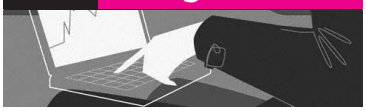
5 Sampling and Sampling Distributions

LEARNING OBJECTIVES

After studying this chapter, you should be able to:

- Take random samples from populations.
- Distinguish between population parameters and sample statistics.
- Apply the central limit theorem.
- Derive sampling distributions of sample means and proportions.
- Explain why sample statistics are good estimators of population parameters.
- Judge one estimator as better than another based on desirable properties of estimators.
- Apply the concept of degrees of freedom.
- Identify special sampling methods.
- Compute sampling distributions and related results using templates.

5–1 Using Statistics



Statistics is a science of *inference*. It is the science of generalization from a *part* (the randomly chosen sample) to the *whole* (the population).[1] Recall from Chapter 1 that the population is the entire collection of measurements in which we are interested, and the sample is a smaller set of measurements selected from the population. A random sample of n elements is a sample selected from the population in such a way that every set of n elements is as likely to be selected as any other set of n elements.[2] It is important that the sample be drawn randomly from the entire population under study. This increases the likelihood that our sample will be truly representative of the population of interest and minimizes the chance of errors. As we will see in this chapter, random sampling also allows us to compute the probabilities of sampling errors, thus providing us with knowledge of the degree of accuracy of our sampling results. The need to sample correctly is best illustrated by the well-known story of the *Literary Digest* (see page 182).

In 1936, the widely quoted *Literary Digest* embarked on the project of predicting the results of the presidential election to be held that year. The magazine boasted it would predict, to within a fraction of the percentage of the votes, the winner of the election—incumbent President Franklin Delano Roosevelt or the Republican governor of Kansas, Alfred M. Landon. The *Digest* tried to gather a sample of staggering proportion—10 million voters! One problem with the survey was that only a fraction of the people sampled, 2.3 million, actually provided the requested information. Should a link have existed between a person's inclination to answer the survey and his or her voting preference, the results of the survey would have been *biased:* slanted toward the voting preference of those who did answer. Whether such a link did exist in the case of the *Digest is not known*. (This problem, *nonresponse bias,* is discussed in Chapter 16.) A very serious problem with the *Digest*'s poll, and one known to have affected the results, is the following.

The sample of voters chosen by the *Literary Digest* was obtained from lists of telephone numbers, automobile registrations, and names of *Digest* readers. Remember that this was 1936—not as many people owned phones or cars as today, and those who did tended to be wealthier and more likely to vote Republican (and the same goes for readers of the *Digest*). The selection procedure for the sample of voters was thus biased (slanted toward one kind of voter) because the sample was not randomly chosen from the entire population of voters. Figure 5–1 demonstrates a correct sampling procedure versus the sampling procedure used by the *Literary Digest*.

As a result of the *Digest* error, the magazine does not exist today; it went bankrupt soon after the 1936 election. Some say that hindsight is useful and that today we know more statistics, making it easy for us to deride mistakes made more than 60 years ago. Interestingly enough, however, the ideas of sampling bias were understood in 1936. A few weeks *before* the election, a small article in *The New York Times* criticized the methodology of the *Digest* poll. Few paid it any attention.

[1]Not all of statistics concerns inferences about populations. One branch of statistics, called *descriptive statistics,* deals with describing data sets—possibly with no interest in an underlying population. The descriptive statistics of Chapter 1, when *not* used for inference, fall in this category.

[2]This is the definition of *simple random sampling,* and we will assume throughout that all our samples are simple random samples. Other methods of sampling are discussed in Chapter 6.

Digest Poll Gives Landon 32 States Landon Leads 4–3 in Last Digest Poll

Final Tabulation Gives Him 370 Electoral Votes to 161 for President Roosevelt

Governor Landon will win the election by an electoral vote of 370 to 161, will carry thirty-two of the forty-eight States, and will lead President Roosevelt about four to three in their share of the popular vote, if the final figures in The Literary Digest poll, made public yesterday, are verified by the count of the ballots next Tuesday.

The New York Times, October 30, 1936. Copyright © 1936 by The New York Times Company. Reprinted by permission.

Roosevelt's Plurality Is 11,000,000 History's Largest Poll 46 States Won by President, Maine and Vermont by Landon Many Phases to Victory

Democratic Landslide Looked Upon as Striking Personal Triumph for Roosevelt

By Arthur Krock

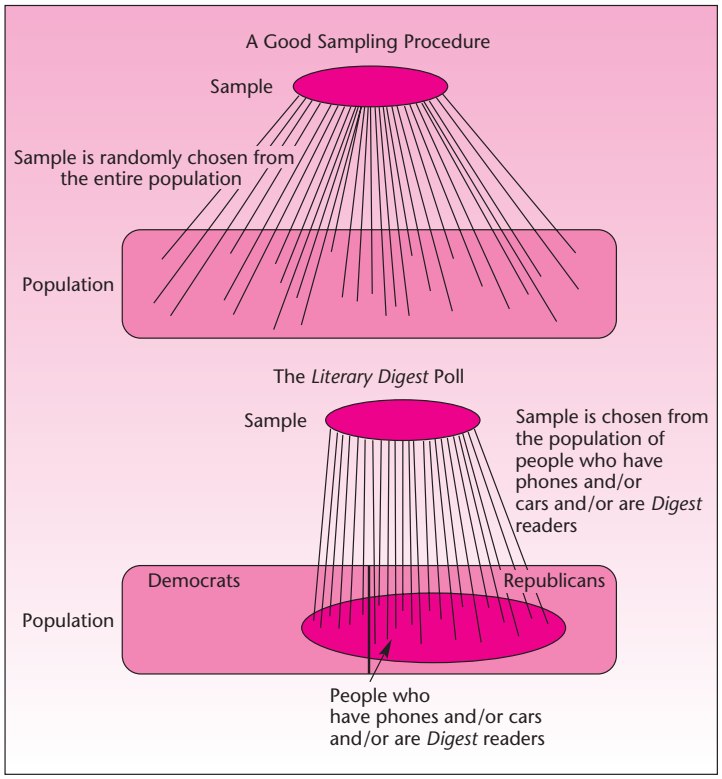
As the count of ballots cast Tuesday in the 1936 Presidential election moved toward completion yesterday, these facts appeared:

Franklin Delano Roosevelt was re-elected President, and John N. Garner Vice President, by the largest popular and electoral majority since the United States became a continental nation–a margin of approximately 11,000,000 plurality of all votes cast, and 523 votes in the electoral college to 8 won by the Republican Presidential candidate, Governor Alfred M. Landon of Kansas. The latter carried only Maine and Vermont of the forty-eight States of the Union . . . .

The New York Times, November 5, 1936. Copyright © 1936 by The New York Times Company. Reprinted by permission.

Sampling is very useful in many situations besides political polling, including business and other areas where we need to obtain information about some population. Our information often leads to a *decision*. There are also situations, as demonstrated by the examples in the introduction to this book, where we are interested in a *process* rather than a single population. One such process is the relationship between advertising and sales. In these more involved situations, we still make the assumption of an underlying population—here, the population of *pairs* of possible advertising and sales values. Conclusions about the process are reached based on information in our data, which are assumed to constitute a random sample from the entire population. The ideas of a population and of a random sample drawn from the population are thus essential to all inferential statistics.

FIGURE 5–1 A Good Sampling Procedure and the One Used by the *Literary Digest*



In statistical inference we are concerned with populations; the samples are of no interest to us in their own right. We wish to use our *known* random sample in the extraction of information about the *unknown* population from which it is drawn. The information we extract is in the form of summary statistics: a sample mean, a sample standard deviation, or other measures computed from the sample. A statistic such as the sample mean is considered an *estimator* of a population *parameter*—the population mean. In the next section, we discuss and define sample estimators and population parameters. Then we explore the relationship between statistics and parameters via the *sampling distribution*. Finally, we discuss desirable properties of statistical estimators.

5–2 Sample Statistics as Estimators of Population Parameters

A population may be a large, sometimes infinite, collection of elements. The population has a *frequency distribution*—the distribution of the frequencies of occurrence of its elements. The population distribution, when stated in relative frequencies, is also the probability distribution of the population. This is so because the relative frequency of a value in the population is also the probability of obtaining the particular value when an element is randomly drawn from the entire population. As with random variables, we may associate with a population its mean and its standard deviation.

In the case of populations, the mean and the standard deviation are called *parameters*. They are denoted by μ and σ, respectively.

A numerical measure of a population is called a **population parameter,** or simply a **parameter.**

Recall that in Chapter 4 we referred to the mean and the standard deviation of a normal probability distribution as the distribution parameters. Here we view parameters as descriptive measures of populations. Inference drawn about a population parameter is based on sample statistics.

A numerical measure of the sample is called a **sample statistic,** or simply a **statistic.**

Population parameters are estimated by sample statistics. When a sample statistic is used to estimate a population parameter, the statistic is called an *estimator* of the parameter.

An **estimator** of a population parameter is a sample statistic used to estimate the parameter. An **estimate** of the parameter is a *particular* numerical value of the estimator obtained by sampling. When a single value is used as an estimate, the estimate is called a **point estimate** of the population parameter.

The sample mean $\overline{X}$ is the sample statistic used as an estimator of the population mean μ. Once we sample from the population and obtain a value of $\overline{X}$ (using equation 1–1), we will have obtained a *particular* sample mean; we will denote this particular value by $\overline{x}$. We may have, for example, $\overline{x} = 12.53$. This value is our estimate of μ. The estimate is a point estimate because it constitutes a single number. In this chapter, every estimate will be a point estimate—a single number that, we hope, lies close to the population parameter it estimates. Chapter 6 is entirely devoted to the concept of an *interval estimate*—an estimate constituting an interval of numbers rather than a single number. An interval estimate is an interval believed likely to contain the unknown population parameter. It conveys more information than just the point estimate on which it is based.

In addition to the sample mean, which estimates the population mean, other statistics are useful. The sample variance S^2 is used as an estimator of the population variance σ^2. A particular estimate obtained will be denoted by s^2. (This estimate is computed from the data using equation 1–3 or an equivalent formula.)

As demonstrated by the political polling example with which we opened this chapter, interest often centers not on a mean or standard deviation of a population, but rather on a population *proportion*. The population proportion parameter is also called a binomial proportion parameter.

The **population proportion** p is equal to the number of elements in the population belonging to the category of interest, divided by the total number of elements in the population.

The population proportion of voters for Governor Landon in 1936, for example, was the number of people who intended to vote for the candidate, divided by the total number of voters. The estimator of the population proportion p is the *sample proportion* $\hat{P}$, defined as the number of *binomial successes* in the sample (i.e., the number of elements in the sample that belong to the category of interest), divided by the

sample size n. A particular estimate of the population proportion p is the sample proportion $\hat{p}$.

The **sample proportion** is

$$\hat{p} = \frac{x}{n} \qquad (5\text{–}1)$$

where x is the number of elements in the sample found to belong to the category of interest and n is the sample size.

Suppose that we want to estimate the proportion of consumers in a certain area who are users of a certain product. The (unknown) population proportion is p. We estimate p by the statistic $\hat{P}$, the sample proportion. Suppose a random sample of 100 consumers in the area reveals that 26 are users of the product. Our point estimate of p is then $\hat{p} = x/n = 26/100 = 0.26$. As another example, let's look at a very important problem, whose seriousness became apparent in early 2007, when more than a dozen dogs and cats in the United States became sick, and some died, after being fed pet food contaminated with an unknown additive originating in China. The culprit was melamine, an artificial additive derived from coal, which Chinese manufacturers have been adding to animal feed, and it was the cause of the death of pets and has even caused problems with the safety of eating farm products.[3] The wider problem of just how this harmful additive ended up in animal feed consumed in the United States is clearly statistical in nature, and it could have been prevented by effective use of sampling. It turned out that in the whole of 2006, Food and Drug Administration (FDA) inspectors sampled only 20,662 shipments out of 8.9 million arriving at American ports.[4] While this sampling percentage is small (about 0.2%), in this chapter you will learn that correct scientific sampling methods do not require larger samples, and good information can be gleaned from random samples of this size when they truly represent the population of all shipments. Suppose that this had indeed been done, and that 853 of the sampled shipments contained melamine. What is the sample estimate of the proportion of all shipments to the United States tainted with melamine? Using equation 5–1, we see that the estimate is $853/20{,}662 = 0.0413$, or about 4.13%.

In summary, we have the following estimation relationships:

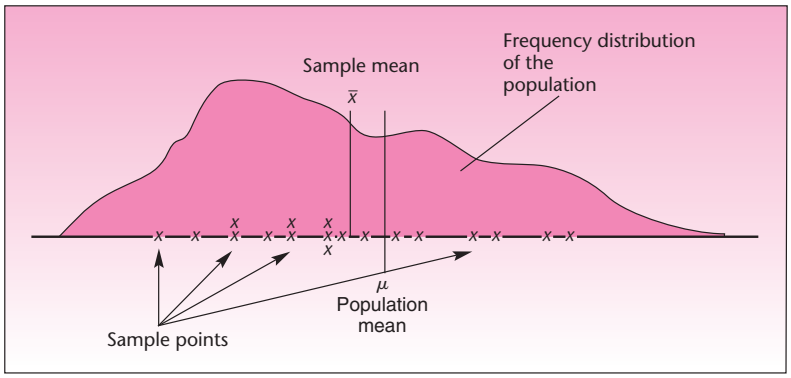
Estimator (Sample Statistic)		Population Parameter
$\bar{X}$	$\xrightarrow{\text{estimates}}$	μ
S^2	$\xrightarrow{\text{estimates}}$	σ^2
$\hat{P}$	$\xrightarrow{\text{estimates}}$	p

Let us consider sampling to estimate the population mean, and let us try to visualize how this is done. Consider a population with a certain frequency distribution. The frequency distribution of the values of the population is the probability distribution of the value of an element in the population, drawn at random. Figure 5–2 shows a frequency distribution of some population and the population mean μ. If we knew the exact frequency distribution of the population, we would be able to determine μ directly in the same way we determine the mean of a random variable when we know its probability distribution. In reality, the frequency distribution of a population is not known; neither is the mean of the population. We try to estimate the population mean by the sample mean, computed from a random sample. Figure 5–2 shows the values of a random sample obtained from the population and the resulting sample mean $\bar{x}$, computed from the data.

