CHAPTER 7

Sampling and Sampling Distributions

Learning Objectives

After mastering the material in this chapter, you will be able to:

- LO7-1 Explain the concept of random sampling and select a random sample.
- LO7-2 Describe and use the sampling distribution of the sample mean.
- LO7-3 Explain and use the Central Limit Theorem.
- LO7-4 Describe and use the sampling distribution of the sample proportion.
- LO7-5 Describe the basic ideas of stratified random, cluster, and systematic sampling (Optional).
- LO7-6 Describe basic types of survey questions, survey procedures, and sources of error (Optional).

Chapter Outline

In Chapter 1 we introduced random sampling. In this chapter we continue our discussion of random sampling by explaining what a random sample is and how to select a random sample. In addition, we discuss two probability distributions that are related to random sampling. To understand these distributions, note that if we select a random sample, then we use the sample mean as the point estimate of the population mean and the sample proportion as the point estimate of the population proportion. Two probability distributions that help us assess how accurate the sample mean and sample proportion are likely to be as point estimates are **the sampling distribution of the sample mean** and **the sampling distribution of the sample proportion.** After discussing random sampling in the first section of this chapter, we consider these sampling distributions in the next two sections. Moreover, using the car mileage case, the e-billing case, and the cheese spread case, we demonstrate how sampling distributions can be used to make statistical inferences.

The discussions of random sampling and of sampling distributions given in the first three sections of this chapter are necessary for understanding the rest of this book. The last three sections of this chapter consider advanced aspects of sampling and are optional. In the first optional section, we discuss three alternatives to random sampling—**stratified random sampling, cluster sampling,** and **systematic sampling.** In the second optional section, we discuss issues related to designing surveys and the errors that can occur in survey sampling. In the last optional section, we derive the mean and variance of the sample mean.

7.1 Random Sampling

LO7-1 Explain the concept of random sampling and select a random sample.

Selecting a random sample from a population is one of the best ways to ensure that the information contained in the sample reflects what is true about the population. To illustrate the idea of a random sample, consider the *cell phone case,* and recall that a bank has 2,136 employees on various 500-minute-per-month calling plans. In order to assess its cellular costs for these 500-minute plans, the bank will analyze in detail the cell phone bills for a random sample of 100 employees on these plans. One intuitive procedure for selecting a random sample of 100 employees from a population of 2,136 employees would begin by numbering the 2,136 employees from 1 to 2,136 and placing 2,136 identical slips of paper numbered from 1 to 2,136 in a suitable container. We would then thoroughly mix the slips of paper in the container and, blindfolded, choose one. The number on the chosen slip of paper would identify the first randomly selected employee. Next, still blindfolded, we would choose another slip of paper from the container. The number on the second slip would identify the second randomly selected employee. Continuing this process, we would select a total of 100 slips of paper from the container. The numbers on the 100 selected slips of paper would identify the 100 employees that make up the random sample.

In practice, numbering 2,136 (or any large number of) slips of paper would be very time consuming, and actual experience has shown that thoroughly mixing slips of paper (or the like) can be difficult. For these reasons, statisticians have developed more efficient and accurate methods for selecting a random sample. To discuss these methods, we let n, which we call the sample size, denote the number of elements in a sample. We then define a random sample of n elements—and explain how to select such a sample—as follows:[1]

1. If we select n elements from a population in such a way that every set of n elements in the population has the same chance of being selected, then the n elements we select are said to be a **random sample.**
2. In order to select a random sample of n elements from a population, we make n *random selections*—one at a time—from the population. On each **random selection,** we give every element remaining in the population for that selection the same chance of being chosen.

In making random selections from a population, we can sample *with or without replacement.* If we **sample with replacement,** we place the element chosen on any particular selection back into the population. Thus, we give this element a chance to be chosen on any succeeding selection. If we **sample without replacement,** we do not place the element chosen on a particular selection back into the population. Thus, we do not give this element a chance to be chosen on any succeeding selection. **It is best to sample without replacement.** Intuitively, this is because

[1]Actually, there are several different kinds of random samples. The type we will define is sometimes called a *simple random sample.* For brevity's sake, however, we will use the term *random sample.*

choosing the sample without replacement guarantees that all of the elements in the sample will be different, and thus we will have the fullest possible look at the population.

The first step in selecting a random sample is to obtain or make a numbered list of the population elements. Then, as illustrated in the following example, we can use a *random number table* or *computer-generated random numbers* to make random selections from the numbered list.

EXAMPLE 7.1 The Cell Phone Case: Reducing Cellular Phone Costs

In order to select a random sample of 100 employees from the population of 2,136 employees on 500-minute-per-month cell phone plans, the bank will make a numbered list of the 2,136 employees on 500-minute plans. The bank can then use a **random number table,** such as Table 7.1(a), to select the random sample. To see how this is done, note that any single-digit number in the table has been chosen in such a way that any of the single-digit numbers between 0 and 9 had the same chance of being chosen. For this reason, we say that any single-digit number in the table is a **random number** between 0 and 9. Similarly, any two-digit number in the table is a random number between 00 and 99, any three digit number in the table is a random number between 000 and 999, and so forth. Note that the table entries are segmented into groups of five to make the table easier to read. Because the total number of cell phone users on the 500-minute plans (2,136) is a four-digit number, we arbitrarily select any set of four digits in the table (we have circled these digits). This number, which is 0511, identifies the first randomly selected user. Then, moving in any direction from the 0511 (up, down, right, or left—it does not matter which), we select additional sets of four digits. These succeeding sets of digits identify additional randomly selected users. Here we arbitrarily move down from 0511 in the table. The first seven sets of four digits we obtain are

0511 7156 0285 4461 3990 4919 1915

(See Table 7.1(a)—these numbers are enclosed in a rectangle.) Because there are no users numbered 7156, 4461, 3990, or 4919 (remember only 2,136 users are on 500-minute plans), we ignore these numbers. This implies that the first three randomly selected users are those numbered 0511, 0285, and 1915. Continuing this procedure, we can obtain the entire random sample of 100 users. Notice that, because we are sampling without replacement, we should ignore any set of four digits previously selected from the random number table.

While using a random number table is one way to select a random sample, this approach has a disadvantage that is illustrated by the current situation. Specifically, since most four-digit random numbers are not between 0001 and 2136, obtaining 100 different, four-digit random numbers between 0001 and 2136 will require ignoring a large number of random numbers in the random number table, and we will in fact need to use a random number table that is larger than

TABLE 7.1 Random Numbers

(a) A portion of a random number table

33276	85590	79936	56865	05859	90106	78188
03427	90511	69445	18663	72695	52180	90322
92737	27156	33488	36320	17617	30015	74952
85689	20285	52267	67689	93394	01511	89868
08178	74461	13916	47564	81056	97735	90707
51259	63990	16308	60756	92144	49442	40719
60268	44919	19885	55322	44819	01188	55157
94904	01915	04146	18594	29852	71585	64951
58586	17752	14513	83149	98736	23495	35749
09998	19509	06691	76988	13602	51851	58104
14346	61666	30168	90229	04734	59193	32812
74103	15227	25306	76468	26384	58151	44592
24200	64161	38005	94342	28728	35806	22851
87308	07684	00256	45834	15398	46557	18510
07351	86679	92420	60952	61280	50001	94953

(b) MINITAB output of 100 different, four-digit random numbers between 1 and 2136

705	1131	169	1703	1709	609
1990	766	1286	1977	222	43
1007	1902	1209	2091	1742	1152
111	69	2049	1448	659	338
1732	1650	7	388	613	1477
838	272	1227	154	18	320
1053	1466	2087	265	2107	1992
582	1787	2098	1581	397	1099
757	1699	567	1255	1959	407
354	1567	1533	1097	1299	277
663	40	585	1486	1021	532
1629	182	372	1144	1569	1981
1332	1500	743	1262	1759	955
1832	378	728	1102	667	1885
514	1128	1046	116	1160	1333
831	2036	918	1535	660	
928	1257	1468	503	468	

Table 7.1(a). Although larger random number tables are readily available in books of mathematical and statistical tables, a good alternative is to use a computer software package, which can generate random numbers that are between whatever values we specify. For example, Table 7.1(b) gives the MINITAB output of 100 different, four-digit random numbers that are between 0001 and 2136 (note that the "leading 0's" are not included in these four-digit numbers). If used, the random numbers in Table 7.1(b) would identify the 100 employees that form the random sample. For example, the first three randomly selected employees would be employees 705, 1990, and 1007. When the number of cellular minutes used by each randomly selected employee is found and recorded, we obtain the sample of cellular usages that has been given in Table 1.4 (see page 9).

To conclude this example, note that computer software packages sometimes generate the same random number twice and thus are sampling with replacement. Because we wished to randomly select 100 employees without replacement, we had MINITAB generate more than 100 (actually, 110) random numbers. We then ignored the repeated random numbers to obtain the 100 different random numbers in Table 7.1(b).

Next, consider the *marketing research case,* and recall that we wish to select a sample of 60 shoppers at a large metropolitan shopping mall on a particular Saturday. Because it is not possible to list and number all of the shoppers who will be at the mall on this Saturday, we cannot select a random sample of these shoppers. However, we can select an *approximately* random sample of these shoppers. To see one way to do this, note that there are 6 ten-minute intervals during each hour, and thus there are 60 ten-minute intervals during the 10-hour period from 10 A.M. to 8 P.M.—the time when the shopping mall is open. Therefore, one way to select an approximately random sample is to choose a particular location at the mall that most shoppers will walk by and then randomly select—at the beginning of each ten-minute period—one of the first shoppers who walks by the location. Here, although we could randomly select one person from any reasonable number of shoppers who walk by, we will (arbitrarily) randomly select one of the first five shoppers who walk by. For example, starting in the upper left-hand corner of Table 7.1(a) and proceeding down the first column, note that the first three random numbers between 1 and 5 are 3, 5, and 1. This implies that (1) at 10 A.M. we would select the 3rd customer who walks by; (2) at 10:10 A.M. we would select the 5th shopper who walks by; (3) at 10:20 A.M. we would select the 1st customer who walks by, and so forth. Furthermore, assume that the composite score ratings of the new bottle design that would be given by all shoppers at the mall on the Saturday are representative of the composite score ratings that would be given by all possible consumers. It then follows that the composite score ratings given by the 60 sampled shoppers can be regarded as an approximately random sample that can be used to make statistical inferences about the population of all possible consumer composite score ratings.

As another example, consider the *car mileage case,* and recall that the automaker has decided to select a sample of 50 cars by randomly selecting one car from the 100 cars produced on each of 50 consecutive production shifts. If we number the 100 cars produced on a particular production shift from 00 to 99, we can randomly select a car from the shift by using a random number table or a computer software package to obtain a random number between 00 and 99. For example, starting in the upper left-hand corner of Table 7.1(a) and proceeding down the first column, we see that the first three random numbers between 00 and 99 are 33, 3, and 92. This implies that we would select car 33 from the first production shift, car 3 from the second production shift, car 92 from the third production shift, and so forth. Moreover, because a new group of 100 cars is produced on each production shift, repeated random numbers would not be discarded. For example, if the 15th and 29th random numbers are both 7, we would select the 7th car from the 15th production shift and the 7th car from the 29th production shift. When the 50 cars are selected and tested as prescribed by the EPA, the sample of 50 mileages that has been given in Table 1.6 (see page 11) is obtained. Furthermore, recall that we waited to randomly select the 50 cars from the 50 production shifts until the midsize car manufacturing process was operating consistently over time and recall that the time series plot in Figure 1.3 (page 11) intuitively verifies that the manufacturing process is producing consistent car mileages over time. It follows that we can regard the 50 mileages in Table 1.6 as an approximately random sample that can be used to make statistical inferences about the population of all possible midsize car mileages. (In Chapter 17 we will discuss more precisely how to assess whether a process is operating consistently over time.)

Random (or approximately random) sampling—as well as the more advanced kinds of sampling discussed in optional Section 7.4—are types of *probability sampling*. In general, **probability sampling** is sampling where we know the chance (or probability) that each element in the population will be included in the sample. If we employ probability sampling, the sample obtained can be used to make valid statistical inferences about the sampled population. However, if we do not employ probability sampling, we cannot make valid statistical inferences.

One type of sampling that is not probability sampling is **convenience sampling,** where we select elements because they are easy or convenient to sample. For example, if we select people to interview because they look "nice" or "pleasant," we are using convenience sampling. Another example of convenience sampling is the use of **voluntary response samples,** which are frequently employed by television and radio stations and newspaper columnists. In such samples, participants self-select—that is, whoever wishes to participate does so (usually expressing some opinion). These samples overrepresent people with strong (usually negative) opinions. For example, the advice columnist Ann Landers once asked her readers, "If you had it to do over again, would you have children?" Of the nearly 10,000 parents who *voluntarily* responded, 70 percent said that they would not. A probability sample taken a few months later found that 91 percent of parents would have children again.

Another type of sampling that is not probability sampling is **judgment sampling,** where a person who is extremely knowledgeable about the population under consideration selects population elements that he or she feels are most representative of the population. Because the quality of the sample depends upon the judgment of the person selecting the sample, it is dangerous to use the sample to make statistical inferences about the population.

To conclude this section, we consider a classic example where two types of sampling errors doomed a sample's ability to make valid statistical inferences. This example occurred prior to the presidential election of 1936, when the *Literary Digest* predicted that Alf Landon would defeat Franklin D. Roosevelt by a margin of 57 percent to 43 percent. Instead, Roosevelt won the election in a landslide. *Literary Digest*'s first error was to send out sample ballots (actually, 10 million ballots) to people who were mainly selected from the *Digest*'s subscription list and from telephone directories. In 1936 the country had not yet recovered from the Great Depression, and many unemployed and low-income people did not have phones or subscribe to the *Digest*. The *Digest*'s sampling procedure excluded these people, who overwhelmingly voted for Roosevelt. Second, only 2.3 million ballots were returned, resulting in the sample being a voluntary response survey. At the same time, George Gallup, founder of the Gallup Poll, was beginning to establish his survey business. He used a probability sample to correctly predict Roosevelt's victory. In optional Section 7.5 we discuss various issues related to designing surveys and more about the errors that can occur in survey samples. Optional Sections 7.4 and 7.5 can now be read at any time and in any order.

Exercises for Section 7.1

CONCEPTS

connect

Companies:

1 Altria Group
2 PepsiCo
3 Coca-Cola
4 Archer Daniels
5 Anheuser-Bush
6 General Mills
7 Sara Lee
8 Coca-Cola Enterprises
9 Reynolds American
10 Kellogg
11 ConAgra Foods
12 HJ Heinz
13 Campbell Soup
14 Pepsi Bottling Group
15 Tyson Foods

7.1 Discuss how we select a random sample.

7.2 Explain why sampling without replacement is preferred to sampling with replacement.

METHODS AND APPLICATIONS

7.3 On the page margin, we list 15 companies that have historically performed well in the food, drink, and tobacco industries. Consider the random numbers given in the random number table of Table 7.1(a) on page 268. Starting in the upper left corner of Table 7.1(a) and moving down the two leftmost columns, we see that the first three two-digit numbers obtained are: 33, 03, and 92. Starting with these three random numbers, and moving down the two leftmost columns of Table 7.1(a) to find more two-digit random numbers, use Table 7.1(a) to randomly select five of these companies to be interviewed in detail about their business strategies. Hint: Note that we have numbered the companies from 1 to 15.

7.4 **THE VIDEO GAME SATISFACTION RATING CASE** VideoGame

A company that produces and markets video game systems wishes to assess its customers' level of satisfaction with a relatively new model, the XYZ-Box. In the six months since the introduction of the model, the company has received 73,219 warranty registrations from purchasers. The company will randomly select 65 of these registrations and will conduct telephone interviews with the purchasers. Assume that the warranty registrations are numbered from 1 to 73,219 in a computer.

