability that your survey will provide a sample proportion within ±.03 of the population proportion?

c. What is the probability that your survey will provide a sample proportion within ±.015 of the population proportion?

38. Forty-two percent of primary care doctors think their patients receive unnecessary medical care (*Reader's Digest,* December 2011/January 2012).

a. Suppose a sample of 300 primary care doctors were taken. Show the sampling distribution of the proportion of the doctors who think their patients receive unnecessary medical care.

b. What is the probability that the sample proportion will be within ±.03 of the population proportion?

c. What is the probability that the sample proportion will be within ±.05 of the population proportion?

d. What would be the effect of taking a larger sample on the probabilities in parts (b) and (c)? Why?

39. In 2008 the Better Business Bureau settled 75% of complaints they received (*USA Today,* March 2, 2009). Suppose you have been hired by the Better Business Bureau to investigate the complaints they received this year involving new car dealers. You plan to select a sample of new car dealer complaints to estimate the proportion of complaints the Better Business Bureau is able to settle. Assume the population proportion of complaints settled for new car dealers is .75, the same as the overall proportion of complaints settled in 2008.

a. Suppose you select a sample of 450 complaints involving new car dealers. Show the sampling distribution of $\bar{p}$.

b. Based upon a sample of 450 complaints, what is the probability that the sample proportion will be within .04 of the population proportion?

c. Suppose you select a sample of 200 complaints involving new car dealers. Show the sampling distribution of $\bar{p}$.

d. Based upon the smaller sample of only 200 complaints, what is the probability that the sample proportion will be within .04 of the population proportion?

e. As measured by the increase in probability, how much do you gain in precision by taking the larger sample in part (b)?

40. The Grocery Manufacturers of America reported that 76% of consumers read the ingredients listed on a product's label. Assume the population proportion is $p = .76$ and a sample of 400 consumers is selected from the population.

a. Show the sampling distribution of the sample proportion $\bar{p}$ where $\bar{p}$ is the proportion of the sampled consumers who read the ingredients listed on a product's label.

b. What is the probability that the sample proportion will be within ±.03 of the population proportion?
c. Answer part (b) for a sample of 750 consumers.

41. The Food Marketing Institute shows that 17% of households spend more than $100 per week on groceries. Assume the population proportion is $p = .17$ and a sample of 800 households will be selected from the population.
a. Show the sampling distribution of $\bar{p}$, the sample proportion of households spending more than $100 per week on groceries.
b. What is the probability that the sample proportion will be within ±.02 of the population proportion?
c. Answer part (b) for a sample of 1600 households.

7.7 Properties of Point Estimators

In this chapter we showed how sample statistics such as a sample mean $\bar{x}$, a sample standard deviation s, and a sample proportion $\bar{p}$ can be used as point estimators of their corresponding population parameters μ, σ, and p. It is intuitively appealing that each of these sample statistics is the point estimator of its corresponding population parameter. However, before using a sample statistic as a point estimator, statisticians check to see whether the sample statistic demonstrates certain properties associated with good point estimators. In this section we discuss three properties of good point estimators: unbiased, efficiency, and consistency.

Because several different sample statistics can be used as point estimators of different population parameters, we use the following general notation in this section.

$$\theta = \text{the population parameter of interest}$$
$$\hat{\theta} = \text{the sample statistic or point estimator of } \theta$$

The notation θ is the Greek letter theta, and the notation $\hat{\theta}$ is pronounced "theta-hat." In general, θ represents any population parameter such as a population mean, population standard deviation, population proportion, and so on; $\hat{\theta}$ represents the corresponding sample statistic such as the sample mean, sample standard deviation, and sample proportion.

Unbiased

If the expected value of the sample statistic is equal to the population parameter being estimated, the sample statistic is said to be an *unbiased estimator* of the population parameter.

UNBIASED

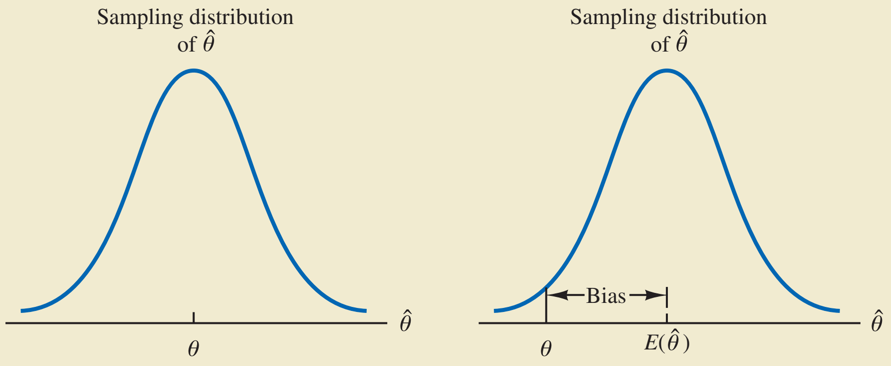
The sample statistic $\hat{\theta}$ is an unbiased estimator of the population parameter θ if

$$E(\hat{\theta}) = \theta$$

where

$$E(\hat{\theta}) = \text{the expected value of the sample statistic } \hat{\theta}$$