[3]Alexei Barrionuevo, "U.S. Says Some Chicken Feed Tainted," *The New York Times,* May 1, 2007, p. C6.

[4]Alexei Barrionuevo, "Food Imports Often Escape Scrutiny," *The New York Times,* May 1, 2007, p. C1.

FIGURE 5–2 **A Population Distribution, a Random Sample from the Population, and Their Respective Means**



In this example, $\bar{x}$ happens to lie close to μ, the population parameter it estimates, although this does not always happen. The sample statistic $\bar{X}$ is a *random variable* whose actual value depends on the particular random sample obtained. The random variable $\bar{X}$ has a relatively high probability of being close to the population mean it estimates, and it has decreasing probabilities of falling farther and farther from the population mean. Similarly, the sample statistic S is a random variable with a relatively high probability of being close to σ, the population parameter it estimates. Also, when sampling for a population proportion p, the estimator $\hat{P}$ has a relatively high probability of being close to p. How high a probability, and how close to the parameter? The answer to this question is the main topic of this chapter, presented in the next section. Before discussing this important topic, we will say a few things about the mechanics of obtaining random samples.

Obtaining a Random Sample

CHAPTER 8

All along we have been referring to random samples. We have stressed the importance of the fact that our sample should always be drawn randomly from the entire population about which we wish to draw an inference. How do we draw a random sample?

To obtain a random sample from the entire population, we need a list of all the elements in the population of interest. Such a list is called a *frame*. The frame allows us to draw elements from the population by randomly generating the numbers of the elements to be included in the sample. Suppose we need a simple random sample of 100 people from a population of 7,000. We make a list of all 7,000 people and assign each person an identification number. This gives us a list of 7,000 numbers—our frame for the experiment. Then we generate by computer or by other means a set of 100 random numbers in the range of values from 1 to 7,000. This procedure gives every set of 100 people in the population an equal chance of being included in the sample.

As mentioned, a computer (or an advanced calculator) may be used for generating random numbers. We will demonstrate an alternative method of choosing random numbers—a random number table. Table 5–1 is a part of such a table. A random number table is given in Appendix C as Table 14. To use the table, we start at any point, pick a number from the table, and continue in the same row or the same column (it does not matter which), systematically picking out numbers with the number of digits appropriate for our needs. If a number is outside our range of required numbers, we ignore it. We also ignore any number already obtained.

For example, suppose that we need a random sample of 10 data points from a population with a total of 600 elements. This means that we need 10 random drawings of

TABLE 5–1 Random Numbers

10480	15011	01536	02011	81647	91646	69179	14194
22368	46573	25595	85393	30995	89198	27982	53402
24130	48360	22527	97265	76393	64809	15179	24830
42167	93093	06243	61680	07856	16376	93440	53537
37570	39975	81837	16656	06121	91782	60468	81305
77921	06907	11008	42751	27756	53498	18602	70659

elements from our frame of 1 through 600. To do this, we note that the number 600 has three digits; therefore, we draw random numbers with three digits. Since our population has only 600 units, however, we ignore any number greater than 600 and take the next number, assuming it falls in our range. Let us decide arbitrarily to choose the first three digits in each set of five digits in Table 5–1; and we proceed by row, starting in the first row and moving to the second row, continuing until we have obtained our 10 required random numbers. We get the following random numbers: 104, 150, 15, 20, 816 (discard), 916 (discard), 691 (discard), 141, 223, 465, 255, 853 (discard), 309, 891 (discard), 279. Our random sample will, therefore, consist of the elements with serial numbers 104, 150, 15, 20, 141, 223, 465, 255, 309, and 279. A similar procedure would be used for obtaining the random sample of 100 people from the population of 7,000 mentioned earlier. Random number tables are included in books of statistical tables.

In many situations obtaining a frame of the elements in the population is impossible. In such situations we may still randomize some aspect of the experiment and thus obtain a random sample. For example, we may randomize the location and the time and date of the collection of our observations, as well as other factors involved. In estimating the average miles-per-gallon rating of an automobile, for example, we may randomly choose the dates and times of our trial runs as well as the particular automobiles used, the drivers, the roads used, and so on.

Other Sampling Methods

Sometimes a population may consist of distinct subpopulations, and including a certain number of samples from each subpopulation may be useful. For example, the students at a university may consist of 54% women and 46% men. We know that men and women may have very different opinions on the topic of a particular survey. Thus having proper representation of men and women in the random sample is desirable. If the total sample size is going to be 100, then a proper representation would mean 54 women and 46 men. Accordingly, the 54 women may be selected at random from a frame of only women students, and the 46 men may be selected similarly. Together they will make up a random sample of 100 with proper representation. This method of sampling is called **stratified sampling**.

In a stratified sampling the population is partitioned into two or more subpopulations called **strata,** and from each stratum a desired number of samples are selected at random.

Each stratum must be distinct in that it differs from other strata in some aspect that is relevant to the sampling experiment. Otherwise, stratification would yield no benefit. Besides sex, another common distinction between strata is their individual variances. For example, suppose we are interested in estimating the average income of all the families in a city. Three strata are possible: high-income, medium-income, and low-income families. High-income families may have a large variance in their incomes, medium-income families a smaller variance, and low-income families the least variance.

Then, by properly representing the three strata in a stratified sampling process, we can achieve a greater accuracy in the estimate than by a regular sampling process.

Sometimes, we may have to deviate from the regular sampling process for practical reasons. For example, suppose we want to find the average opinion of all voters in the state of Michigan on a state legislation issue. Assume that the budget for the sampling experiment is limited. A normal random sampling process will choose voters all over the state. It would be too costly to visit and interview every selected voter. Instead, we could choose a certain number of counties at random and from within the chosen counties select voters at random. This way, the travel will be restricted to chosen counties only. This method of sampling is called **cluster sampling.** Each county in our example is a *cluster.* After choosing a cluster at random if we sample every item or person in that cluster, then the method would be **single-stage cluster sampling.** If we choose a cluster at random and select items or people at random within the chosen clusters, as mentioned in our example, then that is **two-stage cluster sampling. Multistage cluster sampling** is also possible. For example, we might choose counties at random, then choose townships at random within the chosen counties, and finally choose voters at random within the chosen townships.

At times, the frame we have for a sampling experiment may itself be in random order. In such cases we could do a **systematic sampling.** Suppose we have a list of 3,000 customers and the order of customers in the list is random. Assume that we need a random sample of 100 customers. We first note that $3{,}000/100 = 30$. We then pick a number between 1 and 30 at random–say, 18. We select the 18th customer in the list and from there on, we pick every 30th customer in the list. In other words, we pick the 18th, 48th, 78th, and so on. In general, if N is the population size and n is the sample size, let $N/n = k$ where k is a rounded integer. We pick a number at random between 1 and k–say, l. We then pick the kth, $(l + k)$th, $(l + 2k)$th, . . . , items from the frame.

Systematic sampling may also be employed when a frame cannot be prepared. For example, a call center manager may want to select calls at random for monitoring purposes. Here a frame is impossible but the calls can reasonably be assumed to arrive in a random sequence, thus justifying a systematic selection of calls. Starting at a randomly selected time, one may choose every kth call where k depends on the call volume and the sample size desired.

Nonresponse

Nonresponse to sample surveys is one of the most serious problems that occur in practical applications of sampling methodology. The example of polling Jewish people, many of whom do not answer the phone on Saturday, mentioned in the *New York Times* article in 2003 (see Chapter 1), is a case in point. The problem is one of loss of information. For example, suppose that a survey questionnaire dealing with some issue is mailed to a randomly chosen sample of 500 people and that only 300 people respond to the survey. The question is: What can you say about the 200 people who did not respond? This is a very important question, and there is no immediate answer to it, precisely because the people did not respond; we know nothing about them. Suppose that the questionnaire asks for a yes or no answer to a particular public issue over which people have differing views, and we want to estimate the proportion of people who would respond yes. People may have such strong views about the issue that those who would respond no may refuse to respond altogether. In this case, the 200 nonrespondents to our survey will contain a higher proportion of "no" answers than the 300 responses we have. But, again, we would not know about this. The result will be a bias. How can we compensate for such a possible bias?

We may want to consider the population as made up of two *strata:* the respondents' stratum and the nonrespondents' stratum. In the original survey, we managed to sample only the respondents' stratum, and this caused the bias. What we need to do is to obtain a random sample from the nonrespondents' stratum. This is easier said than done. Still, there are ways we can at least reduce the bias and get some idea about the

proportion of "yes" answers in the nonresponse stratum. This entails *callbacks:* returning to the nonrespondents and asking them again. In some mail questionnaires, it is common to send several requests for response, and these reduce the uncertainty. There may, however, be hard-core refusers who just do not want to answer the questionnaire. Such people are likely to have very distinct views about the issue in question, and if you leave them out, there will be a significant bias in your conclusions. In such a situation, gathering a small random sample of the hard-core refusers and offering them some monetary reward for their answers may be useful. In cases where people may find the question embarrassing or may worry about revealing their personal views, a random-response mechanism whereby the respondent randomly answers one of two questions–one the sensitive question, and the other an innocuous question of no relevance–may elicit answers. The interviewer does not know which question any particular respondent answered but does know the probability of answering the sensitive question. This still allows for computation of the aggregated response to the sensitive question while protecting any given respondent's privacy.

PROBLEMS

5–1. Discuss the concepts of a parameter, a sample statistic, an estimator, and an estimate. What are the relations among these entities?

5–2. An auditor selected a random sample of 12 accounts from all accounts receivable of a given firm. The amounts of the accounts, in dollars, are as follows: 87.50, 123.10, 45.30, 52.22, 213.00, 155.00, 39.00, 76.05, 49.80, 99.99, 132.00, 102.11. Compute an estimate of the mean amount of all accounts receivable. Give an estimate of the variance of all the amounts.

5–3. In problem 5–2, suppose the auditor wants to estimate the proportion of all the firm's accounts receivable with amounts over $100. Give a point estimate of this parameter.

5–4. An article in the *New York Times* describes an interesting business phenomenon. The owners of small businesses tend to pay themselves much smaller salaries than they would earn had they been working for someone else.[5] Suppose that a random sample of small business owners' monthly salaries, in dollars, are as follows: 1,000, 1,200, 1,700, 900, 2,100, 2,300, 830, 2,180, 1,300, 3,300, 7,150, 1,500. Compute point estimates of the mean and the standard deviation of the population monthly salaries of small business owners.

5–5. Starbucks regularly introduces new coffee drinks and attempts to evaluate how these drinks fare by estimating the price its franchises can charge for them and sell enough cups to justify marketing the drink.[6] Suppose the following random sample of prices a new drink sells for in New York (in dollars) is available: 4.50, 4.25, 4.10, 4.75, 4.80, 3.90, 4.20, 4.55, 4.65, 4.85, 3.85, 4.15, 4.85, 3.95, 4.30, 4.60, 4.00. Compute the sample estimators of the population mean and standard deviation.

5–6. A market research worker interviewed a random sample of 18 people about their use of a certain product. The results, in terms of Y or N (for Yes, a user of the product, or No, not a user of the product), are as follows: Y N N Y Y Y N Y N Y Y Y N Y N Y Y N. Estimate the population proportion of users of the product.

[5]Eva Tahmincioglu, "When the Boss Is Last in Line for a Paycheck," *The New York Times,* March 22, 2007, p. C5.

[6]Burt Helm, "Saving Starbucks' Soul," *BusinessWeek,* April 9, 2007, p. 56.

5–7. Use a random number table (you may use Table 5–1) to find identification numbers of elements to be used in a random sample of size $n = 25$ from a population of 950 elements.

5–8. Find five random numbers from 0 to 5,600.

5–9. Assume that you have a frame of 40 million voters (something the *Literary Digest* should have had for an unbiased polling). Randomly generate the numbers of five sampled voters.

5–10. Suppose you need to sample the concentration of a chemical in a production process that goes on continuously 24 hours per day, 7 days per week. You need to generate a random sample of six observations of the process over a period of one week. Use a computer, a calculator, or a random number table to generate the six observation times (to the nearest minute).

5–3 Sampling Distributions

The **sampling distribution** of a statistic is the probability distribution of all possible values the statistic may take when computed from random samples of the same size, drawn from a specified population.

CHAPTER 8

Let us first look at the sample mean $\bar{X}$. The sample mean is a random variable. The possible values of this random variable depend on the possible values of the elements in the random sample from which $\bar{X}$ is to be computed. The random sample, in turn, depends on the distribution of the population from which it is drawn. As a random variable, $\bar{X}$ has a *probability distribution*. This probability distribution is the sampling distribution of $\bar{X}$.

The **sampling distribution of $\bar{X}$** is the probability distribution of all possible values the random variable $\bar{X}$ may take when a sample of size n is taken from a specified population.

Let us derive the sampling distribution of $\bar{X}$ in the simple case of drawing a sample of size $n = 2$ items from a population uniformly distributed over the integers 1 through 8. That is, we have a large population consisting of equal proportions of the values 1 to 8. At each draw, there is a 1/8 probability of obtaining any of the values 1 through 8 (alternatively, we may assume there are only eight elements, 1 through 8, and that the sampling is done with replacement). The sample space of the values of the two sample points drawn from this population is given in Table 5–2. This is an example. In real situations, sample sizes are much larger.