Starting in the upper left corner of Table 7.1(a) and moving down the five leftmost columns, we see that the first three five-digit numbers obtained are: 33276, 03427, and 92737. Starting with these three random numbers and moving down the five leftmost columns of Table 7.1(a) to find more five-digit random numbers, use Table 7.1(a) to randomly select the numbers of the first 10 warranty registrations to be included in the sample of 65 registrations.

7.5 THE BANK CUSTOMER WAITING TIME CASE WaitTime

Recall that when the bank manager's new teller system is operating consistently over time, the manager decides to record the waiting times of a sample of 100 customers that need teller service during peak business hours. For each of 100 peak business hours, the first customer that starts waiting for service at or after a randomly selected time during the hour will be chosen. Consider the peak business hours from 2:00 P.M. to 2:59 P.M., from 3:00 P.M. to 3:59 P.M., from 4:00 P.M. to 4:59 P.M., and from 5:00 P.M. to 5:59 P.M. on a particular day. Also, assume that a computer software system generates the following four random numbers between 00 and 59: 32, 00, 18, and 47. This implies that the randomly selected times during the first three peak business hours are 2:32 P.M., 3:00 P.M., and 4:18 P.M. What is the randomly selected time during the fourth peak business hour?

7.6 In an article entitled "Turned Off" in the June 2–4, 1995, issue of *USA Weekend,* Don Olmsted and Gigi Anders reported results of a survey where readers were invited to write in and express their opinions about sex and violence on television. The results showed that 96 percent of respondents were very or somewhat concerned about sex on TV, and 97 percent of respondents were very or somewhat concerned about violence on TV. Do you think that these results could be generalized to all television viewers in 1995? Why or why not?

7.2 The Sampling Distribution of the Sample Mean

LO7-2 Describe and use the sampling distribution of the sample mean.

Introductory ideas and basic properties Suppose that we are about to randomly select a sample of n elements (for example, cars) from a population of elements. Also, suppose that for each sampled element we will measure the value of a characteristic of interest. (For example, we might measure the mileage of each sampled car.) Before we actually select the sample, there are many different samples of n elements and corresponding measurements that we might potentially obtain. Because different samples of measurements generally have different sample means, there are many different sample means that we might potentially obtain. It follows that, *before we draw the sample, the sample mean $\bar{x}$ is a random variable.*

The **sampling distribution of the sample mean** $\bar{x}$ is the probability distribution of the population of all possible sample means that could be obtained from all possible samples of the same size.

In order to illustrate the sampling distribution of the sample mean, we begin with an example that is based on the authors' conversations with University Chrysler/Jeep of Oxford, Ohio. In order to keep the example simple, we have used simplified car mileages to help explain the concepts.

EXAMPLE 7.2 The Car Mileage Case: Estimating Mean Mileage

This is the first year that the automaker has offered its new midsize model for sale to the public. However, last year the automaker made six preproduction cars of this new model. Two of these six cars were randomly selected for testing, and the other four were sent to auto shows at which the new model was introduced to the news media and the public. As is standard industry practice, the automaker did not test the four auto show cars before or during the five months these auto shows were held because testing can potentially harm the appearance of the cars.

In order to obtain a preliminary estimate—to be reported at the auto shows—of the midsize model's combined city and highway driving mileage, the automaker subjected the two cars selected for testing to the EPA mileage test. When this was done, the cars obtained mileages of 30 mpg and 32 mpg. The mean of this sample of mileages is

$$\bar{x} = \frac{30 + 32}{2} = 31 \text{ mpg}$$

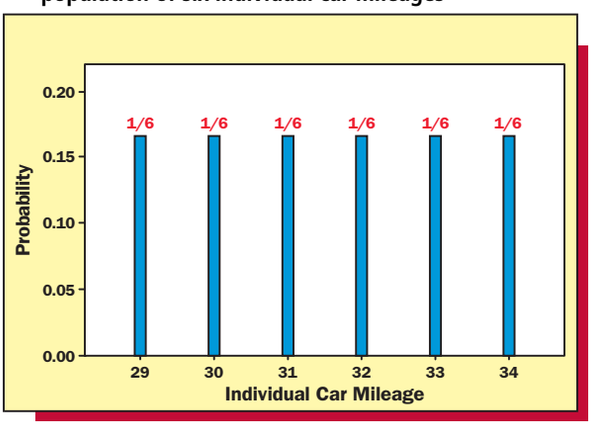
This sample mean is the point estimate of the mean mileage μ for the population of six preproduction cars and is the preliminary mileage estimate for the new midsize model that was reported at the auto shows.

TABLE 7.2 A Probability Distribution Describing the Population of Six Individual Car Mileages

Individual Car Mileage	29	30	31	32	33	34
Probability	1/6	1/6	1/6	1/6	1/6	1/6

FIGURE 7.1 A Comparison of Individual Car Mileages and Sample Means

(a) A graph of the probability distribution describing the population of six individual car mileages



(b) A graph of the probability distribution describing the population of 15 sample means

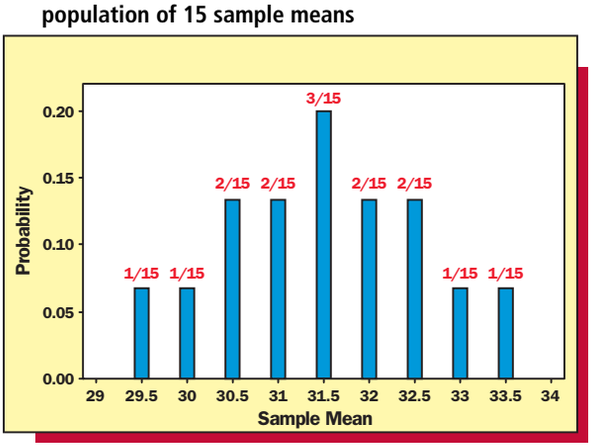


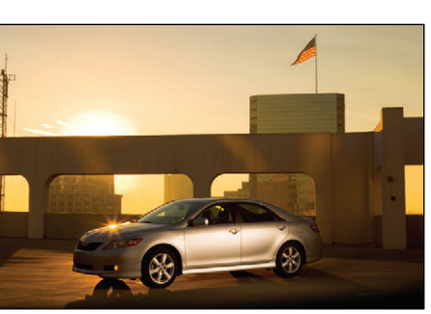
TABLE 7.3 The Population of Sample Means

(a) The population of the 15 samples of $n = 2$ car mileages and corresponding sample means

Sample	Car Mileages	Sample Mean
1	29, 30	29.5
2	29, 31	30
3	29, 32	30.5
4	29, 33	31
5	29, 34	31.5
6	30, 31	30.5
7	30, 32	31
8	30, 33	31.5
9	30, 34	32
10	31, 32	31.5
11	31, 33	32
12	31, 34	32.5
13	32, 33	32.5
14	32, 34	33
15	33, 34	33.5

(b) A probability distribution describing the population of 15 sample means: the sampling distribution of the sample mean

Sample Mean	Frequency	Probability
29.5	1	1/15
30	1	1/15
30.5	2	2/15
31	2	2/15
31.5	3	3/15
32	2	2/15
32.5	2	2/15
33	1	1/15
33.5	1	1/15



When the auto shows were over, the automaker decided to further study the new midsize model by subjecting the four auto show cars to various tests. When the EPA mileage test was performed, the four cars obtained mileages of 29 mpg, 31 mpg, 33 mpg, and 34 mpg. Thus, the mileages obtained by the six preproduction cars were 29 mpg, 30 mpg, 31 mpg, 32 mpg, 33 mpg, and 34 mpg. The probability distribution of this population of six individual car mileages is given in Table 7.2 and graphed in Figure 7.1(a). The mean of the population of car mileages is

$$\mu = \frac{29 + 30 + 31 + 32 + 33 + 34}{6} = 31.5 \text{ mpg}$$

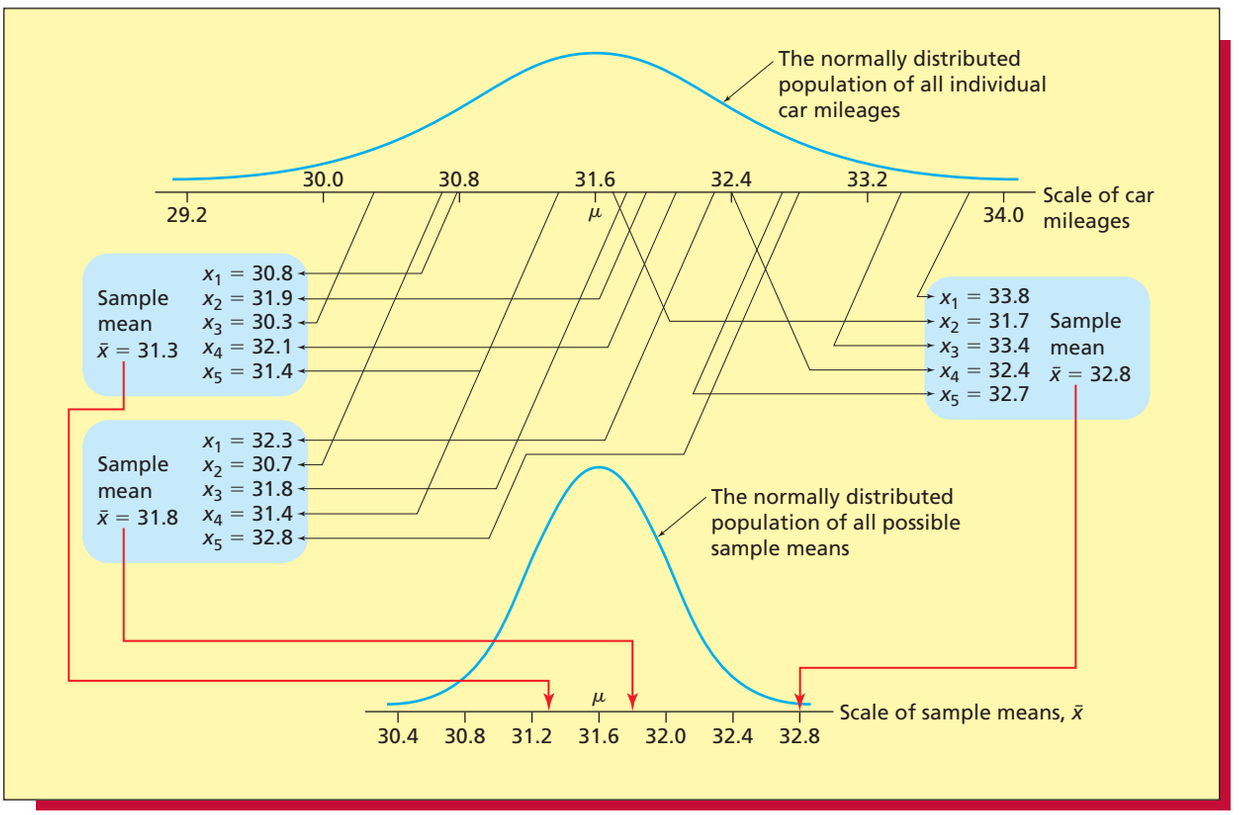
Note that the point estimate $\bar{x} = 31$ mpg that was reported at the auto shows is .5 mpg less than the true population mean μ of 31.5 mpg. Of course, different samples of two cars and corresponding mileages would have given different sample means. There are, in total, 15 samples of two mileages that could have been obtained by randomly selecting two cars from the population of six cars and subjecting the cars to the EPA mileage test. These samples correspond to the 15 combinations of two mileages that can be selected from the six mileages: 29, 30, 31, 32, 33, and 34. The samples are given, along with their means, in Table 7.3(a).

In order to find the probability distribution of the population of sample means, note that different sample means correspond to different numbers of samples. For example, because the sample mean of 31 mpg corresponds to 2 out of 15 samples—the sample (29, 33) and the sample (30, 32)—the probability of obtaining a sample mean of 31 mpg is $2/15$. If we analyze all of the sample means in a similar fashion, we find that the probability distribution of the population of sample means is as given in Table 7.3(b). This distribution is the *sampling distribution of the sample mean.* A graph of this distribution is shown in Figure 7.1(b) and illustrates the accuracies of the different possible sample means as point estimates of the population mean. For example, whereas 3 out of 15 sample means exactly equal the population mean of 31.5 mpg, other sample means differ from the population mean by amounts varying from .5 mpg to 2 mpg.

As illustrated in Example 7.2, one of the purposes of the sampling distribution of the sample mean is to tell us how accurate the sample mean is likely to be as a point estimate of the population mean. Because the population of six individual car mileages in Example 7.2 is small, we were able (after the auto shows were over) to test all six cars, determine the values of the six car mileages, and calculate the population mean mileage. Often, however, the population of individual measurements under consideration is very large—either a large finite population or an infinite population. In this case, it would be impractical or impossible to determine the values of all of the population measurements and calculate the population mean. Instead, we randomly select a sample of individual measurements from the population and use the mean of this sample as the point estimate of the population mean. Moreover, although it would be impractical or impossible to list all of the many (perhaps trillions of) different possible sample means that could be obtained if the sampled population is very large, statisticians know various theoretical properties about the sampling distribution of these sample means. Some of these theoretical properties are intuitively illustrated by the sampling distribution of the 15 sample means in Example 7.2. Specifically, suppose that we will randomly select a sample of n individual measurements from a population of individual measurements having mean μ and standard deviation σ. Then, it can be shown that:

- **In many situations, the distribution of the population of all possible sample means looks, at least roughly, like a normal curve.** For example, consider Figure 7.1. This figure shows that, while the distribution of the population of six individual car mileages is a uniform distribution, the distribution of the population of 15 sample means has a somewhat bell-shaped appearance. Noting, however, that this rough bell-shaped appearance is not extremely close to the appearance of a normal curve, we wish to know when the distribution of all possible sample means is exactly or approximately normally distributed. Answers to this question will begin on the next page.
- **The mean, $\mu_{\bar{x}}$, of the population of all possible sample means is equal to μ, the mean of the population from which we will select the sample.** For example, the mean, $\mu_{\bar{x}}$, of the population of 15 sample means in Table 7.3(a) can be calculated by adding up the 15 sample means, which gives 472.5, and dividing by 15. That is, $\mu_{\bar{x}} = 472.5/15 = 31.5$, which is the same as μ, the mean of the population of six individual car mileages in Table 7.2. Furthermore, because $\mu_{\bar{x}}$ equals μ, we call the sample mean an **unbiased point estimate** of the population mean. This unbiasedness property says that, although most of the possible sample means that we might obtain are either above or below the population mean, there is no systematic tendency for the sample mean to overestimate or underestimate the population mean. That is, although we will randomly select only one sample, the unbiased sample mean is "correct on the average" in all possible samples.
- **The standard deviation, $\sigma_{\bar{x}}$, of the population of all posssible sample means is less than σ, the standard deviation of the population from which we will select the sample.** This is illustrated in Figure 7.1, which shows that the distribution of all possible sample means is less spread out than the distribution of all individual car mileages. Intuitively, we see that $\sigma_{\bar{x}}$ is smaller than σ because each possible sample mean is an average of n measurements (n equals 2 in Table 7.3). Thus, **each sample mean *averages out* high and low sample measurements and can be expected to be closer to the population mean μ than many of the individual population measurements would be.** It follows that the different possible sample means are more closely clustered around μ than are the individual population measurements.