FIGURE 7.10 EXAMPLES OF UNBIASED AND BIASED POINT ESTIMATORS



Parameter θ is located at the mean of the sampling distribution; $E(\hat{\theta}) = \theta$

Panel A: Unbiased Estimator

Parameter θ is not located at the mean of the sampling distribution; $E(\hat{\theta}) \neq \theta$

Panel B: Biased Estimator

Hence, the expected value, or mean, of all possible values of an unbiased sample statistic is equal to the population parameter being estimated.

Figure 7.10 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 $\hat{\theta}$ 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 7.10, $E(\hat{\theta})$ is greater than θ; thus, 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.

In discussing the sampling distributions of the sample mean and the sample proportion, we stated that $E(\bar{x}) = \mu$ and $E(\bar{p}) = p$. Thus, both $\bar{x}$ and $\bar{p}$ are unbiased estimators of their corresponding population parameters μ and p.

In the case of the sample standard deviation s and the sample variance s^2, it can be shown that $E(s^2) = \sigma^2$. Thus, we conclude that the sample variance s^2 is an unbiased estimator of the population variance σ^2. In fact, when we first presented the formulas for the sample variance and the sample standard deviation in Chapter 3, $n - 1$ rather than n was used in the denominator. The reason for using $n - 1$ rather than n is to make the sample variance an unbiased estimator of the population variance.

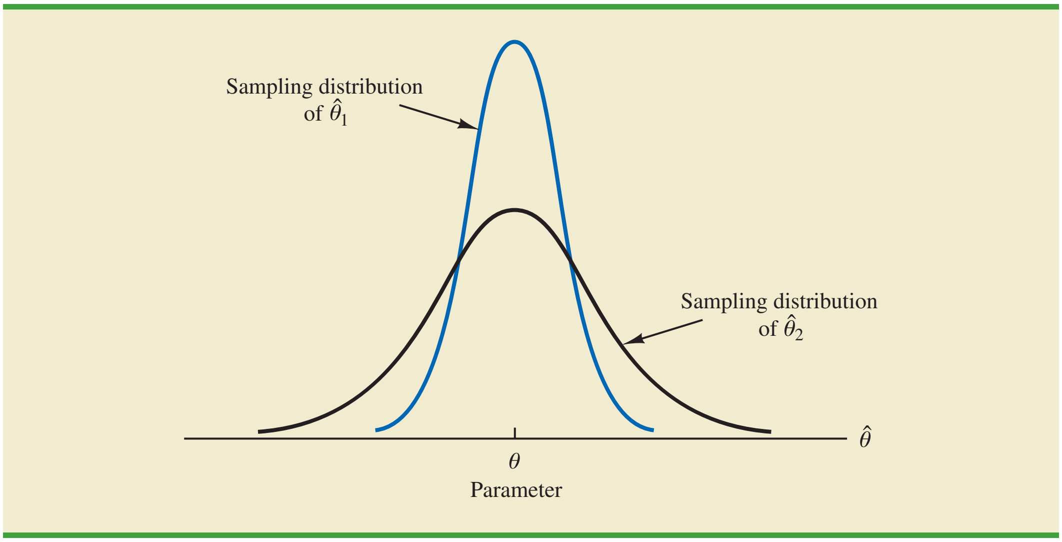
Efficiency

When sampling from a normal population, the standard error of the sample mean is less than the standard error of the sample median. Thus, the sample mean is more efficient than the sample median.

Assume that a simple random sample of n elements can be used to provide two unbiased point estimators of the same population parameter. In this situation, we would prefer to use the point estimator with the smaller standard error, because it tends to provide estimates closer to the population parameter. The point estimator with the smaller standard error is said to have greater **relative efficiency** than the other.

Figure 7.11 shows the sampling distributions of two unbiased point estimators, $\hat{\theta}_1$ and $\hat{\theta}_2$. Note that the standard error of $\hat{\theta}_1$ is less than the standard error of $\hat{\theta}_2$; thus,

FIGURE 7.11 SAMPLING DISTRIBUTIONS OF TWO UNBIASED POINT ESTIMATORS



values of $\hat{\theta}_1$ have a greater chance of being close to the parameter θ than do values of $\hat{\theta}_2$. Because the standard error of point estimator $\hat{\theta}_1$ is less than the standard error of point estimator $\hat{\theta}_2$, $\hat{\theta}_1$ is relatively more efficient than $\hat{\theta}_2$ and is the preferred point estimator.