TABLE 5–2 Possible Values of Two Sample Points from a Uniform Population of the Integers 1 through 8

Second Sample Point	First Sample Point							
	1	2	3	4	5	6	7	8
1	1,1	2,1	3,1	4,1	5,1	6,1	7,1	8,1
2	1,2	2,2	3,2	4,2	5,2	6,2	7,2	8,2
3	1,3	2,3	3,3	4,3	5,3	6,3	7,3	8,3
4	1,4	2,4	3,4	4,4	5,4	6,4	7,4	8,4
5	1,5	2,5	3,5	4,5	5,5	6,5	7,5	8,5
6	1,6	2,6	3,6	4,6	5,6	6,6	7,6	8,6
7	1,7	2,7	3,7	4,7	5,7	6,7	7,7	8,7
8	1,8	2,8	3,8	4,8	5,8	6,8	7,8	8,8

TABLE 5–3 The Sampling Distribution of $\bar{X}$ for a Sample of Size 2 from a Uniformly Distributed Population of the Integers 1 to 8

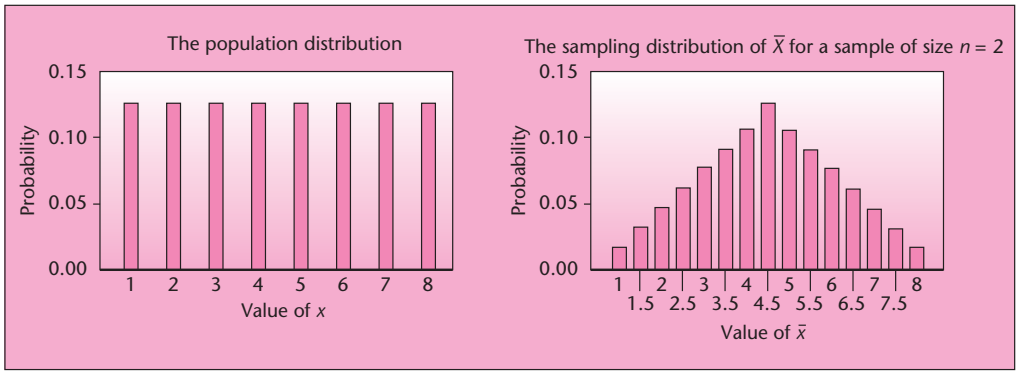
Particular Value $\bar{x}$	Probability of $\bar{x}$	Particular Value $\bar{x}$	Probability of $\bar{x}$
1	1/64	5	7/64
1.5	2/64	5.5	6/64
2	3/64	6	5/64
2.5	4/64	6.5	4/64
3	5/64	7	3/64
3.5	6/64	7.5	2/64
4	7/64	8	1/64
4.5	8/64		1.00

Using the sample space from the table, we will now find all possible values of the sample mean $\bar{X}$ and their probabilities. We compute these probabilities, using the fact that all 64 sample pairs shown are equally likely. This is so because the population is uniformly distributed and because in random sampling each drawing is independent of the other; therefore, the probability of a given pair of sample points is the product $(1/8)(1/8) = 1/64$. From Table 5–2, we compute the sample mean associated with each of the 64 pairs of numbers and find the probability of occurrence of each value of the sample mean. The values and their probabilities are given in Table 5–3. The table thus gives us the sampling distribution of $\bar{X}$ in this particular sampling situation. Verify the values in Table 5–3 using the sample space given in Table 5–2. Figure 5–3 shows the uniform distribution of the population and the sampling distribution of $\bar{X}$, as listed in Table 5–3.

Let us find the mean and the standard deviation of the *population*. We can do this by treating the population as a random variable (the random variable being the value of a single item randomly drawn from the population; each of the values 1 through 8 has a 1/8 probability of being drawn). Using the appropriate equations from Chapter 3, we find $\mu = 4.5$ and $\sigma = 2.29$ (verify these results).

Now let us find the expected value and the standard deviation of the random variable $\bar{X}$. Using the sampling distribution listed in Table 5–3, we find $E(\bar{X}) = 4.5$ and $\sigma_{\bar{x}} = 1.62$ (verify these values by computation). Note that the expected value of $\bar{X}$ is equal to the mean of the population; each is equal to 4.5. The standard deviation of $\bar{X}$, denoted $\sigma_{\bar{x}}$, is equal to 1.62, and the population standard deviation σ is 2.29. But observe an interesting fact: $2.29/\sqrt{2} = 1.62$. The facts we have discovered in this example are not an accident–they hold in all cases. The expected value of the sample

FIGURE 5–3 The Population Distribution and the Sampling Distribution of the Sample Mean



mean $\overline{X}$ is equal to the population mean μ and the standard deviation of $\overline{X}$ is equal to the population standard deviation divided by the square root of the sample size. Sometimes the estimated standard deviation of a statistic is called its *standard error.*

The expected value of the sample mean is[7]

$$E(\overline{X}) = \mu \tag{5–2}$$

The standard deviation of the sample mean is[8]

$$\text{SD}(\overline{X}) = \sigma_{\overline{x}} = \sigma/\sqrt{n} \tag{5–3}$$

We know the two parameters of the sampling distribution of $\overline{X}$: We know the mean of the distribution (the expected value of $\overline{X}$) and we know its standard deviation. What about the shape of the sampling distribution? If the population itself is *normally distributed,* the sampling distribution of $\overline{X}$ is also normal.

When sampling is done from a *normal distribution* with mean μ and standard deviation σ, the sample mean $\overline{X}$ has a **normal sampling distribution:**

$$\overline{X} \sim \text{N}(\mu, \sigma^2/n) \tag{5–4}$$

Thus, when we sample from a normally distributed population with mean μ and standard deviation σ, the sample mean has a normal distribution with the same *center,* μ, as the population but with *width* (standard deviation) that is $1/\sqrt{n}$ the size of the width of the population distribution. This is demonstrated in Figure 5–4, which shows a normal population distribution and the sampling distribution of $\overline{X}$ for different sample sizes.

The fact that the sampling distribution of $\overline{X}$ has mean μ is very important. It means that, *on the average,* the sample mean is equal to the population mean. The distribution of the statistic is *centered* on the parameter to be estimated, and this makes the statistic $\overline{X}$ a good estimator of μ. This fact will become clearer in the next section, where we discuss estimators and their properties. The fact that the standard deviation of $\overline{X}$ is $\sigma/\sqrt{n}$ means that as the sample size *increases,* the standard deviation of $\overline{X}$ *decreases,* making $\overline{X}$ more likely to be close to μ. This is another desirable property of a good estimator, to be discussed later. Finally, when the sampling distribution of $\overline{X}$ is normal, this allows us to compute probabilities that $\overline{X}$ will be within specified distances of μ. What happens in cases where the population itself is *not* normally distributed?

In Figure 5–3, we saw the sampling distribution of $\overline{X}$ when sampling is done from a uniformly distributed population and with a sample of size $n = 2$. Let us now see what happens as we increase the sample size. Figure 5–5 shows results of a simulation giving the sampling distribution of $\overline{X}$ when the sample size is $n = 5$, when the sample size is $n = 20$, and the *limiting* distribution of $\overline{X}$—the distribution of $\overline{X}$ as the sample size increases indefinitely. As can be seen from the figure, the limiting distribution of $\overline{X}$ is, again, the *normal distribution.*

[7]The proof of equation 5–2 relies on the fact that the expected value of the sum of several random variables is equal to the sum of their expected values. Also, from equation 3–6 we know that the expected value of aX, where a is a number, is equal to a times the expected value of X. We also know that the expected value of each element X drawn from the population is equal to μ, the population mean. Using these facts, we find the following: $E(\overline{X}) = E(\Sigma X/n) = (1/n)E(\Sigma X) = (1/n)n\mu = \mu$.

[8]The proof of equation 5–3 relies on the fact that, when several random variables are *independent* (as happens in random sampling), the variance of the sum of the random variables is equal to the sum of their variances. Also, from equation 3–10, we know that the variance of aX is equal to $a^2V(X)$. The variance of each X drawn from the population is equal to σ^2. Using these facts, we find $V(\overline{X}) = V(\Sigma X/n) = (1/n)^2(\Sigma\sigma^2) = (1/n)^2(n\sigma^2) = \sigma^2/n$. Hence, $\text{SD}(\overline{X}) = \sigma/\sqrt{n}$.

FIGURE 5–4 A Normally Distributed Population and the Sampling Distribution of the Sample Mean for Different Sample Sizes

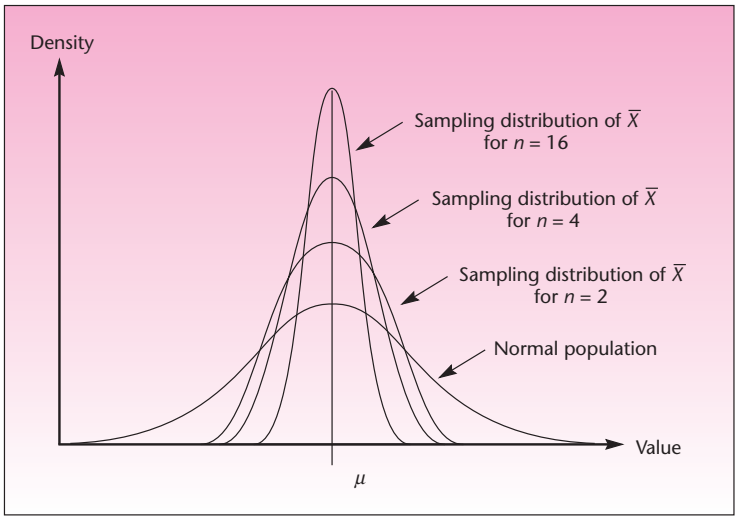
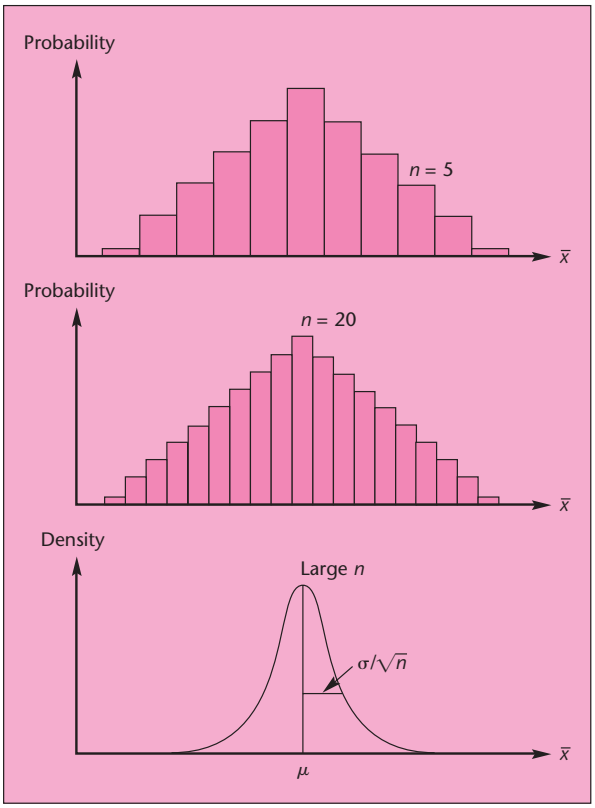


FIGURE 5–5 The Sampling Distribution of $\overline{X}$ as the Sample Size Increases



The Central Limit Theorem

The result we just stated—that the distribution of the sample mean $\bar{X}$ tends to the normal distribution as the sample size increases—is one of the most important results in statistics. It is known as the *central limit theorem*.

The Central Limit Theorem (and additional properties)

When sampling is done from a population with mean μ and finite standard deviation σ, the sampling distribution of the sample mean $\bar{X}$ will tend to a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$ as the sample size n becomes large.

$$\text{For "large enough" } n \qquad \bar{X} \sim N(\mu, \sigma^2/n) \tag{5–5}$$

The central limit theorem is remarkable because it states that the distribution of the sample mean $\bar{X}$ tends to a normal distribution *regardless* of the distribution of the population from which the random sample is drawn. The theorem allows us to make probability statements about the possible range of values the sample mean may take. It allows us to compute probabilities of how far away $\bar{X}$ may be from the population mean it estimates. For example, using our rule of thumb for the normal distribution, we know that the probability that the distance between $\bar{X}$ and μ will be less than $\sigma/\sqrt{n}$ is approximately 0.68. This is so because, as you remember, the probability that the value of a normal random variable will be within 1 standard deviation of its mean is 0.6826; here our normal random variable has mean μ and standard deviation $\sigma/\sqrt{n}$. Other probability statements can be made as well; we will see their use shortly. When is a sample size n "large enough" that we may apply the theorem?

The central limit theorem says that, *in the limit,* as n goes to infinity $(n \to \infty)$, the distribution of $\bar{X}$ becomes a normal distribution (regardless of the distribution of the population). The *rate* at which the distribution approaches a normal distribution does depend, however, on the shape of the distribution of the parent population. If the population itself is normally distributed, the distribution of $\bar{X}$ is normal for *any* sample size n, as stated earlier. On the other hand, for population distributions that are very different from a normal distribution, a relatively large sample size is required to achieve a good normal approximation for the distribution of $\bar{X}$. Figure 5–6 shows several parent population distributions and the resulting sampling distributions of $\bar{X}$ for different sample sizes.

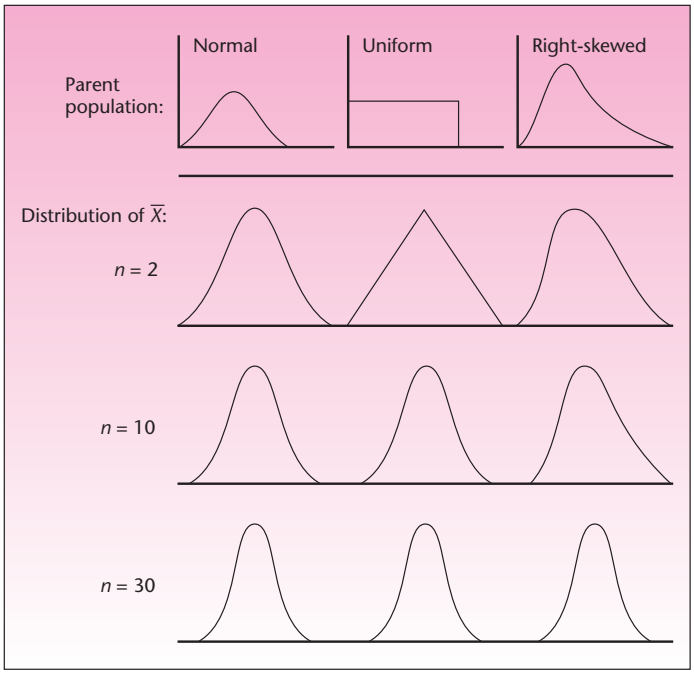
Since we often do not know the shape of the population distribution, some general rule of thumb telling us when a sample is large enough that we may apply the central limit theorem would be useful.