FIGURE 7.2 **The Normally Distributed Population of All Individual Car Mileages and the Normally Distributed Population of All Possible Sample Means**



- **If the population from which we will select the sample is normally distributed, then for any sample size *n* the population of all possible sample means is normally distributed.** For example, consider the population of the mileages of all of the new midsize cars that could potentially be produced by this year's manufacturing process. As discussed in Chapter 1, we consider this population to be an infinite population because the automaker could always make "one more car." Moreover, assume that (as will be verified in a later example) this infinite population of all individual car mileages is normally distributed (see the top curve in Figure 7.2), and assume that the automaker will randomly select a sample of $n = 5$ cars, test them as prescribed by the EPA, and calculate the mean of the resulting sample mileages. It then follows that the population of all possible sample means that the automaker might obtain is normally distributed. This is illustrated in Figure 7.2 (see the bottom curve), which also depicts the unbiasedness of the sample mean $\bar{x}$ as a point estimate of the population mean μ. Specifically, note that the normally distributed population of all possible sample means is centered over μ, the mean of the normally distributed population of all individual car mileages. This says that, although most of the possible sample means that the automaker might obtain are either above or below the true population mean μ, the mean of all of the possible sample means that the automaker might obtain, $\mu_{\bar{x}}$, is equal to μ. To make Figure 7.2 easier to understand, we have assumed that the true value of the population mean mileage μ is 31.6 mpg. Of course, the true value of μ is really unknown. Our objective is to estimate μ, and to do this effectively, it is important to know more about $\sigma_{\bar{x}}$, the standard deviation of the population of all possible sample means. We will see that having a formula for $\sigma_{\bar{x}}$ will help us to choose a sample size n that is likely to make the sample mean $\bar{x}$ an accurate point estimate of the population mean μ. That is, although Figure 7.2 is based on selecting a sample of $n = 5$ car mileages, perhaps we should select a larger sample of, say, 50 or more car mileages. The following summary box gives a formula for $\sigma_{\bar{x}}$ and

also summarizes other previously discussed facts about the probability distribution of the population of all possible sample means.

The Sampling Distribution of $\bar{x}$

Assume that the population from which we will randomly select a sample of n measurements has mean μ and standard deviation σ. Then, the population of all possible sample means

1. Has a normal distribution, if the sampled population has a normal distribution.
2. Has mean $\mu_{\bar{x}} = \mu$.
3. Has standard deviation $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$.

The formula for $\sigma_{\bar{x}}$ in (3) holds exactly if the sampled population is infinite. If the sampled population is finite, this formula holds approximately under conditions to be discussed at the end of this section.

Stated equivalently, the sampling distribution of $\bar{x}$ has mean $\mu_{\bar{x}} = \mu$, has standard deviation $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ (if the sampled population is infinite), and is a normal distribution (if the sampled population has a normal distribution).[2]

The third result in the summary box says that, if the sampled population is infinite, then $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. In words, $\sigma_{\bar{x}}$, the standard deviation of the population of all possible sample means, equals σ, the standard deviation of the sampled population, divided by the square root of the sample size n. It follows that, if the sample size n is greater than 1, then $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ is smaller than σ. This is illustrated in Figure 7.2, where the sample size n is 5. Specifically, note that the normally distributed population of all possible sample means is less spread out than the normally distributed population of all individual car mileages. Furthermore, the formula $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ says that $\sigma_{\bar{x}}$ decreases as n increases. That is, intuitively, when the sample size is larger, each possible sample averages more observations. Therefore, the resulting different possible sample means will differ from each other by less and thus will become more closely clustered around the population mean. It follows that, if we take a larger sample, we are more likely to obtain a sample mean that is near the population mean.

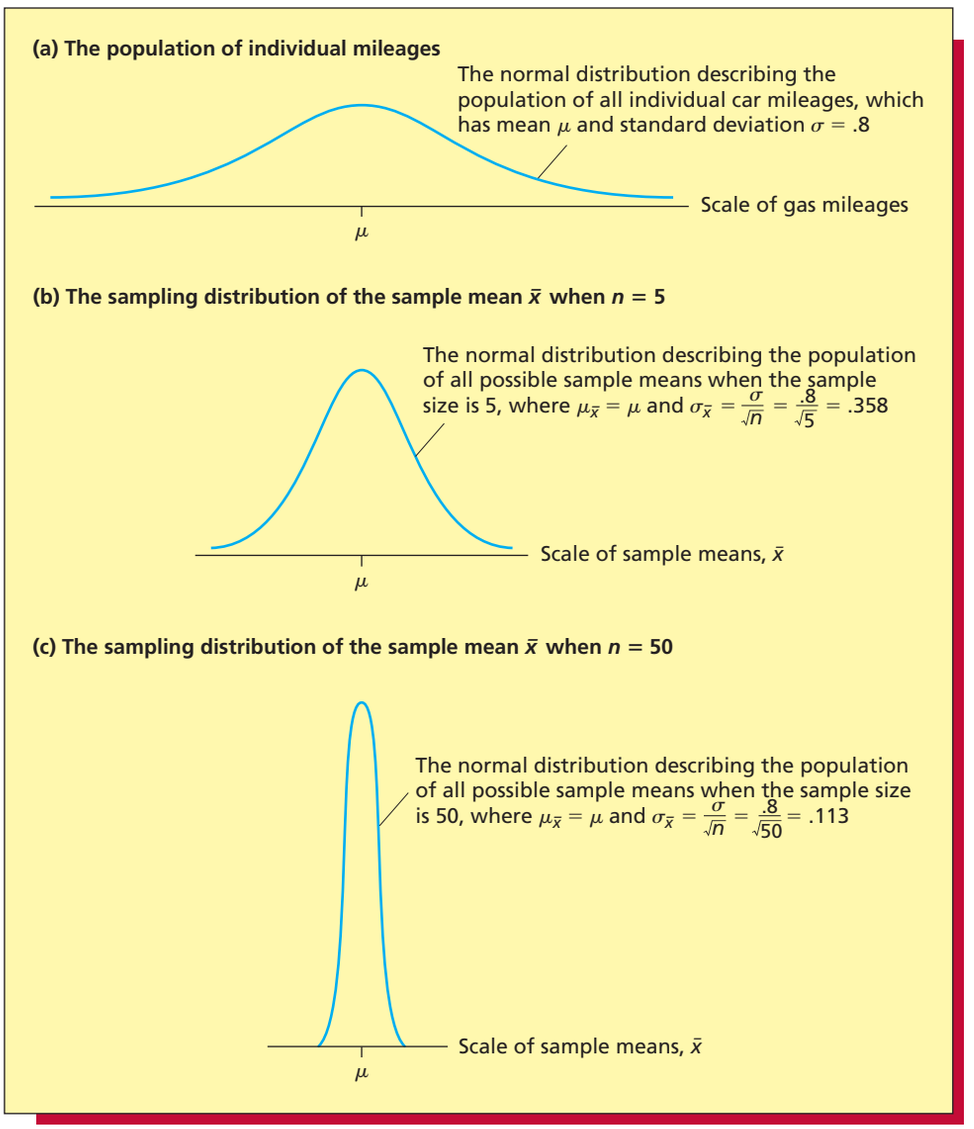
In order to better see how $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ decreases as the sample size n increases, we will compute some values of $\sigma_{\bar{x}}$ in the context of the car mileage case. To do this, we will assume that, although we do not know the true value of the population mean μ, we do know the true value of the population standard deviation σ. Here, knowledge of σ might be based on theory or history related to the population under consideration. For example, because the automaker has been working to improve gas mileages, we cannot assume that we know the true value of the population mean mileage μ for the new midsize model. However, engineering data might indicate that the spread of individual car mileages for the automaker's midsize cars is the same from model to model and year to year. Therefore, if the mileages for previous models had a standard deviation equal to .8 mpg, it might be reasonable to assume that the standard deviation of the mileages for the new model will also equal .8 mpg. Such an assumption would, of course, be questionable, and in most real-world situations there would probably not be an actual basis for knowing σ. However, assuming that σ is known will help us to illustrate sampling distributions, and in later chapters we will see what to do when σ is unknown.

EXAMPLE 7.3 The Car Mileage Case: Estimating Mean Mileage

Part 1: Basic concepts Consider the infinite population of the mileages of all of the new midsize cars that could potentially be produced by this year's manufacturing process. If we assume that this population is normally distributed with mean μ and standard deviation $\sigma = .8$ (see Figure 7.3(a)), and if the automaker will randomly select a sample of n cars and test them as prescribed by the EPA, it follows that the population of all possible sample means is normally distributed with mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \sigma/\sqrt{n} = .8/\sqrt{n}$. In order to show

[2] In optional Section 7.6 we derive the formulas $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \sigma/\sqrt{n}$.

FIGURE 7.3 **A Comparison of (1) the Population of All Individual Car Mileages, (2) the Sampling Distribution of the Sample Mean $\bar{x}$ When $n = 5$, and (3) the Sampling Distribution of the Sample Mean $\bar{x}$ When $n = 50$**



that a larger sample is more likely to give a more accurate point estimate $\bar{x}$ of μ, compare taking a sample of size $n = 5$ with taking a sample of size $n = 50$. If $n = 5$, then

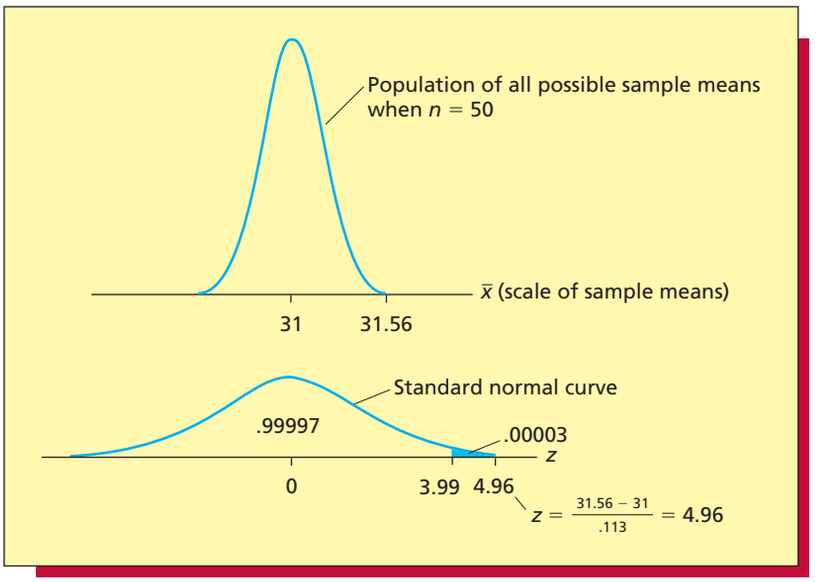
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{.8}{\sqrt{5}} = .358$$

and it follows (by the Empirical Rule) that 95.44 percent of all possible sample means are within plus or minus $2\sigma_{\bar{x}} = 2(.358) = .716$ mpg of the population mean μ. If $n = 50$, then

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{.8}{\sqrt{50}} = .113$$

and it follows that 95.44 percent of all possible sample means are within plus or minus $2\sigma_{\bar{x}} = 2(.113) = .226$ mpg of the population mean μ. Therefore, if $n = 50$, the different possible sample means that the automaker might obtain will be more closely clustered around μ than they will be if $n = 5$ (see Figures 7.3(b) and (c)). This implies that the larger sample of size $n = 50$ is more likely to give a sample mean $\bar{x}$ that is near μ.

FIGURE 7.4 The Probability That $\bar{x} \geq 31.56$ When $\mu = 31$ in the Car Mileage Case



Part 2: Statistical inference Recall from Chapter 3 that the automaker has randomly selected a sample of $n = 50$ mileages, which has mean $\bar{x} = 31.56$. We now ask the following question: If the population mean mileage μ exactly equals 31 mpg (the minimum standard for the tax credit), what is the probability of observing a sample mean mileage that is greater than or equal to 31.56 mpg? To find this probability, recall from Chapter 2 that a histogram of the 50 mileages indicates that the population of all individual mileages is normally distributed. Assuming that the population standard deviation σ is known to equal .8 mpg, it follows that the sampling distribution of the sample mean $\bar{x}$ is a normal distribution, with mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \sigma/\sqrt{n} = .8/\sqrt{50} = .113$. Therefore,

$$P(\bar{x} \geq 31.56 \quad \text{if} \quad \mu = 31) = P\left(z \geq \frac{31.56 - \mu_{\bar{x}}}{\sigma_{\bar{x}}}\right) = P\left(z \geq \frac{31.56 - 31}{.113}\right)$$
$$= P(z \geq 4.96)$$

To find $P(z \geq 4.96)$, notice that the largest z value given in Table A.3 (page 791) is 3.99, which gives a right-hand tail area of .00003. Therefore, because $P(z \geq 3.99) = .00003$, it follows that $P(z \geq 4.96)$ is less than .00003 (see Figure 7.4). The fact that this probability is less than .00003 says that, if μ equals 31, then fewer than 3 in 100,000 of all possible sample means are at least as large as the sample mean $\bar{x} = 31.56$ that we have actually observed. Therefore, if we are to believe that μ equals 31, then we must believe that we have observed a sample mean that can be described as a smaller than 3 in 100,000 chance. Because it is extremely difficult to believe that such a small chance would occur, we have extremely strong evidence that μ does not equal 31 and that μ is, in fact, larger than 31. This evidence would probably convince the federal government that the midsize model's mean mileage μ exceeds 31 mpg and thus that the midsize model deserves the tax credit.

To conclude this subsection, it is important to make two comments. First, the formula $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ follows, in theory, from the formula for $\sigma^2_{\bar{x}}$, the variance of the population of all possible sample means. The formula for $\sigma^2_{\bar{x}}$ is $\sigma^2_{\bar{x}} = \sigma^2/n$. Second, in addition to holding exactly if the sampled population is infinite, **the formula $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ holds approximately if the sampled population is finite and much larger than (say, at least 20 times) the size of the sample.** For example, if we define the population of the mileages of all new midsize cars to be the population of the mileages of all cars that will actually be produced this year, then the population is

finite. However, the population would be very large—certainly at least as large as 20 times any reasonable sample size. For example, if the automaker produces 100,000 new midsize cars this year, and if we randomly select a sample of $n = 50$ of these cars, then the population size of 100,000 is more than 20 times the sample size of 50 (which is 1,000). It follows that, even though the population is finite and thus the formula $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ would not hold exactly, this formula would hold approximately. The exact formula for $\sigma_{\bar{x}}$ when the sampled population is finite is given in a technical note at the end of this section. It is important to use this exact formula if the sampled population is finite and less than 20 times the size of the sample. **However, with the exception of the populations considered in the technical note and in Section 8.5, we will assume that all of the remaining populations to be discussed in this book are either infinite or finite and at least 20 times the size of the sample. Therefore, it will be appropriate to use the formula $\sigma_{\bar{x}} = \sigma/\sqrt{n}$.**

LO7-3 Explain and use the Central Limit Theorem.

Sampling a nonnormally distributed population: The Central Limit Theorem We now consider what can be said about the sampling distribution of $\bar{x}$ when the sampled population is not normally distributed. First, as previously stated, the fact that $\mu_{\bar{x}} = \mu$ is still true. Second, as also previously stated, the formula $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ is exactly correct if the sampled population is infinite and is approximately correct if the sampled population is finite and much larger than (say, at least 20 times as large as) the sample size. Third, an extremely important result called the **Central Limit Theorem** tells us that, **if the sample size n is large, then the sampling distribution of $\bar{x}$ is approximately normal, even if the sampled population is not normally distributed.**

The Central Limit Theorem

If the sample size n is sufficiently large, then the population of all possible sample means is approximately normally distributed (with mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \sigma/\sqrt{n}$), no matter what probability distribution describes the sampled population. Furthermore, the larger the sample size n is, the more nearly normally distributed is the population of all possible sample means.

The Central Limit Theorem is illustrated in Figure 7.5 for several population shapes. Notice that as the sample size increases (from 2 to 6 to 30), the populations of all possible sample means become more nearly normally distributed. This figure also illustrates that, as the sample size increases, the spread of the distribution of all possible sample means decreases (remember that this spread is measured by $\sigma_{\bar{x}}$, which decreases as the sample size increases).

