Jose Luis Pelaez Inc/Getty Images

CHAPTER 5

Sampling Distributions

Introduction

Statistical inference draws conclusions about a population or process from data. It emphasizes substantiating these conclusions via probability calculations because probability allows us to take chance variation into account. We have already examined data and arrived at conclusions many times. How do we move from summarizing a single data set to formal inference involving probability calculations?

The foundation for statistical inference is described in Section 5.1. There, we not only discuss the use of *statistics* as estimates of population *parameters*, but also describe the chance variation of a statistic when the data are produced by random sampling or randomized experimentation.

The *sampling distribution* of a statistic shows how the statistic would vary in identical repeated data collections. That is, the sampling distribution is a probability distribution that answers the question, "What would happen if we did this experiment or sampling many times?" It is these distributions that provide the necessary link between probability and the data in your sample or from your experiment. They are the key to understanding statistical inference.

The last two sections of this chapter study the sampling distributions of two common statistics: the sample mean (for quantitative data) and the sample proportion or count (for categorical data). The general

framework for constructing a sampling distribution is the same for all statistics, so we focus on those statistics commonly used in inference. As part of this study, we revisit the Normal distributions and are introduced to two common discrete probability distributions, the binomial and Poisson distributions.

5.1 Toward Statistical Inference

When you complete this section, you will be able to:

- Identify parameters, populations, statistics, and samples and the relationships among these items.
- Use simulation to study a sampling distribution.
- Interpret and use a sampling distribution to describe a property of a statistic.
- Identify bias in a statistic by examining its sampling distribution and characterize an unbiased estimator of a parameter.
- Describe the relationship between the sample size and the variability of a statistic.
- Identify ways to reduce bias and variability of a statistic.
- Use the margin of error to describe the variability of a statistic.

A market research firm interviews a random sample of 1200 undergraduates enrolled in four-year colleges and universities throughout the United States. One result: the average number of hours spent online weekly is 19.0 hours. That's the truth about the 1200 students in the sample. What is the truth about the millions of undergraduates who make up this population?

Because the sample was chosen at random, it's reasonable to think that these 1200 students represent the entire population fairly well. So the market researchers turn the *fact* that the *sample mean* is $\bar{x} = 19.0$ hours into an *estimate* that the average time spent online weekly in the *population of undergraduates* enrolled in four-year colleges and universities is 19.0 hours.

statistical inference

That's a basic idea in statistics: use a fact about a sample to estimate the truth about the whole population. We call this **statistical inference** because we infer conclusions about the larger population from data on selected individuals.

To think about inference, we must keep straight whether a number describes a sample or a population. Here is the vocabulary we use.

PARAMETERS AND STATISTICS

A **parameter** is a number that describes the **population**. A parameter is a fixed number, but in practice, we do not know its value.

A **statistic** is a number that describes a **sample**. The value of a statistic is known when we have taken a sample, but it can change from sample to sample. We often use a statistic to estimate an unknown parameter.

EXAMPLE 5.1

Understanding the college student market. Since 1987, *Student Monitor* has published an annual market research study that provides clients with information about the college student market. The firm uses a random sample of 1200 students located throughout the United States.[1] One phase of the research focuses on computing and technology. The firm reports that undergraduates spend an average of 19.0 hours per week on the Internet and that 88% own a cell phone.

The sample mean $\bar{x} = 19.0$ hours is a *statistic*. The corresponding *parameter* is the average (call it μ) of all undergraduates enrolled in four-year colleges and universities. Similarly, the **proportion of the sample** who own a cell phone

sample proportion

$$\hat{p} = \frac{1056}{1200} = 0.88 = 88\%$$

population proportion

is a *statistic*. The corresponding *parameter* is the **proportion** (call it p) of all undergraduates at four-year colleges and universities who own a cell phone. We don't know the values of the parameters μ and p, so we use the statistics $\bar{x}$ and $\hat{p}$, respectively, to estimate them.

USE YOUR KNOWLEDGE

5.1 Street harassment. A large-scale survey of 16,607 women from 42 cities around the world reports that 84% of women experience their first street harassment before the age of 17.[2] Describe the statistic, population, and population parameter for this setting.

Sampling variability

If *Student Monitor* took a second random sample of 1200 students, the new sample would have different undergraduates in it. It is almost certain that the sample mean $\bar{x}$ would not again be 19.0. Likewise, we would not expect there to be exactly 1056 students who own a cell phone. In other words, the value of a statistic will vary from sample to sample. This basic fact is called **sampling variability**: the value of a statistic varies in repeated random sampling.

sampling variability

Random samples eliminate any preferences or favoritism from the act of choosing a sample, but they can still be misleading because of this *variability* that results when we choose at random. For example, what if a second random sample of 1200 undergraduates resulted in only 57% of the students owning a cell phone? Do these two results, 88% and 57%, leave you more or less confident in the value of the true population proportion? When sampling variability is too great, we can't trust the results of any one sample.

random digits, p. xxx

We can assess this variability by using the second advantage of random samples (the first advantage being the elimination of *bias*). Specifically, the fact that if we take lots of random samples of the same size from the same population, the variation from sample to sample will follow a predictable pattern. **All of statistical inference is based on one idea: to see how trustworthy a procedure is, ask what would happen if we repeated it many times.**

To understand why sampling variability is not fatal, we ask, "What would happen if we took many samples?" Here's how to answer that question for any statistic:

- Take a large number of random samples of size n from the same population.
- Calculate the statistic for each sample.
- Make a histogram of the values of the statistic.
- Examine the distribution displayed in the histogram for shape, center, and spread, as well as outliers or other deviations.

In practice, it is too expensive to take many samples from a large population such as all undergraduates enrolled in four-year colleges and universities. But we can imitate taking many samples by using random digits from a table or computer software to emulate chance behavior. This is called **simulation**.

simulation

EXAMPLE 5.2

Simulate a random sample. Let's simulate drawing simple random samples (SRSs) of size 100 from the population of undergraduates. Suppose that, in fact, 90% of the population owns a cell phone. Then the true value of the parameter we want to estimate is $p = 0.9$. (Of course, we would not sample in practice if we already knew that $p = 0.9$. We are sampling here to understand how the statistic $\hat{p}$ behaves.)

random digits, p. 179

For cell phone ownership, we can imitate the population by a table of random digits, with each entry standing for a person. Nine of the 10 digits (say, 0 to 8) stand for students who own a cell phone. The remaining digit, 9, stands for those who do not. Because all digits in a random number table are equally likely, this assignment produces a population proportion of cell phone owners equal to $p = 0.9$. We then simulate an SRS of 100 students from the population by taking 100 consecutive digits from Table B. The statistic $\hat{p}$ is the proportion of 0s to 8s in the sample.

Here are the first 100 entries in Table B with digits 0 to 8 highlighted:

19223	95034	**05756**	**28713**	96409	**12531**	**42544**	**82853**
73676	**47150**	99400	**01927**	**27754**	**42648**	**82425**	36290
45467	**71709**	**77558**	**00095**				

There are 90 digits between 0 and 8, so $\hat{p} = 90/100 = 0.90$. We are fortunate here that our estimate is the true population value $p = 0.9$. A second SRS based on the second 100 entries in Table B gives a different result, $\hat{p} = 0.86$. The third SRS gives the result $\hat{p} = 0.92$. All three sample results are different. That's sampling variability.

USE YOUR KNOWLEDGE

5.2 Using a random numbers table. In Example 5.2, we considered $p = 0.9$ and used each entry in Table B as a person for our simulations. Suppose instead that $p = 0.85$. How might we use Table B for simulations in this setting?

Sampling distributions

Simulation is a powerful tool for studying chance variation. Now that we see how simulation works, it is faster to abandon Table B and to use a computer to generate random numbers. This also allows us to study other statistics, such as

the sample mean, when the population cannot be easily imitated by a table of random numbers. We address the sampling distribution of $\bar{x}$ in the next section.

EXAMPLE 5.3

LOOK BACK
histogram, p. 14

Take many random samples. Figure 5.1 illustrates the process of choosing many samples and finding the statistic $\hat{p}$ for each one. Follow the flow of the figure from the population at the left, to choosing an SRS and finding the $\hat{p}$ for this sample, to collecting together the $\hat{p}$'s from many samples. The histogram at the right of the figure shows the distribution of the values of $\hat{p}$ from 1000 separate SRSs of size 100 drawn from a population with $p = 0.9$.

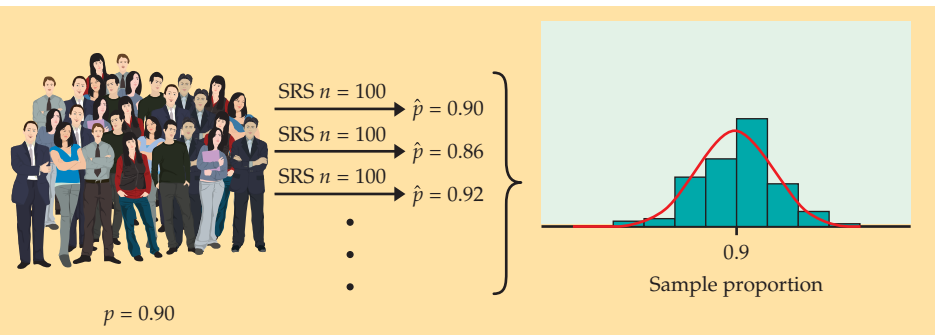

FIGURE 5.1 The results of many SRSs have a regular pattern, Example 5.3. Here we draw 1000 SRSs of size 100 from the same population. The population parameter is $p = 0.9$. The histogram shows the distribution of 1000 sample proportions.

Of course, *Student Monitor* samples 1200 students, not just 100. Figure 5.2 is parallel to Figure 5.1. It shows the process of choosing 1000 SRSs, each of size 1200, from a population in which the true proportion is $p = 0.9$. The 1000 values of $\hat{p}$ from these samples form the histogram at the right of the figure. Figures 5.1 and 5.2 are drawn on the same scale. Comparing them shows what happens when we increase the size of our samples from 100 to 1200. These histograms display the *sampling distribution* of the statistic $\hat{p}$ for two sample sizes.

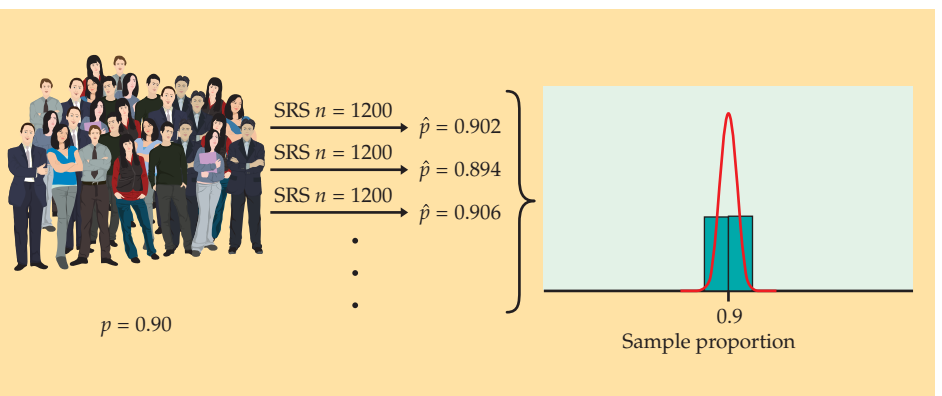

FIGURE 5.2 The distribution of the sample proportion for 1000 SRSs of size 1200 drawn from the same populations as in Figure 5.1. The two histograms have the same scale. The statistic from the larger sample is less variable.

SAMPLING DISTRIBUTION

The **sampling distribution** of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population.

Strictly speaking, the sampling distribution is the ideal pattern that would emerge if we looked at all possible samples of size n (here, 100 or 1200) from our population. A distribution obtained from a fixed number of trials, like the 1000 trials in Figures 5.1 and 5.2, is only an approximation to the sampling distribution. We will see that probability theory, the mathematics of chance behavior, can sometimes describe sampling distributions exactly. The interpretation of a sampling distribution is the same, however, whether we obtain it by simulation or by the mathematics of probability.

USE YOUR KNOWLEDGE

5.3 Poker winnings. Doug plays poker with the same group of friends once a week for three hours. At the end of each night, he records how much he won or lost in an Excel spreadsheet. Does this collection of amounts represent an approximation to a sampling distribution of his weekly winnings? Explain your answer.

We can use the tools of data analysis to describe any distribution. Let's apply those tools to Figures 5.1 and 5.2.

- **Shape:** The histograms look Normal. Figure 5.3 is a Normal quantile plot of the values of $\hat{p}$ for our samples of size 100. It confirms that the distribution in Figure 5.1 is close to Normal. The 1000 values for samples of size 1200 in Figure 5.2 are even closer to Normal. The Normal curves drawn through the histograms describe the overall shape quite well.
- **Center:** In both cases, the values of the sample proportion $\hat{p}$ vary from sample to sample, but the values are centered at 0.9. Recall that $p = 0.9$ is the true population parameter. Some samples have a $\hat{p}$ less than 0.9 and

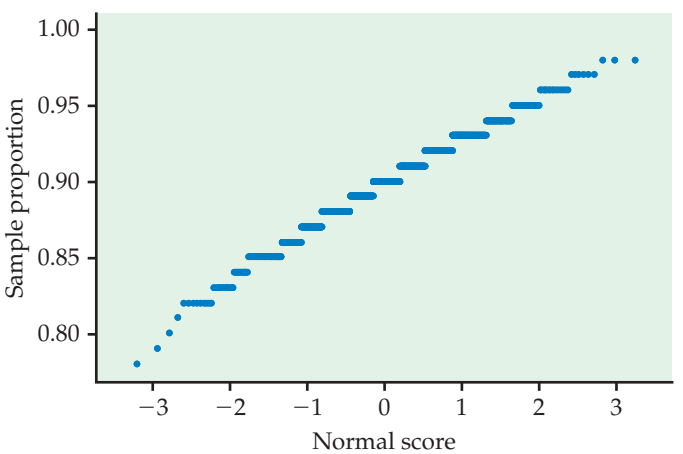

FIGURE 5.3 Normal quantile plot of the sample proportions in Figure 5.1. The distribution is close to Normal except for some clustering due to the fact that the sample proportions from a sample of size 100 can take only values that are a multiple of 0.01.

some greater, but there is no tendency to be always low or always high. That is, $\hat{p}$ has no *bias* as an estimator of p. This is true for both large and small samples. (Want the details? The mean of the 1000 values of $\hat{p}$ is 0.8985 for samples of size 100 and 0.8994 for samples of size 1200. The median value of $\hat{p}$ is exactly 0.9 for samples of both sizes.)

- **Spread:** The values of $\hat{p}$ from samples of size 1200 are much less spread out than the values from samples of size 100. In fact, the standard deviations are 0.0304 for Figure 5.1 and 0.0083 for Figure 5.2.

Although these results describe just two sets of simulations, they reflect facts that are true whenever we use random sampling.

Bias and variability

Our simulations show that a sample of size 1200 will almost always give an estimate $\hat{p}$ that is close to the truth about the population. Figure 5.2 illustrates this fact for just one value of the population proportion ($p = 0.9$), but it is true for any proportion. That is a primary reason *Student Monitor* uses a sample of size of 1200. There is more sampling variability the smaller the sample size. Samples of size 100, for example, might give an estimate of 83% or 97% when the truth is 90%.

Thinking about Figures 5.1 and 5.2 helps us restate the idea of bias when we use a statistic like $\hat{p}$ to estimate a parameter like p. It also reminds us that variability matters as much as bias.

BIAS AND VARIABILITY

Bias concerns the center of the sampling distribution. A statistic used to estimate a parameter is an **unbiased estimator** if the mean of its sampling distribution is equal to the true value of the parameter being estimated.

The **variability of a statistic** is described by the spread of its sampling distribution. This spread is determined by the sampling design and the sample size n. Statistics from larger probability samples have smaller spreads.

The **margin of error** is a numerical measure of the spread of a sampling distribution. It can be used to set bounds on the size of the likely error in using the statistic as an estimator of a population parameter.

We can think of the true value of the population parameter as the bull's-eye on a target and of the sample statistic as an arrow fired at the bull's-eye. Bias and variability describe what happens when an archer fires many arrows at the target. *Bias* means that the aim is off, and the sample values do not center about the population value. Large *variability* means that sample values are widely scattered about the target. In other words, there is a lack of precision, or consistency, among the sample values. Figure 5.4 shows this target illustration of the two types of error.

Notice that small variability (repeated shots are close together) can accompany large bias (the arrows are consistently away from the bull's-eye in one

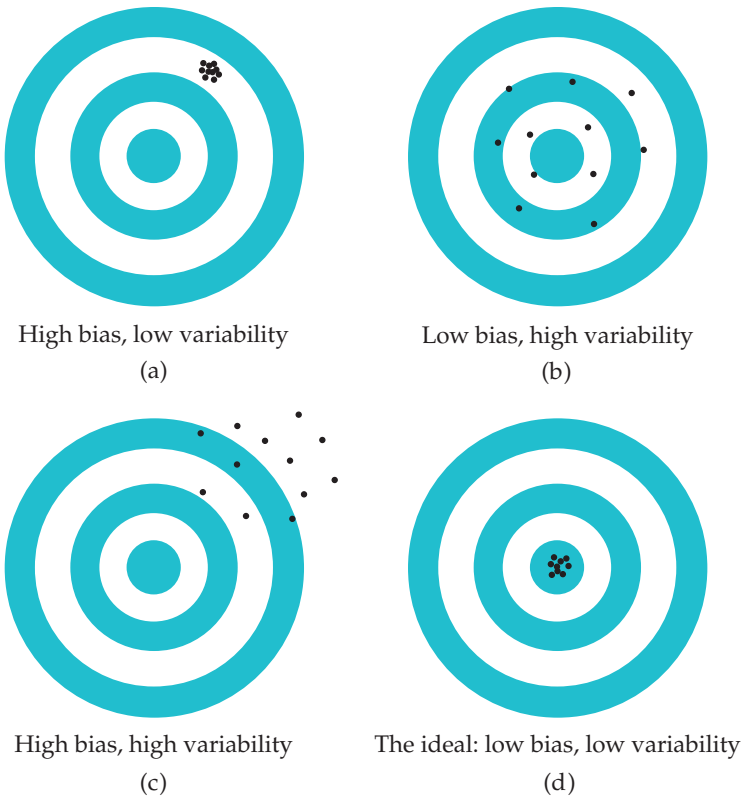

FIGURE 5.4 Bias and variability in shooting arrows at a target. Bias means the shots do not center around the bull's-eye. Variability means that the shots are scattered.

direction). And small bias (the arrows center on the bull's-eye) can accompany large variability (repeated shots are widely scattered). A good sampling scheme, like a good archer, must have both small bias and small variability. Here's how we do this.

MANAGING BIAS AND VARIABILITY

To reduce bias, use random sampling. When we start with a list of the entire population, simple random sampling produces unbiased estimates—the values of a statistic computed from an SRS neither consistently overestimate nor consistently underestimate the value of the population parameter.

To reduce the variability of a statistic from an SRS, use a larger sample. You can make the variability as small as you want by taking a large enough sample.

In practice, the *Student Monitor* takes only one random sample. We don't know how close to the truth the estimate from this one sample is because we don't know what the true population parameter value is. But *large random samples almost always give an estimate that is close to the truth.* Looking at the pattern of many samples when $n = 1200$ shows that we can trust the result of one sample.

Similarly, the Current Population Survey's sample of about 60,000 households estimates the national unemployment rate very accurately. Of course, only probability samples carry this guarantee. Using a probability sampling design and taking care to deal with practical difficulties reduce bias in a sample.