In general, a sample of 30 or more elements is considered **large enough** for the central limit theorem to take effect.

We emphasize that this is a *general,* and somewhat arbitrary, rule. A larger minimum sample size may be required for a good normal approximation when the population distribution is very different from a normal distribution. By the same token, a smaller minimum sample size may suffice for a good normal approximation when the population distribution is close to a normal distribution.

Throughout this book, we will make reference to *small* samples versus *large* samples. By a small sample, we generally mean a sample of fewer than 30 elements. A large sample will generally mean a sample of 30 or more elements. The results we will discuss as applicable for large samples will be more meaningful, however, the larger the sample size. (By the central limit theorem, the larger the sample size, the better the approximation offered by the normal distribution.) The "30 rule" should, therefore, be applied with caution. Let us now look at an example of the use of the central limit theorem.

FIGURE 5–6 The Effects of the Central Limit Theorem: The Distribution of $\bar{X}$ for Different Populations and Different Sample Sizes



EXAMPLE 5–1

Mercury makes a 2.4-liter V-6 engine, the Laser XRi, used in speedboats. The company's engineers believe that the engine delivers an average power of 220 horsepower and that the standard deviation of power delivered is 15 horsepower. A potential buyer intends to sample 100 engines (each engine to be run a single time). What is the probability that the sample mean $\bar{X}$ will be less than 217 horsepower?

Solution

In solving problems such as this one, we use the techniques of Chapter 4. There we used μ as the mean of the normal random variable and σ as its standard deviation. Here our random variable $\bar{X}$ is normal (at least approximately so, by the central limit theorem because our sample size is large) and has mean μ. Note, however, that the standard deviation of our random variable $\bar{X}$ is $\sigma/\sqrt{n}$ and not just σ. We proceed as follows:

$$P(\bar{X} < 217) = P\left(Z < \frac{217 - \mu}{\sigma/\sqrt{n}}\right)$$
$$= P\left(Z < \frac{217 - 220}{15/\sqrt{100}}\right) = P(Z < -2) = 0.0228$$

Thus, if the population mean is indeed $\mu = 220$ horsepower and the standard deviation is $\sigma = 15$ horsepower, the probability that the potential buyer's tests will result in a sample mean less than 217 horsepower is rather small.

FIGURE 5–7 A (Nonnormal) Population Distribution and the Normal Sampling Distribution of the Sample Mean When a Large Sample Is Used

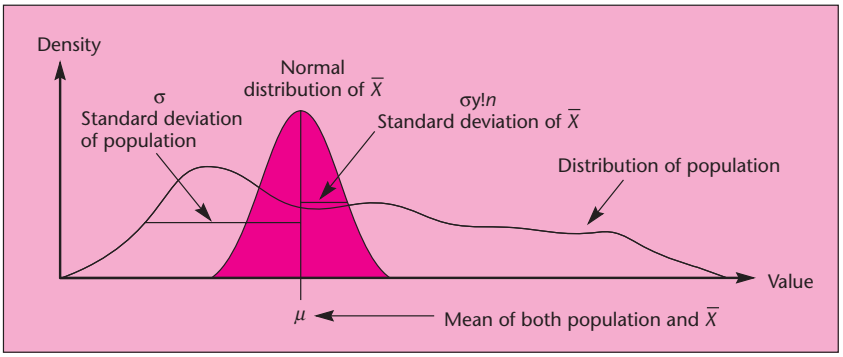


Figure 5–7 should help clarify the distinction between the population distribution and the sampling distribution of $\overline{X}$. The figure emphasizes the three aspects of the central limit theorem:

1. When the sample size is large enough, the sampling distribution of $\overline{X}$ is normal.
2. The expected value of $\overline{X}$ is μ.
3. The standard deviation of $\overline{X}$ is $\sigma/\sqrt{n}$.

The last statement is the key to the important fact that as the sample size increases, the variation of $\overline{X}$ about its mean μ decreases. Stated another way, as we buy *more information* (take a larger sample), our *uncertainty* (measured by the standard deviation) about the parameter being estimated *decreases*.

EXAMPLE 5–2

Eastern-Based Financial Institutions
Second-Quarter EPS and Statistical Summary

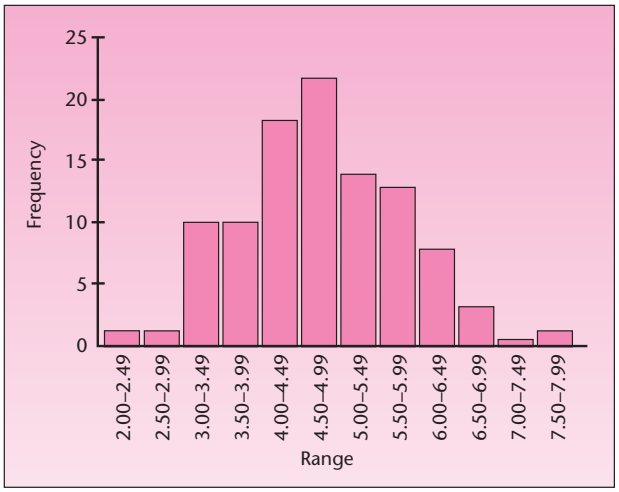
Corporation	EPS ($)
Bank of New York	2.53
Bank Boston	4.38
Banker's Trust NY	7.53
Chase Manhattan	7.53
Citicorp	7.93
Fleet	4.35
MBNA	1.50
Mellon	2.75
JP Morgan	7.25
PNC Bank	3.11
Republic Bank	7.44
State Street Bank	2.04
Summit	3.25

Summary	
Sample size	13
Mean EPS	4.7377
Median EPS	4.3500
Standard deviation	2.4346

This example shows random samples from the data above. Here 100 random samples of five banks each are chosen with replacement. The mean for each sample is computed, and a frequency distribution is drawn. Note the shape of this distribution (Figure 5–8).

Data Set		RS 1	RS 2	RS 3	RS 4	RS 5	RS 6	RS 7	RS 8	RS 9	RS 10	RS 11	RS 12	RS 13	RS 14	RS 15	RS 16	RS 17	RS 18	RS 19	RS 20
		2.53	2.04	2.53	3.25	7.53	4.35	7.93	2.04	7.53	2.04	7.53	3.11	3.25	2.75	7.44	7.44	2.04	2.53	7.93	7.53
2.53		7.53	7.53	2.04	2.75	4.38	1.50	2.53	3.25	3.25	3.25	3.11	7.53	7.53	2.53	4.35	7.53	3.11	4.35	4.38	3.11
4.38		7.93	7.44	7.93	7.93	2.04	7.93	7.53	4.35	2.04	7.53	2.75	1.50	3.25	3.25	7.44	7.53	7.53	4.35	7.25	3.25
7.53		2.53	3.25	4.38	2.04	4.35	7.25	2.75	7.53	3.11	7.93	1.50	7.53	1.50	3.11	2.53	7.53	1.50	1.50	4.35	2.53
7.53		2.75	4.38	4.38	2.53	7.93	3.11	3.25	4.35	2.04	7.53	7.53	2.75	7.53	7.44	3.11	3.25	3.11	4.38	3.25	3.11
7.93	Mean	4.65	4.93	4.25	3.70	5.25	4.83	4.80	4.30	3.59	5.66	4.48	4.48	4.61	3.82	4.97	6.66	3.46	3.42	5.43	3.91
4.35		RS 21	RS 22	RS 23	RS 24	RS 25	RS 26	RS 27	RS 28	RS 29	RS 30	RS 31	RS 32	RS 33	RS 34	RS 35	RS 36	RS 37	RS 38	RS 39	RS 40
1.50		3.11	1.50	2.75	7.53	7.44	7.93	2.53	7.93	7.53	4.38	7.93	7.93	7.44	4.35	7.53	7.93	4.38	4.35	7.44	2.53
2.75		2.04	2.04	7.53	2.04	4.35	1.50	3.11	1.50	7.53	7.53	7.93	7.53	3.25	7.25	1.50	2.75	7.93	3.25	7.53	3.25
7.25		3.25	1.50	2.04	4.38	2.75	7.53	3.25	3.11	4.38	2.53	2.75	4.35	4.38	7.25	4.35	1.50	7.93	3.11	4.35	2.53
3.11		4.38	3.25	7.53	2.53	4.35	2.75	7.25	7.93	7.44	3.11	7.93	7.53	3.25	4.35	4.35	2.04	4.35	1.50	3.25	1.50
7.44		2.75	2.75	7.93	2.75	2.04	2.75	1.50	1.50	3.11	7.44	3.11	3.11	7.44	7.53	7.93	2.04	4.38	2.04	2.53	7.53
2.04	Mean	3.11	2.21	5.56	3.85	4.19	4.49	3.53	4.39	6.00	5.00	5.93	6.09	5.15	6.15	5.13	3.25	5.79	2.85	5.02	3.47
3.25		RS 41	RS 42	RS 43	RS 44	RS 45	RS 46	RS 47	RS 48	RS 49	RS 50	RS 51	RS 52	RS 53	RS 54	RS 55	RS 56	RS 57	RS 58	RS 59	RS 60
		1.50	1.50	2.75	2.75	4.35	7.53	7.44	7.53	4.35	7.44	3.25	2.53	2.53	7.53	7.25	2.75	7.53	1.50	2.75	2.75
		4.38	7.25	7.44	4.35	1.50	7.93	3.25	4.35	3.11	7.25	2.75	7.53	7.53	4.38	7.53	2.04	2.75	1.50	7.93	7.53
		4.38	7.25	1.50	4.35	3.25	3.25	7.25	7.53	7.44	3.11	4.35	2.75	1.50	4.38	1.50	7.53	3.11	2.04	3.11	7.53
		3.11	4.38	2.75	3.11	2.75	7.53	2.04	7.25	4.35	3.11	4.35	7.53	7.53	4.38	7.25	1.50	7.93	7.25	7.93	7.53
		3.25	7.53	2.04	4.38	7.44	2.04	3.11	4.38	3.25	7.53	4.35	1.50	2.04	7.53	3.25	7.93	2.75	2.75	7.25	3.11
	Mean	3.32	5.58	3.30	3.79	3.86	5.66	4.62	6.21	4.50	5.69	3.81	4.37	4.23	5.64	5.36	4.35	4.81	3.01	5.79	5.69
		RS 61	RS 62	RS 63	RS 64	RS 65	RS 66	RS 67	RS 68	RS 69	RS 70	RS 71	RS 72	RS 73	RS 74	RS 75	RS 76	RS 77	RS 78	RS 79	RS 80
		4.38	7.93	3.25	7.53	3.25	2.53	7.25	3.11	7.25	7.53	2.04	7.44	7.25	7.25	7.44	3.25	7.53	7.44	2.53	3.25
		3.25	4.35	7.53	7.44	3.11	7.53	3.11	7.25	7.53	2.75	2.75	7.53	4.38	7.44	7.25	1.50	4.35	4.38	1.50	4.38
		7.93	7.53	3.25	4.35	3.11	7.25	7.25	7.44	7.53	7.53	7.44	4.38	7.25	7.53	2.75	7.25	3.11	1.50	7.53	3.25
		3.25	2.53	7.25	7.44	4.38	2.75	1.50	7.93	3.25	4.38	7.93	3.11	3.11	1.50	3.25	7.25	3.11	7.53	2.53	3.25
		4.35	4.38	3.25	3.25	7.53	4.38	4.38	2.75	7.93	7.25	7.53	7.53	2.04	2.75	3.11	2.04	2.75	2.53	3.25	2.75
	Mean	4.63	5.34	4.91	6.00	4.28	4.89	4.70	5.70	6.70	5.89	5.54	6.00	4.81	5.29	4.76	4.26	4.17	4.68	3.47	3.38
		RS 81	RS 82	RS 83	RS 84	RS 85	RS 86	RS 87	RS 88	RS 89	RS 90	RS 91	RS 92	RS 93	RS 94	RS 95	RS 96	RS 97	RS 98	RS 99	RS 100
		7.53	3.25	7.44	7.93	2.04	7.53	2.75	7.93	7.53	7.25	7.93	7.53	7.53	3.25	2.75	7.93	7.44	2.04	4.35	7.53
		3.25	3.11	7.53	2.04	7.53	7.93	4.38	1.50	4.38	4.38	7.25	7.25	3.11	7.93	3.11	2.04	2.04	7.53	7.93	7.53
		7.25	7.25	7.25	7.93	7.93	3.11	2.75	7.93	4.38	2.75	2.04	7.93	1.50	2.75	2.04	3.25	4.38	7.53	2.75	7.25
		3.11	1.50	7.53	2.04	2.53	3.11	7.25	3.11	2.75	7.53	4.38	7.53	2.04	7.93	4.38	4.35	2.75	7.93	3.25	2.53
		4.38	7.53	2.53	1.50	7.25	4.35	7.44	4.35	7.53	1.50	1.50	7.53	7.25	4.38	7.25	2.75	4.35	7.53	2.53	7.53
	Mean	5.10	4.53	6.46	4.29	5.46	5.21	4.91	4.96	5.31	4.68	4.62	7.55	4.29	5.25	3.91	4.06	4.19	6.51	4.16	6.47

FIGURE 5–8 EPS Mean Distribution—Excel Output



	A	B
1	Distribution	
2		
3	2.00–2.49	1
4	2.50–2.99	1
5	3.00–3.49	10
6	3.50–3.99	10
7	4.00–4.49	18
8	4.50–4.99	21
9	5.00–5.49	14
10	5.50–5.99	13
11	6.00–6.49	8
12	6.50–6.99	3
13	7.00–7.49	0
14	7.50–7.99	1

Figure 5–8 shows a graph of the means of the samples from the banks' data using Excel.