Consistency

A third property associated with good point estimators is **consistency**. Loosely speaking, a point estimator is consistent if the values of the point estimator tend to become closer to the population parameter as the sample size becomes larger. In other words, a large sample size tends to provide a better point estimate than a small sample size. Note that for the sample mean $\bar{x}$, we showed that the standard error of $\bar{x}$ is given by $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. Because $\sigma_{\bar{x}}$ is related to the sample size such that larger sample sizes provide smaller values for $\sigma_{\bar{x}}$, we conclude that a larger sample size tends to provide point estimates closer to the population mean μ. In this sense, we can say that the sample mean $\bar{x}$ is a consistent estimator of the population mean μ. Using a similar rationale, we can also conclude that the sample proportion $\bar{p}$ is a consistent estimator of the population proportion p.

NOTES AND COMMENTS

In Chapter 3 we stated that the mean and the median are two measures of central location. In this chapter we discussed only the mean. The reason is that in sampling from a normal population, where the population mean and population median are identical, the standard error of the median is approximately 25% larger than the standard error of the mean. Recall that in the EAI problem where $n = 30$, the standard error of the mean is $\sigma_{\bar{x}} = 730.3$. The standard error of the median for this problem would be $1.25 \times (730.3) = 913$. As a result, the sample mean is more efficient and will have a higher probability of being within a specified distance of the population mean.

7.8 Other Sampling Methods

This section provides a brief introduction to survey sampling methods other than simple random sampling.

We described simple random sampling as a procedure for sampling from a finite population and discussed the properties of the sampling distributions of $\bar{x}$ and $\bar{p}$ when simple random sampling is used. Other methods such as stratified random sampling, cluster sampling, and systematic sampling provide advantages over simple random sampling in some of these situations. In this section we briefly introduce these alternative sampling methods. A more in-depth treatment is provided in Chapter 22, which is located on the website that accompanies the text.

Stratified Random Sampling

Stratified random sampling works best when the variance among elements in each stratum is relatively small.

In **stratified random sampling**, the elements in the population are first divided into groups called *strata,* such that each element in the population belongs to one and only one stratum. The basis for forming the strata, such as department, location, age, industry type, and so on, is at the discretion of the designer of the sample. However, the best results are obtained when the elements within each stratum are as much alike as possible. Figure 7.12 is a diagram of a population divided into H strata.

After the strata are formed, a simple random sample is taken from each stratum. Formulas are available for combining the results for the individual stratum samples into one estimate of the population parameter of interest. The value of stratified random sampling depends on how homogeneous the elements are within the strata. If elements within strata are alike, the strata will have low variances. Thus relatively small sample sizes can be used to obtain good estimates of the strata characteristics. If strata are homogeneous, the stratified random sampling procedure provides results just as precise as those of simple random sampling by using a smaller total sample size.

Cluster Sampling

Cluster sampling works best when each cluster provides a small-scale representation of the population.

In **cluster sampling**, the elements in the population are first divided into separate groups called *clusters*. Each element of the population belongs to one and only one cluster (see Figure 7.13). A simple random sample of the clusters is then taken. All elements within each sampled cluster form the sample. Cluster sampling tends to provide the best results when the elements within the clusters are not alike. In the ideal case, each cluster is a representative small-scale version of the entire population. The value of cluster sampling depends on how representative each cluster is of the entire population. If all clusters are alike in this regard, sampling a small number of clusters will provide good estimates of the population parameters.