The size of the sample then determines how close to the population truth the sample result is likely to fall. Results from a sample survey usually come with a *margin of error* that sets bounds on the size of the likely error. The margin of error directly reflects the variability of the sample statistic, so it is smaller for larger samples. We will describe the details of its calculation to later chapters.

USE YOUR KNOWLEDGE

5.4 Bigger is better? Radio talk shows often report opinion polls based on tens of thousands of listeners. These sample sizes are typically much larger than those used in opinion polls that incorporate probability sampling. Does a larger sample size mean more trustworthy results? Explain your answer.

5.5 Effect of sample size on the sampling distribution. You are planning an opinion study and are considering taking an SRS of either 200 or 600 people. Explain how the sampling distributions of the population proportion p would differ in terms of center and spread for these two scenarios.

Sampling from large populations

Student Monitor's sample of 1200 students is only about 1 out of every 90,000 undergraduate students in the United States. Does it matter whether we sample 1-in-1000 individuals in the population or 1-in-90,000?

LARGE POPULATIONS DO NOT REQUIRE LARGE SAMPLES

The variability of a statistic from a random sample depends little on the size of the population, as long as the population is at least 20 times larger than the sample.

Why does the size of the population have little influence on the behavior of statistics from random samples? To see why this is plausible, imagine sampling harvested corn by thrusting a scoop into a lot of corn kernels. The scoop doesn't know whether it is surrounded by a bag of corn or by an entire truckload. As long as the corn is well mixed (so that the scoop selects a random sample), the variability of the result depends only on the size of the scoop.

The fact that the variability of sample results is controlled by the size of the sample has important consequences for sampling design. An SRS of size 1200 from the 10.5 million undergraduates gives results as precise as an SRS of size 1200 from the roughly 156,000 inhabitants of San Francisco between the ages of 20 and 29. This is good news for designers of national samples but bad news for those who want accurate information about these citizens of San Francisco. If both use an SRS, both must use the same size sample to obtain equally trustworthy results.

Why randomize?

Why randomize? The act of randomizing guarantees that the results of analyzing our data are subject to the laws of probability. The behavior of statistics is described by a sampling distribution. The form of the distribution is known and, in many cases, is approximately Normal. Often, the center of the distribution lies at the true parameter value so that the notion that randomization eliminates bias is made more explicit. The spread of the distribution describes the variability of the statistic and can be made as small as we wish by choosing a large enough sample. In a randomized experiment, we can reduce variability by choosing larger groups of subjects for each treatment.

These facts are at the heart of formal statistical inference. The remainder of this chapter has much to say in more technical language about sampling distributions. Later chapters describe the way statistical conclusions are based on them. What any user of statistics must understand is that all the technical talk has its basis in a simple question: *What would happen if the sample or the experiment were repeated many times?* The reasoning applies not only to an SRS, but also to the complex sampling designs actually used by opinion polls and other national sample surveys. The same conclusions hold as well for randomized experimental designs. The details vary with the design, but the basic facts are true whenever randomization is used to produce data.

Remember that proper statistical design is not the only aspect of a good sample or experiment. *The sampling distribution shows only how a statistic varies due to the operation of chance in randomization. It reveals nothing about possible bias due to undercoverage or nonresponse in a sample or to lack of realism in an experiment.* The actual error in estimating a parameter by a statistic can be much larger than the sampling distribution suggests. What is worse, there is no way to say how large the added error is. The real world is less orderly than statistics textbooks imply.

In the next two sections, we will study the sampling distributions of two common statistics, the sample mean and the sample proportion. The focus will be on the important features of these distributions so that we can quickly describe and use them in the later chapters on statistical inference. We will see that, in each case, the sampling distribution depends on **both** the population and the way we collect the data from the population.

SECTION 5.1 SUMMARY

- A number that describes a population is a **parameter**. A number that describes a sample (is computed from the sample data) is a **statistic**. The purpose of sampling or experimentation is usually **inference**: use sample statistics to make statements about unknown population parameters.

- A statistic from a probability sample or a randomized experiment has a **sampling distribution** that describes how the statistic varies in repeated data productions. The sampling distribution answers the question "What would happen if we repeated the sample or experiment many times?" Formal statistical inference is based on the sampling distributions of statistics.

- A statistic as an estimator of a parameter may suffer from **bias** or from high **variability**. Bias means that the center of the sampling distribution is not equal to the true value of the parameter. The variability of the statistic

is described by the spread of its sampling distribution. Variability is usually reported by giving a **margin of error** for conclusions based on sample results.

- Properly chosen statistics from randomized data production designs have no bias resulting from the way the sample is selected or the way the experimental units are assigned to treatments. We can reduce the variability of the statistic by increasing the size of the sample or the size of the experimental groups.

SECTION 5.1 EXERCISES

For Exercise 5.1, see page 283; for Exercise 5.2, see page 284; for Exercise 5.3, see page 286; and for Exercises 5.4 and 5.5, see page 289.

5.6 Web polls. If you connect to the website **peopleschoice.com/pca/polls/polls.jsp**, you are given the opportunity to vote on various entertainment questions. Can you apply the ideas about populations and samples to these polls? Explain why or why not.

5.7 What population and sample? Thirty students from your college who are majoring in English are randomly selected to be on a committee to evaluate immediate changes in the statistics requirement for the major. There are 153 English majors at your college. The current rules say that a statistics course is one of three options for a quantitative competency requirement. The proposed change would be to require a statistics course. Each of the committee members is asked to vote Yes or No on the new requirement.

(a) Describe the population for this setting.

(b) What is the sample?

(c) Describe the statistic and how it would be calculated.

(d) What is the population parameter?

(e) Write a short summary based on your answers to parts (a) through (d) using this setting to explain population, sample, parameter, statistic, and the relationships among these items.

5.8 What's wrong? State what is wrong in each of the following scenarios.

(a) A parameter describes a sample.

(b) Bias and variability are two names for the same thing.

(c) Large samples are always better than small samples.

(d) A sampling distribution is something generated by a computer.

5.9 Describe the population and the sample. For each of the following situations, describe the population and the sample.

(a) A survey of 17,096 students in U.S. four-year colleges reported that 19.4% were binge drinkers.

(b) In a study of work stress, 100 restaurant workers were asked about the impact of work stress on their personal lives.

(c) A tract of forest has 584 longleaf pine trees. The diameters of 40 of these trees were measured.

5.10 Is it unbiased? A statistic has a sampling distribution that is somewhat skewed. The mean is 17, the median is 15, the quartiles are 13 and 19.

(a) If the population parameter is 15, is the estimator unbiased?

(b) If the population parameter is 17, is the estimator unbiased?

(c) If the population parameter is 16, is the estimator unbiased?

(d) Write a short summary of your results in parts (a), (b), and (c) and include a discussion of bias and unbiased estimators.

5.11 Constructing a sampling distribution. Refer to Example 5.1 (page 283). Suppose *Student Monitor* also reported that the median number of hours per week spent on the Internet was 12.5 hours.

(a) Explain why we'd expect the population median to be less than the population mean in this setting by drawing the distribution of times spent on the Internet for all undergraduates. This is called the *population distribution*.

(b) Using Figure 5.2 (page 285) as a guide and your distribution from part (a), describe how to approximate the sampling distribution of the sample median in this setting.

5.12 Bias and variability. Figure 5.4 (page 288) shows histograms of four sampling distributions of statistics intended to estimate the same parameter. Label each distribution relative to the others as high or low bias and as high or low variability.

5.13 Constructing sampling distributions. The *Probability* applet simulates tossing a coin, with the advantage that you can choose the true long-term proportion, or probability, of a head. Suppose that we have a population in which proportion $p = 0.4$ (the

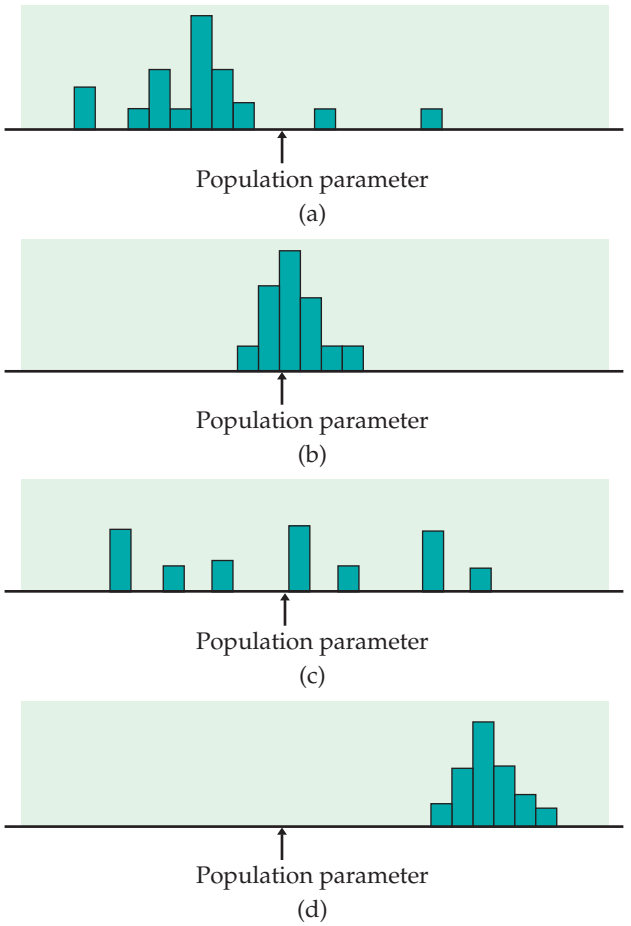

FIGURE 5.5 Determine which of these sampling distributions displays high or low bias and high or low variability, Exercise 5.12.

parameter) plan to vote in the next election. Tossing a coin with probability $p = 0.4$ of a head simulates this situation: each head is a person who plans to vote, and each tail is a person who does not. Set the "Probability of heads" in the applet to 0.4 and the number of tosses to 25. This simulates an SRS of size 25 from this population. By alternating between "Toss" and "Reset," you can take many samples quickly.

(a) Take 50 samples, recording the number of heads in each sample. Make a histogram of the 50 sample proportions (count of heads divided by 25). You are constructing the sampling distribution of this statistic.

(b) Another population contains only 20% who plan to vote in the next election. Take 50 samples of size 25 from this population, record the number in each sample who approve, and make a histogram of the 50 sample proportions.

5.14 Comparing sampling distributions. Refer to the previous exercise.

(a) How do the centers of your two histograms reflect the differing truths about the two populations?

(b) Describe any differences in the shapes of the two histograms. Is one more skewed than the other?

(c) Compare the spreads of the two histograms. For which population is there less sampling variability?

(d) Suppose instead that the population proportions were 0.6 and 0.8, respectively. Describe how the sampling distributions of $\hat{p}$ would differ from those constructed in Exercise 5.13.

5.15 Use the *Simple Random Sample* applet. The *Simple Random Sample* applet can illustrate the idea of a sampling distribution. Form a population labeled 1 to 100. We will choose an SRS of 15 of these numbers. That is, in this exercise, the numbers themselves are the population, not just labels for 100 individuals. The mean of the whole numbers 1 to 100 is 50.5. This is the parameter, the mean of the population.

(a) Use the applet to choose an SRS of size 15. Which 15 numbers were chosen? What is their mean? This is a statistic, the sample mean $\bar{x}$.

(b) Although the population and its mean 50.5 remain fixed, the sample mean changes as we take more samples. Take another SRS of size 15. (Use the "Reset" button to return to the original population before taking the second sample.) What are the 15 numbers in your sample? What is their mean? This is another value of $\bar{x}$.

(c) Take 18 more SRSs from this same population and record their means. You now have 20 values of the sample mean $\bar{x}$ from 20 SRSs of the same size from the same population. Make a histogram of the 20 values and mark the population mean 50.5 on the horizontal axis. Are your 20 sample values roughly centered at the population value? (If you kept going forever, your $\bar{x}$ -values would form the sampling distribution of the sample mean; the population mean would indeed be the center of this distribution.)

5.16 Use the *Simple Random Sample* applet, continued. Refer to the previous exercise.

(a) Suppose instead that a sample size of $n = 10$ was used. Based on what you know about the effect of the sample size on the sampling distribution, which sampling distribution should have the smaller variability?

(b) Repeat the previous exercise using $n = 10$. Did your simulations confirm your answer in part (a)? Explain your answer.

(c) Write a short paragraph about the effect of the sample size on the variability of a sampling distribution using these simulations to illustrate the basic idea.

5.2 The Sampling Distribution of a Sample Mean

When you complete this section, you will be able to:

- Explain the difference between the sampling distribution of $\bar{x}$ and the population distribution.
- Determine the mean and standard deviation of $\bar{x}$ for an SRS of size n from a population with mean μ and standard deviation σ.
- Describe how much larger n has to be for an SRS to reduce the standard deviation of $\bar{x}$ by a certain factor.
- Utilize the central limit theorem to approximate the sampling distribution of $\bar{x}$ and perform probability calculations based on this approximation.

A variety of statistics are used to describe quantitative data. The sample mean, median, and standard deviation are all examples of statistics based on quantitative data. Statistical theory describes the sampling distributions of these statistics. However, the general framework for constructing a sampling distribution is the same for all statistics. In this section, we will concentrate on the sample mean. Because sample means are just averages of observations, they are among the most frequently used statistics.

Suppose that you plan to survey 1000 undergraduates enrolled in four-year U.S. universities about their sleeping habits. The sampling distribution of the average hours of sleep per night describes what this average would be if many simple random samples of 1000 students were drawn from the population of students in the United States. In other words, it gives you an idea of what you are likely to see from your survey. It tells you whether you should expect this average to be near the population mean and whether the variation of the statistic is roughly ± 2 hours or ± 2 minutes.

Before constructing this distribution, however, we need to consider another set of probability distributions that also plays a role in statistical inference. Any quantity that can be measured on each member of a population is described by the distribution of its values for all members of the population. This is the context in which we first met distributions, as density curves that provide models for the overall pattern of data.

LOOK BACK
density curves, p. 51

Imagine choosing one individual at random from a population and measuring a quantity. The quantities obtained from repeated draws of one individual from a population have a probability distribution that is the distribution of the population.

EXAMPLE 5.4

Total sleep time of college students. A recent survey describes the distribution of total sleep time among college students as approximately Normal with a mean of 6.78 hours and standard deviation of 1.24 hours.[3] Suppose that we select a college student at random and obtain his or her sleep time. This result is a random variable X because, prior to the random sampling, we don't know the sleep time. We do know, however, that in repeated sampling, X will have the same $N(6.78, 1.24)$ distribution that describes the pattern of sleep time in the entire population. We call $N(6.78, 1.24)$ the *population distribution*.

POPULATION DISTRIBUTION

The **population distribution** of a variable is the distribution of its values for all members of the population. The population distribution is also the probability distribution of the variable when we choose one individual at random from the population.

In this example, the population of all college students actually exists so that we can, in principle, draw an SRS of students from it. Sometimes, our population of interest does not actually exist. For example, suppose that we are interested in studying final-exam scores in a statistics course, and we have the scores of the 34 students who took the course last semester. For the purposes of statistical inference, we might want to consider these 34 students as part of a hypothetical population of similar students who would take this course. In this sense, these 34 students represent not only themselves, but also a larger population of similar students. The key idea is to think of the observations that you have as coming from a population with a probability distribution.

USE YOUR KNOWLEDGE

5.17 Time spent using apps on a mobile device. Nielsen has installed, with permission, Mobile Netview 3 on approximately 5000 cell phones to gather information on mobile app usage among adults in the United States. Nielsen reported that 18–24 year olds spend an average of 37 hours and 6 minutes a month using mobile apps.[4] State the population that this survey describes, the statistic, and some likely values from the population distribution.

Now that we have made the distinction between the population distributions and sampling distributions, we can proceed with an in-depth study of the sampling distribution of a sample mean $\bar{x}$.

EXAMPLE 5.5

Sample means are approximately Normal. Figure 5.6 illustrates two striking facts about the sampling distribution of a sample mean. Figure 5.6(a) displays the distribution of student visit lengths (in minutes) to a statistics

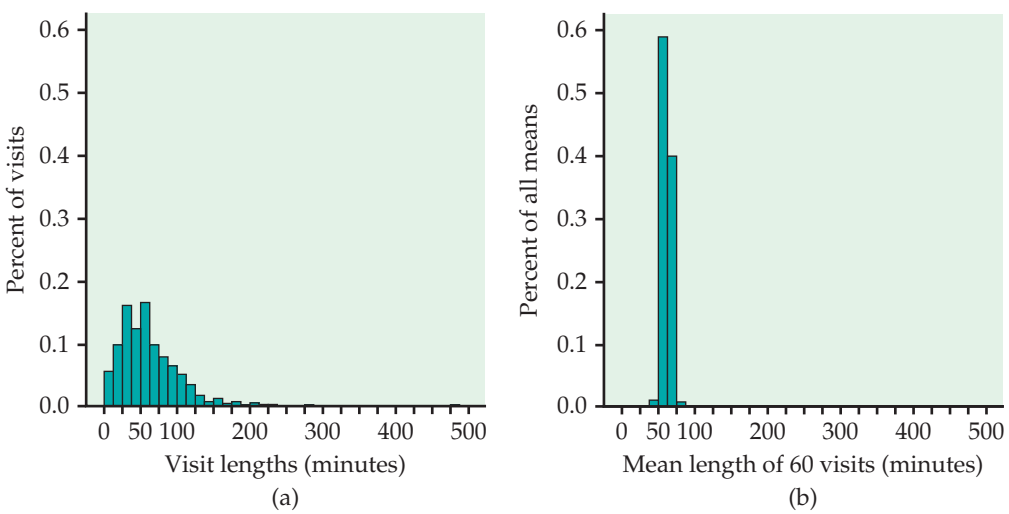

FIGURE 5.6 (a) The distribution of visit lengths to a statistics help room during the school year, Example 5.5. (b) The distribution of the sample means $\bar{x}$ for 500 random samples of size 60 from this population. The scales and histogram classes are exactly the same in both panels.

TABLE 5.1 Length (in Minutes) of 60 Visits to a Statistics Help Room

10	14	15	16	18	20	20	20	23	25
28	30	30	30	30	30	31	33	35	35
46	48	50	50	50	50	51	54	55	55
60	60	60	60	60	60	60	65	65	65
75	77	80	80	84	85	88	98	100	100
105	105	105	115	120	135	135	136	157	210

help room at a large midwestern university. Students visiting the help room were asked to sign in upon arrival and then sign out when leaving. During the school year, there were 1838 visits to the help room but only 1264 recorded visit lengths. This is because many visiting students forgot to sign out. We also omitted a few large outliers (visits lasting more than 10 hours).[5] The distribution is strongly skewed to the right. The population mean is $\mu = 61.28$ minutes.

HELP60

Table 5.1 contains the lengths of a random sample of 60 visits from this population. The mean of these 60 visits is $\bar{x} = 63.45$ minutes. If we were to take another sample of size 60, we would likely get a different value of $\bar{x}$. This is because this new sample would contain a different set of visits. To find the sampling distribution of $\bar{x}$, we take many SRSs of size 60 and calculate $\bar{x}$ for each sample. Figure 5.6(b) is the distribution of the values of $\bar{x}$ for 500 random samples. The scales and choice of classes are exactly the same as in Figure 5.6(a) so that we can make a direct comparison.

The sample means are much less spread out than the individual visit lengths. What is more, the Normal quantile plot in Figure 5.7 confirms that the distribution in Figure 5.6(b) is close to Normal.