The History of the Central Limit Theorem

What we call the central limit theorem actually comprises several theorems developed over the years. The first such theorem was discussed at the beginning of Chapter 4 as the discovery of the normal curve by Abraham De Moivre in 1733. Recall that De Moivre discovered the normal distribution as the *limit* of the binomial distribution. The fact that the normal distribution appears as a limit of the binomial distribution as n increases is a form of the central limit theorem. Around the turn of the twentieth century, Liapunov gave a more general form of the central limit theorem, and in 1922 the final form we use in applied statistics was given by Lindeberg. The proof of the necessary condition of the theorem was given in 1935 by W. Feller [see W. Feller, *An Introduction to Probability Theory and Its Applications* (New York: Wiley, 1971), vol. 2]. A proof of the central limit theorem is beyond the scope of this book, but the interested reader is encouraged to read more about it in the given reference or in other books.

The Standardized Sampling Distribution of the Sample Mean When σ Is Not Known

To use the central limit theorem, we need to know the population standard deviation, σ. When σ is not known, we use its estimator, the sample standard deviation S, in its place. In such cases, the distribution of the standardized statistic

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \qquad (5\text{–}6)$$

(where S is used in place of the unknown σ) is no longer the standard normal distribution. *If the population itself is normally distributed, the statistic in equation 5–6 has a* t *distribution with* n − 1 *degrees of freedom.* The t distribution has wider tails than the standard normal distribution. Values and probabilities of t distributions with different degrees of freedom are given in Table 3 in Appendix C. The t distribution and its uses will be discussed in detail in Chapter 6. The idea of degrees of freedom is explained in section 5–5 of this chapter.

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

The sampling distribution of the sample proportion $\hat{P}$ is based on the binomial distribution with parameters n and p, where n is the sample size and p is the population proportion. Recall that the binomial random variable X counts the number of successes in n trials. Since $\hat{P} = X/n$ and n is fixed (determined before the sampling), the distribution of the number of successes X leads to the distribution of $\hat{P}$.

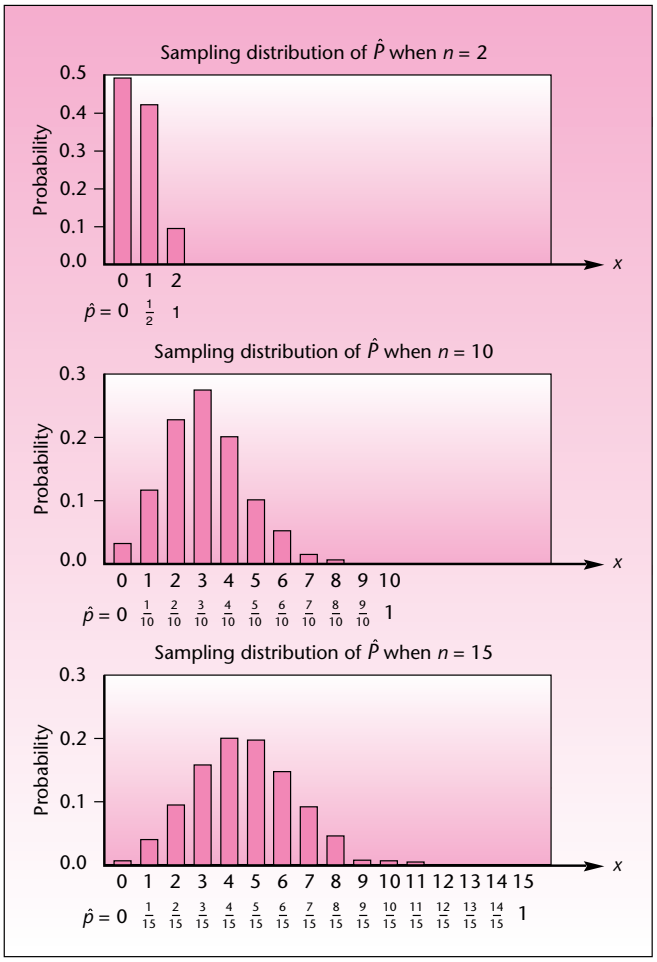
As the sample size increases, the central limit theorem applies here as well. Figure 5–9 shows the effects of the central limit theorem for a binomial distribution with $p = 0.3$. The distribution is skewed to the right for small values of n but becomes more symmetric and approaches the normal distribution as n increases.

We now state the central limit theorem when sampling for the population proportion p.

As the sample size n increases, the sampling distribution of $\hat{P}$ approaches a **normal distribution** with mean p and standard deviation $\sqrt{p(1-p)/n}$.

(The estimated standard deviation of $\hat{P}$ is also called its *standard error.*) In order for us to use the normal approximation for the sampling distribution of $\hat{P}$, the sample size needs to be large. A commonly used rule of thumb says that the normal approximation to the distribution of $\hat{P}$ may be used only if *both* np *and* $n(1-p)$ *are greater than* 5. We demonstrate the use of the theorem with Example 5–3.

FIGURE 5–9 The Sampling Distribution of $\hat{P}$ When $p = 0.3$, as n Increases



EXAMPLE 5–3

In recent years, convertible sport coupes have become very popular in Japan. Toyota is currently shipping Celicas to Los Angeles, where a customizer does a roof lift and ships them back to Japan. Suppose that 25% of all Japanese in a given income and lifestyle category are interested in buying Celica convertibles. A random sample of 100 Japanese consumers in the category of interest is to be selected. What is the probability that at least 20% of those in the sample will express an interest in a Celica convertible?

Solution

We need $P(\hat{P} \geq 0.20)$. Since $np = 100(0.25) = 25$ and $n(1 - p) = 100(0.75) = 75$, both numbers greater than 5, we may use the normal approximation to the distribution of $\hat{P}$. The mean of $\hat{P}$ is $p = 0.25$, and the standard deviation of $\hat{P}$ is $\sqrt{p(1 - p)/n} = 0.0433$. We have

$$P(\hat{P} \geq 0.20) = P\left(Z \geq \frac{0.20 - 0.25}{0.0433}\right) = P(Z \geq -1.15) = 0.8749$$

Sampling distributions are essential to statistics. In the following chapters, we will make much use of the distributions discussed in this section, as well as others that will be introduced as we go along. In the next section, we discuss properties of good estimators.

PROBLEMS

5–11. What is a sampling distribution, and what are the uses of sampling distributions?

5–12. A sample of size $n = 5$ is selected from a population. Under what conditions is the sampling distribution of $\bar{X}$ normal?

5–13. In problem 5–12, suppose the population mean is $\mu = 125$ and the population standard deviation is 20. What are the expected value and the standard deviation of $\bar{X}$?

5–14. What is the most significant aspect of the central limit theorem?

5–15. Under what conditions is the central limit theorem most useful in sampling to estimate the population mean?

5–16. What are the limitations of small samples?

5–17. When sampling is done from a population with population proportion $p = 0.1$, using a sample size $n = 2$, what is the sampling distribution of $\hat{P}$? Is it reasonable to use a normal approximation for this sampling distribution? Explain.

5–18. If the population mean is 1,247, the population variance is 10,000, and the sample size is 100, what is the probability that $\bar{X}$ will be less than 1,230?

5–19. When sampling is from a population with standard deviation $\sigma = 55$, using a sample of size $n = 150$, what is the probability that $\bar{X}$ will be at least 8 units away from the population mean μ?

5–20. The Colosseum, once the most popular monument in Rome, dates from about AD 70. Since then, earthquakes have caused considerable damage to the huge structure, and engineers are currently trying to make sure the building will survive future shocks. The Colosseum can be divided into several thousand small sections. Suppose that the average section can withstand a quake measuring 3.4 on the Richter scale with a standard deviation of 1.5. A random sample of 100 sections is selected and tested for the maximum earthquake force they can withstand. What is the probability that the average section in the sample can withstand an earthquake measuring at least 3.6 on the Richter scale?

5–21. According to *Money,* in the year prior to March 2007, the average return for firms of the S&P 500 was 13.1%.[9] Assume that the standard deviation of returns was 1.2%. If a random sample of 36 companies in the S&P 500 is selected, what is the probability that their average return for this period will be between 12% and 15%?

5–22. An economist wishes to estimate the average family income in a certain population. The population standard deviation is known to be $4,500, and the economist uses a random sample of size $n = 225$. What is the probability that the sample mean will fall within $800 of the population mean?

5–23. When sampling is done for the proportion of defective items in a large shipment, where the population proportion is 0.18 and the sample size is 200, what is the probability that the sample proportion will be at least 0.20?

5–24. A study of the investment industry claims that 58% of all mutual funds outperformed the stock market as a whole last year. An analyst wants to test this claim and obtains a random sample of 250 mutual funds. The analyst finds that only 123

[9]"Market Benchmarks," *Money,* March 2007, p. 128.

of the funds outperformed the market during the year. Determine the probability that another random sample would lead to a sample proportion as low as or lower than the one obtained by the analyst, assuming the proportion of all mutual funds that outperformed the market is indeed 0.58.

5–25. According to a recent article in *Worth,* the average price of a house on Marco Island, Florida, is $2.6 million.[10] Assume that the standard deviation of the prices is $400,000. A random sample of 75 houses is taken and the average price is computed. What is the probability that the sample mean exceeds $3 million?

5–26. It has been suggested that an investment portfolio selected randomly by throwing darts at the stock market page of *The Wall Street Journal* may be a sound (and certainly well-diversified) investment.[11] Suppose that you own such a portfolio of 16 stocks randomly selected from all stocks listed on the New York Stock Exchange (NYSE). On a certain day, you hear on the news that the average stock on the NYSE rose 1.5 points. Assuming that the standard deviation of stock price movements that day was 2 points and assuming stock price movements were normally distributed around their mean of 1.5, what is the probability that the average stock price of your portfolio increased?

5–27. An advertisement for Citicorp Insurance Services, Inc., claims "one person in seven will be hospitalized this year." Suppose you keep track of a random sample of 180 people over an entire year. Assuming Citicorp's advertisement is correct, what is the probability that fewer than 10% of the people in your sample will be found to have been hospitalized (at least once) during the year? Explain.

5–28. Shimano mountain bikes are displayed in chic clothing boutiques in Milan, Italy, and the average price for the bike in the city is $700. Suppose that the standard deviation of bike prices is $100. If a random sample of 60 boutiques is selected, what is the probability that the average price for a Shimano mountain bike in this sample will be between $680 and $720?

5–29. A quality-control analyst wants to estimate the proportion of imperfect jeans in a large warehouse. The analyst plans to select a random sample of 500 pairs of jeans and note the proportion of imperfect pairs. If the actual proportion in the entire warehouse is 0.35, what is the probability that the sample proportion will deviate from the population proportion by more than 0.05?

5–4 Estimators and Their Properties[12]

The sample statistics we discussed—$\bar{X}$, S, and $\hat{P}$—as well as other sample statistics to be introduced later, are used as estimators of population parameters. In this section, we discuss some important properties of good statistical estimators: *unbiasedness, efficiency, consistency,* and *sufficiency.*

CHAPTER 8

An estimator is said to be **unbiased** if its expected value is equal to the population parameter it estimates.

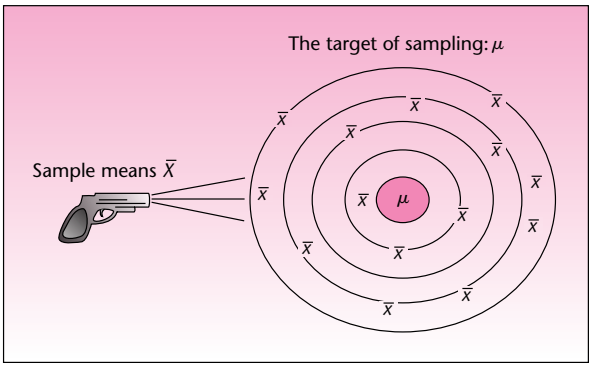
Consider the sample mean $\bar{X}$. From equation 5–2, we know $E(\bar{X}) = \mu$. *The sample mean $\bar{X}$ is, therefore, an unbiased estimator of the population mean* μ. This means that if we sample repeatedly from the population and compute $\bar{X}$ for each of our samples, *in the long run,* the average value of $\bar{X}$ will be the parameter of interest μ. This is an important property of the estimator because it means that there is no systematic *bias* away from the parameter of interest.

[10]Elizabeth Harris, "Luxury Real Estate Investment," *Worth,* April 2007, p. 76.

[11]See the very readable book by Burton G. Malkiel, *A Random Walk Down Wall Street* (New York: W. W. Norton, 2003).

[12]An optional, but recommended, section.

FIGURE 5–10 The Sample Mean $\bar{X}$ as an Unbiased Estimator of the Population Mean μ



If we view the gathering of a random sample and the calculating of its mean as shooting at a target—the target being the population parameter, say, μ—then the fact that $\bar{X}$ is an unbiased estimator of μ means that the device producing the estimates is aiming at the *center* of the target (the parameter of interest), with no systematic deviation away from it.

Any *systematic* deviation of the estimator away from the parameter of interest is called a **bias.**

The concept of unbiasedness is demonstrated for the sample mean $\bar{X}$ in Figure 5–10.

Figure 5–11 demonstrates the idea of a biased estimator of μ. The hypothetical estimator we denote by Y is centered on some point M that lies away from the parameter μ. The distance between the expected value of Y (the point M) and μ is the *bias*.

It should be noted that, in reality, we usually sample *once* and obtain our estimate. The multiple estimates shown in Figures 5–10 and 5–11 serve only as an illustration of the expected value of an estimator as the center of a large collection of the actual estimates that would be obtained in repeated sampling. (Note also that, in reality, the "target" at which we are "shooting" is one-dimensional—on a straight line rather than on a plane.)