FIGURE 7.12 DIAGRAM FOR STRATIFIED RANDOM SAMPLING

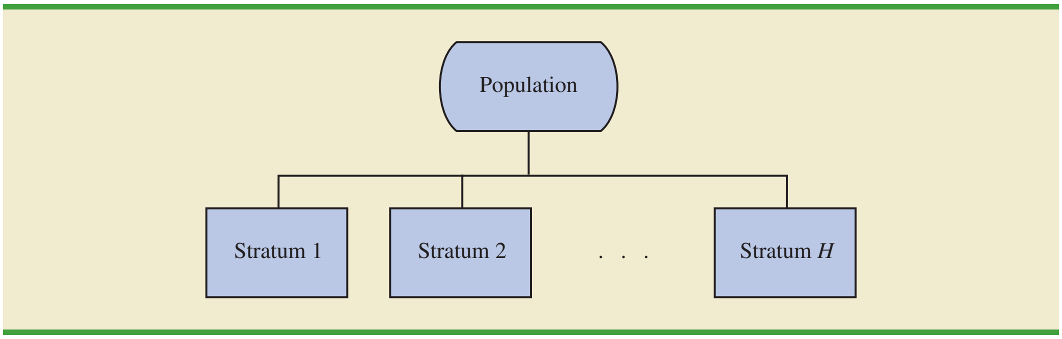
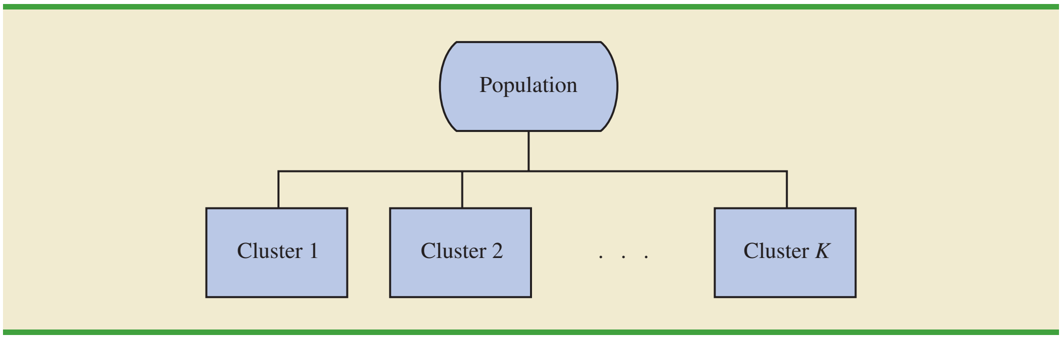


FIGURE 7.13 DIAGRAM FOR CLUSTER SAMPLING



One of the primary applications of cluster sampling is area sampling, where clusters are city blocks or other well-defined areas. Cluster sampling generally requires a larger total sample size than either simple random sampling or stratified random sampling. However, it can result in cost savings because of the fact that when an interviewer is sent to a sampled cluster (e.g., a city-block location), many sample observations can be obtained in a relatively short time. Hence, a larger sample size may be obtainable with a significantly lower total cost.

Systematic Sampling

In some sampling situations, especially those with large populations, it is time-consuming to select a simple random sample by first finding a random number and then counting or searching through the list of the population until the corresponding element is found. An alternative to simple random sampling is **systematic sampling**. For example, if a sample size of 50 is desired from a population containing 5000 elements, we will sample one element for every $5000/50 = 100$ elements in the population. A systematic sample for this case involves selecting randomly one of the first 100 elements from the population list. Other sample elements are identified by starting with the first sampled element and then selecting every 100th element that follows in the population list. 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 usually will be easier to identify in this way than it would be if simple random sampling were used. Because the first element selected is a random choice, a systematic sample is usually assumed to have the properties of a simple random sample. This assumption is especially applicable when the list of elements in the population is a random ordering of the elements.

Convenience Sampling

The sampling methods discussed thus far are referred to as *probability sampling* techniques. Elements selected from the population have a known probability of being included in the sample. The advantage of probability sampling is that the sampling distribution of the appropriate sample statistic generally can be identified. Formulas such as the ones for simple random sampling presented in this chapter can be used to determine the properties of the sampling distribution. Then the sampling distribution can be used to make probability statements about the error associated with using the sample results to make inferences about the population.

Convenience sampling is a *nonprobability sampling* technique. As the name implies, the sample is identified primarily by convenience. Elements are included in the sample without prespecified or known probabilities of being selected. For example, a professor conducting research at a university may use student volunteers to constitute a sample simply because they are readily available and will participate as subjects for little or no cost. Similarly, an inspector may sample a shipment of oranges by selecting oranges haphazardly from among several crates. Labeling each orange and using a probability method of sampling would be impractical. Samples such as wildlife captures and volunteer panels for consumer research are also convenience samples.

Convenience samples have the advantage of relatively easy sample selection and data collection; however, it is impossible to evaluate the "goodness" of the sample in terms of its representativeness of the population. A convenience sample may provide good results or it may not; no statistically justified procedure allows a probability analysis and inference about the quality of the sample results. Sometimes researchers apply statistical methods designed for probability samples to a convenience sample, arguing that the convenience sample can be treated as though it were a probability sample. However, this argument cannot be supported, and we should be cautious in interpreting the results of convenience samples that are used to make inferences about populations.

Judgment Sampling

One additional nonprobability sampling technique is **judgment sampling**. In this approach, the person most knowledgeable on the subject of the study selects elements of the population that he or she feels are most representative of the population. Often this method is a relatively easy way of selecting a sample. For example, a reporter may sample two or three senators, judging that those senators reflect the general opinion of all senators. However, the quality of the sample results depends on the judgment of the person selecting the sample. Again, great caution is warranted in drawing conclusions based on judgment samples used to make inferences about populations.