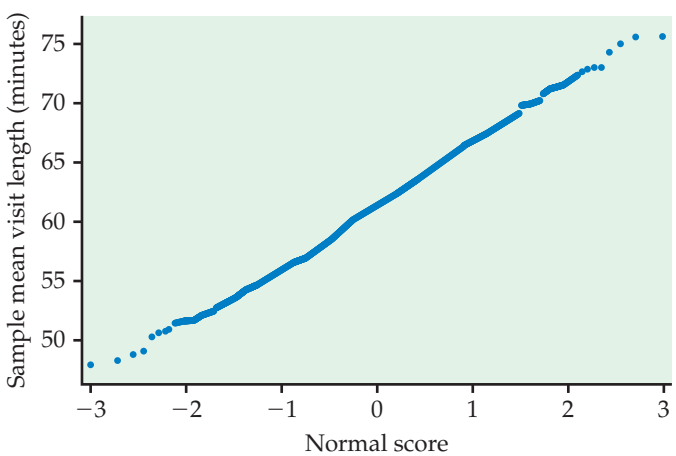

FIGURE 5.7 Normal quantile plot of the 500 sample means in Figure 5.6(b). The distribution is close to Normal.

This example illustrates two important facts about sample means that we will discuss in this section.

FACTS ABOUT SAMPLE MEANS

1. Sample means are less variable than individual observations.
2. Sample means are more Normal than individual observations.

These two facts contribute to the popularity of sample means in statistical inference.

The mean and standard deviation of $\bar{x}$

The sample mean $\bar{x}$ from a sample or an experiment is an estimate of the mean μ of the underlying population. The sampling distribution of $\bar{x}$ is determined by

- the design used to produce the data,
- the sample size n, and
- the population distribution.

Select an SRS of size n from a population, and measure a variable X on each individual in the sample. The n measurements are values of n random variables $X_1, X_2, \ldots, X_n$. A single X_i is a measurement on one individual selected at random from the population and, therefore, has the distribution of the population. If the population is large relative to the sample, we can consider $X_1, X_2, \ldots, X_n$ to be independent random variables, each having the same distribution. This is our probability model for measurements on each individual in an SRS.

The sample mean of an SRS of size n is

$$\bar{x} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

rules for means, p. 254

If the population has mean μ, then μ is the mean of the distribution of each observation X_i. To get the mean of $\bar{x}$, we use the rules for means of random variables. Specifically,

$$\begin{aligned}\mu_{\bar{x}} &= \frac{1}{n}(\mu_{X_1} + \mu_{X_2} + \cdots + \mu_{X_n}) \\ &= \frac{1}{n}(\mu + \mu + \cdots + \mu) = \mu\end{aligned}$$

rules for variances, p. 258

That is, *the mean of $\bar{x}$ is the same as the mean of the population.* The sample mean $\bar{x}$ is, therefore, an unbiased estimator of the unknown population mean μ.

The observations are independent, so the addition rule for variances also applies:

$$\begin{aligned}\sigma^2_{\bar{x}} &= \left(\frac{1}{n}\right)^2(\sigma^2_{X_1} + \sigma^2_{X_2} + \cdots + \sigma^2_{X_n}) \\ &= \left(\frac{1}{n}\right)^2(\sigma^2 + \sigma^2 + \cdots + \sigma^2) \\ &= \frac{\sigma^2}{n}\end{aligned}$$

With n in the denominator, the variability of $\bar{x}$ about its mean decreases as the sample size grows. Thus, a sample mean from a large sample will usually be very close to the true population mean μ. Here is a summary of these facts.

MEAN AND STANDARD DEVIATION OF A SAMPLE MEAN

Let $\bar{x}$ be the mean of an SRS of size n from a population having mean μ and standard deviation σ. The mean and standard deviation of $\bar{x}$ are

$$\mu_{\bar{x}} = \mu$$
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

How precisely does a sample mean $\bar{x}$ estimate a population mean μ? Because the values of $\bar{x}$ vary from sample to sample, we must give an answer in terms of the sampling distribution. We know that $\bar{x}$ is an unbiased estimator of μ, so its values in repeated samples are not systematically too high or too low. Most samples will give an $\bar{x}$-value close to μ if the sampling distribution is concentrated close to its mean μ. So the precision of estimation depends on the spread of the sampling distribution.

Because the standard deviation of $\bar{x}$ is $\sigma/\sqrt{n}$, the standard deviation of the statistic decreases in proportion to the square root of the sample size. This means, for example, that a sample size must be multiplied by 4 in order to divide the statistic's standard deviation in half. By comparison, a sample size must be multiplied by 100 in order to reduce the standard deviation by a factor of 10.

EXAMPLE 5.6

Standard deviations for sample means of visit lengths. The standard deviation of the population of visit lengths in Figure 5.6(a) (page 294) is $\sigma = 41.84$ minutes. The length of a single visit will often be far from the population mean. If we choose an SRS of 15 visits, the standard deviation of their mean length is

$$\sigma_{\bar{x}} = \frac{41.84}{\sqrt{15}} = 10.80 \text{ minutes}$$

Averaging over more visits reduces the variability and makes it more likely that $\bar{x}$ is close to μ. Our sample size of 60 visits is 4 times 15, so the standard deviation will be half as large:

$$\sigma_{\bar{x}} = \frac{41.84}{\sqrt{60}} = 5.40 \text{ minutes}$$

USE YOUR KNOWLEDGE

5.18 Find the mean and the standard deviation of the sampling distribution. Compute the mean and standard deviation of the sampling distribution of the sample mean when you plan to take an SRS of size 64 from a population with mean 44 and standard deviation 16.

5.19 The effect of increasing the sample size. In the setting of the previous exercise, repeat the calculations for a sample size of 576. Explain the effect of the sample size increase on the mean and standard deviation of the sampling distribution.

The central limit theorem

We have described the center and spread of the probability distribution of a sample mean $\bar{x}$, but not its shape. The shape of the distribution of $\bar{x}$ depends on the shape of the population distribution. Here is one important case: if the population distribution is Normal, then so is the distribution of the sample mean.

SAMPLING DISTRIBUTION OF A SAMPLE MEAN

If a population has the $N(\mu, \sigma)$ distribution, then the sample mean $\bar{x}$ of n independent observations has the $N(\mu, \sigma/\sqrt{n})$ distribution.

This is a somewhat special result. Many population distributions are not Normal. The help room visit lengths in Figure 5.6(a), for example, are strongly skewed. Yet Figures 5.6(b) and 5.7 show that means of samples of size 60 are close to Normal.

One of the most famous facts of probability theory says that, for large sample sizes, the distribution of $\bar{x}$ is close to a Normal distribution. This is true no matter what shape the population distribution has, as long as the population has a finite standard deviation σ. This is the **central limit theorem**. It is much more useful than the fact that the distribution of $\bar{x}$ is exactly Normal if the population is exactly Normal.

central limit theorem

CENTRAL LIMIT THEOREM

Draw an SRS of size n from any population with mean μ and finite standard deviation σ. When n is large, the sampling distribution of the sample mean $\bar{x}$ is approximately Normal:

$$\bar{x} \text{ is approximately } N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

EXAMPLE 5.7

How close will the sample mean be to the population mean? With the Normal distribution to work with, we can better describe how precisely a random sample of 60 visits estimates the mean length of all visits to the help room. The population standard deviation for the 1264 visits in the population of Figure 5.6(a) is $\sigma = 41.84$ minutes. From Example 5.6 we know $\sigma_{\bar{x}} = 5.4$ minutes. By the 95 part of the 68–95–99.7 rule, about 95% of all samples will have mean $\bar{x}$ within two standard deviations of μ, that is, within ± 10.8 minutes of μ.

68–95–99.7 rule, p. 57

USE YOUR KNOWLEDGE

5.20 Use the 68–95–99.7 rule. You take an SRS of size 64 from a population with mean 82 and standard deviation 24. According to the central limit theorem, what is the approximate sampling distribution of the sample mean? Use the 95 part of the 68–95–99.7 rule to describe the variability of $\bar{x}$.

For the sample size of $n = 60$ in Example 5.7, the sample mean is not very precise. The population of help room visit lengths is very spread out, so the sampling distribution of $\bar{x}$ has a large standard deviation.

EXAMPLE 5.8

How can we reduce the standard deviation? In the setting of Example 5.7, if we want to reduce the standard deviation of $\bar{x}$ by a factor of 2, we must take a sample four times as large, $n = 4 \times 60$, or 240. Then

$$\sigma_{\bar{x}} = \frac{41.84}{\sqrt{240}} = 2.70 \text{ minutes}$$

For samples of size 240, about 95% of the sample means will be within twice 2.70, or 5.40 minutes, of the population mean μ.

finite population correction factor

The standard deviation computed in Example 5.8 is actually too large. This is due to the fact that the population size, $N = 1264$, is not at least 20 times larger than the sample size, $n = 240$. In these settings, it is better to adjust the standard deviation of $\bar{x}$ to reflect only the variance remaining in the population that is not in the sample. This is done by multiplying the unadjusted standard deviation by the **finite population correction factor**. This quantity is $\sqrt{\dfrac{N-n}{N-1}}$ and moves the standard deviation of $\bar{x}$ toward 0 as n moves toward N. Applying this correction to Example 5.8, the standard deviation of $\bar{x}$ is reduced 10% to

$$\frac{41.84}{\sqrt{240}}\sqrt{\frac{1264-240}{1264-1}} = 2.43 \text{ minutes}$$

Thus, for samples of size 240, about 95% of the sample means will be within twice 2.43, or 4.86 minutes, of the population mean μ, rather than the 5.40 minutes reported in Example 5.8.

USE YOUR KNOWLEDGE

5.21 The effect of increasing the sample size. In the setting of Exercise 5.20, suppose that we increase the sample size to 2304. Use the 95 part of the 68–95–99.7 rule to describe the variability of this sample mean. Compare your results with those you found in Exercise 5.20.

Example 5.8 reminds us that if the population is very spread out, the $\sqrt{n}$ in the formula for the deviation of $\bar{x}$ implies that very large samples are needed to estimate the population mean precisely. The main point of the example, however, is that the central limit theorem allows us to use Normal probability calculations to answer questions about sample means even when the population distribution is not Normal.

How large a sample size n is needed for $\bar{x}$ to be close to Normal depends on the population distribution. More observations are required if the shape of the population distribution is far from Normal. For the very skewed visit length population, samples of size 60 are large enough. Further study would be needed to see if the distribution of $\bar{x}$ is close to Normal for smaller samples like $n = 20$ or $n = 40$. Here is a more detailed study of another skewed distribution.

EXAMPLE 5.9

exponential distribution

The central limit theorem in action. Figure 5.8 shows the central limit theorem in action for another very non-Normal population. Figure 5.8(a) displays the density curve of a single observation from the population. The distribution is strongly right-skewed, and the most probable outcomes are near 0. The mean μ of this distribution is 1, and its standard deviation σ is also 1. This particular continuous distribution is called an **exponential distribution**. Exponential distributions are used as models for how long an iPhone will function properly and for the time between snaps you receive on Snapchat.

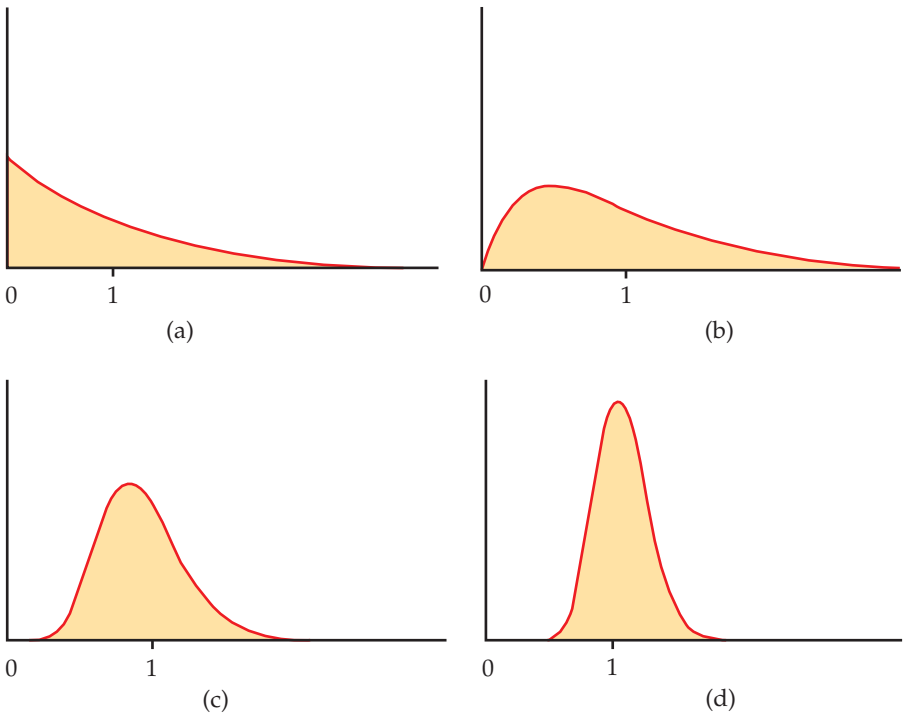

FIGURE 5.8 The central limit theorem in action: the sampling distribution of sample means from a strongly non-Normal population becomes more Normal as the sample size increases, Example 5.9. (a) The distribution of 1 observation. (b) The distribution of $\bar{x}$ for 2 observations. (c) The distribution of $\bar{x}$ for 10 observations. (d) The distribution of $\bar{x}$ for 25 observations.

Figures 5.8(b), (c), and (d) are the density curves of the sample means of 2, 10, and 25 observations from this population. As n increases, the shape becomes more Normal. The mean remains at $\mu = 1$, but the standard deviation decreases, taking the value $1/\sqrt{n}$. The density curve for 10 observations is still somewhat skewed to the right but already resembles a Normal curve having $\mu = 1$ and $\sigma = 1/\sqrt{10} = 0.32$. The density curve for $n = 25$ is yet more Normal. The contrast between the shape of the population distribution and of the distribution of the mean of 10 or 25 observations is striking.

You can also use the *Central Limit Theorem* applet to study the sampling distribution of $\bar{x}$. From one of three population distributions, 10,000 SRSs of a user-specified sample size n are generated, and a histogram of the sample means is constructed. You can then compare this estimated sampling distribution with the Normal curve that is based on the central limit theorem.

EXAMPLE 5.10

Using the *Central Limit Theorem* applet. In Example 5.9, we considered sample sizes of $n = 2$, 10, and 25 from an exponential distribution. Figure 5.9 shows a screenshot of the *Central Limit Theorem* applet for the exponential distribution when $n = 10$. The mean and standard deviation of this sampling distribution are 1 and $1/\sqrt{10} = 0.316$, respectively. From the 10,000 SRSs, the mean is estimated to be 1.001 and the estimated standard deviation is 0.319. These are both quite close to the true values. In Figure 5.8(c), we saw that the density curve for 10 observations is still somewhat skewed to the right. We can see this same behavior in Figure 5.9 when we compare the histogram with the Normal curve based on the central limit theorem.

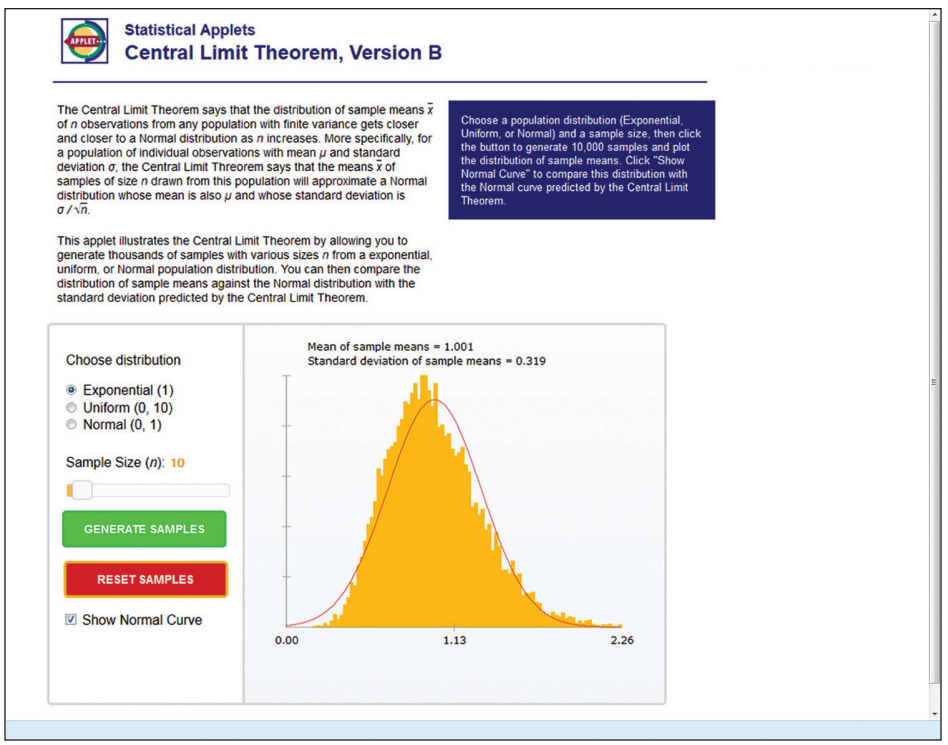

FIGURE 5.9 Screenshot of the *Central Limit Theorem* applet for the exponential distribution when $n = 10$, Example 5.10.

Try using the applet for the other sample sizes in Example 5.9. You should get histograms shaped like the density curves shown in Figure 5.8. You can also consider other sample sizes by sliding n from 1 to 100. As you increase n, the shape of the histogram moves closer to the Normal curve that is based on the central limit theorem.

USE YOUR KNOWLEDGE

5.22 Use the *Central Limit Theorem* applet. Let's consider the uniform distribution between 0 and 10. For this distribution, all intervals of the same length between 0 and 10 are equally likely. This distribution has a mean of 5 and standard deviation of 2.89.

(a) Approximate the population distribution by setting $n = 1$ and clicking the "Generate samples" button.

(b) What are your estimates of the population mean and population standard deviation based on the 10,000 SRSs? Are these population estimates close to the true values?

(c) Describe the shape of the histogram and compare it with the Normal curve.

5.23 Use the *Central Limit Theorem* applet again. Refer to the previous exercise. In the setting of Example 5.9, let's approximate the sampling distribution for samples of size $n = 2$, 10, and 25 observations.

(a) For each sample size, compute the mean and standard deviation of $\bar{x}$.

(b) For each sample size, use the applet to approximate the sampling distribution. Report the estimated mean and standard deviation. Are they close to the true values calculated in part (a)?

(c) For each sample size, compare the shape of the sampling distribution with the Normal curve based on the central limit theorem.

(d) For this population distribution, what sample size do you think is needed to make you feel comfortable using the central limit theorem to approximate the sampling distribution of $\bar{x}$? Explain your answer.

Now that we know that the sampling distribution of the sample mean $\bar{x}$ is approximately Normal for a sufficiently large n, let's consider some probability calculations.

EXAMPLE 5.11

Time between snaps. Snapchat has more than 100 million daily users sending well over 400 million snaps a day.[6] Suppose that the time X between snaps received is governed by the exponential distribution with mean $\mu = 15$ minutes and standard deviation $\sigma = 15$ minutes. You record the next 50 times between snaps. What is the probability that their average exceeds 13 minutes?

The central limit theorem says that the sample mean time $\bar{x}$ (in minutes) between snaps has approximately the Normal distribution with mean equal to the population mean $\mu = 15$ minutes and standard deviation

$$\frac{\sigma}{\sqrt{50}} = \frac{15}{\sqrt{50}} = 2.12 \text{ minutes}$$

The sampling distribution of $\bar{x}$ is, therefore, approximately $N(15,2.12)$. Figure 5.10 shows this Normal curve (solid) and also the actual density curve of $\bar{x}$ (dashed).