How large must the sample size be for the sampling distribution of $\bar{x}$ to be approximately normal? In general, the more skewed the probability distribution of the sampled population, the larger the sample size must be for the population of all possible sample means to be approximately normally distributed. For some sampled populations, particularly those described by symmetric distributions, the population of all possible sample means is approximately normally distributed for a fairly small sample size. In addition, studies indicate that, **if the sample size is at least 30, then for most sampled populations the population of all possible sample means is approximately normally distributed.** In this book, whenever the sample size n is at least 30, we will assume that the sampling distribution of $\bar{x}$ is approximately a normal distribution. Of course, if the sampled population is exactly normally distributed, the sampling distribution of $\bar{x}$ is exactly normal for any sample size.

We can see the shapes of sampling distributions such as those illustrated in Figure 7.5 by using computer simulation. Specifically, for a population having a particular probability distribution, we can have the computer draw a given number of samples of n observations, compute the mean of each sample, and arrange the sample means into a histogram. To illustrate this, consider the upper portion of Figure 7.6, which shows the exponential distribution describing the hospital emergency room interarrival times discussed in Example 6.11 (page 253). Figure 7.6(a) gives the results of a simulation in which MINITAB randomly selected 1,000 samples of $n = 5$ interarrival times from this exponential distribution, calculated the mean of each sample, and arranged the 1,000 sample means into a histogram. Figure 7.6(b) gives the results of a simulation in which

FIGURE 7.5 The Central Limit Theorem Says That the Larger the Sample Size Is, the More Nearly Normally Distributed Is the Population of All Possible Sample Means

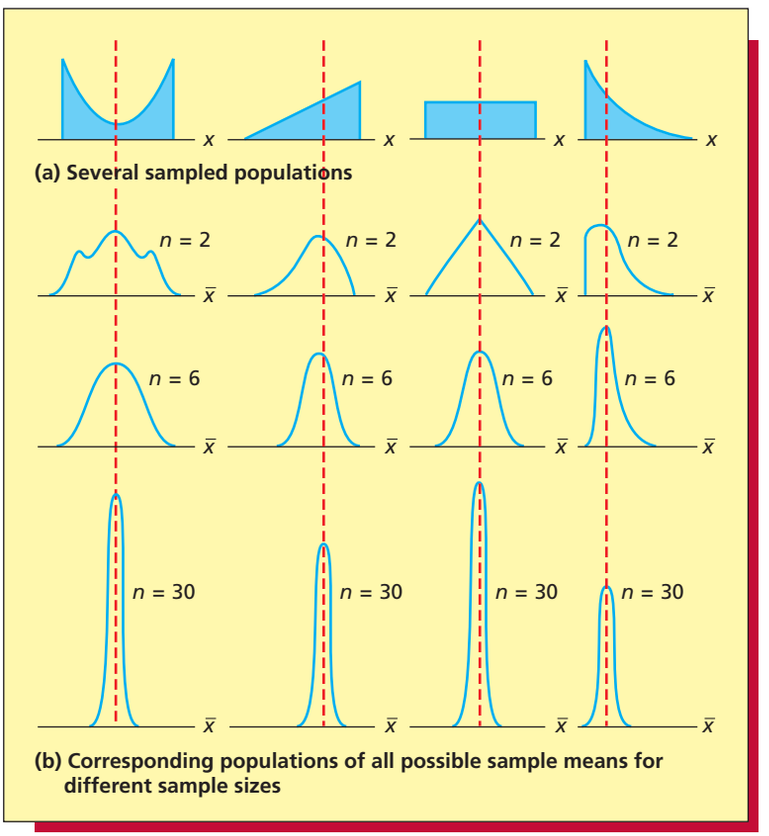
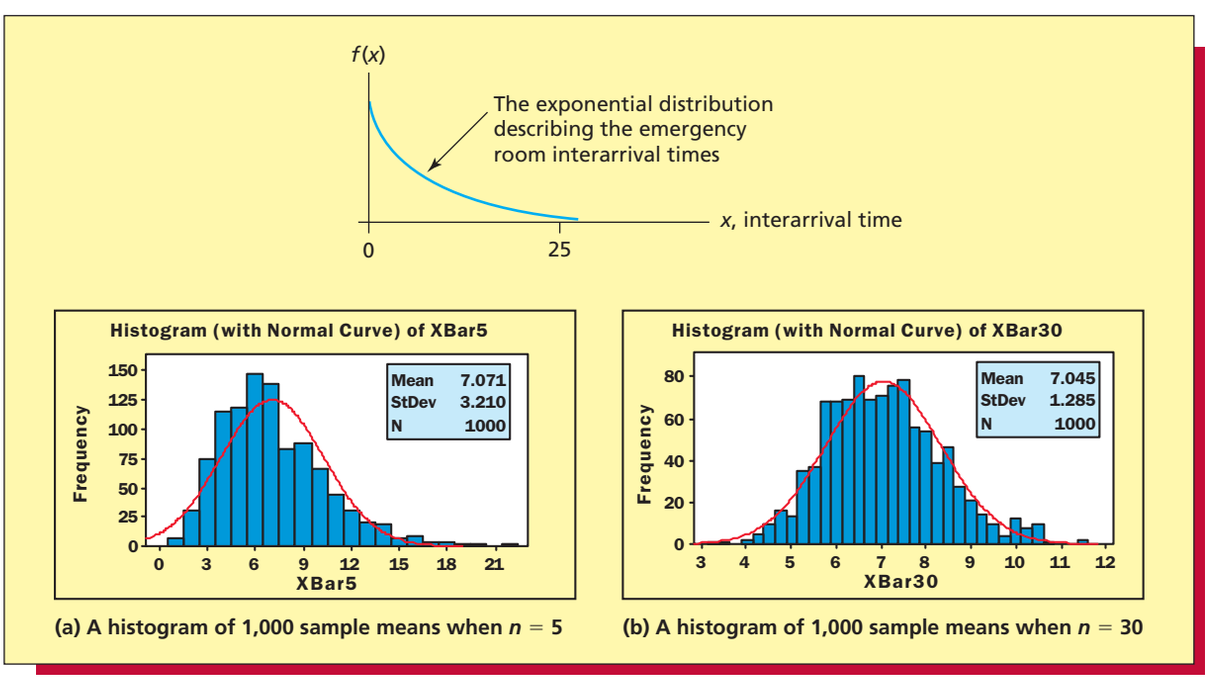


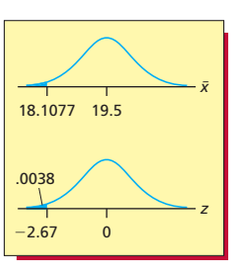
FIGURE 7.6 Simulating the Sampling Distribution of the Sample Mean When Sampling from an Exponential Distribution



MINITAB randomly selected 1,000 samples of $n = 30$ interarrival times from the exponential distribution, calculated the mean of each sample, and arranged the 1,000 sample means into a histogram. Note that, whereas the histogram in Figure 7.6(a) is somewhat skewed to the right, the histogram in Figure 7.6(b) appears approximately bell-shaped. Therefore, we might conclude that when we randomly select a sample of n observations from an exponential distribution, the sampling distribution of the sample mean is somewhat skewed to the right when $n = 5$ and is approximately normal when $n = 30$.

EXAMPLE 7.4 The e-billing Case: Reducing Mean Bill Payment Time

Recall that a management consulting firm has installed a new computer-based electronic billing system in a Hamilton, Ohio, trucking company. Because of the previously discussed advantages of the new billing system, and because the trucking company's clients are receptive to using this system, the management consulting firm believes that the new system will reduce the mean bill payment time by more than 50 percent. The mean payment time using the old billing system was approximately equal to, but no less than, 39 days. Therefore, if μ denotes the new mean payment time, the consulting firm believes that μ will be less than 19.5 days. To assess whether μ is less than 19.5 days, the consulting firm has randomly selected a sample of $n = 65$ invoices processed using the new billing system and has determined the payment times for these invoices. The mean of the 65 payment times is $\bar{x} = 18.1077$ days, which is less than 19.5 days. Therefore, we ask the following question: If the population mean payment time is 19.5 days, what is the probability of observing a sample mean payment time that is less than or equal to 18.1077 days? To find this probability, recall from Chapter 2 that a histogram of the 65 payment times indicates that the population of all payment times is skewed with a tail to the right. However, the Central Limit Theorem tells us that, because the sample size $n = 65$ is large, the sampling distribution of $\bar{x}$ is approximately a normal distribution with mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. Moreover, whereas this is the first time that the consulting firm has implemented an electronic billing system for a trucking company, the firm has installed electronic billing systems for clients in other industries. Analysis of results from these installations shows that, although the population mean payment time μ varies from company to company, the population standard deviation σ of payment times is the same for different applications and equals 4.2 days. Assuming that σ also equals 4.2 days for the trucking company, it follows that $\sigma_{\bar{x}}$ equals $4.2/\sqrt{65} = .5209$ and that



$$P(\bar{x} \leq 18.1077 \text{ if } \mu = 19.5) = P\left(z \leq \frac{18.1077 - 19.5}{.5209}\right) = P(z \leq -2.67)$$

which is the area under the standard normal curve to the left of −2.67. The normal table tells us that this area equals .0038. This probability says that, if μ equals 19.5, then only .0038 of all possible sample means are at least as small as the sample mean $\bar{x} = 18.1077$ that we have actually observed. Therefore, if we are to believe that μ equals 19.5, we must believe that we have observed a sample mean that can be described as a 38 in 10,000 chance. It is very difficult to believe that such a small chance would occur, so we have very strong evidence that μ does not equal 19.5 and is, in fact, less than 19.5. We conclude that the new billing system has reduced the mean bill payment time by more than 50 percent.

Unbiasedness and minimum-variance estimates

Recall that a sample statistic is any descriptive measure of the sample measurements. For instance, the sample mean $\bar{x}$ is a statistic, and so are the sample median, the sample variance s^2, and the sample standard deviation s. Not only do different samples give different values of $\bar{x}$, different samples also give different values of the median, s^2, s, or any other statistic. It follows that, *before we draw the sample, any sample statistic is a random variable,* and

The **sampling distribution of a sample statistic** is the probability distribution of the population of all possible values of the sample statistic.

In general, we wish to estimate a population parameter by using a sample statistic that is what we call an *unbiased point estimate* of the parameter.

A sample statistic is an **unbiased point estimate** of a population parameter if the mean of the population of all possible values of the sample statistic equals the population parameter.

For example, we use the sample mean $\bar{x}$ as the point estimate of the population mean μ because **$\bar{x}$ is an unbiased point estimate of μ.** That is, $\mu_{\bar{x}} = \mu$. In words, the average of all the different possible sample means (that we could obtain from all the different possible samples) equals μ.

Although we want a sample statistic to be an unbiased point estimate of the population parameter of interest, we also want the statistic to have a small standard deviation (and variance). That is, we wish the different possible values of the sample statistic to be closely clustered around the population parameter. If this is the case, when we actually randomly select one sample and compute the sample statistic, its value is likely to be close to the value of the population parameter. Furthermore, some general results apply to estimating the mean μ of a normally distributed population. In this situation, it can be shown that both the sample mean and the sample median are unbiased point estimates of μ. In fact, there are many unbiased point estimates of μ. However, it can be shown that the variance of the population of all possible sample means is smaller than the variance of the population of all possible values of any other unbiased point estimate of μ. For this reason, **we call the sample mean a minimum-variance unbiased point estimate of μ.** When we use the sample mean as the point estimate of μ, we are more likely to obtain a point estimate close to μ than if we used any other unbiased sample statistic as the point estimate of μ. This is one reason why we use the sample mean as the point estimate of the population mean.

We next consider estimating the population variance σ^2. It can be shown that if the sampled population is infinite, then **s^2 is an unbiased point estimate of σ^2.** That is, the average of all the different possible sample variances that we could obtain (from all the different possible samples) is equal to σ^2. This is why we use a divisor equal to $n - 1$ rather than n when we estimate σ^2. It can be shown that, if we used n as the divisor when estimating σ^2, we would not obtain an unbiased point estimate of σ^2. When the population is finite, s^2 may be regarded as an approximately unbiased estimate of σ^2 as long as the population is fairly large (which is usually the case).

It would seem logical to think that, because s^2 is an unbiased point estimate of σ^2, s should be an unbiased point estimate of σ. This seems plausible, but it is not the case. There is no easy way to calculate an unbiased point estimate of σ. Because of this, the usual practice is to use s as the point estimate of σ (even though it is not an unbiased estimate).

This ends our discussion of the theory of point estimation. It suffices to say that in this book we estimate population parameters by using sample statistics that statisticians generally agree are best. Whenever possible, these sample statistics are unbiased point estimates and have small variances.

Technical Note: If we randomly select a sample of size n without replacement from a finite population of size N, then it can be shown that $\sigma_{\bar{x}} = (\sigma/\sqrt{n})\sqrt{(N-n)/(N-1)}$, where the quantity $\sqrt{(N-n)/(N-1)}$ is called the **finite population multiplier.** If the size of the sampled population is at least 20 times the size of the sample (that is, **if $N \geq 20n$**), **then the finite population multiplier is approximately equal to one, and $\sigma_{\bar{x}}$ approximately equals $\sigma/\sqrt{n}$.** However, if the population size N is smaller than 20 times the size of the sample, then the finite population multiplier is substantially less than one, and we must include this multiplier in the calculation of $\sigma_{\bar{x}}$. For instance, in Example 7.2, where the standard deviation σ of the population of $N = 6$ car mileages can be calculated to be 1.7078, and where $N = 6$ is only three times the sample size $n = 2$, it follows that

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \left(\frac{1.7078}{\sqrt{2}}\right)\sqrt{\frac{6-2}{6-1}} = 1.2076(.8944) = 1.08$$

We will see how this formula can be used to make statistical inferences in Section 8.5.

Exercises for Section 7.2

CONCEPTS

connect

7.7 Suppose that we will randomly select a sample of four measurements from a larger population of measurements. The sampling distribution of the sample mean $\bar{x}$ is the probability distribution of a population. In your own words, describe the elements in this population.

7.8 What does the Central Limit Theorem tell us about the sampling distribution of the sample mean?

METHODS AND APPLICATIONS

7.9 Suppose that we will take a random sample of size n from a population having mean μ and standard deviation σ. For each of the following situations, find the mean, variance, and standard deviation of the sampling distribution of the sample mean $\bar{x}$:

a $\mu = 10, \quad \sigma = 2, \quad n = 25$

b $\mu = 500, \quad \sigma = .5, \quad n = 100$

c $\mu = 3, \quad \sigma = .1, \quad n = 4$

d $\mu = 100, \quad \sigma = 1, \quad n = 1{,}600$

7.10 For each situation in Exercise 7.9, find an interval that contains (approximately or exactly) 99.73 percent of all the possible sample means. In which cases must we assume that the population is normally distributed? Why?

7.11 Suppose that we will randomly select a sample of 64 measurements from a population having a mean equal to 20 and a standard deviation equal to 4.

a Describe the shape of the sampling distribution of the sample mean $\bar{x}$. Do we need to make any assumptions about the shape of the population? Why or why not?

b Find the mean and the standard deviation of the sampling distribution of the sample mean $\bar{x}$.

c Calculate the probability that we will obtain a sample mean greater than 21; that is, calculate $P(\bar{x} > 21)$. Hint: Find the z value corresponding to 21 by using $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$ because we wish to calculate a probability about $\bar{x}$. Then sketch the sampling distribution and the probability.

d Calculate the probability that we will obtain a sample mean less than 19.385; that is, calculate $P(\bar{x} < 19.385)$.

THE GAME SHOW CASE

Exercises 7.12 through 7.16 are based on the following situation.