NOTES AND COMMENTS

We recommend using probability sampling methods when sampling from finite populations: simple random sampling, stratified random sampling, cluster sampling, or systematic sampling. For these methods, formulas are available for evaluating the "goodness" of the sample results in terms of the closeness of the results to the population parameters being estimated. An evaluation of the goodness cannot be made with convenience or judgment sampling. Thus, great care should be used in interpreting the results based on nonprobability sampling methods.

Summary

In this chapter we presented the concepts of sampling and sampling distributions. We demonstrated how a simple random sample can be selected from a finite population and how a random sample can be collected from an infinite population. The data collected from such samples can be used to develop point estimates of population parameters. Because different samples provide different values for the point estimators, point estimators such as $\bar{x}$ and $\bar{p}$ 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 $\bar{x}$ and the sample proportion $\bar{p}$.

In considering the characteristics of the sampling distributions of $\bar{x}$ and $\bar{p}$, we stated that $E(\bar{x}) = \mu$ and $E(\bar{p}) = p$. Thus $\bar{x}$ and $\bar{p}$ are unbiased estimators. After developing the standard deviation or standard error formulas for these estimators, we described the conditions necessary for the sampling distributions of $\bar{x}$ and $\bar{p}$ to follow a normal distribution. Other sampling methods including stratified random sampling, cluster sampling, systematic sampling, convenience sampling, and judgment sampling were discussed.

Glossary

Sampled population The population from which the sample is taken.

Frame A listing of the elements the sample will be selected from.

Parameter A numerical characteristic of a population, such as a population mean μ, a population standard deviation σ, a population proportion p, and so on.

Simple random sample 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.

Sampling without replacement Once an element has been included in the sample, it is removed from the population and cannot be selected a second time.

Sampling with replacement Once an element has been included in the sample, it is returned to the population. A previously selected element can be selected again and therefore may appear in the sample more than once.

Random sample A random sample from an infinite population is a sample selected such that the following conditions are satisfied: (1) Each element selected comes from the same population; (2) each element is selected independently.

Sample statistic A sample characteristic, such as a sample mean $\bar{x}$, a sample standard deviation s, a sample proportion $\bar{p}$, and so on. The value of the sample statistic is used to estimate the value of the corresponding population parameter.

Point estimator The sample statistic, such as $\bar{x}$, s, or $\bar{p}$, that provides the point estimate of the population parameter.

Point estimate The value of a point estimator used in a particular instance as an estimate of a population parameter.

Target population The population for which statistical inferences such as point estimates are made. It is important for the target population to correspond as closely as possible to the sampled population.

Sampling distribution A probability distribution consisting of all possible values of a sample statistic.

Unbiased A property of a point estimator that is present when the expected value of the point estimator is equal to the population parameter it estimates.

Finite population correction factor The term $\sqrt{(N - n)/(N - 1)}$ that is used in the formulas for $\sigma_{\bar{x}}$ and $\sigma_{\bar{p}}$ whenever a finite population, rather than an infinite population, is being sampled. The generally accepted rule of thumb is to ignore the finite population correction factor whenever $n/N \leq .05$.

Standard error The standard deviation of a point estimator.

Central limit theorem A theorem that enables one to use the normal probability distribution to approximate the sampling distribution of $\bar{x}$ whenever the sample size is large.

Relative efficiency Given two unbiased point estimators of the same population parameter, the point estimator with the smaller standard error is more efficient.

Consistency A property of a point estimator that is present whenever larger sample sizes tend to provide point estimates closer to the population parameter.

Stratified random sampling A probability sampling method in which the population is first divided into strata and a simple random sample is then taken from each stratum.

Cluster sampling A probability sampling method in which the population is first divided into clusters and then a simple random sample of the clusters is taken.

Systematic sampling A probability sampling method in which we randomly select one of the first k elements and then select every kth element thereafter.

Convenience sampling A nonprobability method of sampling whereby elements are selected for the sample on the basis of convenience.

Judgment sampling A nonprobability method of sampling whereby elements are selected for the sample based on the judgment of the person doing the study.

Key Formulas

Expected Value of $\bar{x}$

$$E(\bar{x}) = \mu \qquad \textbf{(7.1)}$$

Standard Deviation of $\bar{x}$ (Standard Error)

Finite Population *Infinite Population*

$$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}\left(\frac{\sigma}{\sqrt{n}}\right) \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad \textbf{(7.2)}$$

Expected Value of $\bar{p}$

$$E(\bar{p}) = p \qquad \textbf{(7.4)}$$

Standard Deviation of $\bar{p}$ (Standard Error)

Finite Population *Infinite Population*

$$\sigma_{\bar{p}} = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{p(1-p)}{n}} \qquad \sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} \qquad \textbf{(7.5)}$$

Supplementary Exercises

42. *U.S. News & World Report* publishes comprehensive information on America's best colleges (*America's Best Colleges,* 2009 ed.). Among other things, they provide a listing of their 133 best national universities. You would like to take a sample of these universities for a follow-up study on their students. Begin at the bottom of the third column of random digits in Table 7.1. Ignoring the first two digits in each five-number group and using the three-digit random numbers beginning with 959 read *up* the column to identify the number (from 1 to 133) of the first seven universities to be included in a simple random sample. Continue by starting at the bottom of the fourth and fifth columns and reading up if necessary.