The probability we want is $P(\bar{x} > 13.0)$. This is the area to the right of 13 under the solid Normal curve in Figure 5.10. A Normal distribution calculation gives

$$\begin{aligned} P(\bar{x} > 13.0) &= P\left(\frac{\bar{x} - 15}{2.12} > \frac{13.0 - 15}{2.12}\right) \\ &= P(Z > -0.94) = 0.8264 \end{aligned}$$

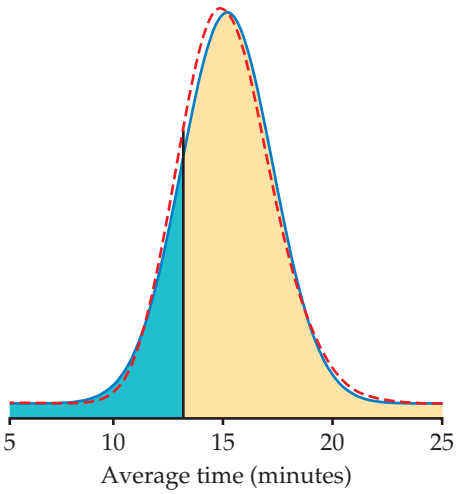

FIGURE 5.10 The exact distribution (dashed) and the Normal approximation from the central limit theorem (solid) for the average time between snaps received, Example 5.11.

The exactly correct probability is the area under the dashed density curve in the figure. It is 0.8265. The central limit theorem Normal approximation is off by only about 0.0001.

We can also use this sampling distribution to talk about the total time between the 1st and 51st snap received.

EXAMPLE 5.12

Convert the results to the total time. There are 50 time intervals between the 1st and 51st snap. According to the central limit theorem calculations in Example 5.11,

$$P(\bar{x} > 13.0) = 0.8264$$

We know that the sample mean is the total time divided by 50, so the event $\{\bar{x} > 13.0\}$ is the same as the event $\{50\bar{x} > 50(13.0)\}$. We can say that the probability is 0.8264 that the total time is $50(13.0) = 650$ minutes (10.8 hours) or greater.

USE YOUR KNOWLEDGE

5.24 Find a probability. Refer to Example 5.11. Find the probability that the mean time between snaps is less than 15 minutes. The exact probability is 0.5188. Compare your answer with the exact one.

Figure 5.11 summarizes the facts about the sampling distribution of $\bar{x}$ in a way that emphasizes the big idea of a sampling distribution. The general framework for constructing the sampling distribution of $\bar{x}$ is shown on the left.

- Take many random samples of size n from a population with mean μ and standard deviation σ.
- Find the sample mean $\bar{x}$ for each sample.
- Collect all the $\bar{x}$'s and display their distribution.

The sampling distribution of $\bar{x}$ is shown on the right. Keep this figure in mind as you go forward.

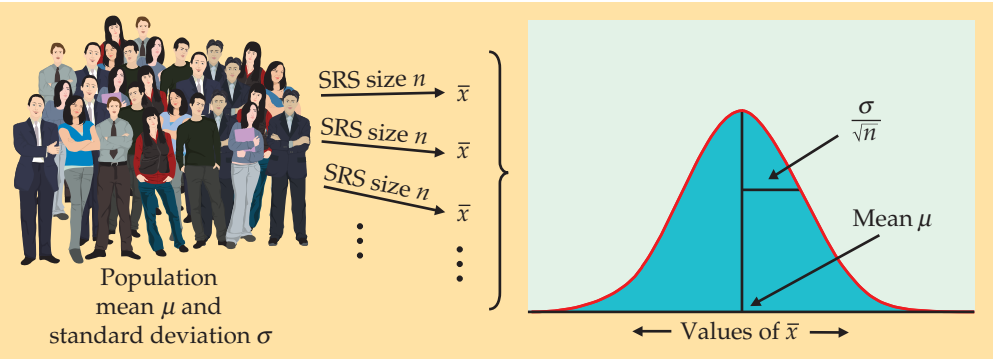

FIGURE 5.11 The sampling distribution of a sample mean $\bar{x}$ has mean μ and standard deviation $\sigma/\sqrt{n}$. The sampling distribution is Normal if the population distribution is Normal; it is approximately Normal for large samples in any case.

A few more facts

The central limit theorem is the big fact of probability theory in this section. Here are three additional facts related to our investigations that will be useful in describing methods of inference in later chapters.

The fact that the sample mean of an SRS from a Normal population has a Normal distribution is a special case of a more general fact: **any linear combination of independent Normal random variables is also Normally distributed**. That is, if X and Y are independent Normal random variables and a and b are any fixed numbers, $aX + bY$ is also Normally distributed, and this is true for any number of Normal random variables. In particular, the sum or difference of independent Normal random variables has a Normal distribution. The mean and standard deviation of $aX + bY$ are found as usual from the rules for means and variances. These facts are often used in statistical calculations. Here is an example.

rules for means, p. 254
rules for variances, p. 258

EXAMPLE 5.13

Getting to and from campus. You live off campus and take the shuttle, provided by your apartment complex, to and from campus. Your time on the shuttle in minutes varies from day to day. The time going to campus X has the $N(20,4)$ distribution, and the time returning from campus Y varies according to the $N(18, 8)$ distribution. If they vary independently, what is the probability that you will be on the shuttle for less time going to campus?

The difference in times $X - Y$ is Normally distributed, with mean and variance

$$\mu_{X-Y} = \mu_X - \mu_Y = 20 - 18 = 2$$
$$\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y = 4^2 + 8^2 = 80$$

Because $\sqrt{80} = 8.94$, $X - Y$ has the $N(2, 8.94)$ distribution. Figure 5.12 illustrates the probability computation:

$$\begin{aligned} P(X < Y) &= P(X - Y < 0) \\ &= P\left(\frac{(X - Y) - 2}{8.94} < \frac{0 - 2}{8.94}\right) \\ &= P(Z < -0.22) = 0.4129 \end{aligned}$$

Although, on average, it takes longer to go to campus than return, the trip to campus will take less time on roughly two of every five days.

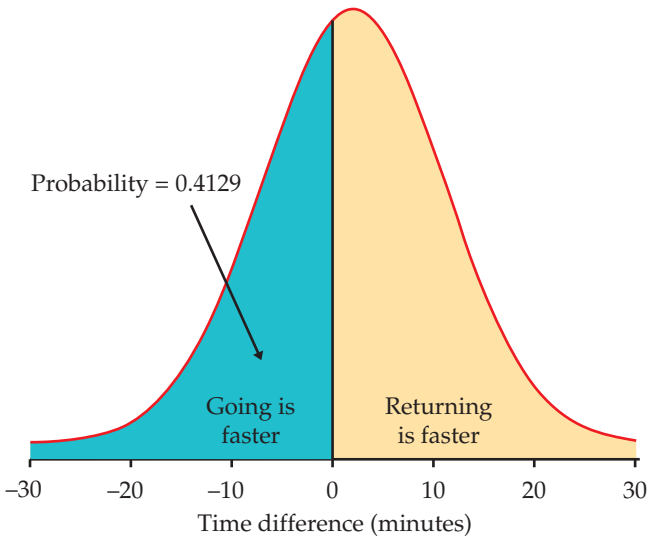

FIGURE 5.12 The Normal probability calculation, Example 5.13. The difference in times going to campus and returning from campus ($X - Y$) is Normal with mean 2 minutes and standard deviation 8.94 minutes.

The second useful fact is that **more general versions of the central limit theorem say that the distribution of a sum or average of many small random quantities is close to Normal**. This is true even if the quantities are not independent (as long as they are not too highly correlated) and even if they have different distributions (as long as no single random quantity is so large that it dominates the others). These more general versions of the central limit theorem suggest why the Normal distributions are common models for observed data. Any variable that is a sum of many small random influences will have approximately a Normal distribution.

Finally, **the central limit theorem also applies to discrete random variables**. An average of discrete random variables will never result in a continuous sampling distribution, but the Normal distribution often serves as a good approximation. In Section 5.3, we will discuss the sampling distribution and Normal approximation for counts and proportions. This Normal approximation is just an example of the central limit theorem applied to these discrete random variables.

BEYOND THE BASICS

Weibull Distributions

Our discussion of sampling distributions so far has concentrated on the Normal model to approximate the sampling distribution of the sample mean $\bar{x}$. This model is important in statistical practice because of the central limit theorem and the fact that sample means are among the most frequently used statistics. Simplicity also contributes to its popularity. The parameter μ is easy to understand, and to estimate it, we use a statistic $\bar{x}$ that is also easy to understand and compute.

There are, however, many other probability distributions that are used to model data in various circumstances. The time that a product, such as a computer hard drive, lasts before failing rarely has a Normal distribution. Earlier, we mentioned the use of the exponential distribution to model time to failure. Another class of continuous distributions, the **Weibull distributions**, is more commonly used in these situations.

Weibull distributions

EXAMPLE 5.14

Weibull density curves. Figure 5.13 shows the density curves of three members of the Weibull family. Each describes a different type of distribution for the time to failure of a product.

1. The top curve in Figure 5.13 is a model for *infant mortality*. This describes products that often fail immediately, prior to delivery to the customer. However, if the product does not fail right away, it will likely last a long time. For products like this, a manufacturer might test them and ship only the ones that do not fail immediately.
2. The middle curve in Figure 5.13 is a model for *early failure*. These products do not fail immediately, but many fail early in their lives after they are in the hands of customers. This is disastrous—the product or the process that makes it must be changed at once.
3. The bottom curve in Figure 5.13 is a model for *old-age wear-out*. Most of these products fail only when they begin to wear out, and then many fail at about the same age.

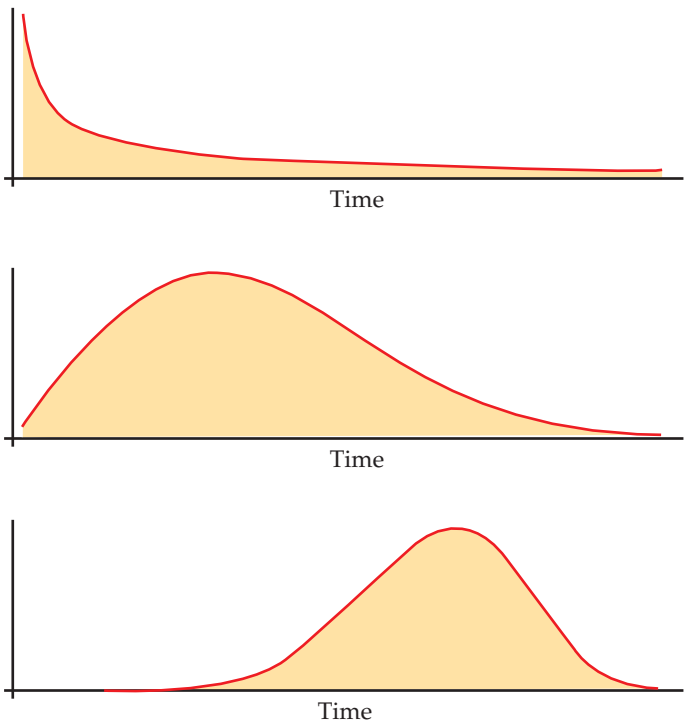

FIGURE 5.13 Density curves for three members of the Weibull family of distributions, Example 5.14.

A manufacturer certainly wants to know to which of these classes a new product belongs. To find out, engineers operate a random sample of products until they fail. From the failure time data, we can estimate the parameter (called the "shape parameter") that distinguishes among the three Weibull distributions in Figure 5.13. The shape parameter has no simple definition like that of a population proportion or mean, and it cannot be estimated by a simple statistic such as $\hat{p}$ or $\bar{x}$.

Two things save the situation. First, statistical theory provides general approaches for finding good estimates of any parameter. These general methods not only tell us how to use $\bar{x}$ in the Normal settings, but also how to

estimate the Weibull shape parameter. Second, software can calculate the estimate from data even though there is no algebraic formula that we can write for the estimate. Statistical practice often relies on both mathematical theory and methods of computation more elaborate than the ones we will meet in this book. Fortunately, big ideas such as sampling distributions carry over to more complicated situations.[7]

SECTION 5.2 SUMMARY

- The **population distribution** of a variable is the distribution of its values for all members of the population.
- The **sample mean** $\bar{x}$ of an SRS of size n drawn from a large population with mean μ and standard deviation σ has a sampling distribution with mean and standard deviation

$$\mu_{\bar{x}} = \mu$$
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

The sample mean $\bar{x}$ is an unbiased estimator of the population mean μ and is less variable than a single observation. The standard deviation decreases in proportion to the square root of the sample size n. This means that to reduce the standard deviation by a factor of C, we need to increase the sample size by a factor of C^2.

- The **central limit theorem** states that, for large n, the sampling distribution of $\bar{x}$ is approximately $N(\mu, \sigma/\sqrt{n})$ for any population with mean μ and finite standard deviation σ. This allows us to approximate probability calculations of $\bar{x}$ using the Normal distribution.
- Linear combinations of independent Normal random variables have Normal distributions. In particular, if the population has a Normal distribution, so does $\bar{x}$.

SECTION 5.2 EXERCISES

For Exercise 5.17, see page 294; for Exercises 5.18 and 5.19, see page 297; for Exercise 5.20, see page 298; for Exercise 5.21, see page 299; for Exercises 5.22 and 5.23, see pages 301–302; and for Exercise 5.24, see page 303.

5.25 What is wrong? Explain what is wrong in each of the following statements.

(a) If the population standard deviation is 10, then the standard deviation of $\bar{x}$ for an SRS of 10 observations is $10/10 = 1$.

(b) When taking SRSs from a population, larger sample sizes will result in larger standard deviations of $\bar{x}$.

(c) For an SRS from a population, both the mean and the standard deviation of $\bar{x}$ depend on the sample size n.

(d) The larger the population size N, the larger the sample size n needs to be for a desired standard deviation of $\bar{x}$.

5.26 What is wrong? Explain what is wrong in each of the following statements.

(a) The central limit theorem states that for large n, the population mean μ is approximately Normal.

(b) For large n, the distribution of observed values will be approximately Normal.

(c) For sufficiently large n, the 68–95–99.7 rule says that $\bar{x}$ should be within $\mu \pm 2\sigma$ about 95% of the time.

(d) As long as the sample size n is less than half the population size N, the standard deviation of $\bar{x}$ is $\sigma/\sqrt{n}$.

5.27 Generating a sampling distribution. Let's illustrate the idea of a sampling distribution in the case of a very small sample from a very small population. The population is the 10 scholarship players currently on your women's basketball team. For convenience, the 10 players have been labeled with the integers

0 to 9. For each player, the total amount of time spent (in minutes) on Twitter during the last week is recorded in the following table.

Player	0	1	2	3	4	5	6	7	8	9
Total time (min)	98	63	137	210	52	88	151	133	105	168

The parameter of interest is the average amount of time on Twitter. The sample is an SRS of size $n = 3$ drawn from this population of players. Because the players are labeled 0 to 9, a single random digit from Table B chooses one player for the sample.

(a) Find the mean for the 10 players in the population. This is the population mean μ.

(b) Use Table B to draw an SRS of size 3 from this population. (*Note*: You may sample the same player's time more than once.) Write down the three times in your sample and calculate the sample mean $\bar{x}$. This statistic is an estimate of μ.

(c) Repeat this process nine more times using different parts of Table B. Make a histogram of the 10 values of $\bar{x}$. You are approximating the sampling distribution of $\bar{x}$.

(d) Is the center of your histogram close to μ? Explain why you'd expect it to get closer to μ the more times you repeated this sampling process.

5.28 Total sleep time of college students. In Example 5.4 (page 293), the total sleep time per night among college students was approximately Normally distributed with mean $\mu = 6.78$ hours and standard deviation $\sigma = 1.24$ hours. You plan to take an SRS of size $n = 120$ and compute the average total sleep time.

(a) What is the standard deviation for the average time?

(b) Use the 95 part of the 68–95–99.7 rule to describe the variability of this sample mean.

(c) What is the probability that your average will be below 6.9 hours?

5.29 Determining sample size. Refer to the previous exercise. You want to use a sample size such that about 95% of the averages fall within ±5 minutes (0.08 hour) of the true mean $\mu = 6.78$.

(a) Based on your answer to part (b) in Exercise 5.28, should the sample size be larger or smaller than 120? Explain.

(b) What standard deviation of $\bar{x}$ do you need such that approximately 95% of all samples will have a mean within 5 minutes of μ?

(c) Using the standard deviation you calculated in part (b), determine the number of students you need to sample.

5.30 Music file size on a tablet PC. A tablet PC contains 3217 music files. The distribution of file size is highly skewed with many small file sizes. Assume that the standard deviation for this population is 3.25 megabytes (MB).

(a) What is the standard deviation of the average file size when you take an SRS of 25 files from this population?

(b) How many files would you need to sample if you wanted the standard deviation of $\bar{x}$ to be no larger than 0.50 MB?

5.31 Bottling an energy drink. A bottling company uses a filling machine to fill cans with an energy drink. The cans are supposed to contain 250 milliliters (ml). The machine, however, has some variability, so the standard deviation of the volume is $\sigma = 0.4$ ml. A sample of five cans is inspected each hour for process control purposes, and records are kept of the sample mean volume. If the process mean is exactly equal to the target value, what is the mean and standard deviation of the numbers recorded?

5.32 Average file size on a tablet. Refer to Exercise 5.30. Suppose that the true mean file size of the music and video files on the tablet is 2.35 MB and you plan to take an SRS of $n = 50$ files.

(a) Explain why it may be reasonable to assume that the average $\bar{x}$ is approximately Normal even though the population distribution is highly skewed.

(b) Sketch the approximate Normal curve for the sample mean, making sure to specify the mean and standard deviation.

(c) What is the probability that your sample mean will differ from the population mean by more than 0.15 MB?

5.33 Can volumes. Averages are less variable than individual observations. It is reasonable to assume that the can volumes in Exercise 5.31 vary according to a Normal distribution. In that case, the mean $\bar{x}$ of an SRS of cans also has a Normal distribution.

(a) Make a sketch of the Normal curve for a single can. Add the Normal curve for the mean of an SRS of five cans on the same sketch.

(b) What is the probability that the volume of a single randomly chosen can differs from the target value by 0.1 ml or more?

(c) What is the probability that the mean volume of an SRS of five cans differs from the target value by 0.1 ml or more?

5.34 Number of friends on Facebook. To commemorate Facebook's 10-year milestone, Pew Research reported several facts about Facebook obtained from its Internet Project survey. One was that the average adult user of Facebook has 338 friends. This population distribution

takes only integer values, so it is certainly not Normal. It is also highly skewed to the right, with a reported median of 200 friends.[8] Suppose that $\sigma = 380$ and you take an SRS of 80 adult Facebook users.

(a) For your sample, what are the mean and standard deviation of $\bar{x}$, the mean number of friends per adult user?

(b) Use the central limit theorem to find the probability that the average number of friends for 80 Facebook users is greater than 350.

(c) What are the mean and standard deviation of the total number of friends in your sample?

(d) What is the probability that the total number of friends among your sample of 80 Facebook users is greater than 28,000?

5.35 Cholesterol levels of teenagers. A study of the health of teenagers plans to measure the blood cholesterol level of an SRS of 13- to 16-year olds. The researchers will report the mean $\bar{x}$ from their sample as an estimate of the mean cholesterol level μ in this population.

(a) Explain to someone who knows no statistics what it means to say that $\bar{x}$ is an "unbiased" estimator of μ.

(b) The sample result $\bar{x}$ is an unbiased estimator of the population truth μ no matter what size SRS the study chooses. Explain to someone who knows no statistics why a large sample gives more trustworthy results than a small sample.

5.36 Grades in a math course. Indiana University posts the grade distributions for its courses online.[9] Students in one section of Math 118 in the fall semester received 18% A's, 31% B's, 26% C's, 13% D's and 12% F's.