FIGURE 5–11 An Example of a Biased Estimator of the Population Mean μ

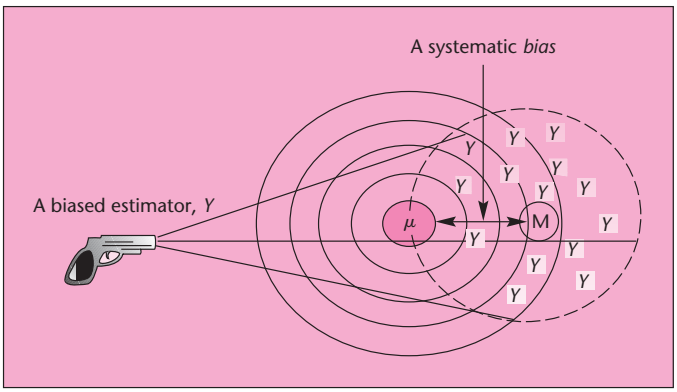
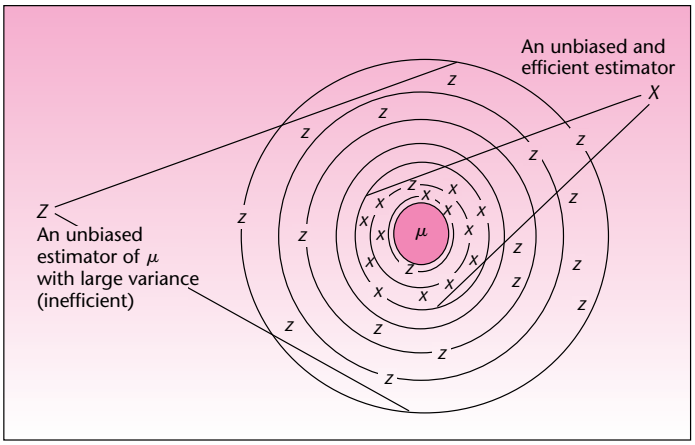


FIGURE 5–12 Two Unbiased Estimators of μ, Where the Estimator *X* Is Efficient Relative to the Estimator *Z*



The next property of good estimators we discuss is *efficiency*.

An estimator is **efficient** if it has a relatively small variance (and standard deviation).

Efficiency is a relative property. We say that one estimator is efficient *relative* to another. This means that the estimator has a smaller variance (also a smaller standard deviation) than the other. Figure 5–12 shows two hypothetical unbiased estimators of the population mean μ. The two estimators, which we denote by $\overline{X}$ and Z, are unbiased: Their distributions are centered at μ. The estimator $\overline{X}$, however, is more efficient than the estimator Z because it has a smaller variance than that of Z. This is seen from the fact that repeated estimates produced by Z have a larger spread about their mean μ than repeated estimates produced by $\overline{X}$.

Another desirable property of estimators is *consistency*.

An estimator is said to be **consistent** if its probability of being close to the parameter it estimates increases as the sample size increases.

The sample mean $\overline{X}$ is a consistent estimator of μ. This is so because the standard deviation of $\overline{X}$ is $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. As the sample size n increases, the standard deviation of $\overline{X}$ decreases and, hence, the probability that $\overline{X}$ will be close to its expected value μ increases.

We now define a fourth property of good estimators: *sufficiency*.

An estimator is said to be **sufficient** if it contains all the information in the data about the parameter it estimates.

Applying the Concepts of Unbiasedness, Efficiency, Consistency, and Sufficiency

We may evaluate possible estimators of population parameters based on whether they possess important properties of estimators and thus choose the best estimator to be used.

For a *normally distributed population,* for example, both the sample mean and the sample median are *unbiased* estimators of the population mean μ. The sample mean, however, is more *efficient* than the sample median. This is so because the variance of the sample median happens to be 1.57 times as large as the variance of the sample

mean. In addition, the sample mean is a *sufficient* estimator because in computing it we use the *entire* data set. The sample median is not sufficient; it is found as the point in the middle of the data set, regardless of the exact magnitudes of all other data elements. The sample mean $\bar{X}$ is the *best* estimator of the population mean μ, because it is unbiased and has the smallest variance of all unbiased estimators of μ. The sample mean is also *consistent*. (Note that while the sample mean is best, the sample median is sometimes used because it is more resistant to extreme observations.)

The sample proportion $\hat{P}$ is the best estimator of the population proportion p. Since $E(\hat{P}) = p$, the estimator $\hat{P}$ is unbiased. It also has the smallest variance of all unbiased estimators of p.

What about the sample variance S^2? The sample variance, as defined in equation 1–3, is an unbiased estimator of the population variance σ^2. Recall equation 1–3:

$$S^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1}$$

Dividing the sum of squared deviations in the equation by n rather than by $n - 1$ seems logical because we are seeking the *average* squared deviation from the sample mean. We have n deviations from the mean, so why not divide by n? It turns out that if we were to divide by n rather than by $n - 1$, our estimator of σ^2 would be biased. Although the bias becomes small as n increases, we will always use the statistic given in equation 1–3 as an estimator of σ^2. The reason for dividing by $n - 1$ rather than n will become clearer in the next section, when we discuss the concept of degrees of freedom.

Note that while S^2 is an unbiased estimator of the population variance σ^2, the sample standard deviation S (the square root of S^2) is *not* an unbiased estimator of the population standard deviation σ. Still, we will use S as our estimator of the population standard deviation, ignoring the small bias that results and relying on the fact that S^2 is the unbiased estimator of σ^2.

PROBLEMS

5–30. Suppose that you have two statistics A and B as possible estimators of the same population parameter. Estimator A is unbiased, but has a large variance. Estimator B has a small bias, but has only one-tenth the variance of estimator A. Which estimator is better? Explain.

5–31. Suppose that you have an estimator with a relatively large bias. The estimator is consistent and efficient, however. If you had a generous budget for your sampling survey, would you use this estimator? Explain.

5–32. Suppose that in a sampling survey to estimate the population variance, the biased estimator (with n instead of $n - 1$ in the denominator of equation 1–3) was used instead of the unbiased one. The sample size used was $n = 100$, and the estimate obtained was 1,287. Can you find the value of the unbiased estimate of the population variance?

5–33. What are the advantages of a sufficient statistic? Can you think of a possible disadvantage of sufficiency?

5–34. Suppose that you have two biased estimators of the same population parameter. Estimator A has a bias equal to $1/n$ (that is, the mean of the estimator is $1/n$ unit away from the parameter it estimates), where n is the sample size used. Estimator B has a bias equal to 0.01 (the mean of the estimator is 0.01 unit away from the parameter of interest). Under what conditions is estimator A better than B?

5–35. Why is consistency an important property?

5–5 Degrees of Freedom

Suppose you are asked to choose 10 numbers. You then have the freedom to choose 10 numbers as you please, and we say you have 10 **degrees of freedom.** But suppose a condition is imposed on the numbers. The condition is that the sum of all the numbers you choose must be 100. In this case, you cannot choose all 10 numbers as you please. After you have chosen the ninth number, let's say the sum of the nine numbers is 94. Your tenth number then has to be 6, and you have no choice. Thus you have only 9 degrees of freedom. In general, if you have to choose n numbers, and a condition on their total is imposed, you will have only $(n - 1)$ degrees of freedom.

As another example, suppose that I wrote five checks last month, and the total amount of these checks is \$80. Now if I know that the first four checks were for \$30, \$20, \$15, and \$5, then I don't need to be told that the fifth check was for \$10. I can simply deduce this information by subtraction of the other four checks from \$80. My degrees of freedom are thus four, and not five.

In Chapter 1, we saw the formula for the sample variance

$$S^2 = \text{SSD}/(n - 1)$$

where SSD is the sum of squared deviations from the sample mean. In particular, note that SSD is to be divided by $(n - 1)$ rather than n. The reason concerns the degrees of freedom for the deviations. A more complex case of degrees of freedom occurs in the use of a technique called ANOVA, which is discussed in Chapter 9. In the following paragraphs, we shall see the details of these cases.

We first note that in the calculation of SSD, the deviations are taken from the sample mean $\bar{x}$ and not from the population mean μ. The reason is simple: While sampling, almost always, the population mean μ is not known. Not knowing the population mean, we take the deviations from the sample mean. But this introduces a downward bias in the deviations. To see the bias, refer to Figure 5–13, which shows the deviation of a sample point x from the sample mean and from the population mean.

It can be seen from Figure 5–13 that for sample points that fall to the right of the midpoint between μ and $\bar{x}$, the deviation from the sample mean will be smaller than the deviation from the population mean. Since the sample mean is where the sample points gravitate, a majority of the sample points are expected to fall to the right of the midpoint. Thus, overall, the deviations will have a downward bias.

To compensate for the downward bias, we use the concept of degrees of freedom. Let the population be a uniform distribution of the values $\{1, 2, \ldots, 10\}$. The mean of this population is 5.5. Suppose a random sample of size 10 is taken from this population. Assume that we are told to take the deviations from this population mean.

FIGURE 5–13 Deviations from the Population Mean and the Sample Mean

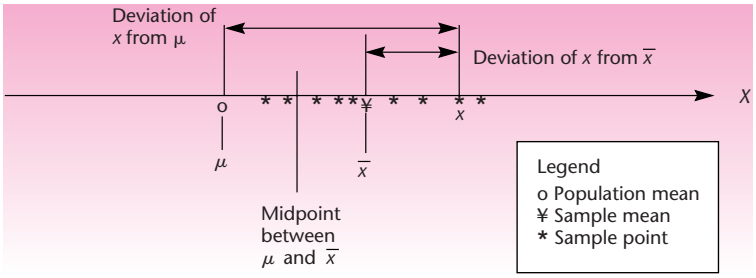


FIGURE 5–14 SSD and df

df = 10

	Sample	Deviation from	Deviation	Deviation Squared
1	10	5.5	4.5	20.25
2	3	5.5	−2.5	6.25
3	2	5.5	−3.5	12.25
4	6	5.5	0.5	0.25
5	1	5.5	−4.5	20.25
6	9	5.5	3.5	12.25
7	6	5.5	0.5	0.25
8	4	5.5	−1.5	2.25
9	10	5.5	4.5	20.25
10	7	5.5	1.5	2.25
			SSD	96.5

In Figure 5–14, the Sample column shows the sampled values. The calculation of SSD is shown taking deviations from the population mean of 5.5. The SSD works out to 96.5. *Since we had no freedom in taking the deviations, all the 10 deviations are completely left to chance.* Hence we say that the deviations have 10 degrees of freedom.

Suppose we do not know the population mean and are told that we can take the deviation from any number we choose. The best number to choose then is the sample mean, which will minimize the SSD (see problem 1–85). Figure 5–15*a* shows the calculation of SSD where the deviations are taken from the sample mean of 5.8. Because of the downward bias, the SSD has decreased to 95.6. The SSD would decrease further if we were allowed to select two different numbers from which the deviations are taken. Suppose we are allowed to use one number for the first five data points and another for the next five. Our best choices are the average of the first five numbers, 4.4, and the average of next five numbers, 7.2. Only these choices will minimize the SSD. The minimized SSD works out to 76, as seen in Figure 5–15*b*.

We can carry this process further. If we were allowed 10 different numbers from which the deviations are taken, then we could reduce the SSD all the way to zero.

FIGURE 5–15 SSD and df (*continued*)

df = 10 − 1 = 9

	Sample	Deviation from	Deviation	Deviation Squared
1	10	5.8	4.2	17.64
2	3	5.8	−2.8	7.84
3	2	5.8	−3.8	14.44
4	6	5.8	0.2	0.04
5	1	5.8	−4.8	23.04
6	9	5.8	3.2	10.24
7	6	5.8	0.2	0.04
8	4	5.8	−1.8	3.24
9	10	5.8	4.2	17.64
10	7	5.8	1.2	1.44
			SSD	95.6

(a)

df = 10 − 2 = 8

Sample	Deviation from	Deviation	Deviation Squared
10	4.4	5.6	31.36
3	4.4	−1.4	1.96
2	4.4	−2.4	5.76
6	4.4	1.6	2.56
1	4.4	−3.4	11.56
9	7.2	1.8	3.24
6	7.2	−1.2	1.44
4	7.2	−3.2	10.24
10	7.2	2.8	7.84
7	7.2	−0.2	0.04
		SSD	76

(b)

FIGURE 5–16 SSD and df (*continued*)

df = 10 − 10 = 0

	Sample	Deviation from	Deviation	Deviation Squared
1	10	10	0	0
2	3	3	0	0
3	2	2	0	0
4	6	6	0	0
5	1	1	0	0
6	9	9	0	0
7	6	6	0	0
8	4	4	0	0
9	10	10	0	0
10	7	7	0	0
			SSD	0

How? See Figure 5–16. We choose the 10 numbers equal to the 10 sample points (which in effect are 10 means). In the case of Figure 5–15*a*, we had one choice, and this takes away 1 degree of freedom from the deviations. The df of SSD is then declared as $10 - 1 = 9$. In Figure 5–15*b*, we had two choices and this took away 2 degrees of freedom from the deviations. Thus the df of SSD is $10 - 2 = 8$. In Figure 5–16, the df of SSD is $10 - 10 = 0$.

In every one of these cases, *dividing the SSD by only its corresponding df will yield an unbiased estimate of the population variance* σ^2. Hence the concept of the degrees of freedom is important. This also explains the denominator of $(n - 1)$ in the formula for sample variance S^2. For the case in Figure 5–15*a*, SSD/df $= 95.6/9 = 10.62$, and this is an unbiased estimate of the population variance.