43. The latest available data showed health expenditures were \$8086 per person in the United States or 17.6% of gross domestic product (Centers for Medicare & Medicaid Services website, April 1, 2012). Use \$8086 as the population mean and suppose a survey research firm will take a sample of 100 people to investigate the nature of their health expenditures. Assume the population standard deviation is \$2500.
    a. Show the sampling distribution of the mean amount of health care expenditures for a sample of 100 people.

b. What is the probability the sample mean will be within ± $200 of the population mean?

c. What is the probability the sample mean will be greater than $9000? If the survey research firm reports a sample mean greater than $9000, would you question whether the firm followed correct sampling procedures? Why or why not?

44. Foot Locker uses sales per square foot as a measure of store productivity. Sales are currently running at an annual rate of $406 per square foot (*The Wall Street Journal,* March 7, 2012). You have been asked by management to conduct a study of a sample of 64 Foot Locker stores. Assume the standard deviation in annual sales per square foot for the population of all 3400 Foot Locker stores is $80.

a. Show the sampling distribution of $\bar{x}$, the sample mean annual sales per square foot for a sample of 64 Foot Locker stores.

b. What is the probability that the sample mean will be within $15 of the population mean?

c. Suppose you find a sample mean of $380. What is the probability of finding a sample mean of $380 or less? Would you consider such a sample to be an unusually low performing group of stores?

45. The mean television viewing time for Americans is 15 hours per week (*Money,* November 2003). Suppose a sample of 60 Americans is taken to further investigate viewing habits. Assume the population standard deviation for weekly viewing time is $\sigma = 4$ hours.

a. What is the probability the sample mean will be within 1 hour of the population mean?

b. What is the probability the sample mean will be within 45 minutes of the population mean?

46. After deducting grants based on need, the average cost to attend the University of Southern California (USC) is $27,175 (*U.S. News & World Report, America's Best Colleges,* 2009 ed.). Assume the population standard deviation is $7400. Suppose that a random sample of 60 USC students will be taken from this population.

a. What is the value of the standard error of the mean?

b. What is the probability that the sample mean will be more than $27,175?

c. What is the probability that the sample mean will be within $1000 of the population mean?

d. How would the probability in part (c) change if the sample size were increased to 100?

47. Three firms carry inventories that differ in size. Firm A's inventory contains 2000 items, firm B's inventory contains 5000 items, and firm C's inventory contains 10,000 items. The population standard deviation for the cost of the items in each firm's inventory is $\sigma = 144$. A statistical consultant recommends that each firm take a sample of 50 items from its inventory to provide statistically valid estimates of the average cost per item. Managers of the small firm state that because it has the smallest population, it should be able to make the estimate from a much smaller sample than that required by the larger firms. However, the consultant states that to obtain the same standard error and thus the same precision in the sample results, all firms should use the same sample size regardless of population size.

a. Using the finite population correction factor, compute the standard error for each of the three firms given a sample of size 50.

b. What is the probability that for each firm the sample mean $\bar{x}$ will be within ± 25 of the population mean μ?

48. 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 used in this survey?

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

49. A production process is checked periodically by a quality control inspector. The inspector selects simple random samples of 30 finished products and computes the sample mean product weights $\bar{x}$. If test results over a long period of time show that 5% of the $\bar{x}$ values are over 2.1 pounds and 5% are under 1.9 pounds, what are the mean and the standard deviation for the population of products produced with this process?

50. About 28% of private companies are owned by women (*The Cincinnati Enquirer,* January 26, 2006). Answer the following questions based on a sample of 240 private companies.
    a. Show the sampling distribution of $\bar{p}$, the sample proportion of companies that are owned by women.
    b. What is the probability the sample proportion will be within ±.04 of the population proportion?
    c. What is the probability the sample proportion will be within ±.02 of the population proportion?

51. A market research firm conducts telephone surveys with a 40% 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 $150/400 = .375$?

52. 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 ad. Unfortunately, click fraud—the practice of someone clicking on an ad solely for the purpose of driving up advertising revenue—has become a problem. Forty percent of advertisers claim they have been a victim of click fraud (*BusinessWeek,* March 13, 2006). Suppose a simple random sample of 380 advertisers will be taken to learn more about how they are affected by this practice.
    a. What is the probability that the sample proportion will be within ±.04 of the population proportion experiencing click fraud?
    b. What is the probability that the sample proportion will be greater than .45?