(a) Using the common scale A = 4, B = 3, C = 2, D = 1, F = 0, take X to be the grade of a randomly chosen Math 118 student. Use the definitions of the mean (page 28) and standard deviation (page 38) for discrete random variables to find the mean μ and the standard deviation σ of grades in this course.

(b) Math 118 is a large enough course that we can take the grades of an SRS of 25 students and not worry about the finite population correction factor. If $\bar{x}$ is the average of these 25 grades, what are the mean and standard deviation of $\bar{x}$?

(c) What is the probability that a randomly chosen Math 118 student gets a B or better, $P(X \geq 3)$?

(d) What is the approximate probability $P(\bar{x} \geq 3)$ that the grade point average for 25 randomly chosen Math 118 students is B or better?

5.37 Monitoring the emerald ash borer. The emerald ash borer is a beetle that poses a serious threat to ash trees. Purple traps are often used to detect or monitor populations of this pest. In the counties of your state where the beetle is present, thousands of traps are used to monitor the population. These traps are checked periodically. The distribution of beetle counts per trap is discrete and strongly skewed. A majority of traps have no beetles, and only a few will have more than two beetles. For this exercise, assume that the mean number of beetles trapped is 0.4 with a standard deviation of 0.9.

(a) Suppose that your state does not have the resources to check all the traps, so it plans to check only an SRS of $n = 100$ traps. What are the mean and standard deviation of the average number of beetles $\bar{x}$ in 100 traps?

(b) Use the central limit theorem to find the probability that the average number of beetles in 100 traps is greater than 0.5.

(c) Do you think it is appropriate in this situation to use the central limit theorem? Explain your answer.

5.38 Risks and insurance. The idea of insurance is that we all face risks that are unlikely but carry high cost. Think of a fire destroying your home. So we form a group to share the risk: we all pay a small amount, and the insurance policy pays a large amount to those few of us whose homes burn down. An insurance company looks at the records for millions of homeowners and sees that the mean loss from fire in a year is $\mu = \$500$ per house and that the standard deviation of the loss is $\sigma = \$10{,}000$. (The distribution of losses is extremely right-skewed: most people have \$0 loss, but a few have large losses.) The company plans to sell fire insurance for \$500 plus enough to cover its costs and profit.

(a) Explain clearly why it would be unwise to sell only 100 policies. Then explain why selling many thousands of such policies is a safe business.

(b) Suppose the company sells the policies for \$600. If the company sells 50,000 policies, what is the approximate probability that the average loss in a year will be greater than \$600?

5.39 Weights of airline passengers. In 2005, the Federal Aviation Administration (FAA) updated its passenger weight standards to an average of 190 pounds in the summer (195 in the winter). This includes clothing and carry-on baggage. The FAA, however, did not specify a standard deviation. A reasonable standard deviation is 35 pounds. Weights are not Normally distributed, especially when the population includes both men and women, but they are not very non-Normal. A commuter plane carries 25 passengers. What is the approximate probability that, in the summer, the total weight of the passengers exceeds 5200 pounds? (*Hint*: To apply the central limit theorem, restate the problem in terms of the mean weight.)

5.40 Iron depletion without anemia and physical performance. Several studies have shown a link

between iron depletion without anemia (IDNA) and physical performance. In one recent study, the physical performance of 24 female collegiate rowers with IDNA was compared with 24 female collegiate rowers with normal iron status.[10] Several different measures of physical performance were studied, but we'll focus here on training-session duration. Assume that training-session duration of female rowers with IDNA is Normally distributed, with mean 58 minutes and standard deviation 11 minutes. Training-session duration of female rowers with normal iron status is Normally distributed, with mean 69 minutes and standard deviation 18 minutes.

(a) What is the probability that the mean duration of the 24 rowers with IDNA exceeds 63 minutes?

(b) What is the probability that the mean duration of the 24 rowers with normal iron status is less than 63 minutes?

(c) What is the probability that the mean duration of the 24 rowers with IDNA is greater than the mean duration of the 24 rowers with normal iron status?

5.41 Treatment and control groups. The previous exercise illustrates a common setting for statistical inference. This exercise gives the general form of the sampling distribution needed in this setting. We have a sample of n observations from a treatment group and an independent sample of m observations from a control group. Suppose that the response to the treatment has the $N(\mu_X, \sigma_X)$ distribution and that the response of control subjects has the $N(\mu_Y, \sigma_Y)$ distribution. Inference about the difference $\mu_Y - \mu_X$ between the population means is based on the difference $\bar{y} - \bar{x}$ between the sample means in the two groups.

(a) Under the assumptions given, what is the distribution of $\bar{y}$? Of $\bar{x}$?

(b) What is the distribution of $\bar{y} - \bar{x}$?

5.42 Investments in two funds. Jennifer invests her money in a portfolio that consists of 70% Fidelity Spartan 500 Index Fund and 30% Fidelity Diversified International Fund. Suppose that, in the long run, the annual real return X on the 500 Index Fund has mean 10% and standard deviation 15%, the annual real return Y on the Diversified International Fund has mean 9% and standard deviation 19%, and the correlation between X and Y is 0.6.

(a) The return on Jennifer's portfolio is $R = 0.7X + 0.3Y$. What are the mean and standard deviation of R?

(b) The distribution of returns is typically roughly symmetric but with more extreme high and low observations than a Normal distribution. The average return over a number of years, however, is close to Normal. If Jennifer holds her portfolio for 20 years, what is the approximate probability that her average return is less than 5%?

(c) The calculation you just made is not overly helpful because Jennifer isn't really concerned about the mean return $\overline{R}$. To see why, suppose that her portfolio returns 12% this year and 6% next year. The mean return for the two years is 9%. If Jennifer starts with $1000, how much does she have at the end of the first year? At the end of the second year? How does this amount compare with what she would have if both years had the mean return, 9%? Over 20 years, there may be a large difference between the ordinary mean $\overline{R}$ and the *geometric mean*, which reflects the fact that returns in successive years multiply rather than add.

5.3 Sampling Distributions for Counts and Proportions

When you complete this section, you will be able to:

- Determine when a count X can be modeled using the binomial distribution.
- Determine when the sampling distribution of a count can be modeled using the binomial distribution.
- Calculate the mean and standard deviation of X when it has the $B(n, p)$ distribution.
- Explain the difference between the sampling distribution of a count X and the sampling distribution of the sample proportion $\hat{p} = X/n$.
- Determine when one can approximate the sampling distribution of a count using the Normal distribution.
- Determine when one can approximate the sampling distribution of the sample proportion using the Normal distribution.
- Use the Normal approximation for counts and proportions to perform probability calculations about the statistics.

categorical variable, p. 3

In the previous section, we discussed the probability distribution of the sample mean, which meant a focus on population values that were quantitative. We will now shift our focus to population values that are categorical. Counts and proportions are common discrete statistics that describe categorical data.

In Section 5.1 (pages 284–287), we discussed the use of simulation to study the sampling distribution of the sample proportion. In this section, we will use probability theory to more precisely describe the sampling distributions of the sample count and proportion. Let's start with an example.

EXAMPLE 5.15

Istockphoto/Thinkstock

Work hours make it difficult to spend time with children. A sample survey asks 1006 British parents whether they think long working hours are making it difficult to spend enough time with their children.[11] We would like to view the responses of these parents as representative of a larger population of British parents who hold similar beliefs. That is, we will view the responses of the sampled parents as an SRS from a population.

When there are only two possible outcomes for a random variable, we can summarize the results by giving the count for one of the possible outcomes. We let n represent the sample size, and we use X to represent the random variable that gives the count for the outcome of interest.

EXAMPLE 5.16

The random variable of interest. In this sample survey of British parents, $n = 1006$. The parents in the sample were asked if they agree with the statement "These days, long working hours make it difficult for parents to spend enough time with their children." The variable X is the number of parents who agreed with the statement. In this case, $X = 755$.

In our example, we chose the random variable X to be the number of parents who think that long working hours make it difficult to spend enough time with their children. We could have chosen X to be the number of parents who do not think that long working hours make it difficult to spend enough time with their children. The choice is yours. Often, we make the choice based on how we would like to describe the results in a summary. Which choice do you prefer in this case?

sample proportion, p. 199

When a random variable has only two possible outcomes, it is more common to use the sample proportion $\hat{p} = X/n$ as the summary rather than the count X.

EXAMPLE 5.17

The sample proportion. The sample proportion of parents surveyed who think that long working hours make it difficult to spend enough time with their children is

$$\hat{p} = \frac{755}{1006} = 0.75$$

Notice that this summary takes into account the sample size n. We need to know n in order to properly interpret the meaning of the random variable X. For example, the conclusion we would draw about parent opinions in this survey would be quite different if we had observed $X = 755$ from a sample twice as large, $n = 2012$.

USE YOUR KNOWLEDGE

5.43 Sexual harassment in middle school. A survey of 1391 students in grades 5 to 8 reports that 26% of the students say they have encountered some type of sexual harassment while at school.[12] Give the sample size n, the count X, and the sample proportion $\hat{p}$ for this survey.

5.44 High school graduates who took a statistics course. In a random sample of $n = 4012$ high school graduates, 10.8% reported that they had taken a statistics course.[13] Give the sample size n, the count X, and the sample proportion $\hat{p}$ for this setting.

5.45 Use of the Internet to find a place to live. A poll of 1500 college students asked whether or not they have used the Internet to find a place to live sometime within the past year. There were 1234 students who answered Yes; the other 266 answered No.

(a) What is the sample size n?

(b) Choose one of the two possible outcomes to define the random variable, X. Give a reason for your choice.

(c) What is the value of the count X?

(d) Find the sample proportion, $\hat{p}$.

Just like the sample mean, sample counts and sample proportions are commonly used statistics, and understanding their sampling distributions is important for statistical inference. These statistics, however, are discrete random variables, so their sampling distributions introduce us to a new family of probability distributions.

The binomial distributions for sample counts

The distribution of a count X depends on how the data are produced. Here is a simple but common situation.

THE BINOMIAL SETTING

1. There is a fixed number of observations n.
2. The n observations are all independent.
3. Each observation falls into one of just two categories, which for convenience we call "success" and "failure."
4. The probability of a success, call it p, is the same for each observation.

Think of tossing a coin n times as an example of the binomial setting. Each toss gives either heads or tails, and the outcomes of successive tosses are independent. If we call heads a success, then p is the probability of a head and

remains the same as long as we toss the same coin. The number of heads we count is a random variable X. The distribution of X (and, more generally, the distribution of the count of successes in any binomial setting) is completely determined by the number of observations n and the success probability p.

BINOMIAL DISTRIBUTIONS

The distribution of the count X of successes in the binomial setting is called the **binomial distribution** with parameters n and p. The parameter n is the number of observations, and p is the probability of a success on any one observation. The possible values of X are the whole numbers from 0 to n. As an abbreviation, we say that the distribution of X is $B(n, p)$.

The binomial distributions are an important class of discrete probability distributions. Later in this section, we will learn how to assign probabilities to outcomes and how to find the mean and standard deviation of binomial distributions. *The most important skill for using binomial distributions is the ability to recognize situations to which they do and do not apply.* This can be done by checking all the facets of the binomial setting.

EXAMPLE 5.18

Binomial examples? **(a)** Genetics says that children receive genes from their parents independently. Each child of a particular pair of parents has probability 0.25 of having type O blood. If these parents have three children, the number who have type O blood is the count X of successes in three independent trials with probability 0.25 of a success on each trial. So X has the $B(3, 0.25)$ distribution.

(b) Engineers define reliability as the probability that an item will perform its function under specific conditions for a specific period of time. Replacement heart valves made of animal tissue, for example, have probability 0.77 of performing well for 15 years.[14] The probability of failure within 15 years is, therefore, 0.23. It is reasonable to assume that valves in different patients fail (or not) independently of each other. The number of patients in a group of 500 who will need another valve replacement within 15 years has the $B(500, 0.23)$ distribution.

(c) A multicenter trial is designed to assess a new surgical procedure. A total of 540 patients will undergo the procedure, and the count of patients X who suffer a major adverse cardiac event (MACE) within 30 days of surgery will be recorded. Because these patients will receive this procedure from different surgeons at different hospitals, it may not be true that the probability of a MACE is the same for each patient. Thus, X may not have the binomial distribution.

USE YOUR KNOWLEDGE

5.46 Genetics and blood types. Genetics says that children receive genes from each of their parents independently. Suppose that each child of a particular pair of parents has probability 0.375 of having type AB blood. If these parents have three children, what is the distribution of the number who have type AB blood? Explain your answer.

5.47 Tossing a coin. Suppose you plan to toss a coin 20 times and record X, the number of heads that you observe. If the coin is fair ($p = 0.5$), what is the distribution of X? Also, explain why this distribution is also the sampling distribution of X.

Binomial distributions in statistical sampling

The binomial distributions are important in statistics when we wish to make inferences about the proportion p of "successes" in a population. Here is a typical example.

EXAMPLE 5.19

D. Hurst/Alamy

Audits of financial records. The financial records of businesses are often audited by state tax authorities to test compliance with tax laws. Suppose that for one retail business, 800 of the 10,000 sales are incorrectly classified as subject to state sales tax. It would be too time-consuming for authorities to examine all these sales. Instead, an auditor examines an SRS of sales records. Is the count X of misclassified records in an SRS of 150 records a binomial random variable?

Choosing an SRS from a population is not quite a binomial setting. Removing one record in Example 5.19 changes the proportion of bad records in the remaining population, so the state of the second record chosen is not independent of the first. Because the population is large, however, removing a few items has a very small effect on the composition of the remaining population. Successive inspection results are very nearly independent. The population proportion of misclassified records is

$$p = \frac{800}{10{,}000} = 0.08$$

If the first record chosen is bad, the proportion of bad records remaining is 799/9999 = 0.079908. If the first record is good, the proportion of bad records left is 800/9999 = 0.080008. These proportions are so close to 0.08 that, for practical purposes, we can act as if removing one record has no effect on the proportion of misclassified records remaining. We act as if the count X of misclassified sales records in the audit sample has the binomial distribution $B(150, 0.08)$.

Populations like the one described in Example 5.19 often contain a relatively small number of items with very large values. For this example, these values would be very large sale amounts and likely represent an important group of items to the auditor. An SRS taken from such a population will likely include very few items of this type. Therefore, it is common to use a stratified sample in settings like this. Strata are defined based on dollar value of the sale, and within each stratum, an SRS is taken. The results are then combined to obtain an estimate for the entire population.

LOOK BACK
stratified random sample, p. 194

SAMPLING DISTRIBUTION OF A COUNT

A population contains proportion p of successes. If the population is much larger than the sample, the count X of successes in an SRS of size n has approximately the binomial distribution $B(n, p)$.

The accuracy of this approximation improves as the size of the population increases relative to the size of the sample. As a rule of thumb, we will use the binomial sampling distribution for counts when the population is at least 20 times as large as the sample.

Finding binomial probabilities

We will later give a formula for the probability that a binomial random variable takes any of its values. In practice, you will rarely have to use this formula for calculations because some calculators and most statistical software packages will calculate binomial probabilities for you.

EXAMPLE 5.20

Probabilities for misclassified sales records. In the audit setting of Example 5.19, what is the probability that the audit finds exactly 10 misclassified sales records? What is the probability that the audit finds no more than 10 misclassified records? Figure 5.14 shows the output from one statistical software system. You see that if the count X has the $B(150, 0.08)$ distribution,

$$P(X = 10) = 0.106959$$
$$P(X \leq 10) = 0.338427$$

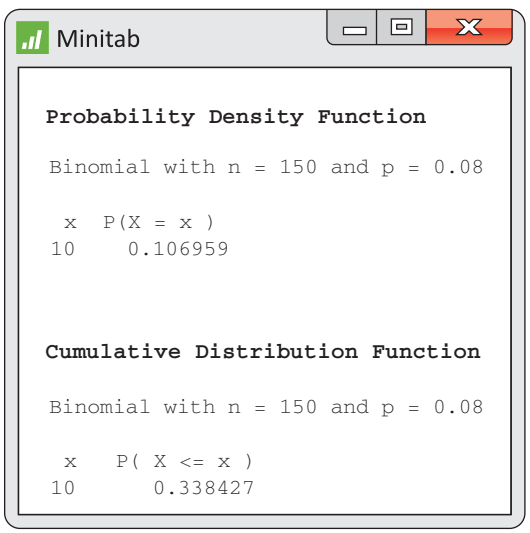

FIGURE 5.14 Binomial probabilities, Example 5.20: output from the Minitab statistical software.

It was easy to request these calculations in the software's menus. For the TI-83/84 calculator, the functions binompdf and binomcdf would be used. In R, the functions dbinom and pbinom would be used. Typically, the output supplies more decimal places than we need and uses labels that may not be helpful (for example, "Probability Density Function" when the distribution is discrete, not continuous). But, as usual with software, we can ignore distractions and find the results we need.

If you do not have suitable computing facilities, you can still shorten the work of calculating binomial probabilities for some values of n and p by looking up probabilities in Table C in the back of this book. The entries in the table are the probabilities $P(X = k)$ of individual outcomes for a binomial random variable X.

EXAMPLE 5.21

The probability histogram. Suppose that the audit in Example 5.19 chose just 15 sales records. What is the probability that no more than one of the 15 is misclassified? The count X of misclassified records in the sample has approximately the $B(15, 0.08)$ distribution. Figure 5.15 is a probability histogram for this distribution. The distribution is strongly skewed. Although X can take any whole-number value from 0 to 15, the probabilities of values larger than 5 are so small that they do not appear in the histogram.

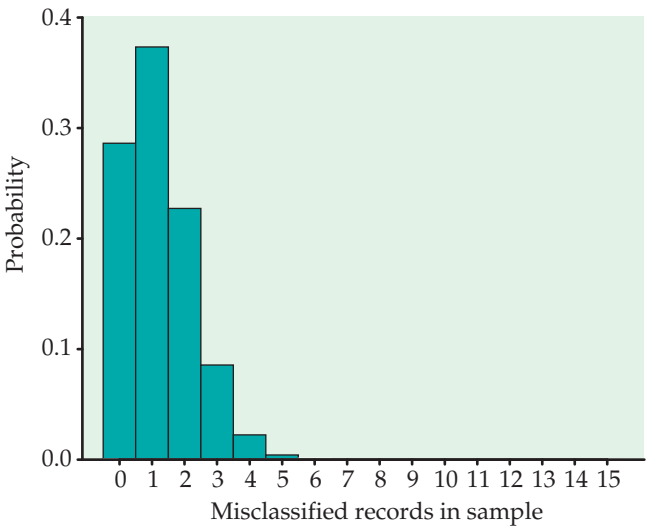

FIGURE 5.15 Probability histogram for the binomial distribution with $n = 15$ and $p = 0.08$, Example 5.21.

		p
n	k	.08
15	0	.2863
	1	.3734
	2	.2273
	3	.0857
	4	.0223
	5	.0043
	6	.0006
	7	.0001
	8	
	9	

We want to calculate

$$P(X \le 1) = P(X = 0) + P(X = 1)$$

when X has the $B(15, 0.08)$ distribution. To use Table C for this calculation, look opposite $n = 15$ and under $p = 0.08$. The entries in the rows for each k are $P(X = k)$. Blank cells in the table are 0 to four decimal places. You see that

$$\begin{aligned} P(X \le 1) &= P(X = 0) + P(X = 1) \\ &= 0.2863 + 0.3734 = 0.6597 \end{aligned}$$

About two-thirds of all samples will contain no more than 1 bad record. In fact, almost 29% of the samples will contain no bad records. The sample of size 15 cannot be trusted to provide adequate evidence about misclassified sales records. A larger number of observations is needed.

The values of p that appear in Table C are all 0.5 or smaller. When the probability of a success is greater than 0.5, restate the problem in terms of the number of failures. The probability of a failure is less than 0.5 when the probability of a success exceeds 0.5. When using the table, always stop to ask whether you must count successes or failures.