Congratulations! You have just won the question-and-answer portion of a popular game show and will now be given an opportunity to select a grand prize. The game show host shows you a large revolving drum containing four identical white envelopes that have been thoroughly mixed in the drum. Each of the envelopes contains one of four checks made out for grand prizes of 20, 40, 60, and 80 thousand dollars. Usually, a contestant reaches into the drum, selects an envelope, and receives the grand prize in the envelope. Tonight, however, is a special night. You will be given the choice of either selecting one envelope or selecting two envelopes and receiving the average of the grand prizes in the two envelopes. If you select one envelope, the probability is $1/4$ that you will receive any one of the individual grand prizes 20, 40, 60, and 80 thousand dollars. To see what could happen if you select two envelopes, do Exercises 7.12 through 7.16.

7.12 There are six combinations, or samples, of two grand prizes that can be randomly selected from the four grand prizes 20, 40, 60, and 80 thousand dollars. Four of these samples are (20, 40), (20, 60), (20, 80), and (40, 60). Find the other two samples.

7.13 Find the mean of each sample in Exercise 7.12.

7.14 Find the probability distribution of the population of six sample mean grand prizes.

7.15 If you select two envelopes, what is the probability that you will receive a sample mean grand prize of at least 50 thousand dollars?

7.16 Compare the probability distribution of the four individual grand prizes with the probability distribution of the six sample mean grand prizes. Would you select one or two envelopes? Why? Note: There is no one correct answer. It is a matter of opinion.

7.17 THE BANK CUSTOMER WAITING TIME CASE DS WaitTime

Recall that the bank manager wants to show that the new system reduces typical customer waiting times to less than six minutes. One way to do this is to demonstrate that the mean of the population of all customer waiting times is less than 6. Letting this mean be μ, in this exercise we wish to investigate whether the sample of 100 waiting times provides evidence to support the claim that μ is less than 6.

For the sake of argument, we will begin by assuming that μ equals 6, and we will then attempt to use the sample to contradict this assumption in favor of the conclusion that μ is less than 6.

Recall that the mean of the sample of 100 waiting times is $\bar{x} = 5.46$ and assume that σ, the standard deviation of the population of all customer waiting times, is known to be 2.47.

a Consider the population of all possible sample means obtained from random samples of 100 waiting times. What is the shape of this population of sample means? That is, what is the shape of the sampling distribution of $\bar{x}$? Why is this true?

b Find the mean and standard deviation of the population of all possible sample means when we assume that μ equals 6.

c The sample mean that we have actually observed is $\bar{x} = 5.46$. Assuming that μ equals 6, find the probability of observing a sample mean that is less than or equal to $\bar{x} = 5.46$.

d If μ equals 6, what percentage of all possible sample means are less than or equal to 5.46? Because we have actually observed a sample mean of $\bar{x} = 5.46$, is it more reasonable to believe that (1) μ equals 6 and we have observed one of the sample means that is less than or equal to 5.46 when μ equals 6, or (2) that we have observed a sample mean less than or equal to 5.46 because μ is less than 6? Explain. What do you conclude about whether the new system has reduced the typical customer waiting time to less than six minutes?

7.18 THE VIDEO GAME SATISFACTION RATING CASE VideoGame

Recall that a customer is considered to be very satisfied with his or her XYZ Box video game system if the customer's composite score on the survey instrument is at least 42. One way to show that customers are typically very satisfied is to show that the mean of the population of all satisfaction ratings is at least 42. Letting this mean be μ, in this exercise we wish to investigate whether the sample of 65 satisfaction ratings provides evidence to support the claim that μ exceeds 42 (and, therefore, is at least 42).

For the sake of argument, we begin by assuming that μ equals 42, and we then attempt to use the sample to contradict this assumption in favor of the conclusion that μ exceeds 42. Recall that the mean of the sample of 65 satisfaction ratings is $\bar{x} = 42.95$, and assume that σ, the standard deviation of the population of all satisfaction ratings, is known to be 2.64.

a Consider the sampling distribution of $\bar{x}$ for random samples of 65 customer satisfaction ratings. Use the properties of this sampling distribution to find the probability of observing a sample mean greater than or equal to 42.95 when we assume that μ equals 42.

b If μ equals 42, what percentage of all possible sample means are greater than or equal to 42.95? Because we have actually observed a sample mean of $\bar{x} = 42.95$, is it more reasonable to believe that (1) μ equals 42 and we have observed a sample mean that is greater than or equal to 42.95 when μ equals 42, or (2) that we have observed a sample mean that is greater than or equal to 42.95 because μ is greater than 42? Explain. What do you conclude about whether customers are typically very satisfied with the XYZ Box video game system?

7.19 In an article in the *Journal of Management,* Joseph Martocchio studied and estimated the costs of employee absences. Based on a sample of 176 blue-collar workers, Martocchio estimated that the mean amount of paid time lost during a three-month period was 1.4 days per employee with a standard deviation of 1.3 days. Martocchio also estimated that the mean amount of unpaid time lost during a three-month period was 1.0 day per employee with a standard deviation of 1.8 days.

Suppose we randomly select a sample of 100 blue-collar workers. Based on Martocchio's estimates:

a What is the probability that the average amount of paid time lost during a three-month period for the 100 blue-collar workers will exceed 1.5 days? Assume σ equals 1.3 days.

b What is the probability that the average amount of unpaid time lost during a three-month period for the 100 blue-collar workers will exceed 1.5 days? Assume σ equals 1.8 days.

c Suppose we randomly select a sample of 100 blue-collar workers, and suppose the sample mean amount of unpaid time lost during a three-month period actually exceeds 1.5 days. Would it be reasonable to conclude that the mean amount of unpaid time lost has increased above the previously estimated 1.0 day? Explain. Assume σ still equals 1.8 days.

7.20 When a pizza restaurant's delivery process is operating effectively, pizzas are delivered in an average of 45 minutes with a standard deviation of 6 minutes. To monitor its delivery process, the restaurant randomly selects five pizzas each night and records their delivery times.

a For the sake of argument, assume that the population of all delivery times on a given evening is normally distributed with a mean of $\mu = 45$ minutes and a standard deviation of $\sigma = 6$ minutes. (That is, we assume that the delivery process is operating effectively.) Find the mean and the standard deviation of the population of all possible sample means, and calculate an interval containing 99.73 percent of all possible sample means.

b Suppose that the mean of the five sampled delivery times on a particular evening is $\bar{x} = 55$ minutes. Using the interval that you calculated in *a*, what would you conclude about whether the restaurant's delivery process is operating effectively? Why?

LO7-4 Describe and use the sampling distribution of the sample proportion.

7.3 The Sampling Distribution of the Sample Proportion

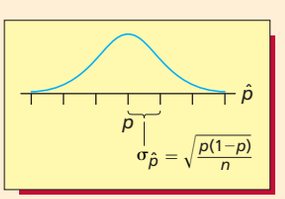
A food processing company markets a soft cheese spread that is sold in a plastic container with an "easy pour" spout. Although this spout works extremely well and is popular with consumers, it is expensive to produce. Because of the spout's high cost, the company has developed a new, less expensive spout. While the new, cheaper spout may alienate some purchasers, a company study shows that its introduction will increase profits if fewer than 10 percent of the cheese spread's current purchasers are lost. That is, if we let p be the true proportion of all current purchasers who would stop buying the cheese spread if the new spout were used, profits will increase as long as p is less than .10.

Suppose that (after trying the new spout) 63 of 1,000 randomly selected purchasers say that they would stop buying the cheese spread if the new spout were used. The point estimate of the population proportion p is the sample proportion $\hat{p} = 63/1{,}000 = .063$. This sample proportion says that we estimate that 6.3 percent of all current purchasers would stop buying the cheese spread if the new spout were used. Because $\hat{p}$ equals .063, we have some evidence that the population proportion p is less than .10. In order to determine the strength of this evidence, we need to consider the sampling distribution of $\hat{p}$. In general, assume that we will randomly select a sample of n elements from a population, and assume that a proportion p of all the elements in the population fall into a particular category (for instance, the category of consumers who would stop buying the cheese spread). Before we actually select the sample, there are many different samples of n elements that we might potentially obtain. The number of elements that fall into the category in question will vary from sample to sample, so the sample proportion of elements falling into the category will also vary from sample to sample. For example, if three possible random samples of 1,000 soft cheese spread purchasers had, respectively, 63, 58, and 65 purchasers say that they would stop buying the cheese spread if the new spout were used, then the sample proportions given by the three samples would be $\hat{p} = 63/1000 = .063$, $\hat{p} = 58/1000 = .058$, and $\hat{p} = 65/1000 = .065$. In general, before we randomly select the sample, there are many different possible sample proportions that we might obtain, and thus the sample proportion $\hat{p}$ is a random variable. In the following box we give the properties of the probability distribution of this random variable, which is called **the sampling distribution of the sample proportion $\hat{p}$.**

The Sampling Distribution of the Sample Proportion $\hat{p}$

The population of all possible sample proportions

1. Approximately has a normal distribution, if the sample size n is large.
2. Has mean $\mu_{\hat{p}} = p$.
3. Has standard deviation $\sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$.



Stated equivalently, the sampling distribution of $\hat{p}$ has mean $\mu_{\hat{p}} = p$, has standard deviation $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$, and is approximately a normal distribution (if the sample size n is large).

Property 1 in the box says that, if n is large, then the population of all possible sample proportions approximately has a normal distribution. Here, it can be shown that ***n* should be considered large if both *np* and *n*(1 − *p*) are at least 5.**[3] Property 2, which says that $\mu_{\hat{p}} = p$, is valid for any sample size and tells us that $\hat{p}$ is an unbiased estimate of p. That is, although the sample proportion $\hat{p}$ that we calculate probably does not equal p, the average of all the different sample proportions that we could have calculated (from all the different possible samples) is equal to p. Property 3, which says that

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

[3]Some statisticians suggest using the more conservative rule that both np and $n(1-p)$ must be at least 10.

is exactly correct if the sampled population is infinite and is approximately correct if the sampled population is finite and much larger than (say, at least 20 times as large as) the sample size. Property 3 tells us that the standard deviation of the population of all possible sample proportions decreases as the sample size increases. That is, the larger n is, the more closely clustered are all the different sample proportions around the true population proportion. Finally, note that the formula for $\sigma_{\hat{p}}$ follows, in theory, from the formula for $\sigma^2_{\hat{p}}$, the variance of the population of all possible sample proportions. The formula for $\sigma^2_{\hat{p}}$ is $\sigma^2_{\hat{p}} = p(1 - p)/n$.

EXAMPLE 7.5 The Cheese Spread Case: Improving Profitability

In the cheese spread situation, the food processing company must decide whether p, the proportion of all current purchasers who would stop buying the cheese spread if the new spout were used, is less than .10. In order to do this, remember that when 1,000 purchasers of the cheese spread are randomly selected, 63 of these purchasers say they would stop buying the cheese spread if the new spout were used. Noting that the sample proportion $\hat{p} = .063$ is less than .10, we ask the following question. If the true population proportion is .10, what is the probability of observing a sample proportion that is less than or equal to .063?

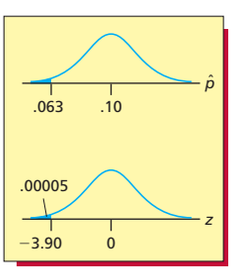
If p equals .10, we can assume that the sampling distribution of $\hat{p}$ is approximately a normal distribution because both $np = 1{,}000(.10) = 100$ and $n(1 - p) = 1{,}000(1 - .10) = 900$ are at least 5. Furthermore, the mean and standard deviation of the sampling distribution of $\hat{p}$ are $\mu_{\hat{p}} = p = .10$ and

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1 - p)}{n}} = \sqrt{\frac{(.10)(.90)}{1{,}000}} = .0094868$$

Therefore,

$$P(\hat{p} \leq .063 \text{ if } p = .10) = P\left(z \leq \frac{.063 - \mu_{\hat{p}}}{\sigma_{\hat{p}}}\right) = P\left(z \leq \frac{.063 - .10}{.0094868}\right)$$

$$= P(z \leq -3.90)$$



which is the area under the standard normal curve to the left of −3.90. The normal table tells us that this area equals .00005. This probability says that, if p equals .10, then only 5 in 100,000 of all possible sample proportions are at least as small as the sample proportion $\hat{p} = .063$ that we have actually observed. That is, if we are to believe that p equals .10, we must believe that we have observed a sample proportion that can be described as a 5 in 100,000 chance. It follows that we have extremely strong evidence that p does not equal .10 and is, in fact, less than .10. In other words, we have extremely strong evidence that fewer than 10 percent of current purchasers would stop buying the cheese spread if the new spout were used. It seems that introducing the new spout will be profitable.

Exercises for Section 7.3

CONCEPTS

7.21 What population is described by the sampling distribution of $\hat{p}$?

7.22 Suppose that we will randomly select a sample of n elements from a population and that we will compute the sample proportion $\hat{p}$ of these elements that fall into a category of interest. If we consider the sampling distribution of $\hat{p}$:

- **a** If the sample size n is large, the sampling distribution of $\hat{p}$ is approximately a normal distribution. What condition must be satisfied to guarantee that n is large enough to say that $\hat{p}$ is normally distributed?
- **b** Write formulas that express the central tendency and variability of the population of all possible sample proportions. Explain what each of these formulas means in your own words.

7.23 Describe the effect of increasing the sample size on the population of all possible sample proportions.

METHODS AND APPLICATIONS

7.24 In each of the following cases, determine whether the sample size n is large enough to say that the sampling distribution of $\hat{p}$ is a normal distribution.

a $p = .4$, $n = 100$ **d** $p = .8$, $n = 400$
b $p = .1$, $n = 10$ **e** $p = .98$, $n = 1{,}000$
c $p = .1$, $n = 50$ **f** $p = .99$, $n = 400$

7.25 In each of the following cases, find the mean, variance, and standard deviation of the sampling distribution of the sample proportion $\hat{p}$.

a $p = .5$, $n = 250$ **c** $p = .8$, $n = 400$
b $p = .1$, $n = 100$ **d** $p = .98$, $n = 1{,}000$

7.26 For each situation in Exercise 7.25, find an interval that contains approximately 95.44 percent of all the possible sample proportions.

7.27 Suppose that we will randomly select a sample of $n = 100$ elements from a population and that we will compute the sample proportion $\hat{p}$ of these elements that fall into a category of interest. If the true population proportion p equals .9:

a Describe the shape of the sampling distribution of $\hat{p}$. Why can we validly describe the shape?
b Find the mean and the standard deviation of the sampling distribution of $\hat{p}$.

7.28 For the situation in Exercise 7.27, calculate the following probabilities. In each case sketch the sampling distribution and the probability.

a $P(\hat{p} \geq .96)$
b $P(.855 \leq \hat{p} \leq .945)$
c $P(\hat{p} \leq .915)$

7.29 A past issue of *The Journal News* (Hamilton, Ohio) reported on a study conducted by the Kaiser Family Foundation regarding parents' use of television set V-chips for controlling their childrens' TV viewing. The study asked parents who own TVs equipped with V-chips whether they use the devices to block programs with objectionable content.

a Suppose that we wish to use the study results to justify the claim that fewer than 20 percent of parents who own TV sets with V-chips use the devices. The study actually found that 17 percent of the parents polled used their V-chips. If the poll surveyed 1,000 parents, and if for the sake of argument we assume that 20 percent of parents who own V-chips actually use the devices (that is, $p = .2$), calculate the probability of observing a sample proportion of .17 or less. That is, calculate $P(\hat{p} \leq .17)$.
b Based on the probability you computed in part *a*, would you conclude that fewer than 20 percent of parents who own TV sets equipped with V-chips actually use the devices? Explain.

7.30 On February 8, 2002, the Gallup Organization released the results of a poll concerning American attitudes toward the 19th Winter Olympic Games in Salt Lake City, Utah. The poll results were based on telephone interviews with a randomly selected national sample of 1,011 adults, 18 years and older, conducted February 4–6, 2002.

a Suppose we wish to use the poll's results to justify the claim that more than 30 percent of Americans (18 years or older) say that figure skating is their favorite Winter Olympic event. The poll actually found that 32 percent of respondents reported that figure skating was their favorite event.[4] If, for the sake of argument, we assume that 30 percent of Americans (18 years or older) say figure skating is their favorite event (that is, $p = .3$), calculate the probability of observing a sample proportion of .32 or more; that is, calculate $P(\hat{p} \geq .32)$.
b Based on the probability you computed in part *a*, would you conclude that more than 30 percent of Americans (18 years or older) say that figure skating is their favorite Winter Olympic event?