53. The proportion of individuals insured by the All-Driver Automobile Insurance Company who received at least one traffic ticket during a five-year period is .15.
    a. Show the sampling distribution of $\bar{p}$ if a random sample of 150 insured individuals is used to estimate the proportion having received at least one ticket.
    b. What is the probability that the sample proportion will be within ±.03 of the population proportion?

54. Lori Jeffrey is a successful sales representative for a major publisher of college textbooks. Historically, Lori obtains a book adoption on 25% of her sales calls. Viewing her sales calls for one month as a sample of all possible sales calls, assume that a statistical analysis of the data yields a standard error of the proportion of .0625.
    a. How large was the sample used in this analysis? That is, how many sales calls did Lori make during the month?
    b. Let $\bar{p}$ indicate the sample proportion of book adoptions obtained during the month. Show the sampling distribution of $\bar{p}$.
    c. Using the sampling distribution of $\bar{p}$, compute the probability that Lori will obtain book adoptions on 30% or more of her sales calls during a one-month period.

Appendix 7.1 The Expected Value and Standard Deviation of $\bar{x}$

In this appendix we present the mathematical basis for the expressions for $E(\bar{x})$, the expected value of $\bar{x}$ as given by equation (7.1), and $\sigma_{\bar{x}}$, the standard deviation of $\bar{x}$ as given by equation (7.2).

Expected Value of $\bar{x}$

Assume a population with mean μ and variance σ^2. A simple random sample of size n is selected with individual observations denoted $x_1, x_2, \ldots, x_n$. A sample mean $\bar{x}$ is computed as follows.

$$\bar{x} = \frac{\Sigma x_i}{n}$$

With repeated simple random samples of size n, $\bar{x}$ is a random variable that assumes different numerical values depending on the specific n items selected. The expected value of the random variable $\bar{x}$ is the mean of all possible $\bar{x}$ values.

$$\begin{aligned}\text{Mean of } \bar{x} = E(\bar{x}) &= E\left(\frac{\Sigma x_i}{n}\right)\\ &= \frac{1}{n}[E(x_1 \quad x_2 \quad \cdots \quad x_n)]\\ &= \frac{1}{n}[E(x_1) \quad E(x_2) \quad \cdots \quad E(x_n)]\end{aligned}$$

For any x_i we have $E(x_i) = \mu$; therefore we can write

$$\begin{aligned}E(\bar{x}) &= \frac{1}{n}(\mu \quad \mu \quad \cdots \quad \mu)\\ &= \frac{1}{n}(n\mu) = \mu\end{aligned}$$

This result shows that the mean of all possible $\bar{x}$ values is the same as the population mean μ. That is, $E(\bar{x}) = \mu$.

Standard Deviation of $\bar{x}$

Again assume a population with mean μ, variance σ^2, and a sample mean given by

$$\bar{x} = \frac{\Sigma x_i}{n}$$

With repeated simple random samples of size n, we know that $\bar{x}$ is a random variable that takes different numerical values depending on the specific n items selected. What follows is the derivation of the expression for the standard deviation of the $\bar{x}$ values, $\sigma_{\bar{x}}$, for the case of an infinite population. The derivation of the expression for $\sigma_{\bar{x}}$ for a finite population when sampling is done without replacement is more difficult and is beyond the scope of this text.

Returning to the infinite population case, recall that a simple random sample from an infinite population consists of observations $x_1, x_2, \ldots, x_n$ that are independent. The following two expressions are general formulas for the variance of random variables.

$$Var(ax) = a^2\, Var(x)$$

where a is a constant and x is a random variable, and

$$Var(x \quad y) = Var(x) \quad Var(y)$$

where x and y are *independent* random variables. Using the two preceding equations, we can develop the expression for the variance of the random variable $\bar{x}$ as follows.

$$Var(\bar{x}) = Var\left(\frac{\Sigma x_i}{n}\right) = Var\left(\frac{1}{n}\Sigma x_i\right)$$

Then, with $1/n$ a constant, we have

$$\begin{aligned} Var(\bar{x}) &= \left(\frac{1}{n}\right)^2 Var(\Sigma x_i) \\ &= \left(\frac{1}{n}\right)^2 Var(x_1 \quad x_2 \quad \cdots \quad x_n) \end{aligned}$$

In the infinite population case, the random variables $x_1, x_2, \ldots, x_n$ are independent, which enables us to write

$$Var(\bar{x}) = \left(\frac{1}{n}\right)^2 [Var(x_1) \quad Var(x_2) \quad \cdots \quad Var(x_n)]$$

For any x_i, we have $Var(x_i) = \sigma^2$; therefore we have

$$Var(\bar{x}) = \left(\frac{1}{n}\right)^2 (\sigma^2 \quad \sigma^2 \quad \cdots \quad \sigma^2)$$