EXAMPLE 5.22

Falling asleep in class. In the survey of 4513 college students described in Example 5.4 (page 293), 46% of the respondents reported falling asleep in class due to poor sleep. You randomly sample 10 students in your dormitory, and eight state that they fell asleep in class during the last week due to poor sleep. Relative to the survey results, is this an unusually high number of students?

To answer this question, assume that the students' actions (falling asleep or not) are independent, with the probability of falling asleep equal to 0.46. This independence assumption may not be reasonable if the students study and socialize together or if there is a loud student in the dormitory who keeps everyone up. We'll assume this is not an issue here, so the number X of students who fell asleep in class out of 10 students has the $B(10, 0.46)$ distribution.

We want the probability of classifying at least eight students as having fallen asleep in class. Using software, we find

$$\begin{aligned} P(X \geq 8) &= P(X = 8) + P(X = 9) + P(X = 10) \\ &= 0.0263 + 0.0050 + 0.0004 = 0.0317 \end{aligned}$$

We would expect to find eight or more students falling asleep in class about 3% of the time or in fewer than one of every 30 surveys of 10 students. This is a pretty rare outcome and falls outside the range of the usual chance variation due to random sampling.

USE YOUR KNOWLEDGE

5.48 Free-throw shooting. April is a college basketball player who makes 80% of her free throws. In a recent game, she had 10 free throws and missed three of them. How unusual is this outcome? Using software, calculator, or Table C, compute $1 - P(X \leq 2)$, where X is the number of free throws missed in 10 shots. Explain your answer.

5.49 Find the probabilities.

(a) Suppose that X has the $B(8, 0.3)$ distribution. Use software, calculator, or Table C to find $P(X = 0)$ and $P(X \geq 6)$.

(b) Suppose that X has the $B(8, 0.7)$ distribution. Use software, calculator, or Table C to find $P(X = 8)$ and $P(X \leq 2)$.

(c) Explain the relationship between your answers to parts (a) and (b) of this exercise.

Binomial mean and standard deviation

If a count X has the $B(n, p)$ distribution, what are the mean μ_X and the standard deviation σ_X? We can guess the mean. If we expect 46% of the students to have fallen asleep in class due to poor sleep, the mean number in 10 students should be 46% of 10, or 4.6. That's μ_X when X has the $B(10, 0.46)$ distribution.

Intuition suggests more generally that the mean of the $B(n, p)$ distribution should be np. Can we show that this is correct and also obtain a short formula for the standard deviation? Because binomial distributions are discrete probability distributions, we could find the mean and variance by using the definitions in Section 4.4. Here is an easier way.

LOOK BACK
means and variances of random variables, pp. 246, 255

A binomial random variable X is the count of successes in n independent observations that each have the same probability p of success. Let the random variable S_i indicate whether the ith observation is a success or failure by taking the values $S_i = 1$ if a success occurs and $S_i = 0$ if the outcome is a failure. The S_i are independent because the observations are, and each S_i has the same simple distribution:

Outcome	1	0
Probability	p	$1 - p$

mean and variance of a discrete random variable, p. 236

From the definition of the mean of a discrete random variable, we know that the mean of each S_i is

$$\mu_S = (1)(p) + (0)(1 - p) = p$$

Similarly, the definition of the variance shows that $\sigma_S^2 = p(1 - p)$. Because each S_i is 1 for a success and 0 for a failure, to find the total number of successes X we add the S_i's:

$$X = S_1 + S_2 + \cdots + S_n$$

Apply the addition rules for means and variances to this sum. To find the mean of X we add the means of the S_i's:

$$\begin{aligned}\mu_X &= \mu_{S1} + \mu_{S2} + \cdots + \mu_{Sn} \\ &= n\mu_S = np\end{aligned}$$

Similarly, the variance is n times the variance of a single S, so that $\sigma_X^2 = np(1 - p)$. The standard deviation σ_X is the square root of the variance. Here is the result.

BINOMIAL MEAN AND STANDARD DEVIATION

If a count X has the binomial distribution $B(n, p)$, then

$$\mu_X = np$$
$$\sigma_X = \sqrt{np(1 - p)}$$

EXAMPLE 5.23

The Helsinki Heart Study. The Helsinki Heart Study asked whether the anticholesterol drug gemfibrozil reduces heart attacks. In planning such an experiment, the researchers must be confident that the sample sizes are large enough to enable them to observe enough heart attacks. The Helsinki study planned to give gemfibrozil to about 2000 men aged 40 to 55 and a placebo to another 2000. The probability of a heart attack during the five-year period of the study for men this age is about 0.04. What are the mean and standard deviation of the number of heart attacks that will be observed in one group if the treatment does not change this probability?

There are 2000 independent observations, each having probability $p = 0.04$ of a heart attack. The count X of heart attacks has the $B(2000, 0.04)$ distribution, so that

$$\mu_X = np = (2000)(0.04) = 80$$
$$\sigma_X = \sqrt{np(1 - p)} = \sqrt{(2000)(0.04)(0.96)} = 8.76$$

The expected number of heart attacks is large enough to permit conclusions about the effectiveness of the drug. In fact, there were 84 heart attacks among the 2035 men actually assigned to the placebo, quite close to the mean. The gemfibrozil group of 2046 men suffered only 56 heart attacks. This is evidence that the drug reduces the chance of a heart attack. In a later chapter, we will learn how to determine if this is strong enough evidence to conclude the drug is effective.

USE YOUR KNOWLEDGE

5.50 Free-throw shooting. Refer to Exercise 5.48. If April takes 85 free throws in the upcoming season, what are the mean and standard deviation of the number of free throws made?

5.51 Find the mean and standard deviation

(a) Suppose that X has the $B(8, 0.3)$ distribution. Compute the mean and standard deviation of X.

(b) Suppose that X has the $B(8, 0.7)$ distribution. Compute the mean and standard deviation of X.

(c) Explain the relationship between your answers to parts (a) and (b) of this exercise.

Sample proportions

population proportion, p. 199

What proportion of a company's sales records have an incorrect sales tax classification? What percent of adults favor stronger laws restricting firearms? In statistical sampling, we often want to estimate the proportion p of "successes" in a population. Our estimator is the sample proportion of successes:

$$\hat{p} = \frac{\text{count of successes in sample}}{\text{size of sample}} = \frac{X}{n}$$

Be sure to distinguish between the proportion $\hat{p}$ and the count X. The count takes whole-number values between 0 and n, but a proportion is always a number between 0 and 1. In the binomial setting, the count X has a binomial distribution. The proportion $\hat{p}$ does *not* have a binomial distribution. We can, however, do probability calculations about $\hat{p}$ by restating them in terms of the count X and using binomial methods. In Example 5.11 (pages 302–303), we took a similar approach for the sum, restating the problem in terms of the sample mean and then using the Normal distribution to calculate the probability.

EXAMPLE 5.24

NetPhotos/Alamy

Shopping online. A survey by the Consumer Reports National Research Center revealed that 84% of all respondents were very satisfied with their online shopping experience.[15] It was also reported, however, that people over the age of 40 were generally more satisfied than younger respondents. You decide to take a nationwide random sample of 2500 college students and ask if they agree or disagree that "I am very satisfied with my online shopping experience." Suppose that 60% of all college students would agree if asked this question. What is the probability that the sample proportion who agree is at least 58%?

The count X who agree has the binomial distribution $B(2500, 0.6)$. The sample proportion $\hat{p} = X/2500$ does *not* have a binomial distribution because it is not a count. But we can translate any question about a sample proportion $\hat{p}$ into a question about the count X. Because 58% of 2500 is 1450,

$$\begin{aligned} P(\hat{p} \geq 0.58) &= P(X \geq 1450) \\ &= P(X = 1450) + P(X = 1451) + \cdots + P(X = 2500) \end{aligned}$$

This is a rather elaborate calculation. We must add more than 1000 binomial probabilities. Software tells us that $P(\hat{p} \geq 0.58) = 0.9802$. But what do we do if we don't have access to software?

rules for means, p. 254
rules for variances, p. 258

As a first step, find the mean and standard deviation of a sample proportion. We know the mean and standard deviation of a sample count, so apply the rules from Section 4.4 for the mean and variance of a constant times a random variable. Here is the result.

MEAN AND STANDARD DEVIATION OF A SAMPLE PROPORTION

Let $\hat{p}$ be the sample proportion of successes in an SRS of size n drawn from a large population having population proportion p of successes. The mean and standard deviation of $\hat{p}$ are

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

The formula for $\sigma_{\hat{p}}$ is exactly correct in the binomial setting. It is approximately correct for an SRS from a large population. We will use it when the population is at least 20 times as large as the sample.

Let's now use these formulas to calculate the mean and standard deviation for Example 5.24.

EXAMPLE 5.25

The mean and the standard deviation. The mean and standard deviation of the proportion of the survey respondents in Example 5.24 who are satisfied with their online clothes-shopping experience are

$$\mu_{\hat{p}} = p = 0.6$$
$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(0.6)(0.4)}{2500}} = 0.0098$$

USE YOUR KNOWLEDGE

5.52 Find the mean and the standard deviation. If we toss a fair coin 150 times, the number of heads is a random variable that is binomial.

(a) Find the mean and the standard deviation of the sample proportion of heads.

(b) Is your answer to part (a) the same as the mean and the standard deviation of the sample count of heads in 150 throws? Explain your answer.

The fact that the mean of $\hat{p}$ is p states in statistical language that the sample proportion $\hat{p}$ in an SRS is an *unbiased estimator* of the population proportion p. When a sample is drawn from a new population having a different

value of the population proportion p, the sampling distribution of the unbiased estimator $\hat{p}$ changes so that its mean moves to the new value of p. We observed this fact empirically in Section 5.1 and have now verified it from the laws of probability.

The variability of $\hat{p}$ about its mean, as described by the variance or standard deviation, gets smaller as the sample size increases. So a sample proportion from a large sample will usually lie quite close to the population proportion p. We observed this in the simulation experiment on page 284 in Section 5.1. Now we have discovered exactly how the variability decreases: the standard deviation is $\sqrt{p(1-p)/n}$. Similar to what we observed in the previous section, the $\sqrt{n}$ in the denominator means that the sample size must be multiplied by 4 if we wish to divide the standard deviation in half.

Normal approximation for counts and proportions

Using simulation, we discovered in Section 5.1 that the sampling distribution of a sample proportion $\hat{p}$ is close to Normal. Now we know that the distribution of $\hat{p}$ is that of a binomial count divided by the sample size n. This seems at first to be a contradiction. To clear up the matter, look at Figure 5.16. This is a probability histogram of the exact distribution of the proportion of very satisfied shoppers $\hat{p}$, based on the binomial distribution $B(2500, 0.6)$. There are hundreds of narrow bars, one for each of the 2501 possible values of $\hat{p}$. Most have probabilities too small to show in a graph. *The probability histogram looks very Normal!* In fact, both the count X and the sample proportion $\hat{p}$ are approximately Normal in large samples.

We also know this to be true as a result of the central limit theorem discussed in the previous section (page 298). Recall that we can consider the count X as a sum

$$X = S_1 + S_2 + \cdots + S_n$$

of independent random variables S_i that take the value 1 if a success occurs on the ith trial and the value 0 otherwise. The proportion of successes $\hat{p} = X/n$

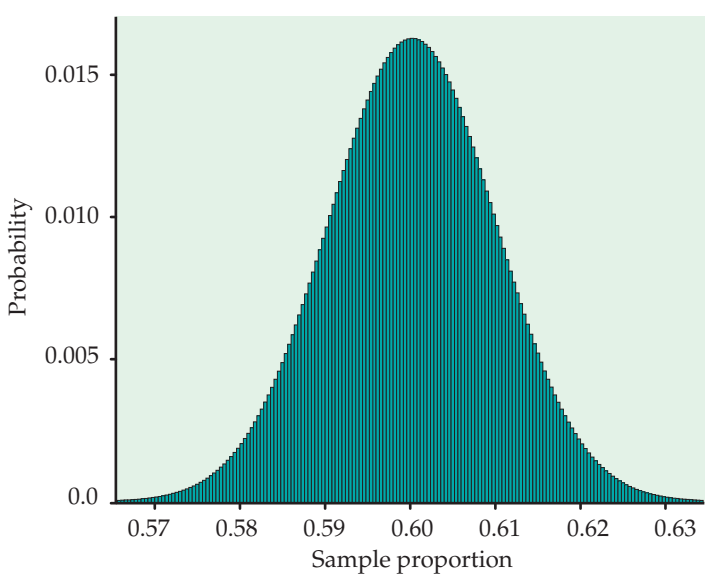

FIGURE 5.16 Probability histogram of the sample proportion $\hat{p}$ based on a binomial count with $n = 2500$ and $p = 0.6$. The distribution is very close to Normal.

can then be thought of as the sample mean of the S_i and, like all sample means, is approximately Normal when n is large. Given that $\hat{p}$ is approximately Normal, the count will also be approximately Normal because it is just a constant n times $\hat{p}$, an approximately Normal random variable.

NORMAL APPROXIMATION FOR COUNTS AND PROPORTIONS

Draw an SRS of size n from a large population having population proportion p of successes. Let X be the count of successes in the sample and $\hat{p} = X/n$ be the sample proportion of successes. When n is large, the sampling distributions of these statistics are approximately Normal:

$$X \text{ is approximately } N\left(np, \sqrt{np(1-p)}\right)$$

$$\hat{p} \text{ is approximately } N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

As a rule of thumb, we will use this approximation for values of n and p that satisfy $np \geq 10$ and $n(1-p) \geq 10$.

These Normal approximations are easy to remember because they say that $\hat{p}$ and X are Normal, with their usual means and standard deviations. Whether or not you use the Normal approximations should depend on how accurate your calculations need to be. For most statistical purposes, great accuracy is not required. Our "rule of thumb" for use of the Normal approximations reflects this judgment.

The accuracy of the Normal approximations improves as the sample size n increases. They are most accurate for any fixed n when p is close to 1/2, and least accurate when p is near 0 or 1. You can compare binomial distributions with their Normal approximations by using the *Normal Approximation to Binomial* applet. This applet allows you to change n or p while watching the effect on the binomial probability histogram and the Normal curve that approximates it.

Figure 5.17 summarizes the distribution of a sample proportion in a form that emphasizes the big idea of a sampling distribution. Just as with Figure 5.11 (page 294), the general framework for constructing a sampling distribution is shown on the left.

- Take many random samples of size n from a population that contains proportion p of successes.

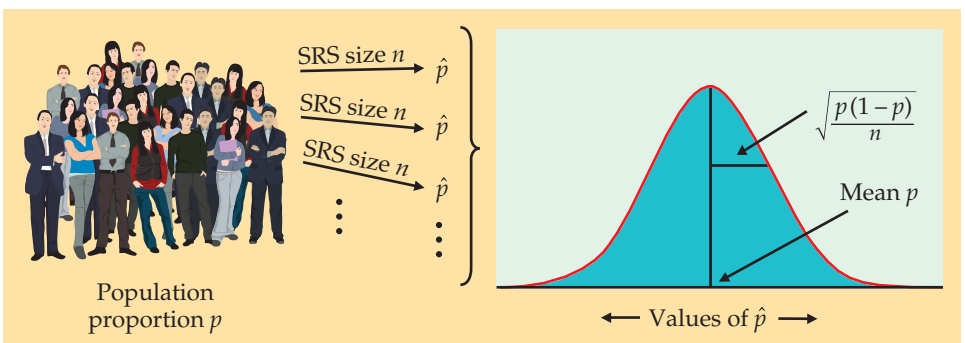

FIGURE 5.17 The sampling distribution of a sample proportion $\hat{p}$ is approximately Normal with mean p and standard deviation $\sqrt{p(1-p)/n}$.

- Find the sample proportion $\hat{p}$ for each sample.
- Collect all the $\hat{p}$'s and display their distribution.

The sampling distribution of $\hat{p}$ is shown on the right. Keep this figure in mind as you move toward statistical inference.

EXAMPLE 5.26

Compare the Normal approximation with the exact calculation. Let's compare the Normal approximation for the calculation of Example 5.24 with the exact calculation from software. We want to calculate $P(\hat{p} \geq 0.58)$ when the sample size is $n = 2500$ and the population proportion is $p = 0.6$. Example 5.25 shows that

$$\mu_{\hat{p}} = p = 0.6$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = 0.0098$$

Act as if $\hat{p}$ were Normal with mean 0.6 and standard deviation 0.0098. The approximate probability, as illustrated in Figure 5.18, is

$$P(\hat{p} \geq 0.58) = P\left(\frac{\hat{p} - 0.6}{0.0098} \geq \frac{0.58 - 0.6}{0.0098}\right)$$
$$\doteq P(Z \geq -2.04) = 0.9793$$

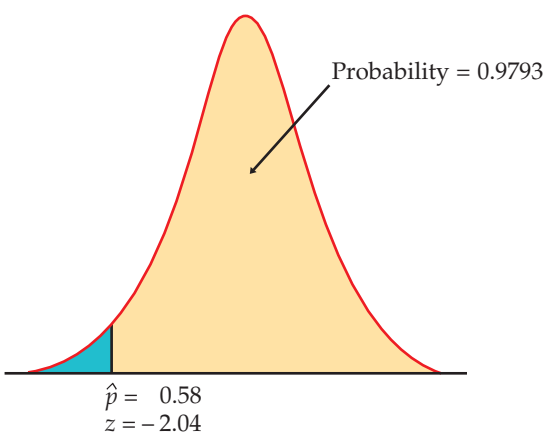

FIGURE 5.18 The Normal probability calculation, Example 5.26.

That is, about 98% of all samples have a sample proportion that is at least 0.58. Because the sample was large, this Normal approximation is quite accurate. It misses the software value 0.9802 by only 0.0009.

EXAMPLE 5.27

Using the Normal approximation. The audit described in Example 5.19 (page 314) examined an SRS of 150 sales records for compliance with sales tax laws. In fact, 8% of all the company's sales records have an incorrect

sales tax classification. The count X of bad records in the sample has approximately the $B(150, 0.08)$ distribution.

According to the Normal approximation to the binomial distributions, the count X is approximately Normal with mean and standard deviation

$$\mu_X = np = (150)(0.08) = 12$$
$$\sigma_X = \sqrt{np(1-p)} = \sqrt{(150)(0.08)(0.92)} = 3.3226$$

The Normal approximation for the probability of no more than 10 misclassified records is the area to the left of $X = 10$ under the Normal curve. Using Table A,

$$P(X \le 10) = P\left(\frac{X - 12}{3.3226} \le \frac{10 - 12}{3.3226}\right)$$
$$\doteq P(Z \le -0.60) = 0.2743$$

Software tells us that the actual binomial probability that no more than 10 of the records in the sample are misclassified is $P(X \le 10) = 0.3384$. The Normal approximation is only roughly accurate. Because $np = 12$, this combination of n and p is close to the border of the values for which we are willing to use the approximation.

The distribution of the count of bad records in a sample of 15 is distinctly non-Normal, as Figure 5.15 (page 316) showed. When we increase the sample size to 150, however, the shape of the binomial distribution becomes roughly Normal. Figure 5.19 displays the probability histogram of the binomial distribution with the density curve of the approximating Normal distribution superimposed. Both distributions have the same mean and standard deviation, and both the area under the histogram and the area under the curve are 1. The Normal curve fits the histogram reasonably well. Look closely: the histogram is slightly skewed to the right, a property that the symmetric Normal curve can't match.