7.31 *Quality Progress,* February 2005, reports on improvements in customer satisfaction and loyalty made by Bank of America. A key measure of customer satisfaction is the response (on a scale from 1 to 10) to the question: "Considering all the business you do with Bank of America, what is your overall satisfaction with Bank of America?" Here, a response of 9 or 10 represents "customer delight."

a Historically, the percentage of Bank of America customers expressing customer delight has been 48%. Suppose that we wish to use the results of a survey of 350 Bank of America customers to justify the claim that more than 48% of all current Bank of America customers would express customer delight. The survey finds that 189 of 350 randomly selected Bank of America customers express customer delight. If, for the sake of argument, we assume that the proportion of customer delight is $p = .48$, calculate the probability of observing a sample proportion greater than or equal to $189/350 = .54$. That is, calculate $P(\hat{p} \geq .54)$.

[4] **Source:** The Gallup Organization, www.gallup.com/poll/releases/, February 13, 2002.

b Based on the probability you computed in part *a*, would you conclude that more than 48 percent of current Bank of America customers express customer delight? Explain.

7.32 Again consider the survey of 350 Bank of America customers discussed in Exercise 7.31, and assume that 48% of Bank of America customers would currently express customer delight. That is, assume $p = .48$. Find:

a The probability that the sample proportion obtained from the sample of 350 Bank of America customers would be within three percentage points of the population proportion. That is, find $P(.45 \leq \hat{p} \leq .51)$.

b The probability that the sample proportion obtained from the sample of 350 Bank of America customers would be within six percentage points of the population proportion. That is, find $P(.42 \leq \hat{p} \leq .54)$.

7.33 Based on your results in Exercise 7.32, would it be reasonable to state that the survey's "margin of error" is ± 3 percentage points? ± 6 percentage points? Explain.

7.34 An article in *Fortune* magazine discussed "outsourcing." According to the article, outsourcing is "the assignment of critical, but noncore, business functions to outside specialists." This allows a company to immediately bring operations up to best-in-world standards while avoiding huge capital investments. The article included the results of a poll of business executives addressing the benefits of outsourcing.

a Suppose we wish to use the poll's results to justify the claim that fewer than 20 percent of business executives feel that the benefits of outsourcing are either "less or much less than expected." The poll actually found that 15 percent of the respondents felt that the benefits of outsourcing were either "less or much less than expected." If 1,000 randomly selected business executives were polled, and if for the sake of argument, we assume that 20 percent of all business executives feel that the benefits of outsourcing are either less or much less than expected (that is, $p = .20$), calculate the probability of observing a sample proportion of .15 or less. That is, calculate $P(\hat{p} \leq .15)$.

b Based on the probability you computed in part *a*, would you conclude that fewer than 20 percent of business executives feel that the benefits of outsourcing are either "less or much less than expected"? Explain.

7.35 *Fortune* magazine reported the results of a survey on executive training that was conducted by the Association of Executive Search Consultants. The survey showed that 75 percent of 300 polled CEOs believe that companies should have "fast-track training programs" for developing managerial talent.

a Suppose we wish to use the results of this survey to justify the claim that more than 70 percent of CEOs believe that companies should have fast-track training programs. Assuming that the 300 surveyed CEOs were randomly selected, and assuming, for the sake of argument, that 70 percent of CEOs believe that companies should have fast-track training programs (that is, $p = .70$), calculate the probability of observing a sample proportion of .75 or more. That is, calculate $P(\hat{p} \geq .75)$.

b Based on the probability you computed in part *a*, would you conclude that more than 70 percent of CEOs believe that companies should have fast-track training programs? Explain.

7.4 Stratified Random, Cluster, and Systematic Sampling (Optional)

LO7-5 Describe the basic ideas of stratified random, cluster, and systematic sampling (Optional).

Random sampling is not the only kind of sampling. Methods for obtaining a sample are called **sampling designs,** and the sample we take is sometimes called a **sample survey.** In this section we explain three sampling designs that are alternatives to random sampling—**stratified random sampling, cluster sampling,** and **systematic sampling.**

One common sampling design involves separately sampling important groups within a population. Then, the samples are combined to form the entire sample. This approach is the idea behind **stratified random sampling.**

In order to select a **stratified random sample,** we divide the population into nonoverlapping groups of similar elements (people, objects, etc.). These groups are called **strata.** Then a random sample is selected from each stratum, and these samples are combined to form the full sample.

It is wise to stratify when the population consists of two or more groups that differ with respect to the variable of interest. For instance, consumers could be divided into strata based on gender, age, ethnic group, or income.

As an example, suppose that a department store chain proposes to open a new store in a location that would serve customers who live in a geographical region that consists of (1) an industrial city, (2) a suburban community, and (3) a rural area. In order to assess the potential profitability of the proposed store, the chain wishes to study the incomes of all households in the region. In addition, the chain wishes to estimate the proportion and the total number of households whose members would be likely to shop at the store. The department store chain feels that the industrial city, the suburban community, and the rural area differ with respect to income and the store's potential desirability. Therefore, it uses these subpopulations as strata and takes a stratified random sample.

Taking a stratified sample can be advantageous because such a sample takes advantage of the fact that elements in the same stratum are similar to each other. It follows that a stratified sample can provide more accurate information than a random sample of the same size. As a simple example, if all of the elements in each stratum were exactly the same, then examining only one element in each stratum would allow us to describe the entire population. Furthermore, stratification can make a sample easier (or possible) to select. Recall that, in order to take a random sample, we must have a list, or **frame** of all of the population elements. Although a frame might not exist for the overall population, a frame might exist for each stratum. For example, suppose nearly all the households in the department store's geographical region have telephones. Although there might not be a telephone directory for the overall geographical region, there might be separate telephone directories for the industrial city, the suburb, and the rural area. For more discussion of stratified random sampling, see Section 8.5 and Mendenhall, Schaeffer, and Ott (1986).

Sometimes it is advantageous to select a sample in stages. This is a common practice when selecting a sample from a very large geographical region. In such a case, a frame often does not exist. For instance, there is no single list of all registered voters in the United States. There is also no single list of all households in the United States. In this kind of situation, we can use **multistage cluster sampling.** To illustrate this procedure, suppose we wish to take a sample of registered voters from all registered voters in the United States. We might proceed as follows:

Stage 1: Randomly select a sample of counties from all of the counties in the United States.

Stage 2: Randomly select a sample of townships from each county selected in Stage 1.

Stage 3: Randomly select a sample of voting precincts from each township selected in Stage 2.

Stage 4: Randomly select a sample of registered voters from each voting precinct selected in Stage 3.

We use the term *cluster sampling* to describe this type of sampling because at each stage we "cluster" the voters into subpopulations. For instance, in Stage 1 we cluster the voters into counties, and in Stage 2 we cluster the voters in each selected county into townships. Also, notice that the random sampling at each stage can be carried out because there are lists of (1) all counties in the United States, (2) all townships in each county, (3) all voting precincts in each township, and (4) all registered voters in each voting precinct.

As another example, consider sampling the households in the United States. We might use Stages 1 and 2 above to select counties and townships within the selected counties. Then, if there is a telephone directory of the households in each township, we can randomly sample households from each selected township by using its telephone directory. Because *most* households today have telephones, and telephone directories are readily available, most national polls are now conducted by telephone. Further, polling organizations have recognized that many households are giving up landline phones, and have developed ways to sample households that only have cell phones.

It is sometimes a good idea to combine stratification with multistage cluster sampling. For example, suppose a national polling organization wants to estimate the proportion of all registered voters who favor a particular presidential candidate. Because the presidential preferences of voters might tend to vary by geographical region, the polling organization might divide the United States into regions (say, Eastern, Midwestern, Southern, and Western regions). The polling organization might then use these regions as strata, and might take a multistage cluster sample from each stratum (region).

The analysis of data produced by multistage cluster sampling can be quite complicated. For a more detailed discussion of cluster sampling, see Mendenhall, Schaeffer, and Ott (1986).

In order to select a random sample, we must number the elements in a frame of all the population elements. Then we use a random number table (or a random number generator on a computer) to make the selections. However, numbering all the population elements can be quite time-consuming. Moreover, random sampling is used in the various stages of many complex sampling designs (requiring the numbering of numerous populations). Therefore, it is useful to have an alternative to random sampling. One such alternative is called **systematic sampling.** In order to systematically select a sample of n elements without replacement from a frame of N elements, we divide N by n and round the result down to the nearest whole number. Calling the rounded result ℓ, we then randomly select one element from the first ℓ elements in the frame—this is the first element in the systematic sample. The remaining elements in the sample are obtained by selecting every ℓth element following the first (randomly selected) element. For example, suppose we wish to sample a population of $N = 14{,}327$ allergists to investigate how often they have prescribed a particular drug during the last year. A medical society has a directory listing the 14,327 allergists, and we wish to draw a systematic sample of 500 allergists from this frame. Here we compute $14{,}327/500 = 28.654$, which is 28 when rounded down. Therefore, we number the first 28 allergists in the directory from 1 to 28, and we use a random number table to randomly select one of the first 28 allergists. Suppose we select allergist number 19. We interview allergist 19 and every 28th allergist in the frame thereafter, so we choose allergists 19, 47, 75, and so forth until we obtain our sample of 500 allergists. In this scheme, we must number the first 28 allergists, but we do not have to number the rest because we can "count off" every 28th allergist in the directory. Alternatively, we can measure the approximate amount of space in the directory that it takes to list 28 allergists. This measurement can then be used to select every 28th allergist.

Exercises for Section 7.4

CONCEPTS

connect

7.36 When is it appropriate to use stratified random sampling? What are strata, and how should strata be selected?

7.37 When is cluster sampling used? Why do we describe this type of sampling by using the term *cluster?*

7.38 Explain how to take a systematic sample of 100 companies from the 1,853 companies that are members of an industry trade association.

7.39 Explain how a stratified random sample is selected. Discuss how you might define the strata to survey student opinion on a proposal to charge all students a $100 fee for a new university-run bus system that will provide transportation between off-campus apartments and campus locations.

7.40 Marketing researchers often use city blocks as clusters in cluster sampling. Using this fact, explain how a market researcher might use multistage cluster sampling to select a sample of consumers from all cities having a population of more than 10,000 in a large state having many such cities.

7.5 More about Surveys and Errors in Survey Sampling (Optional)

LO7-6 Describe basic types of survey questions, survey procedures, and sources of error (Optional).

We have seen in Section 1.2 that people in surveys are asked questions about their behaviors, opinions, beliefs, and other characteristics. In this section we discuss various issues related to designing surveys and the errors that can occur in survey sampling.

Types of survey questions Survey instruments can use **dichotomous** ("yes or no"), **multiple-choice,** or **open-ended** questions. Each type of question has its benefits and drawbacks. Dichotomous questions are usually clearly stated, can be answered quickly, and yield data that are

easily analyzed. However, the information gathered may be limited by this two-option format. If we limit voters to expressing support or disapproval for stem-cell research, we may not learn the nuanced reasoning that voters use in weighing the merits and moral issues involved. Similarly, in today's heterogeneous world, it would be unusual to use a dichotomous question to categorize a person's religious preferences. Asking whether respondents are Christian or non-Christian (or to use any other two categories like Jewish or non-Jewish; Muslim or non-Muslim) is certain to make some people feel their religion is being slighted. In addition, this is a crude way and unenlightening way to learn about religious preferences.

Multiple-choice questions can assume several different forms. Sometimes respondents are asked to choose a response from a list (for example, possible answers to the religion question could be Jewish; Christian; Muslim; Hindu; Agnostic; or Other). Other times, respondents are asked to choose an answer from a numerical range. We could ask the question:

"In your opinion, how important are SAT scores to a college student's success?"

Not important at all 1 2 3 4 5 Extremely important

These numerical responses are usually summarized and reported in terms of the average response, whose size tells us something about the perceived importance. The Zagat restaurant survey (www.zagat.com) asks diners to rate restaurants' food, décor, and service, each on a scale of 1 to 30 points, with a 30 representing an incredible level of satisfaction. Although the Zagat scale has an unusually wide range of possible ratings, the concept is the same as in the more common 5-point scale.

Open-ended questions typically provide the most honest and complete information because there are no suggested answers to divert or bias a person's response. This kind of question is often found on instructor evaluation forms distributed at the end of a college course. College students at Georgetown University are asked the open-ended question, "What comments would you give to the instructor?" The responses provide the instructor feedback that may be missing from the initial part of the teaching evaluation survey, which consists of numerical multiple-choice ratings of various aspects of the course. While these numerical ratings can be used to compare instructors and courses, there are no easy comparisons of the diverse responses instructors receive to the open-ended question. In fact, these responses are often seen only by the instructor and are useful, constructive tools for the teacher despite the fact they cannot be readily summarized.

Survey questionnaires must be carefully constructed so they do not inadvertently bias the results. Because survey design is such a difficult and sensitive process, it is not uncommon for a pilot survey to be taken before a lot of time, effort, and financing go into collecting a large amount of data. Pilot surveys are similar to the beta version of a new electronic product; they are tested out with a smaller group of people to work out the "kinks" before being used on a larger scale. Determination of the sample size for the final survey is an important process for many reasons. If the sample size is too large, resources may be wasted during the data collection. On the other hand, not collecting enough data for a meaningful analysis will obviously be detrimental to the study. Fortunately, there are several formulas that will help decide how large a sample should be, depending on the goal of the study and various other factors.

Types of surveys There are several different survey types, and we will explore just a few of them. The **phone survey** is particularly well-known (and often despised). A phone survey is inexpensive and usually conducted by callers who have very little training. Because of this and the impersonal nature of the medium, the respondent may misunderstand some of the questions. A further drawback is that some people cannot be reached and that others may refuse to answer some or all of the questions. Phone surveys are thus particularly prone to have a low **response rate.**

The **response rate** is the proportion of all people whom we attempt to contact that actually respond to a survey. A low response rate can destroy the validity of a survey's results.

The popular television sitcom *Seinfeld* parodied the difficulties of collecting data through a phone survey. After receiving several calls from telemarketers, Jerry replied in exasperation:

> "I'm sorry; I'm a little tied up now. Give me your home number and I'll call you back later. Oh! You don't like being called at home? Well, now you know how I feel."

Numerous complaints have been filed with the Federal Trade Commission (FTC) about the glut of marketing and survey telephone calls to private residences. The National Do Not Call Registry

was created as the culmination of a comprehensive, three-year review of the Telemarketing Sales Rule (TSR) (www.ftc.gov/donotcall/). This legislation allows people to enroll their phone numbers on a website so as to prevent most marketers from calling them.

Self-administered surveys, or **mail surveys,** are also very inexpensive to conduct. However, these also have their drawbacks. Often, recipients will choose not to reply unless they receive some kind of financial incentive or other reward. Generally, after an initial mailing, the response rate will fall between 20 and 30 percent (www.pra.ca/resources/rates.pdf). Response rates can be raised with successive follow-up reminders, and after three contacts, they might reach between 65 and 75 percent. Unfortunately, the entire process can take significantly longer than a phone survey would.

Web-based surveys have become increasingly popular, but they suffer from the same problems as mail surveys. In addition, as with phone surveys, respondents may record their true reactions incorrectly because they have misunderstood some of the questions posed.