With n values of σ^2 in this expression, we have

$$Var(\bar{x}) = \left(\frac{1}{n}\right)^2 (n\sigma^2) = \frac{\sigma^2}{n}$$

Taking the square root provides the formula for the standard deviation of $\bar{x}$.

$$\sigma_{\bar{x}} = \sqrt{Var(\bar{x})} = \frac{\sigma}{\sqrt{n}}$$

Appendix 7.2 Random Sampling with Minitab

If a list of the elements in a population is available in a Minitab file, Minitab can be used to select a simple random sample. For example, a list of the top 100 metropolitan areas in the United States and Canada is provided in column 1 of the data set MetAreas (*Places Rated Almanac—The Millennium Edition 2000*). Column 2 contains the overall rating of each metropolitan area. The first 10 metropolitan areas in the data set and their corresponding ratings are shown in Table 7.6.

Suppose that you would like to select a simple random sample of 30 metropolitan areas in order to do an in-depth study of the cost of living in the United States and Canada. The following steps can be used to select the sample.

Step 1. Select the **Calc** pull-down menu
Step 2. Choose **Random Data**
Step 3. Choose **Sample From Columns**

TABLE 7.6 OVERALL RATING FOR THE FIRST 10 METROPOLITAN AREAS IN THE DATA SET METAREAS

WEB file
MetAreas

Metropolitan Area	Rating
Albany, NY	64.18
Albuquerque, NM	66.16
Appleton, WI	60.56
Atlanta, GA	69.97
Austin, TX	71.48
Baltimore, MD	69.75
Birmingham, AL	69.59
Boise City, ID	68.36
Boston, MA	68.99
Buffalo, NY	66.10

Step 4. When the Sample From Columns dialog box appears:
Enter 30 in the **Number of rows to sample** box
Enter C1 C2 in the **From columns** box below
Enter C3 C4 in the **Store samples in** box
Step 5. Click **OK**

The random sample of 30 metropolitan areas appears in columns C3 and C4.

Appendix 7.3 Random Sampling with Excel

If a list of the elements in a population is available in an Excel file, Excel can be used to select a simple random sample. For example, a list of the top 100 metropolitan areas in the United States and Canada is provided in column A of the data set MetAreas (*Places Rated Almanac—The Millennium Edition 2000*). Column B contains the overall rating of each metropolitan area. The first 10 metropolitan areas in the data set and their corresponding ratings are shown in Table 7.6. Assume that you would like to select a simple random sample of 30 metropolitan areas in order to do an in-depth study of the cost of living in the United States and Canada.

The rows of any Excel data set can be placed in a random order by adding an extra column to the data set and filling the column with random numbers using the =RAND() function. Then, using Excel's sort ascending capability on the random number column, the rows of the data set will be reordered randomly. The random sample of size *n* appears in the first *n* rows of the reordered data set.

In the MetAreas data set, labels are in row 1 and the 100 metropolitan areas are in rows 2 to 101. The following steps can be used to select a simple random sample of 30 metropolitan areas.

Step 1. Enter =RAND() in cell C2
Step 2. Copy cell C2 to cells C3:C101
Step 3. Select any cell in Column C
Step 4. Click the **Home** tab on the Ribbon
Step 5. In the **Editing** group, click **Sort & Filter**
Step 6. Click **Sort Smallest to Largest**

The random sample of 30 metropolitan areas appears in rows 2 to 31 of the reordered data set. The random numbers in column C are no longer necessary and can be deleted if desired.

Appendix 7.4 Random Sampling with StatTools

WEB file

MetAreas

If a list of the elements in a population is available in an Excel file, StatTools Random Sample Utility can be used to select a simple random sample. For example, a list of the top 100 metropolitan areas in the United States and Canada is provided in column A of the data set MetAreas (*Places Rated Almanac—The Millennium Edition 2000*). Column B contains the overall rating of each metropolitan area. Assume that you would like to select a simple random sample of 30 metropolitan areas in order to do an in-depth study of the cost of living in the United States and Canada.

Begin by using the Data Set Manager to create a StatTools data set for these data using the procedure described in the appendix to Chapter 1. The following steps will generate a simple random sample of 30 metropolitan areas.

Step 1. Click the **StatTools** tab on the Ribbon
Step 2. In the **Data Group** click **Data Utilities**
Step 3. Choose the **Random Sample** option
Step 4. When the StatTools—Random Sample Utility dialog box appears:
In the **Variables** section:
Select **Metropolitan Area**
Select **Rating**
In the **Options** section:
Enter 1 in the **Number of Samples** box
Enter 30 in the **Sample Size** box
Click **OK**

The random sample of 30 metropolitan areas will appear in columns A and B of the worksheet entitled Random Sample.