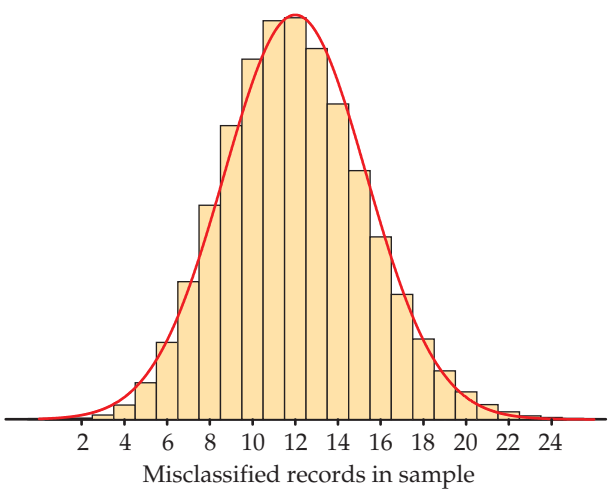

FIGURE 5.19 Probability histogram and Normal approximation for the binomial distribution with $n = 150$ and $p = 0.08$, Example 5.27.

USE YOUR KNOWLEDGE

5.53 Use the Normal approximation. Suppose that we toss a fair coin 150 times. Use the Normal approximation to find the probability that the sample proportion of heads is

(a) between 0.4 and 0.6.

(b) between 0.45 and 0.55.

The continuity correction

Figure 5.20 illustrates an idea that greatly improves the accuracy of the Normal approximation to binomial probabilities. The binomial probability $P(X \le 10)$ is the area of the histogram bars for values 0 to 10. The bar for $X = 10$ actually extends from 9.5 to 10.5. Because the discrete binomial distribution puts probability only on whole numbers, the probabilities $P(X \le 10)$ and $P(X \le 10.5)$ are the same. The Normal distribution spreads probability continuously, so these two Normal probabilities are different. The Normal approximation is more accurate if we consider $X = 10$ to extend from 9.5 to 10.5, matching the bar in the probability histogram.

The event $\{X \le 10\}$ includes the outcome $X = 10$. Figure 5.20 shades the area under the Normal curve that matches all the histogram bars for outcomes 0 to 10, bounded on the right not by 10, but by 10.5. So $P(X \le 10)$ is calculated as $P(X \le 10.5)$. On the other hand, $P(X < 10)$ excludes the outcome $X = 10$, so we exclude the entire interval from 9.5 to 10.5 and calculate $P(X \le 9.5)$ from the Normal table. Here is the result of the Normal calculation in Example 5.27 improved in this way:

$$\begin{aligned} P(X \le 10) &= P(X \le 10.5) \\ &= P\left(\frac{X - 12}{3.3226} \le \frac{10.5 - 12}{3.3226}\right) \\ &\doteq P(Z \le -0.45) = 0.3264 \end{aligned}$$

The improved approximation misses the binomial probability by only 0.012. Acting as though a whole number occupies the interval from 0.5 below to 0.5 above the number is called the **continuity correction** to the Normal

continuity correction

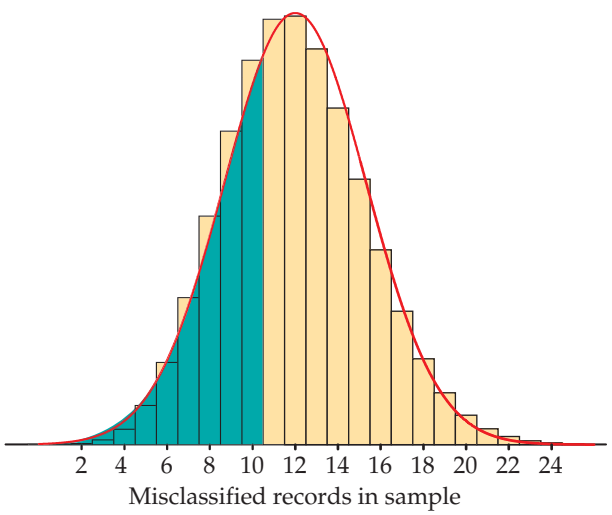

FIGURE 5.20 Area under the Normal approximation curve for the probability in Example 5.27.

approximation. If you need accurate values for binomial probabilities, try to use software to do exact calculations. If no software is available, use the continuity correction unless n is very large. Because most statistical purposes do not require extremely accurate probability calculations, we do not emphasize use of the continuity correction.

Binomial formula

We can find a formula for the probability that a binomial random variable takes any value by adding probabilities for the different ways of getting exactly that many successes in n observations. Here is the example we will use to show the idea.

EXAMPLE 5.28

Blood types of children. Each child born to a particular set of parents has probability 0.25 of having blood type O. If these parents have five children, what is the probability that exactly two of them have type O blood?

The count of children with type O blood is a binomial random variable X with $n = 5$ tries and probability $p = 0.25$ of a success on each try. We want $P(X = 2)$.

Because the method doesn't depend on the specific example, we will use "S" for success and "F" for failure. In Example 5.28, "S" would stand for type O blood. Do the work in two steps.

Step 1: Find the probability that a specific two of the five tries give successes—say, the first and the third. This is the outcome SFSFF. The multiplication rule for independent events tells us that

$$\begin{aligned} P(\text{SFSFF}) &= P(\text{S})P(\text{F})P(\text{S})P(\text{F})P(\text{F}) \\ &= (0.25)(0.75)(0.25)(0.75)(0.75) \\ &= (0.25)^2(0.75)^3 \end{aligned}$$

Step 2: Observe that the probability of *any one* arrangement of two S's and three F's has this same probability. That's true because we multiply together 0.25 twice and 0.75 three times whenever we have two S's and three F's. The probability that $X = 2$ is the probability of getting two S's and three F's in any arrangement whatsoever. Here are all the possible arrangements:

SSFFF SFSFF SFFSF SFFFS FSSFF
FSFSF FSFFS FFSSF FFSFS FFFSS

There are 10 of them, all with the same probability. The overall probability of two successes is, therefore,

$$P(X = 2) = 10(0.25)^2(0.75)^3 = 0.2637$$

The pattern of this calculation works for any binomial probability. To use it, we need to be able to count the number of arrangements of k successes in n observations without actually listing them. We use the following fact to do the counting.

BINOMIAL COEFFICIENT

The number of ways of arranging k successes among n observations is given by the **binomial coefficient**

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

for $k = 0, 1, 2, \ldots, n$.

factorial

The formula for binomial coefficients uses the **factorial** notation. The factorial $n!$ for any positive whole number n is

$$n! = n \times (n-1) \times (n-2) \times \cdots \times 3 \times 2 \times 1$$

Also, $0! = 1$. Notice that the larger of the two factorials in the denominator of a binomial coefficient will cancel much of the $n!$ in the numerator. For example, the binomial coefficient we need for Example 5.28 is

$$\begin{aligned}\binom{5}{2} &= \frac{5!}{2!3!} \\ &= \frac{(5)(4)(3)(2)(1)}{(2)(1) \times (3)(2)(1)} \\ &= \frac{(5)(4)}{(2)(1)} = \frac{20}{2} = 10\end{aligned}$$

This agrees with our previous calculation.

The notation $\binom{n}{k}$ *is not related to the fraction* $\frac{n}{k}$. A helpful way to remember its meaning is to read it as "binomial coefficient n choose k." Binomial coefficients have many uses in mathematics, but we are interested in them only as an aid to finding binomial probabilities. The binomial coefficient $\binom{n}{k}$ counts the number of ways in which k successes can be distributed among n observations. The binomial probability $P(X = k)$ is this count multiplied by the probability of any specific arrangement of the k successes. Here is the formula we seek.

BINOMIAL PROBABILITY

If X has the binomial distribution $B(n, p)$ with n observations and probability p of success on each observation, the possible values of X are $0, 1, 2, \ldots, n$. If k is any one of these values, the **binomial probability** is

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Here is an example of the use of the binomial probability formula.

EXAMPLE 5.29

Using the binomial probability formula. The number X of misclassified sales records in the auditor's sample in Example 5.21 (page 316) has the $B(15,0.08)$ distribution. The probability of finding no more than one misclassified record is

$$\begin{aligned} P(X \le 1) &= P(X = 0) + P(X = 1) \\ &= \binom{15}{0}(0.08)^0(0.92)^{15} + \binom{15}{1}(0.08)^1(0.92)^{14} \\ &= \frac{15!}{0!15!}(1)(0.2863) + \frac{15!}{1!14!}(0.08)(0.3112) \\ &= (1)(1)(0.2863) + (15)(0.08)(0.3112) \\ &= 0.2863 + 0.3734 = 0.6597 \end{aligned}$$

The calculation used the facts that $0! = 1$ and that $a^0 = 1$ for any number $a \neq 0$. The result agrees with that obtained from Table C in Example 5.21.

USE YOUR KNOWLEDGE

5.54 An unfair coin. A coin is slightly bent, and as a result, the probability of a head is 0.53. Suppose that you toss the coin five times.

(a) Use the binomial formula to find the probability of three or more heads.

(b) Compare your answer with the one that you would obtain if the coin were fair.

The Poisson distributions

A count X has a binomial distribution when it is produced under the binomial setting. If one or more facets of this setting do not hold, the count X will have a different distribution. In this subsection, we discuss one of these distributions.

Frequently, we meet counts that are open-ended; that is, they are not based on a fixed number of n observations: the number of customers at a popular café between 12:00 P.M. and 1:00 P.M.; the number of dings on your car door; the number of reported pedestrian/bicyclist collisions on campus during the academic year. These are all counts that could be 0, 1, 2, 3, and so on indefinitely.

The Poisson distribution is another model for a count and can often be used in these open-ended situations. The count represents the number of events (call them "successes") that occur in some fixed unit of measure such as a period of time or region of space. The Poisson distribution is appropriate under the following conditions.

THE POISSON SETTING

1. The number of successes that occur in two nonoverlapping units of measure are **independent**.
2. The probability that a success will occur in a unit of measure is the same for all units of equal size and is proportional to the size of the unit.
3. The probability that more than one event occurs in a unit of measure is negligible for very small-sized units. In other words, the events occur one at a time.

For binomial distributions, the important quantities were n, the fixed number of observations, and p, the probability of success on any given observation. For Poisson distributions, the only important quantity is the mean number of successes μ occurring per unit of measure.

POISSON DISTRIBUTION

The distribution of the count X of successes in the Poisson setting is the **Poisson distribution** with **mean** μ. The parameter μ is the mean number of successes per unit of measure. The possible values of X are the whole numbers 0, 1, 2, 3, … . If k is any whole number, then*

$$P(X = k) = \frac{e^{-\mu}\mu^k}{k!}$$

The **standard deviation** of the distribution is $\sqrt{\mu}$.

EXAMPLE 5.30

Number of dropped calls. Suppose that the number of dropped calls on your cell phone varies, with an average of 2.1 calls per day. If we assume that the Poisson setting is reasonable for this situation, we can model the daily count of dropped calls X using the Poisson distribution with $\mu = 2.1$. What is the probability of having no more than two dropped calls tomorrow?

We can calculate $P(X \le 2)$ either using software or the Poisson probability formula. Using the probability formula:

$$\begin{aligned} P(X \le 2) &= P(X = 0) + P(X = 1) + P(X = 2) \\ &= \frac{e^{-2.1}(2.1)^0}{0!} + \frac{e^{-2.1}(2.1)^1}{1!} + \frac{e^{-2.1}(2.1)^2}{2!} \\ &= 0.1225 + 0.2572 + 0.2700 \\ &= 0.6497 \end{aligned}$$

Using the R software, the probability is

```
dpois(0,2.1) + dpois(1,2.1) + dpois(2,2.1)
[1] 0.6496314
```

These two answers differ slightly due to roundoff error in the hand calculation. There is roughly a 65% chance that you will have no more than two dropped calls tomorrow.

Similar to the binomial, Poisson probability calculations are rarely done by hand if the event includes numerous possible values for X. Most software provides functions to calculate $P(X = k)$ and the cumulative probabilities of the form $P(X \le k)$. These cumulative probability calculations make solving many problems less tedious. Here's an example.

*The e in the Poisson probability formula is a mathematical constant equal to 2.71828 to six decimal places. Many calculators have an e^x function.

EXAMPLE 5.31

Counting software remote users. Your university supplies online remote access to various software programs used in courses. Suppose that the number of students remotely accessing these programs in any given hour can be modeled by a Poisson distribution with $\mu = 17.2$. What is the probability that more than 25 students will remotely access these programs in the next hour?

Calculating this probability requires two steps.

1. Write $P(X > 25)$ as an expression involving a cumulative probability:

$$P(X > 25) = 1 - P(X \leq 25)$$

2. Obtain $P(X \leq 25)$ and subtract the value from 1. Again using R,

```
1- ppois(25,17.2)
[1] 0.02847261
```

The probability that more than 25 students will use this remote access in the next hour is only 0.028. Relying on software to get the cumulative probability is much quicker and less prone to error than the method of Example 5.30. For this case, that method would involve determining 26 probabilities and then summing their values.

Under the Poisson setting, this probability of 0.028 applies not only to the next hour, but also to any other hour in the future. The probability does not change because the units of measure are the same size and nonoverlapping.

USE YOUR KNOWLEDGE

5.55 Number of aphids. The milkweed aphid is a common pest to many ornamental plants. Suppose that the number of aphids on a shoot of a Mexican butterfly weed follows a Poisson distribution with $\mu = 4.4$ aphids.

(a) What is the probability of observing exactly five aphids on a shoot?

(b) What is the probability of observing five or fewer aphids on a shoot?

5.56 Number of aphids, continued. Refer to the previous exercise.

(a) What proportion of shoots would you expect to have no aphids present?

(b) If you do not observe any aphids on a shoot, is the probability that a nearby shoot has no aphids smaller than, equal to, or larger than your answer in part (a)? Explain your reasoning.

If we add counts from successive nonoverlapping areas of equal size, we are just counting the successes in a larger area. That count still meets the conditions of the Poisson setting. However, because our unit of measure has doubled, the mean of this new count is twice as large. Put more formally, if X is a Poisson random variable with mean μ_X and Y is a Poisson random variable with mean μ_Y and Y is independent of X, then $X + Y$ is a Poisson random variable with mean $\mu_X + \mu_Y$. This fact means that we can combine areas or look at a portion of an area and still use Poisson distributions to model the count.

EXAMPLE 5.32

Number of potholes. The Automobile Association (AA) in Britain had member volunteers make a 60-minute, two-mile walk around their neighborhoods and survey the condition of their roads and sidewalks. One outcome was the number of potholes, defined as being at least 2 inches deep and at least 6 inches in diameter, in their roads.[16] It was reported that Scotland averages 8.9 potholes per mile of road and London averages 4.9 potholes per mile of road. Suppose that the number of potholes per mile in each of these two regions follow the Poisson distribution. Then

- The number of potholes per 20 miles of road in Scotland is a Poisson random variable with mean $20 \times 8.9 = 178$.
- The number of potholes per half mile of road in London is a Poisson random variable with mean $0.5 \times 4.9 = 2.45$.
- The number of potholes per 500 miles of road in Scotland is a Poisson random variable with mean $500 \times 8.9 = 4450$.
- If we examined 2 miles of road in Scotland and 5 miles of road in London, the total number of potholes would be a Poisson random variable with mean $2 \times 8.9 + 5 \times 4.9 = 42.3$.

When the mean of the Poisson distribution is large, it may be difficult to calculate Poisson probabilities using a calculator or software. Fortunately, when μ is large, Poisson probabilities can be approximated using the Normal distribution with mean μ and standard deviation $\sqrt{\mu}$. Here is an example.

EXAMPLE 5.33

Number of snaps received. In Example 5.11 (pages 302–303), it was reported that Snapchat has more than 100 million daily users who send over 400 million snaps a day. Suppose that the number of snaps you receive per day follows a Poisson distribution with mean 12. What is the probability that, over a week, you would receive more than 100 snaps?

To answer this using software, we first compute the mean number of snaps sent per week. Because there are seven days in a week, the mean is $7 \times 12 = 84$. Plugging this into R tells us that there is slightly less than an 4% chance of receiving this many snaps:

```
1-ppois(100,84)
[1] 0.03891883
```

For the Normal approximation we compute

$$\begin{aligned} P(X > 100) &= P\left(\frac{X - 84}{\sqrt{84}} > \frac{100 - 84}{\sqrt{84}}\right) \\ &= P(Z > 1.75) \\ &= 1 - P(Z < 1.75) \\ &= 1 - 0.9599 = 0.0401 \end{aligned}$$

The approximation is quite accurate, differing from the actual probability by only 0.0012.

While the Normal approximation is adequate for many practical purposes, we recommend using statistical software when possible so you can get exact Poisson probabilities.

There is one other approximation associated with the Poisson distribution that is worth mentioning. It is related to the binomial distribution. Previously, we recommended using the Normal distribution to approximate the binomial distribution when n and p satisfy $np \geq 10$ and $n(1 - p) \geq 10$. In cases where n is large but p is so small that $np < 10$, the Poisson distribution with $\mu = np$ yields more accurate results. For example, suppose that you wanted to calculate $P(X \leq 2)$ when X has the $B(1000, .001)$ distribution. Using R, the actual binomial probability and the Poisson approximation are

```
pbinom(2,1000,.001)      ppois(2,1)
[1] 0.9197907            [1] 0.9196986
```

The Poisson approximation gives a very accurate probability calculation for the binomial distribution in this case.

SECTION 5.3 SUMMARY

- A **count** X of successes has the **binomial distribution** $B(n, p)$ in the **binomial setting**: there are n trials, all independent, each resulting in a success or a failure, and each having the same probability p of a success.
- The binomial distribution $B(n, p)$ is a good approximation to the **sampling distribution of the count of successes** in an SRS of size n from a large population containing proportion p of successes. We will use this approximation when the population is at least 20 times larger than the sample.
- The **sample proportion** of successes $\hat{p} = X/n$ is an estimator of the population proportion p. It does not have a binomial distribution, but we can do probability calculations about $\hat{p}$ by restating them in terms of X.
- **Binomial probabilities** are most easily found by software. There is an exact formula that is practical for calculations when n is small. Table C contains binomial probabilities for some values of n and p. For large n, you can use the Normal approximation.
- The mean and standard deviation of a **binomial count** X and a **sample proportion** $\hat{p} = X/n$ are

$$\mu_X = np \qquad \mu_{\hat{p}} = p$$

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

The sample proportion $\hat{p}$ is, therefore, an unbiased estimator of the population proportion p.

- The **Normal approximation** to the binomial distribution says that if X is a count having the $B(n,p)$ distribution, then when n is large,

$$X \text{ is approximately } N\left(np, \sqrt{np(1-p)}\right)$$

$$\hat{p} \text{ is approximately } N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

We will use this approximation when $np \geq 10$ and $n(1 - p) \geq 10$. It allows us to approximate probability calculations about X and $\hat{p}$ using the Normal distribution.

- The **continuity correction** improves the accuracy of the Normal approximations.
- The exact **binomial probability formula** is

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

where the possible values of X are $k = 0,1, \ldots, n$. The binomial probability formula uses the **binomial coefficient**

$$\binom{n}{k} = \frac{n!}{k!(n - k)!}$$

- Here the **factorial** $n!$ is

$$n! = n \times (n - 1) \times (n - 2) \times \cdots \times 3 \times 2 \times 1$$

for positive whole numbers n and $0! = 1$. The binomial coefficient counts the number of ways of distributing k successes among n trials.

- A count X of successes has a **Poisson distribution** in the **Poisson setting**: the number of successes that occur in two nonoverlapping units of measure are independent; the probability that a success will occur in a unit of measure is the same for all units of equal size and is proportional to the size of the unit; the probability that more than one event occurs in a unit of measure is negligible for very small-sized units. In other words, the events occur one at a time.
- If X has the Poisson distribution with mean μ, then the standard deviation of X is $\sqrt{\mu}$, and the possible values of X are the whole numbers 0, 1, 2, 3, and so on.
- The **Poisson probability** that X takes any of these values is

$$P(X = k) = \frac{e^{-\mu}\mu^k}{k!} \qquad k = 0, 1, 2, 3, \ldots$$

Sums of independent Poisson random variables also have the Poisson distribution. For example, in a Poisson model with mean μ per unit of measure, the count of successes in a units is a Poisson random variable with mean $a\mu$.