We can now summarize how the number of degrees of freedom is determined. If we take a sample of size n and take the deviations from the (known) population mean, then the deviations, and therefore the SSD, will have df $= n$. But if we take the deviations from the sample mean, then the deviations, and therefore the SSD, will have df $= n - 1$. If we are allowed to take the deviations from $k\,(\leq n)$ different numbers that we choose, then the deviations, and therefore the SSD, will have df $= n - k$. While choosing each of the k numbers, we should choose the mean of the sample points to which that number applies. The case of $k > 1$ will be seen in Chapter 9, "Analysis of Variance."

EXAMPLE 5–4

A sample of size 10 is given below. We are to choose three different numbers from which the deviations are to be taken. The first number is to be used for the first five sample points; the second number is to be used for the next three sample points; and the third number is to be used for the last two sample points.

	Sample
1	93
2	97
3	60
4	72
5	96
6	83
7	59
8	66
9	88
10	53

1. What three numbers should we choose to minimize the SSD?
2. Calculate the SSD with the chosen numbers.
3. What is the df for the calculated SSD?
4. Calculate an unbiased estimate of the population variance.

Solution

1. We choose the means of the corresponding sample points: 83.6, 69.33, 70.5.
2. SSD = 2030.367. See the spreadsheet calculation below.
3. df = 10 − 3 = 7.
4. An unbiased estimate of the population variance is SSD/df = 2030.367/7 = 290.05.

	Sample	Mean	Deviation	Deviation Squared
1	93	83.6	9.4	88.36
2	97	83.6	13.4	179.56
3	60	83.6	−23.6	556.96
4	72	83.6	−11.6	134.56
5	96	83.6	12.4	153.76
6	83	69.33	13.6667	186.7778
7	59	69.33	−10.3333	106.7778
8	66	69.33	−3.33333	11.11111
9	88	70.5	17.5	306.25
10	53	70.5	−17.5	306.25
			SSD	2030.367
			SSD/df	290.0524

PROBLEMS

5–36. Three random samples of sizes, 30, 48, and 32, respectively, are collected, and the three sample means are computed. What is the total number of degrees of freedom for deviations from the means?

5–37. The data points in a sample of size 9 are 34, 51, 40, 38, 47, 50, 52, 44, 37.

a. If you can take the deviations of these data from any number you select, and you want to minimize the sum of the squared deviations (SSD), what number would you select? What is the minimized SSD? How many degrees of freedom are associated with this SSD? Calculate the mean squared deviation (MSD) by dividing the SSD by its degrees of freedom. (This MSD is an unbiased estimate of population variance.)

b. If you can take the deviations from three different numbers you select, and the first number is to be used with the first four data points to get the deviations, the second with the next three data points, and the third with the last two data points, what three numbers would you select? What is the minimized SSD? How many degrees of freedom are associated with this SSD? Calculate MSD.

c. If you can select nine different numbers to be used with each of the nine data points, what numbers would you select? What is the minimized SSD? How many degrees of freedom are associated with this SSD? Does MSD make sense in this case?

d. If you are told that the deviations are to be taken with respect to 50, what is the SSD? How many degrees of freedom are associated with this SSD? Calculate MSD.

5–38. Your bank sends you a summary statement, giving the average amount of all checks you wrote during the month. You have a record of the amounts of 17 out of the 19 checks you wrote during the month. Using this and the information provided by the bank, can you figure out the amounts of the two missing checks? Explain.

5–39. In problem 5–38, suppose you know the amounts of 18 of the 19 checks you wrote and the average of all the checks. Can you figure out the amount of the missing check? Explain.

5–40. You are allowed to take the deviations of the data points in a sample of size n, from k numbers you select, in order to calculate the sum of squared deviations (SSD). You select them to minimize SSD. How many degrees of freedom are associated with this SSD? As k increases, what happens to the degrees of freedom? What happens to SSD? What happens to MSD = SSD/df(SSD)?

5–6 Using the Computer

Using Excel for Generating Sampling Distributions

Figure 5–17 shows the template that can be used to calculate the sampling distribution of a sample mean. It is largely the same as the normal distribution template. The additional items are the population distribution entries at the top. To use the template, enter the population mean and standard deviation in cells B5 and C5. Enter the sample size in cell B8. In the drop-down box in cell I4, select Yes or No to answer the question "Is the population normally distributed?" The sample mean will follow

FIGURE 5–17 The Template for Sampling Distribution of a Sample Mean [Sampling Distribution.xls; Sheet: X-bar]

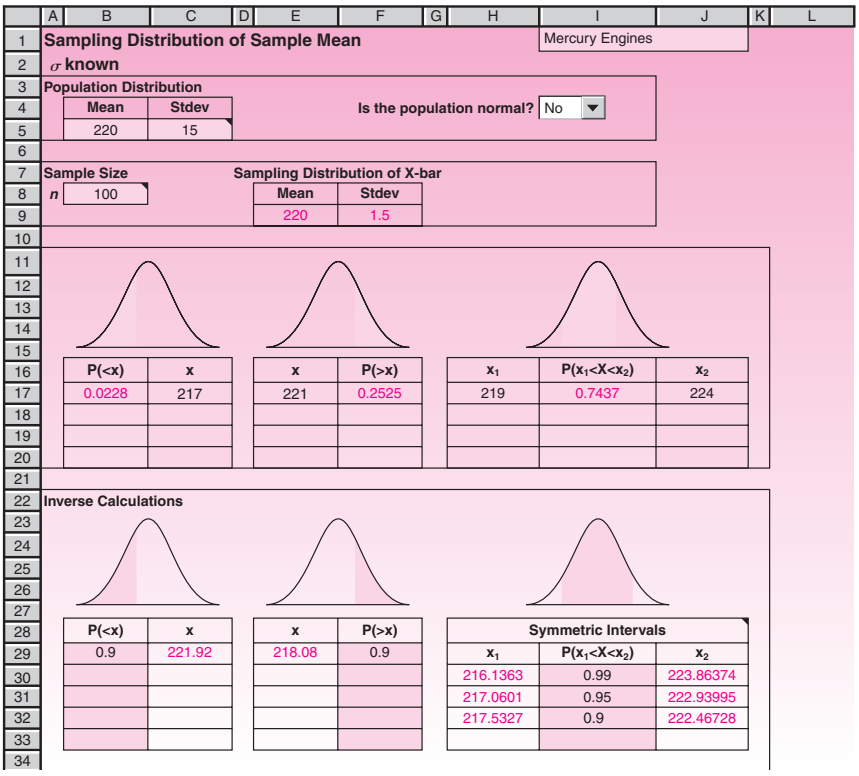
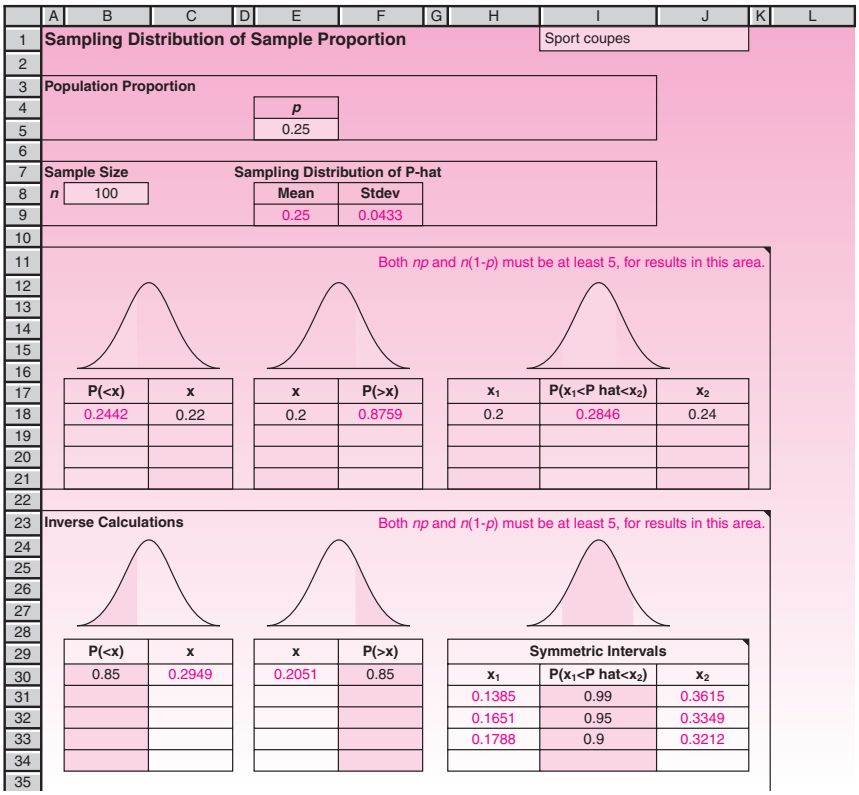


FIGURE 5–18 The Template for Sampling Distribution of a Sample Proportion [Sampling Distribution.xls; Sheet: P-hat]



a normal distribution if either the population is normally distributed or the sample size is at least 30. Only in such cases should this template be used. In other cases, a warning message—"Warning: The sampling distribution cannot be approximated as normal. Results appear anyway"—will appear in cell A10.

To solve Example 5–1, enter the population mean 220 in cell B5 and the population standard deviation 15 in cell C5. Enter the sample size 100 in cell B8. To find the probability that the sample mean will be less than 217, enter 217 in cell C17. The answer 0.0228 appears in cell B17.

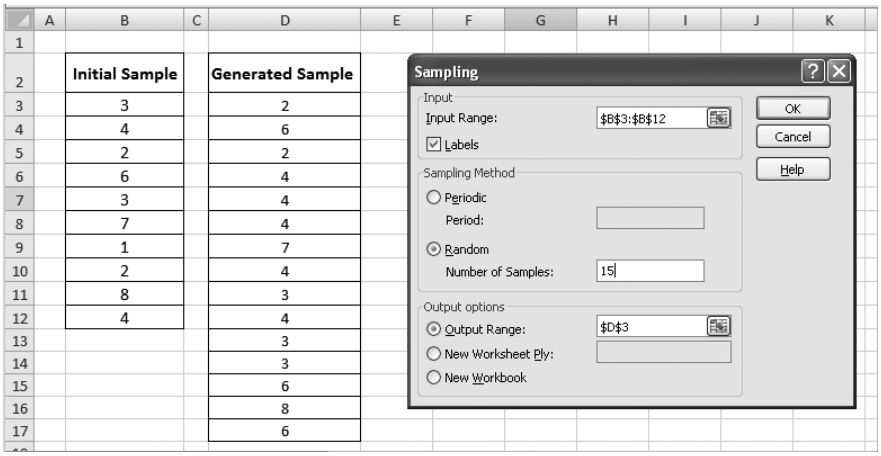
Figure 5–18 shows the template that can be used to calculate the sampling distribution of a sample proportion. To use the template, enter the population proportion in cell E5 and the sample size in cell B8.

To solve Example 5–3, enter the population proportion 0.25 in cell E5 and the sample size 100 in cell B8. Enter the value 0.2 in cell E17 to get the probability of the sample proportion being more than 0.2 in cell F17. The answer is 0.8749.

In addition to the templates discussed above, you can use Excel statistical tools to develop a variety of statistical analyses.

The **Sampling** analysis tool of Excel creates a sample from a population by treating the input range as a population. You can also create a sample that contains only the values from a particular part of a cycle if you believe that the input data is periodic. The Sampling analysis tool is accessible via Data Analysis in the Analysis group on the Data tab. If the Data Analysis command is not available, you need to load the Analysis ToolPack add-in program as described in Chapter 1.

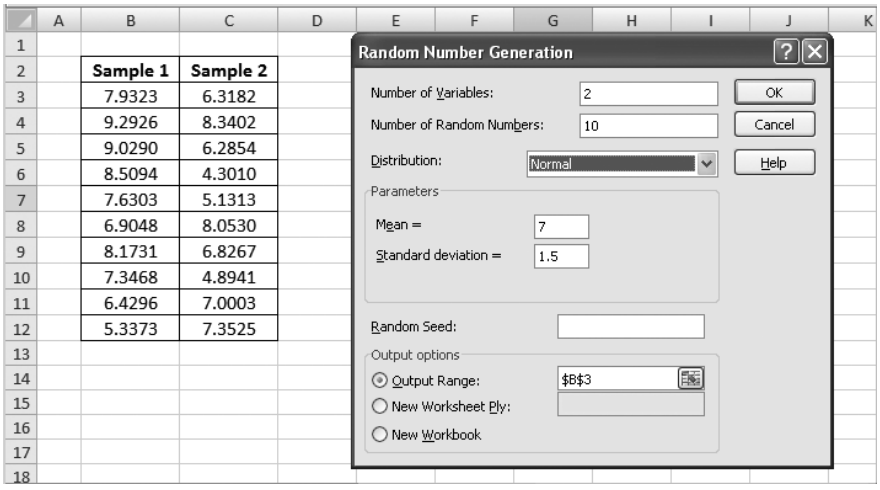
FIGURE 5–19 Generating a Random Sample by Excel



As an example, imagine you have a sample of size 10 from a population and you wish to generate another sample of size 15 from this population. You can start by choosing Sampling from Data Analysis. The Sampling dialog box will appear as shown in Figure 5–19. Specify the input range which represents your initial sample, cells B3 to B12. In the Sampling Method section you can indicate that you need a random sample of size 15. Determine the output range in the Output Options section. In Figure 5–19 the output has been placed in the column labeled Generated Sample starting from cell D3.