A personal interview provides more control over the survey process. People selected for interviews are more likely to respond because the questions are being asked by someone face-to-face. Questions are less likely to be misunderstood because the people conducting the interviews are typically trained employees who can clear up any confusion arising during the process. On the other hand, interviewers can potentially "lead" a respondent by body language which signals approval or disapproval of certain sorts of answers. They can also prompt certain replies by providing too much information. **Mall surveys** are examples of personal interviews. Interviewers approach shoppers as they pass by and ask them to answer the survey questions. Response rates around 50 percent are typical (http://en.wikipedia.org/wiki/Statistical_survey#Survey_methods). Personal interviews are more costly than mail or phone surveys. Obviously, the objective of the study will be important in deciding upon the survey type employed.

Errors occurring in surveys In general, the goal of a survey is to obtain accurate information from a group, or sample, that is representative of the entire population of interest. We are trying to estimate some aspect (numerical descriptor) of the entire population from a subset of the population. This is not an easy task, and there are many pitfalls. First and foremost, the *target population* must be well defined and a *sample frame* must be chosen.

The **target population** is the entire population of interest to us in a particular study.

Are we intending to estimate the average starting salary of students graduating from any college? Or from four year colleges? Or from business schools? Or from a particular business school?

The **sample frame** is a list of sampling elements (people or things) from which the sample will be selected. It should closely agree with the target population.

Consider a study to estimate the average starting salary of students who have graduated from the business school at Miami University of Ohio over the last five years; the target population is obviously that particular group of graduates. A sample frame could be the Miami University Alumni Association's roster of business school graduates for the past five years. Although it will not be a perfect replication of the target population, it is a reasonable frame.

We now discuss two general classes of survey errors: **errors of non-observation** and **errors of observation.** From the sample frame, units are randomly chosen to be part of the sample. Simply by virtue of the fact that we are taking a sample instead of a census, we are susceptible to *sampling error.*

Sampling error is the difference between a numerical descriptor of the population and the corresponding descriptor of the sample.

Sampling error occurs because our information is incomplete. We observe only the portion of the population included in the sample while the remainder is obscured. Suppose, for example, we wanted to know about the heights of 13-year-old boys. There is extreme variation in boys' heights at this age. Even if we could overcome the logistical problems of choosing a random sample of 20 boys, there is nothing to guarantee the sample will accurately reflect heights at this age. By sheer luck of the draw, our sample could include a higher proportion of tall boys than appears in the population. We would then overestimate average height at this age (to the chagrin of the shorter boys). Although samples tend to look more similar to their parent populations as the sample sizes increase, we should always keep in mind that sample characteristics and population characteristics are not the same.

If a sample frame is not identical to the target population, we will suffer from an *error of coverage*.

Undercoverage occurs when some population elements are excluded from the process of selecting the sample.

Undercoverage was part of the problem dooming the *Literary Digest* Poll of 1936. Although millions of Americans were included in the poll, the large sample size could not rescue the poll results. The sample represented those who could afford phone service and magazine subscriptions in the lean Depression years, but in excluding everyone else, it failed to yield an honest picture of the entire American populace. Undercoverage often occurs when we do not have a complete, accurate list of all the population units. If we select our sample from an incomplete list, like a telephone directory or a list of all Internet subscribers in a region, we automatically eliminate those who cannot afford phone or Internet service. Even today, 7 to 8 percent of the people in the United States do not own telephones. Low-income people are often underrepresented in surveys. If underrepresented groups differ from the rest of the population with respect to the characteristic under study, the survey results will be biased.

Often, pollsters cannot find all the people they intend to survey, and sometimes people who are found will refuse to answer the questions posed. Both of these are examples of the **nonresponse** problem. Unfortunately, there may be an association between how difficult it is to find and elicit responses from people and the type of answers they give.

Nonresponse occurs whenever some of the individuals who were supposed to be included in the sample are not.

For example, universities often conduct surveys to learn how graduates have fared in the workplace. The alumnus who has risen through the corporate ranks is more likely to have a current address on file with his alumni office and to be willing to share career information than a classmate who has foundered professionally. We should be politely skeptical about reports touting the average salaries of graduates of various university programs. In some surveys, 35 percent or more of the selected individuals cannot be contacted—even when several callbacks are made. In such cases, other participants are often substituted for those who cannot be contacted. If the substitutes and the originally selected participants differ with respect to the characteristic under study, the survey will be biased. Furthermore, people who will answer highly sensitive, personal, or embarrassing questions might be very different from those who will not.

As discussed in Section 1.2, the opinions of those who bother to complete a voluntary response survey may be dramatically different from those who do not. (Recall the Ann Landers question about having children.) The viewer voting on the popular television show *American Idol* is another illustration of **selection bias,** because only those who are interested in the outcome of the show will bother to phone in or text message their votes. The results of the voting are not representative of the performance ratings the country would give as a whole.

Errors of observation occur when data values are recorded incorrectly. Such errors can be caused by the data collector (the interviewer), the survey instrument, the respondent, or the data collection process. For instance, the manner in which a question is asked can influence the response. Or, the order in which questions appear on a questionnaire can influence the survey results. Or, the data collection method (telephone interview, questionnaire, personal interview, or direct observation) can influence the results. A **recording error** occurs when either the respondent or interviewer incorrectly marks an answer. Once data are collected from a survey, the results are often entered into a computer for statistical analysis. When transferring data from a survey form to a spreadsheet program like Excel, MINITAB, or MegaStat, there is potential for entering them incorrectly. Before the survey is administered, the questions need to be very carefully worded so that there is little chance of misinterpretation. A poorly framed question might yield results that lead to unwarranted decisions. Scaled questions are particularly susceptible to this type of error. Consider the question "How would you rate this course?" Without a proper explanation, the respondent may not know whether "1" or "5" is the best.

If the survey instrument contains highly sensitive questions and respondents feel compelled to answer, they may not tell the truth. This is especially true in personal interviews. We then have what is called **response bias.** A surprising number of people are reluctant to be candid about what they like to read or watch on television. People tend to over-report "good" activities like reading

respected newspapers and underreport their "bad" activities like delighting in the *National Enquirer*'s stories of alien abductions and celebrity meltdowns. Imagine, then, the difficulty in getting honest answers about people's gambling habits, drug use, or sexual histories. Response bias can also occur when respondents are asked slanted questions whose wording influences the answer received. For example, consider the following question:

Which of the following best describes your views on gun control?

1. The government should take away our guns, leaving us defenseless against heavily armed criminals.
2. We have the right to keep and bear arms.

This question is biased toward eliciting a response against gun control.

Exercises for Section 7.5

CONCEPTS

7.41 Explain:
- **a** Three types of surveys and discuss their advantages and disadvantages.
- **b** Three types of survey questions and discuss their advantages and disadvantages.

7.42 Explain each of the following terms:
- **a** Undercoverage **b** Nonresponse **c** Response bias

7.43 A market research firm sends out a Web-based survey to assess the impact of advertisements placed on a search engine's results page. About 65% of the surveys were answered and sent back. What types of errors are possible in this scenario?

7.6 Derivation of the Mean and the Variance of the Sample Mean (Optional)

Before we randomly select the sample values $x_1, x_2, \ldots, x_n$ from a population having mean μ and variance σ^2, we note that, for $i = 1, 2, \ldots, n$, the ith sample value x_i is a random variable that can potentially be any of the values in the population. Moreover, it can be proven (and is intuitive) that

1. The mean (or expected value) of x_i, denoted μ_{x_i}, is μ, the mean of the population from which x_i will be randomly selected. That is, $\mu_{x_1} = \mu_{x_2} = \cdots = \mu_{x_n} = \mu$.
2. The variance of x_i, denoted $\sigma^2_{x_i}$, is σ^2, the variance of the population from which x_i will be randomly selected. That is, $\sigma^2_{x_1} = \sigma^2_{x_2} = \cdots = \sigma^2_{x_n} = \sigma^2$.

If we consider the sample mean $\bar{x} = \sum_{i=1}^{n} x_i/n$, then we can prove that $\mu_{\bar{x}} = \mu$ by using the following two properties of the mean discussed in Section 5.6:

Property 1: If a is a fixed number, $\mu_{ax} = a\mu_x$

Property 2: $\mu_{(x_1 + x_2 + \cdots + x_n)} = \mu_{x_1} + \mu_{x_2} + \cdots + \mu_{x_n}$

The proof that $\mu_{\bar{x}} = \mu$ is as follows:

$$\mu_{\bar{x}} = \mu\left(\sum_{i=1}^{n} x_i/n\right)$$

$$= \frac{1}{n}\mu\left(\sum_{i=1}^{n} x_i\right) \qquad \text{(see Property 1)}$$

$$= \frac{1}{n}\mu_{(x_1 + x_2 + \cdots + x_n)}$$

$$= \frac{1}{n}(\mu_{x_1} + \mu_{x_2} + \cdots + \mu_{x_n}) \qquad \text{(see Property 2)}$$

$$= \frac{1}{n}(\mu + \mu + \cdots + \mu) = \frac{n\mu}{n} = \mu$$

We can prove that $\sigma^2_{\bar{x}} = \sigma^2/n$ by using the following two properties of the variance discussed in Section 5.6:

Property 3: If a is a fixed number, $\sigma^2_{ax} = a^2\sigma^2_x$

Property 4: If $x_1, x_2, \cdots, x_n$ are statistically independent, $\sigma^2_{(x_1+x_2+\cdots+x_n)} = \sigma^2_{x_1} + \sigma^2_{x_2} + \cdots + \sigma^2_{x_n}$

The proof that $\sigma^2_{\bar{x}} = \sigma^2/n$ is as follows:

$$\sigma^2_{\bar{x}} = \sigma^2\left(\sum_{i=1}^{n} x_i/n\right) = \left(\frac{1}{n}\right)^2 \sigma^2\left(\sum_{i=1}^{n} x_i\right) \qquad \text{(see Property 3)}$$

$$= \frac{1}{n^2}\sigma^2_{(x_1+x_2+\cdots+x_n)}$$

$$= \frac{1}{n^2}(\sigma^2_{x_1} + \sigma^2_{x_2} + \cdots + \sigma^2_{x_n}) \qquad \text{(see Property 4)}$$

$$= \frac{1}{n^2}(\sigma^2 + \sigma^2 + \cdots + \sigma^2) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Note that we can use Property 4 if $x_1, x_2, \cdots, x_n$ are independent random variables. In general, $x_1, x_2, \cdots, x_n$ are independent if we are drawing these sample values from an infinite population. When we select a sample from an infinite population, a population value obtained on one selection can also be obtained on any other selection. This is because, since the population is infinite, there are an infinite number of repetitions of each population value. Therefore, because a value obtained on one selection is not precluded from being obtained on any other selection, the selections and thus $x_1, x_2, \cdots, x_n$ are statistically independent. Furthermore, this statistical independence approximately holds if the population size is much larger than (say, at least 20 times as large as) the sample size. Therefore, in this case $\sigma^2_{\bar{x}} = \sigma^2/n$ is approximately correct.

Chapter Summary

We began this chapter by defining a random sample and by explaining how to use a **random number table** or **computer-generated random numbers** to select a **random sample.** We then discussed **sampling distributions.** A **sampling distribution** is the probability distribution that describes the population of all possible values of a sample statistic. In this chapter we studied the properties of two important sampling distributions—the sampling distribution of the sample mean, $\bar{x}$, and the sampling distribution of the sample proportion, $\hat{p}$.

Because different samples that can be randomly selected from a population give different sample means, there is a population of sample means corresponding to a particular sample size. The probability distribution describing the population of all possible sample means is called the **sampling distribution of the sample mean, $\bar{x}$.** We studied the properties of this sampling distribution when the sampled population is and is not normally distributed. We found that, when the sampled population has a normal distribution, then the sampling distribution of the sample mean is a normal distribution. Furthermore, the **Central Limit Theorem** tells us that, if the sampled population is not normally distributed, then the sampling distribution of the sample mean is approximately a normal distribution when the sample size is large (at least 30). We also saw that the mean of the sampling distribution of $\bar{x}$ always equals the mean of the sampled population, and we presented formulas for the variance and the standard deviation of this sampling distribution. Finally, we explained that the sample mean is a **minimum-variance unbiased point estimate** of the mean of a normally distributed population.

We also studied the properties of the **sampling distribution of the sample proportion $\hat{p}$.** We found that, if the sample size is large, then this sampling distribution is approximately a normal distribution, and we gave a rule for determining whether the sample size is large. We found that the mean of the sampling distribution of $\hat{p}$ is the population proportion p, and we gave formulas for the variance and the standard deviation of this sampling distribution.

Throughout our discussions of sampling distributions, we demonstrated that knowing the properties of sampling distributions can help us make statistical inferences about population parameters. In fact, we will see that the properties of various sampling distributions provide the foundation for most of the techniques to be discussed in future chapters.

We concluded this chapter with three optional sections. In the first optional section, we discussed some advanced sampling designs. Specifically, we introduced **stratified random sampling,** in which we divide a population into groups **(strata)** and then select a random sample from each group. We also introduced **multistage cluster sampling,** which involves selecting a sample in stages, and we explained how to select a **systematic sample.** In the second optional section, we discussed more about surveys, as well as some potential problems that can occur when conducting a sample survey—**undercoverage, nonresponse, response bias,** and **slanted questions.** In the last optional section, we derived the mean and the variance of the sampling distribution of the sample mean $\bar{x}$.

Glossary of Terms

Central Limit Theorem: A theorem telling us that when the sample size n is sufficiently large, then the population of all possible sample means is approximately normally distributed no matter what probability distribution describes the sampled population. (page 278)

cluster sampling (multistage cluster sampling): A sampling design in which we sequentially cluster population elements into subpopulations. (page 288)

convenience sampling: Sampling where we select elements because they are easy or convenient to sample. (page 270)

errors of non-observation: Sampling error related to population elements that are not observed. (page 291)

errors of observation: Sampling error that occurs when the data collected in a survey differs from the truth. (page 292)

judgment sampling: Sampling where an expert selects population elements that he/she feels are representative of the population. (page 270)

minimum-variance unbiased point estimate: An unbiased point estimate of a population parameter having a variance that is smaller than the variance of any other unbiased point estimate of the parameter. (page 281)

nonresponse: A situation in which population elements selected to participate in a survey do not respond to the survey instrument. (page 292)

probability sampling: Sampling where we know the chance (probability) that each population element will be included in the sample. (page 270)

random number table: A table containing random digits that is often used to select a random sample. (page 268)

random sample: A sample selected in such a way that every set of n elements in the population has the same chance of being selected. (page 267)

response bias: Bias in the results obtained when carrying out a statistical study that is related to how survey participants answer the survey questions. (page 292)

response rate: The proportion of all people whom we attempt to contact that actually respond to a survey. (page 290)

sample frame: A list of sampling elements from which a sample will be selected. It should closely agree with the target population. (page 291)

sampling distribution of a sample statistic: The probability distribution of the population of all possible values of the sample statistic. (page 280)

sampling distribution of the sample mean $\bar{x}$: The probability distribution of the population of all possible sample means obtained from samples of a particular size n. (page 271)

sampling distribution of the sample proportion $\hat{p}$: The probability distribution of the population of all possible sample proportions obtained from samples of a particular size n. (page 284)

sampling error: The difference between the value of a sample statistic and the population parameter; it occurs because not all of the elements in the population have been measured. (page 291)

sampling with replacement: A sampling procedure in which we place any element that has been chosen back into the population to give the element a chance to be chosen on succeeding selections. (page 267)

sampling without replacement: A sampling procedure in which we do not place previously selected elements back into the population and, therefore, do not give these elements a chance to be chosen on succeeding selections. (page 267)

selection bias: Bias in the results obtained when carrying out a statistical study that is related to how survey participants are selected. (page 292)

strata: The subpopulations in a stratified sampling design. (page 287)

stratified random sampling: A sampling design in which we divide a population into nonoverlapping subpopulations and then select a random sample from each subpopulation (stratum). (page 287)

systematic sample: A sample taken by moving systematically through the population. For instance, we might randomly select one of the first 200 population elements and then systematically sample every 200th population element thereafter. (page 289)

target population: The entire population of interest in a statistical study. (page 291)

unbiased point estimate: A sample statistic is an unbiased point estimate of a population parameter if the mean of the population of all possible values of the sample statistic equals the population parameter. (page 281)

undercoverage: A situation in sampling in which some groups of population elements are underrepresented. (page 292)

voluntary response sample: Sampling in which the sample participants self-select. (page 270)

Important Results and Formulas

The sampling distribution of the sample mean: pages 271 and 275
  when a population is normally distributed (page 275)
Central Limit Theorem (page 278)
The sampling distribution of the sample proportion: page 284

Supplementary Exercises

7.44 A company that sells and installs custom designed home theatre systems claims to have sold 977 such systems last year. In order to assess whether these claimed sales are valid, an accountant numbers the company's sales invoices from 1 to 977 and plans to select a random sample of 50 sales invoices. The accountant will then contact the purchasers listed on the 50 sampled sales invoices and determine whether the sales amounts on the invoices are correct. Starting in the upper left-hand corner of Table 7.1(a) (see page 268), determine which 50 of the 977 sales invoices should be included in the random sample. Note: There are many possible answers to this exercise.