Another very useful tool of Excel is the **Random Number Generation** analysis tool, which fills a range with independent random numbers that are drawn from one of several distributions. Start by choosing the Random Number Generation analysis tool from Data Analysis in the Analysis group on the Data tab. Then the Random Number Generation dialog box will appear as shown in Figure 5–20. The number of

FIGURE 5–20 Generating Random Samples from Specific Distributions



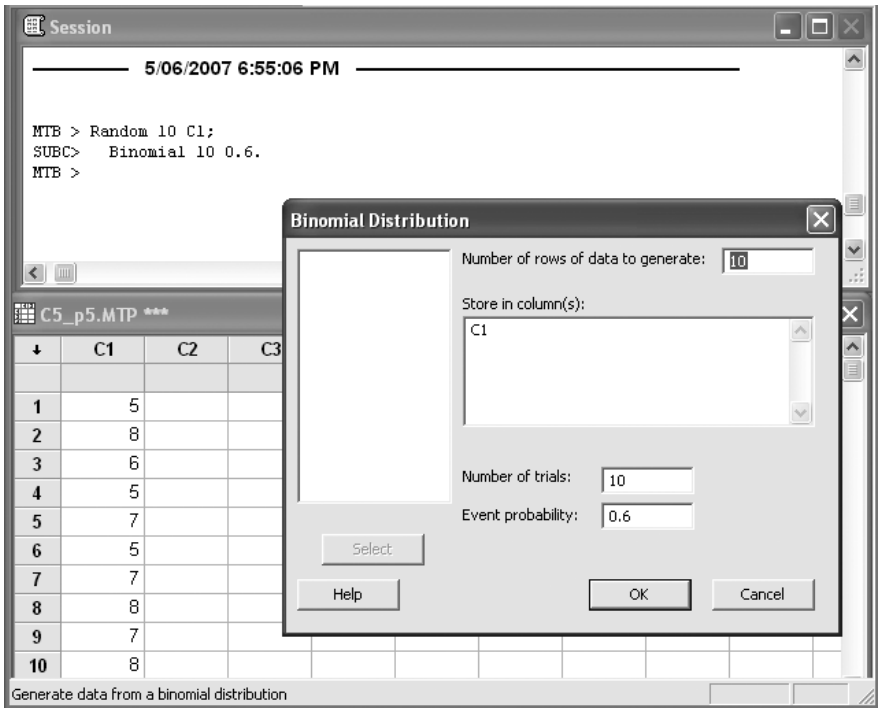
variables and number of random numbers at each set are defined by the values 2 and 10, respectively. The type of distribution and its parameters are defined in the next section. Define the output range in the Output Options. The two sets of random numbers are labeled Sample 1 and Sample 2 in Figure 5–20.

Using MINITAB for Generating Sampling Distributions

In this section we will illustrate how to use the Random Number Generation tool of MINITAB for simulating sampling distributions. To develop a random sample from a specific distribution you have to start by choosing Calc ▶ Random Data from the menu. You will observe a list of all distributions. Let's start by generating a random sample of size 10 from a binomial distribution with parameters 10 and 0.6 for number of trials and event probability, respectively. After choosing Calc ▶ Random Data ▶ Binomial from the menu, the Binomial Distribution dialog box will appear as shown in Figure 5–21. You need to specify the size of your sample as the number of rows of data to generate. As can be seen, the number 10 has been entered in the corresponding edit box. Specify the name of the column that will store the generated random numbers. Define the parameters of the binomial distribution in the next section. Then press the OK button. The generated binomial random numbers as well as corresponding Session commands will appear as shown in Figure 5–21.

MINITAB also enables you to generate a sample with an arbitrary size from a specific sample space with or without replacement. You need to specify the members of your sample space in a column. Imagine we need to generate a sample of size 8 from a sample space that has been defined in the first column. Start by choosing Calc ▶ Random Data ▶ Sample Form Columns from the menu bar. You need to specify

FIGURE 5–21 Using MINITAB for Generating Sampling Distributions



the size of your sample, the column that contains your sample space, and the column that will store the generated random numbers. You can also specify that the sampling occurs with or without replacement.

5–7 Summary and Review of Terms

In this chapter, we saw how samples are randomly selected from populations for the purpose of drawing inferences about **population parameters.** We saw how **sample statistics** computed from the data—the sample mean, the sample standard deviation, and the sample proportion—are used as **estimators** of population parameters. We presented the important idea of a **sampling distribution** of a statistic, the probability distribution of the values the statistic may take. We saw how the **central limit theorem** implies that the sampling distributions of the sample mean and the sample proportion approach normal distributions as the sample size increases. Sampling distributions of estimators will prove to be the key to the construction of confidence intervals in the following chapter, as well as the key to the ideas presented in later chapters. We also presented important properties we would like our estimators to possess: **unbiasedness, efficiency, consistency,** and **sufficiency.** Finally, we discussed the idea of **degrees of freedom.**

ADDITIONAL PROBLEMS

5–41. Suppose you are sampling from a population with mean $\mu = 1{,}065$ and standard deviation $\sigma = 500$. The sample size is $n = 100$. What are the expected value and the variance of the sample mean $\overline{X}$?

5–42. Suppose you are sampling from a population with population variance $\sigma^2 = 1{,}000{,}000$. You want the standard deviation of the sample mean to be at most 25. What is the minimum sample size you should use?

5–43. When sampling is from a population with mean 53 and standard deviation 10, using a sample of size 400, what are the expected value and the standard deviation of the sample mean?

5–44. When sampling is for a population proportion from a population with actual proportion $p = 0.5$, using a sample of size $n = 120$, what is the standard deviation of our estimator $\hat{P}$?

5–45. What are the expected value and the standard deviation of the sample proportion $\hat{P}$ if the true population proportion is 0.2 and the sample size is $n = 90$?

5–46. For a fixed sample size, what is the value of the true population proportion p that maximizes the variance of the sample proportion $\hat{P}$? (*Hint:* Try several values of p on a grid between 0 and 1.)

5–47. The average value of \$1.00 in euros in early 2007 was 0.76.[13] If $\sigma = 0.02$ and $n = 30$, find $P(0.72 < \overline{X} < 0.82)$.

5–48. In problem 5–41, what is the probability that the sample mean will be at least 1,000? Do you need to use the central limit theorem to answer this question? Explain.

5–49. In problem 5–43, what is the probability that the sample mean will be between 52 and 54?

5–50. In problem 5–44, what is the probability that the sample proportion will be at least 0.45?

[13]From "Foreign Exchange," *The New York Times,* May 2, 2007, p. C16.

5–51. Searches at Switzerland's 406 commercial banks turned up only $3.3 million in accounts belonging to Zaire's deposed president, Mobutu Sese Seko. The Swiss banks had been asked to look a little harder after finding nothing at all the first time round.

a. If President Mobutu's money was distributed in *all* 406 banks, how much was found, on average, per bank?

b. If a random sample of 16 banks was first selected in a preliminary effort to estimate how much money was in all banks, then assuming that amounts were normally distributed with standard deviation of $2,000, what was the probability that the mean of this sample would have been less than $7,000?

5–52. The proportion of defective microcomputer disks of a certain kind is believed to be anywhere from 0.06 to 0.10. The manufacturer wants to draw a random sample and estimate the proportion of all defective disks. How large should the sample be to ensure that the standard deviation of the estimator is *at most* 0.03?

5–53. Explain why we need to draw random samples and how such samples are drawn. What are the properties of a (simple) random sample?

5–54. Explain the idea of a bias and its ramifications.

5–55. Is the sample median a biased estimator of the population mean? Why do we usually prefer the sample mean to the sample median as an estimator for the population mean? If we use the sample median, what must we assume about the population? Compare the two estimators.

5–56. Explain why the sample variance is defined as the sum of squared deviations from the sample mean, divided by $n - 1$ and not by n.

5–57. Residential real estate in New York rents for an average of $44 per square foot, for a certain segment of the market.[14] If the population standard deviation is $7, and a random sample of 50 properties is chosen, what is the probability that the sample average will be below $35?

5–58. In problem 5–57, give 0.95 probability bounds on the value of the sample mean that would be obtained. Also give 0.90 probability bounds on the value of the sample mean.

5–59. According to *Money,* the average U.S. government bond fund earned 3.9% over the 12 months ending in February 2007.[15] Assume a standard deviation of 0.5%. What is the probability that the average earning in a random sample of 25 bonds exceeded 3.0%?

5–60. You need to fill in a table of five rows and three columns with numbers. All the row totals and column totals are given to you, and the numbers you fill in must add to these given totals. How many degrees of freedom do you have?

5–61. Thirty-eight percent of all shoppers at a large department store are holders of the store's charge card. If a random sample of 100 shoppers is taken, what is the probability that at least 30 of them will be found to be holders of the card?

5–62. When sampling is from a normal population with an unknown variance, is the sampling distribution of the sample mean normal? Explain.

5–63. When sampling is from a normal population with a known variance, what is the smallest sample size required for applying a normal distribution for the sample mean?

5–64. Which of the following estimators are unbiased estimators of the appropriate population parameters: $\overline{X}$, $\hat{P}$, S^2, S? Explain.

[14] "Square Feet," *The New York Times,* May 2, 2007, p. C7.

[15] "Money Benchmarks," *Money,* March 2007, p. 130.

5–65. Suppose a new estimator for the population mean is discovered. The new estimator is unbiased and has variance equal to σ^2/n^2. Discuss the merits of the new estimator compared with the sample mean.

5–66. Three independent random samples are collected, and three sample means are computed. The total size of the combined sample is 124. How many degrees of freedom are associated with the deviations from the sample means in the combined data set? Explain.

5–67. Discuss, in relative terms, the sample size needed for an application of a normal distribution for the sample mean when sampling is from each of the following populations. (Assume the population standard deviation is known in each case.)

a. A normal population

b. A mound-shaped population, close to normal

c. A discrete population consisting of the values 1,006, 47, and 0, with equal frequencies

d. A slightly skewed population

e. A highly skewed population

5–68. When sampling is from a normally distributed population, is there an advantage to taking a large sample? Explain.

5–69. Suppose that you are given a new sample statistic to serve as an estimator of some population parameter. You are unable to assume any theoretical results such as the central limit theorem. Discuss how you would empirically determine the sampling distribution of the new statistic.

5–70. Recently, the federal government claimed that the state of Alaska had overpaid 20% of the Medicare recipients in the state. The director of the Alaska Department of Health and Social Services planned to check this claim by selecting a random sample of 250 recipients of Medicare checks in the state and determining the number of overpaid cases in the sample. Assuming the federal government's claim is correct, what is the probability that less than 15% of the people in the sample will be found to have been overpaid?

5–71. A new kind of alkaline battery is believed to last an average of 25 hours of continuous use (in a given kind of flashlight). Assume that the population standard deviation is 2 hours. If a random sample of 100 batteries is selected and tested, is it likely that the average battery in the sample will last less than 24 hours of continuous use? Explain.

5–72. Häagen-Dazs ice cream produces a frozen yogurt aimed at health-conscious ice cream lovers. Before marketing the product in 2007, the company wanted to estimate the proportion of grocery stores currently selling Häagen-Dazs ice cream that would sell the new product. If 60% of the grocery stores would sell the product and a random sample of 200 stores is selected, what is the probability that the percentage in the sample will deviate from the population percentage by no more than 7 percentage points?

5–73. Japan's birthrate is believed to be 1.57 per woman. Assume that the population standard deviation is 0.4. If a random sample of 200 women is selected, what is the probability that the sample mean will fall between 1.52 and 1.62?

5–74. The Toyota Prius uses both gasoline and electric power. Toyota claims its mileage per gallon is 52. A random sample of 40 cars is taken and each sampled car is tested for its fuel efficiency. Assuming that 52 miles per gallon is the population mean and 2.4 miles per gallon is the population standard deviation, calculate the probability that the sample mean will be between 52 and 53.

5–75. A bank that employs many part-time tellers is concerned about the increasing number of errors made by the tellers. To estimate the proportion of errors made

in a day, a random sample of 400 transactions on a particular day was checked. The proportion of the transactions with errors was computed. If the true proportion of transactions that had errors was 6% that day, what is the probability that the estimated proportion is less than 5%?

5–76. The daily number of visitors to a Web site follows a normal distribution with mean 15,830 and standard deviation 458. The average number of visitors on 10 randomly chosen days is computed. What is the probability that the estimated average exceeds 16,000?

5–77. According to *BusinessWeek,* profits in the energy sector have been rising, with one company averaging $3.42 monthly per share.[16] Assume this is an average from a population with standard deviation of $1.5. If a random sample of 30 months is selected, what is the probability that its average will exceed $4.00?

[16]Gene G. Marcial, "Tremendous Demand for Superior Energy Services," *BusinessWeek,* March 26, 2007, p. 132.

CASE 6 Acceptance Sampling of Pins

A company supplies pins in bulk to a customer. The company uses an automatic lathe to produce the pins. Factors such as vibration, temperature, and wear and tear affect the pins, so that the lengths of the pins made by the machine are normally distributed with a mean of 1.008 inches and a standard deviation of 0.045 inch. The company supplies the pins in large batches to a customer. The customer will take a random sample of 50 pins from the batch and compute the sample mean. If the sample mean is within the interval 1.000 inch ± 0.010 inch, then the customer will buy the whole batch.

1. What is the probability that a batch will be acceptable to the consumer? Is the probability large enough to be an acceptable level of performance?

To improve the probability of acceptance, the production manager and the engineers discuss adjusting the population mean and standard deviation of the lengths of the pins.

2. If the lathe can be adjusted to have the mean of the lengths at any desired value, what should it be adjusted to? Why?
3. Suppose the mean cannot be adjusted, but the standard deviation can be reduced. What maximum value of the standard deviation would make 90% of the parts acceptable to the consumer? (Assume the mean continues to be 1.008 inches.)
4. Repeat part 3 with 95% and 99% of the pins acceptable.
5. In practice, which one do you think is easier to adjust, the mean or the standard deviation? Why?

The production manager then considers the costs involved. The cost of resetting the machine to adjust the population mean involves the engineers' time and the cost of production time lost. The cost of reducing the population standard deviation involves, in addition to these costs, the cost of overhauling the machine and reengineering the process.

6. Assume it costs 150x^2$ to decrease the standard deviation by $(x/1{,}000)$ inch. Find the cost of reducing the standard deviation to the values found in parts 3 and 4.
7. Now assume that the mean has been adjusted to the best value found in part 2 at a cost of $80. Calculate the reduction in standard deviation necessary to have 90%, 95%, and 99% of the parts acceptable. Calculate the respective costs, as in part 6.
8. Based on your answers to parts 6 and 7, what are your recommended mean and standard deviation to which the machine should be adjusted?