7.45 THE TRASH BAG CASE TrashBag

Recall that the trash bag manufacturer has concluded that its new 30-gallon bag will be the strongest such bag on the market if its mean breaking strength is at least 50 pounds. In order to provide statistical evidence that the mean breaking strength of the new bag is at least 50 pounds, the manufacturer randomly selects a sample of n bags and calculates the mean $\bar{x}$ of the breaking strengths of these bags. If the sample mean so obtained is at least 50 pounds, this provides some evidence that the mean breaking strength of all new bags is at least 50 pounds.

Suppose that (unknown to the manufacturer) the breaking strengths of the new 30-gallon bag are normally distributed with a mean of $\mu = 50.6$ pounds and a standard deviation of $\sigma = 1.62$ pounds.

a Find an interval containing 95.44 percent of all possible sample means if the sample size employed is $n = 5$.

b Find an interval containing 95.44 percent of all possible sample means if the sample size employed is $n = 40$.

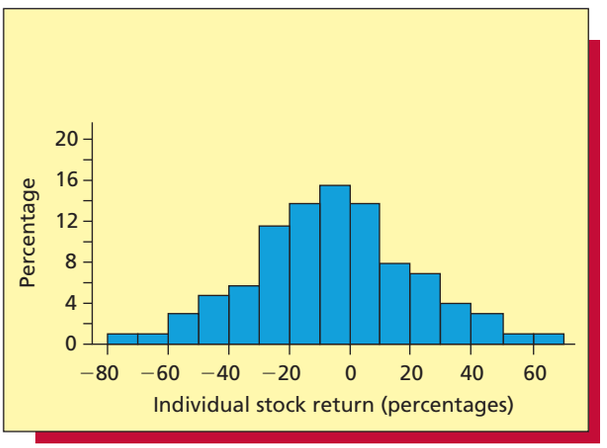
c If the trash bag manufacturer hopes to obtain a sample mean that is at least 50 pounds (so that it can provide evidence that the population mean breaking strength of the new bags is at least 50), which sample size ($n = 5$ or $n = 40$) would be best? Explain why.

7.46 THE STOCK RETURN CASE

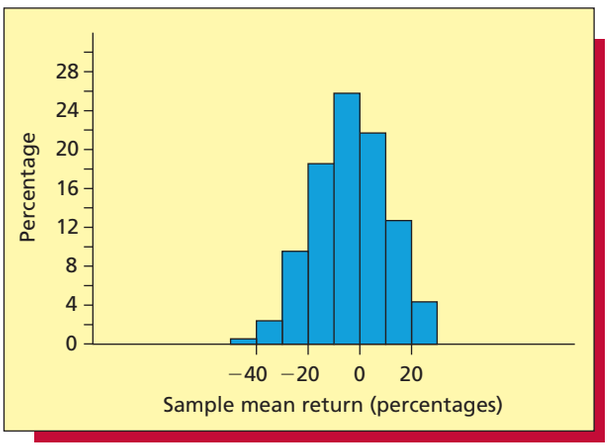
The year 1987 featured extreme volatility on the stock market, including a loss of over 20 percent of the market's value on a single day. Figure 7.7(a) shows the percent frequency histogram of the percentage returns for the entire year 1987 for the population of all 1,815 stocks listed on the New York Stock Exchange. The mean and the standard deviation of the population of percentage returns are -3.5 percent and 26 percent, respectively. Consider drawing a random sample of $n = 5$ stocks from the population of 1,815 stocks and calculating the mean return, $\bar{x}$, of the sampled stocks. If we use a computer, we can generate all the different samples of five stocks that can be obtained (there are trillions of such samples) and calculate the corresponding sample mean returns. A percent frequency histogram describing the population of all possible sample mean returns is given in Figure 7.7(b). Comparing Figures 7.7(a) and (b), we see that, although the histogram of individual stock returns and the histogram of sample mean returns are both bell-shaped and centered over the same mean of -3.5 percent, the histogram of sample mean returns looks *less spread out* than the histogram of individual returns. A sample of 5 stocks is a portfolio of stocks, where the average return of the 5 stocks is the portfolio's return if we invest equal amounts of money in each of the 5 stocks. Because the sample mean returns are less spread out than the individual stock returns, we have illustrated that diversification reduces risk. Find the standard deviation of the population of all sample mean returns, and assuming that this population is normally distributed, find an interval that contains 95.44 percent of all sample mean returns.

FIGURE 7.7 The New York Stock Exchange in 1987: A Comparison of Individual Stock Returns and Sample Mean Returns

(a) The percent frequency histogram describing the population of individual stock returns



(b) The percent frequency histogram describing the population of all possible sample mean returns when $n = 5$



Source: Figure 7.7 is adapted with permission from John K. Ford, "A Method for Grading 1987 Stock Recommendations," *The American Association of Individual Investors Journal,* March 1988, pp. 16–17.

7.47 Suppose that we wish to assess whether more than 60 percent of all U.S. households in a particular income class bought life insurance last year. That is, we wish to assess whether p, the proportion of all U.S. households in the income class that bought life insurance last year, exceeds .60. Assume that an insurance survey is based on 1,000 randomly selected U.S. households in the income class and that 640 of these households bought life insurance last year.

a Assuming that p equals .60 and the sample size is 1,000, what is the probability of observing a sample proportion that is at least .64?

b Based on your answer in part a, do you think more than 60 percent of all U.S. households in the income class bought life insurance last year? Explain.

7.48 A computer supply house receives a large shipment of flash drives each week. Past experience has shown that the number of flaws (bad sectors) per flash drive is either 0, 1, 2, or 3 with probabilities .65, .2, .1, and .05, respectively.

a Calculate the mean and standard deviation of the number of flaws per flash drive.

b Suppose that we randomly select a sample of 100 flash drives. Describe the shape of the sampling distribution of the sample mean $\bar{x}$. Then compute the mean and the standard deviation of the sampling distribution of $\bar{x}$.

c Sketch the sampling distribution of the sample mean $\bar{x}$ and compare it to the distribution describing the number of flaws on a single flash drive.

d The supply house's managers are worried that the flash drives being received have an excessive number of flaws. Because of this, a random sample of 100 flash drives is drawn from each shipment and the shipment is rejected (sent back to the supplier) if the average number of flaws per flash drive for the 100 sample drives is greater than .75. Suppose that the mean number of flaws per flash drive for this week's entire shipment is actually .55. What is the probability that this shipment will be rejected and sent back to the supplier?

7.49 Each day a manufacturing plant receives a large shipment of drums of Chemical ZX-900. These drums are supposed to have a mean fill of 50 gallons, while the fills have a standard deviation known to be .6 gallon.

a Suppose that the mean fill for the shipment is actually 50 gallons. If we draw a random sample of 100 drums from the shipment, what is the probability that the average fill for the 100 drums is between 49.88 gallons and 50.12 gallons?

b The plant manager is worried that the drums of Chemical ZX-900 are underfilled. Because of this, she decides to draw a sample of 100 drums from each daily shipment and will reject the shipment (send it back to the supplier) if the average fill for the 100 drums is less than 49.85 gallons. Suppose that a shipment that actually has a mean fill of 50 gallons is received. What is the probability that this shipment will be rejected and sent back to the supplier?

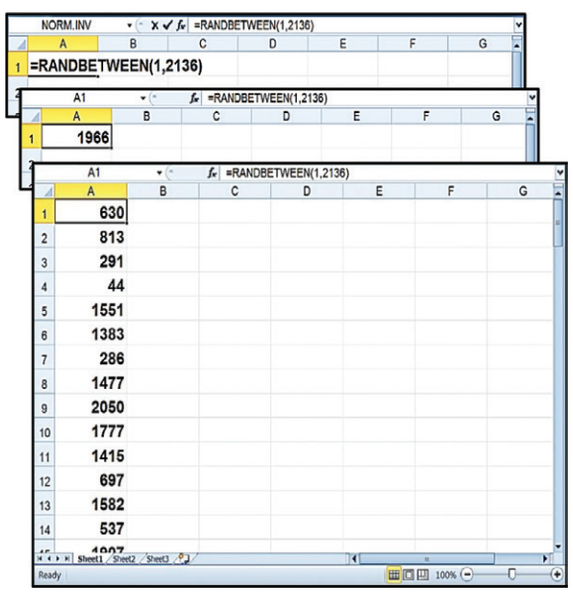
Appendix 7.1 ■ Generating Random Numbers Using Excel

To create 100 random numbers between 1 and 2136 similar to those in Table 7.1(b) on page 268.

- Type the cell formula

  =RANDBETWEEN(1,2136)

  into cell A1 of the Excel worksheet and press the enter key. This will generate a random integer between 1 and 2136, which will be placed in cell A1.
- Using the mouse, copy the cell formula for cell A1 down through cell A100. This will generate 100 random numbers between 1 and 2136 in cells A1 through A100 (note that the random number in cell A1 will change when this is done—this is not a problem).
- The random numbers are generated with replacement. Repeated numbers would be skipped if the random numbers were being used to sample without replacement.

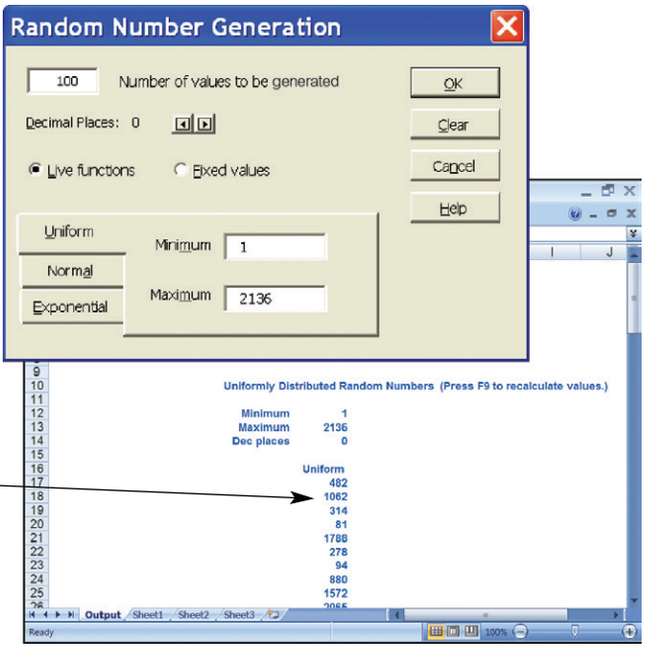


Appendix 7.2 ■ Generating Random Numbers Using MegaStat

To create 100 random numbers between 1 and 2136 similar to those in Table 7.1(b) on page 268:

- Select **Add-Ins : MegaStat : Generate Random Numbers…**
- In the Random Number Generation dialog box, enter 100 into the "Number of values to be generated" window.
- Click the right arrow button to select 0 Decimal Places.
- Select the Uniform tab, and enter 1 into the Minimum box and enter 2136 into the Maximum box.
- Click OK in the Random Number Generation dialog box.

The 100 random numbers will be placed in the Output Sheet. These numbers are generated with replacement. Repeated numbers would be skipped for random sampling without replacement.

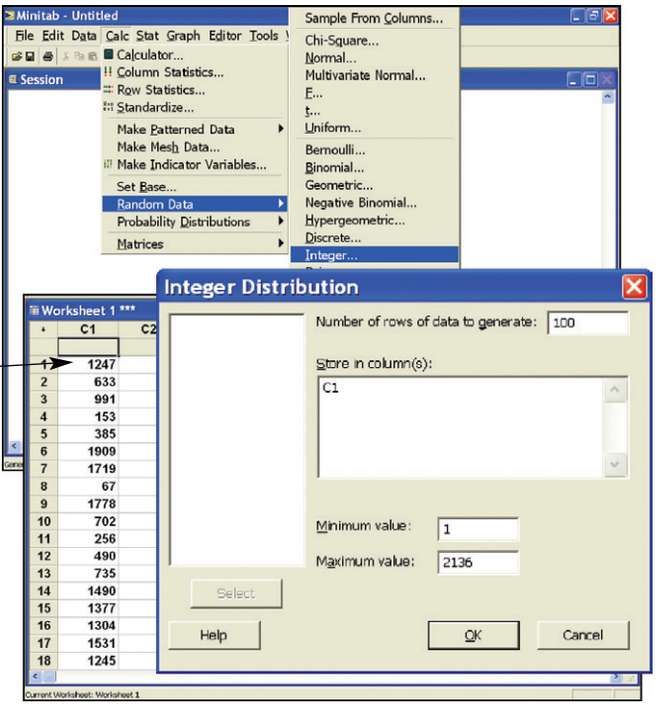


Appendix 7.3 ■ Generating Random Numbers and Simulating Sampling Distributions Using MINITAB

To create 100 random numbers between 1 and 2136 similar to those in Table 7.1(b) on page 268:

- Select **Calc : Random Data : Integer**
- In the Integer Distribution dialog box, enter 100 into the "Number of rows of data to generate" window.
- Enter C1 into the "Store in column(s)" window.
- Enter 1 into the Minimum value box and enter 2136 into the Maximum value box.
- Click OK in the Integer Distribution dialog box.

The 100 random numbers will be placed in the Data Window in column C1. These numbers are generated with replacement. Repeated numbers would be skipped if the random numbers are being used to sample without replacement.



Histogram of sample means from an exponential distribution similar to Figure 7.6(a) on page 279:

In this example we construct a histogram of 1000 sample means from exponential samples of size 5.

- Select **Calc : Random Data : Exponential.**
- In the Exponential Distribution dialog box, enter 1000 into the "Number of rows of data to generate:" window.
- Enter C1-C5 in the "Store in column(s):" window to request 1000 values per column in columns C1 to C5.
- Be sure that 0.0 is the entry in the Threshold window.
- Enter 7 in the Scale window. This specifies the mean of the exponential distribution when the Threshold equals 0.
- Click OK in the Exponential Distribution dialog box. The 1000 exponential samples of size 5 will be generated in rows 1 through 1000.
- Select **Calc : Row Statistics.**
- In the Row Statistics dialog box, under "Statistic" select the Mean option.
- Enter C1-C5 in the "Input variables" window.
- Enter XBar5 in the "Store result in" window.
- Click OK in the Row Statistics dialog box to compute the means for the 1000 samples of size 5.
- Select **Stat : Basic Statistics : Display Descriptive Statistics.**
- In the Display Descriptive Statistics dialog box, enter XBar5 into the Variables window.
- Click on the Graphs… button.
- In the "Display Descriptive Statistics—Graphs" dialog box, check the "Histogram of data, with normal curve" checkbox.
- Click OK in the "Display Descriptive Statistics—Graphs" dialog box.
- Click OK in the Display Descriptive Statistics dialog box.
- The histogram will appear in a graphics window.