SECTION 5.3 EXERCISES

For Exercises 5.43, 5.44, and 5.45, see page 312; for Exercises 5.46 and 5.47, see page 313; for Exercises 5.48 and 5.49, see page 317; for Exercises 5.50 and 5.51, see page 319; for Exercise 5.52, see page 320; for Exercise 5.53, see page 325; for Exercise 5.54, see page 328; and for Exercises 5.55 and 5.56, see page 330.

Most binomial probability calculations required in these exercises can be done by using Table C or the Normal approximation. Your instructor may request that you use the binomial probability formula or software. In exercises requiring the Normal approximation, you should use the continuity correction if you studied that topic.

5.57 What is wrong? Explain what is wrong in each of the following scenarios.

(a) If you toss a fair coin four times and a head appears each time, then the next toss is more likely to be a tail than a head.

(b) If you toss a fair coin four times and observe the pattern HTHT, then the next toss is more likely to be a head than a tail.

(c) The quantity $\hat{p}$ is one of the parameters for a binomial distribution.

(d) The binomial distribution can be used to model the daily number of pedestrian/cyclist near-crash events on campus.

5.58 What is wrong? Explain what is wrong in each of the following scenarios.

(a) In the binomial setting, X is a proportion.

(b) The variance for a binomial count is $\sqrt{p(1-p)/n}$.

(c) The Normal approximation to the binomial distribution is always accurate when n is greater than 1000.

(d) We can use the binomial distribution to approximate the sampling distribution of $\hat{p}$ when we draw an SRS of size $n = 50$ students from a population of 500 students.

5.59 Should you use the binomial distribution? In each of the following situations, is it reasonable to use a binomial distribution for the random variable X? Give reasons for your answer in each case. If a binomial distribution applies, give the values of n and p.

(a) A poll of 200 college students asks whether or not you usually feel irritable in the morning. X is the number who reply that they do usually feel irritable in the morning.

(b) You toss a fair coin until a head appears. X is the count of the number of tosses that you make.

(c) Most calls made at random by sample surveys don't succeed in talking with a person. Of calls to New York City, only one-twelfth succeed. A survey calls 500 randomly selected numbers in New York City. X is the number of times that a person is reached.

(d) You deal 10 cards from a shuffled deck of standard playing cards and count the number X of black cards.

5.60 Should you use the binomial distribution? In each of the following situations, is it reasonable to use a binomial distribution for the random variable X? Give reasons for your answer in each case.

(a) In a random sample of students in a fitness study, X is the mean daily exercise time of the sample.

(b) A manufacturer of running shoes picks a random sample of 20 shoes from the production of shoes each day for a detailed inspection. X is the number of pairs of shoes with a defect.

(c) A nutrition study chooses an SRS of college students. They are asked whether or not they usually eat at least five servings of fruits or vegetables per day. X is the number who say that they do.

(d) X is the number of days during the school year when you skip a class.

5.61 Stealing from a store. A survey of more than 20,000 U.S. high school students revealed that 20% of the students say that they stole something from a store in the past year.[17] This is down 7% from the last survey, which was performed two years earlier. You decide to take a random sample of 10 high school students from your city and ask them this question.

(a) If the high school students in your city match this 20% rate, what is the distribution of the number of students who say that they stole something from a store in the past year? What is the distribution of the number of students who do not say that they stole something from a store in the past year?

(b) What is the probability that four or more of the 10 students in your sample say that they stole something from a store in the past year?

5.62 Illegal downloading. New regulations in Canada require all Internet service providers (ISPs) to send a notice to subscribers who are downloading files illegally asking them to stop. This "notice and notice" system was already in place with Rogers Cable. That company says that prior to these new regulations, 67% of its subscribers who received a notice did not reoffend.[18] Consider a random sample of 50 of these Rogers subscribers who received a first notice.

(a) What is the distribution of the number X of subscribers who reoffend? Explain your answer.

(b) What is the probability that at least 18 of the 50 subscribers in your sample reoffend?

5.63 Stealing from a store, continued. Refer to Exercise 5.61.

(a) What is the expected number of students in your sample who say that they stole something from a store in the past year? What is the expected number of students who do not say that they stole? You should see that these two means add to 10, the total number of students.

(b) What is the standard deviation σ of the number of students in your sample who say that they stole something?

(c) Suppose that you live in a city where only 10% of the high school students say that they stole something from a store in the past year. What is σ in this case? What is σ if $p = 0.01$? What happens to the standard deviation of a binomial distribution as the probability of a success gets close to 0?

5.64 Illegal downloading, continued. Refer to Exercise 5.62. Given the new regulations, suppose that 75% of the Canadian ISP subscribers will not reoffend after receiving a notice.

(a) If you choose at random 15 subscribers who received a notice, what is the mean of the count X who will not reoffend? What is the mean of the proportion $\hat{p}$ in your sample who will not reoffend?

(b) Repeat the calculations in part (a) for samples of size 150 and 1500. What happens to the mean count of

successes as the sample size increases? What happens to the mean proportion of successes?

5.65 More on illegal downloading. Consider the settings of Exercises 5.62 and 5.64.

(a) Using the 67% rate of Rogers subscribers prior to the new regulations, what is the smallest number m out of $n = 15$ Canadian ISP subscribers who receive a notice such that $P(X \geq m)$ is no larger than 0.05? You might consider m or more subscribers as evidence that the rate in your sample is larger than 67%.

(b) Now using the 75% rate of Canadian ISP subscribers after the new regulations and your answer to part (a), what is $P(X \geq m)$? This represents the chance of obtaining enough evidence given that the rate is 75%.

(c) If you were to increase the sample size from $n = 15$ to $n = 100$ and repeat parts (a) and (b), would you expect the probability in part (b) to increase or decrease? Explain your answer.

5.66 Attitudes toward drinking and studies of behavior. Some of the methods in this section are approximations rather than exact probability results. We have given rules of thumb for safe use of these approximations.

(a) You are interested in attitudes toward drinking among the 75 members of a fraternity. You choose 30 members at random to interview. One question is "Have you had five or more drinks at one time during the last week?" Suppose that, in fact, 30% of the 75 members would say Yes. Explain why you *cannot* safely use the $B(30, 0.3)$ distribution for the count X in your sample who say Yes.

(b) The National AIDS Behavioral Surveys found that 0.2% (that's 0.002 as a decimal fraction) of adult heterosexuals had both received a blood transfusion and had a sexual partner from a group at high risk of AIDS. Suppose that this national proportion holds for your region. Explain why you *cannot* safely use the Normal approximation for the sample proportion who fall in this group when you interview an SRS of 1000 adults.

5.67 Random digits. Each entry in a table of random digits like Table B has probability 0.1 of being any given digit, and digits are independent of each other.

(a) What is the probability that a group of six digits from the table will contain at least one digit greater than 5?

(b) What is the mean number of digits greater than 5 in lines 40 digits long?

5.68 Use the *Probability* applet. The *Probability* applet simulates tosses of a coin. You can choose the number of tosses n and the probability p of a head. You can therefore use the applet to simulate binomial random variables.

The count of misclassified sales records in Example 5.21 (page 316) has the binomial distribution with $n = 15$ and $p = 0.08$. Set these values for the number of tosses and probability of heads in the applet. Table C shows that the probability of getting a sample with exactly 0 misclassified records is 0.2863. This is the long-run proportion of samples with no bad records. Click "Toss" and "Reset" repeatedly to simulate 25 samples of 15 tosses. Record the number of bad records (the count of heads) in each of the 25 samples.

(a) What proportion of the 25 samples had exactly 0 bad records? Do you think this sample proportion is close to the probability?

(b) Remember that this probability of 0.2863 tells us only what happens in the long run. Here we're considering only 25 samples. If X is the number of samples out of 25 with exactly 0 misclassified records, what is the distribution of X?

(c) Explain how to use the distribution in part (b) to describe the sampling distribution of $\hat{p}$ in part (a).

5.69 Cyberbullying. An online survey, in partnership with Habbo, was conducted to study cyberbullying among 13- to 25-year-olds in the United Kingdom. It was reported that 62% of the young people had received nasty private messages on a smartphone social network app.[19] You randomly sample four young people from the United Kingdom and ask them if they've received nasty messages. Let X be the number who say Yes.

(a) What are n and p in the binomial distribution of X?

(b) Find the probability of each possible value of X, and draw a probability histogram for this distribution.

(c) Find the mean number of positive responders and mark the location of this value on your histogram.

5.70 The ideal number of children. "What do you think is the ideal number of children for a family to have?" A Gallup Poll asked this question of 1020 randomly chosen adults. Slightly less than half (48%) thought that a total of two children was ideal.[20] Suppose that $p = 0.48$ is exactly true for the population of all adults. Gallup announced a margin of error of ± 4 percentage points for this poll. What is the probability that the sample proportion $\hat{p}$ for an SRS of size $n = 1020$ falls between 0.44 and 0.52? You see that it is likely, but not certain, that polls like this give results that are correct within their margin of error. We say more about margins of error in Chapter 6.

5.71 Cyberbullying, continued. Refer to Exercise 5.69. Assume instead that that you sample $n = 500$ young people from the United Kingdom.

(a) What is the probability that the sample proportion $\hat{p}$ of those who received nasty messages is between 0.59 and 0.65 if the population proportion is $p = 0.62$?

(b) What is the probability that the sample proportion $\hat{p}$ is between 0.87 and 0.93 if the population proportion is $p = 0.90$?

(c) Using the results from parts (a) and (b), how does the probability that $\hat{p}$ falls within ± 0.03 of the true p change as p gets closer to 1?

5.72 How do the results depend on the sample size? Return to the Gallup Poll setting of Exercise 5.70. We are supposing that the proportion of all adults who think that having two children is ideal is $p = 0.48$. What is the probability that a sample proportion $\hat{p}$ falls between 0.44 and 0.52 (that is, within ± 4 percentage points of the true p) if the sample is an SRS of size $n = 300$? Of size $n = 5000$? Combine these results with your work in Exercise 5.70 to make a general statement about the effect of larger samples in a sample survey.

5.73 Shooting free throws. Since the mid-1960s, the overall free-throw percent at all college levels, for both men and women, has remained pretty consistent. For men, players have been successful on roughly 69% of these free throws, with the season percent never falling below 67% or above 70%.[21] Assume that 300,000 free throws will be attempted in the upcoming season.

(a) What are the mean and standard deviation of $\hat{p}$ if the population proportion is $p = 0.69$?

(b) Using the 68–95–99.7 rule, we expect $\hat{p}$ to fall between what two percents about 95% of the time?

(c) Given the width of the interval in part (b) and the range of season percents, do you think that it is reasonable to assume that the population proportion has been the same over the last 50 seasons? Explain your answer.

5.74 Online learning. The U.S. Department of Education released a report on online learning stating that blended instruction, a combination of conventional face-to-face and online instruction, appears more effective in terms of student performance than conventional teaching.[22] You decide to poll incoming students at your institution to see if they prefer courses that blend face-to-face instruction with online components. In an SRS of 400 incoming students, you find that 373 prefer this type of course.

(a) What is the sample proportion of incoming students at your school who prefer this type of blended instruction?

(b) Assume the population proportion for all students nationwide is 85%. Assuming this is true for your insitution too, what is the standard deviation of $\hat{p}$?

(c) Using the 68–95–99.7 rule, you would expect $\hat{p}$ to fall between what two percents about 95% of the time?

(d) Based on your result in part (a), do you think that the incoming students at your institution prefer this type of instruction more, less, or about the same as students nationally? Explain your answer.

5.75 Binge drinking. The Centers for Disease Control and Prevention finds that 28% of people aged 18 to 24 years binge drank. Those who binge drank averaged 9.3 drinks per episode and 4.2 episodes per month. The study took a sample of over 18,000 people aged 18 to 24 years, so the population proportion of people who binge drank is very close to $p = 0.28$.[23] The administration of your college surveys an SRS of 200 students and finds that 56 binge drink.

(a) What is the sample proportion of students at your college who binge drink?

(b) If, in fact, the proportion of all students on your campus who binge drink is the same as the national 28%, what is the probability that the proportion in an SRS of 200 students is as large or larger than the result of the administration's sample?

(c) A writer for the student paper says that the percent of students who binge brink is higher on your campus than nationally. Write a short letter to the editor explaining why the survey does not support this conclusion.

5.76 How large a sample is needed? The changing probabilities you found in Exercises 5.70 and 5.72 are due to the fact that the standard deviation of the sample proportion $\hat{p}$ gets smaller as the sample size n increases. If the population proportion is $p = 0.48$, how large a sample is needed to reduce the standard deviation of $\hat{p}$ to $\sigma_{\hat{p}} = 0.005$? (The 68–95–99.7 rule then says that about 95% of all samples will have $\hat{p}$ within 0.01 of the true p.)

5.77 A test for ESP. In a test for ESP (extrasensory perception), the experimenter looks at cards that are hidden from the subject. Each card contains either a star, a circle, a wave, or a square. As the experimenter looks at each of 20 cards in turn, the subject names the shape on the card.

(a) If a subject simply guesses the shape on each card, what is the probability of a successful guess on a single card? Because the cards are independent, the count of successes in 20 cards has a binomial distribution.

(b) What is the probability that a subject correctly guesses at least 10 of the 20 shapes?

(c) In many repetitions of this experiment with a subject who is guessing, how many cards will the subject guess correctly on the average? What is the standard deviation of the number of correct guesses?

(d) A standard ESP deck actually contains 25 cards. There are five different shapes, each of which appears on five cards. The subject knows that the deck has this makeup. Is a binomial model still appropriate for the count of correct guesses in one pass through this deck? If so, what are n and p? If not, why not?

5.78 Admitting students to college. A selective college would like to have an entering class of 1000 students.

Because not all students who are offered admission accept, the college admits more than 1000 students. Past experience shows that about 83% of the students admitted will accept. The college decides to admit 1200 students. Assuming that students make their decisions independently, the number who accept has the $B(1200, 0.83)$ distribution. If this number is less than 1000, the college will admit students from its waiting list.

(a) What are the mean and the standard deviation of the number X of students who accept?

(b) Use the Normal approximation to find the probability that at least 800 students accept.

(c) The college does not want more than 1000 students. What is the probability that more than 1000 will accept?

(d) If the college decides to decrease the number of admission offers to 1150, what is the probability that more than 1000 will accept?

5.79 Is the ESP result better than guessing? When the ESP study of Exercise 5.77 discovers a subject whose performance appears to be better than guessing, the study continues at greater length. The experimenter looks at many cards bearing one of five shapes (star, square, circle, wave, and cross) in an order determined by random numbers. The subject cannot see the experimenter as the experimenter looks at each card in turn, in order to avoid any possible nonverbal clues. The answers of a subject who does not have ESP should be independent observations, each with probability 1/5 of success. We record 900 attempts.

(a) What are the mean and the standard deviation of the count of successes?

(b) What are the mean and the standard deviation of the proportion of successes among the 900 attempts?

(c) What is the probability that a subject without ESP will be successful in at least 24% of 900 attempts?

(d) The researcher considers evidence of ESP to be a proportion of successes so large that there is only probability 0.01 that a subject could do this well or better by guessing. What proportion of successes must a subject have to meet this standard? (Example 1.45, on pages 65–66, shows how to do an inverse calculation for the Normal distribution that is similar to the type required here.)

5.80 Show that these facts are true. Use the definition of binomial coefficients to show that each of the following facts is true. Then restate each fact in words in terms of the number of ways that k successes can be distributed among n observations.

(a) $\binom{n}{n} = 1$ for any whole number $n \geq 1$.

(b) $\binom{n}{n-1} = n$ for any whole number $n \geq 1$.

(c) $\binom{n}{k} = \binom{n}{n-k}$ for any n and k with $k \leq n$.

5.81 English Premier League Goals. The total number of goals scored per soccer match in the English Premier League (EPL) often follows the Poisson distribution. In one recent season, the average number of goals scored per match (over 380 games played) was 2.768. Compute the following probabilities.

(a) What is the probability that three or more goals are scored in a game?

(b) What is the probability that a game will end in a 0–0 tie?

(c) Explain why you cannot compute the probability that a game will end in a 1–1 tie but can provide an upper bound on this probability.

5.82 Number of colony-forming units. In microbiology, colony-forming units (CFUs) are used to measure the number of microorganisms present in a sample. To determine the number of CFUs, the sample is prepared, spread uniformly on an agar plate, and then incubated at some suitable temperature. Suppose that the number of CFUs that appear after incubation follows a Poisson distribution with $\mu = 15$.

(a) If the area of the agar plate is 75 square centimeters (cm^2), what is the probability of observing fewer than 4 CFUs in a 25 cm^2 area of the plate?

(b) If you were to count the total number of CFUs in five plates, what is the probability you would observe more than 90 CFUs? Use the Poisson distribution to obtain this probability.

(c) Repeat the probability calculation in part (b), but now use the Normal approximation. How close is your answer to your answer in part (b)?

5.83 Metal fatigue. Metal fatigue refers to the gradual weakening and eventual failure of metal that undergoes cyclic loads. The wings of an aircraft, for example, are subject to cyclic loads when in the air, and cracks can form. It is thought that these cracks start at large particles found in the metal. Suppose that the number of particles large enough to initiate a crack follows a Poisson distribution with mean $\mu = 0.5$ per square centimeter (cm^2).

(a) What is the mean of the Poisson distribution if we consider a 100 cm^2 area?

(b) Using the Normal approximation, what is the probability that this section has more than 60 of these large particles?

CHAPTER 5 EXERCISES

5.84 The cost of Internet access. In Canada, households spent an average of $80.63 CDN monthly for high-speed broadband access.[24] Assume that the standard deviation is $27.32. If you ask an SRS of 500 Canadian households with high-speed broadband access how much they pay, what is the probability that the average amount will exceed $85?

5.85 Dust in coal mines. A laboratory weighs filters from a coal mine to measure the amount of dust in the mine atmosphere. Repeated measurements of the weight of dust on the same filter vary Normally with standard deviation $\sigma = 0.09$ milligram (mg) because the weighing is not perfectly precise. The dust on a particular filter actually weighs 137 mg.

(a) The laboratory reports the mean of three weighings of this filter. What is the distribution of this mean?

(b) What is the probability that the laboratory reports a weight of 140 mg or higher for this filter?

5.86 The effect of sample size on the standard deviation. Assume that the standard deviation in a very large population is 100.

(a) Calculate the standard deviation for the sample mean for samples of size 1, 4, 25, 100, 250, 500, 1000, and 5000.

(b) Graph your results with the sample size on the x axis and the standard deviation on the y axis.

(c) Summarize the relationship between the sample size and the standard deviation that your graph shows.

5.87 Marks per round in cricket. Cricket is a dart game that uses the numbers 15 to 20 and the bull's-eye. Each time you hit one of these regions, you score either 0, 1, 2 or 3 marks. Thus, in a round of three throws, a person can score 0 to 9 marks. Lex plans to play 20 games. Her distribution of marks per round is discrete and strongly skewed. A majority of her rounds result in 0, 1, or 2 marks and only a few are more than 4 marks. Assume that her mean is 2.07 marks per round with a standard deviation of 2.11.

(a) Her 20 games involve 140 rounds of three throws each. What are the mean and standard deviation of the average number of marks $\bar{x}$ in 140 rounds?

(b) Using the central limit theorem, what is the probability that she averages fewer than 2 marks per round?

(c) Do you think that the central limit theorem can be used in this setting? Explain your answer.

5.88 Common last names. The U.S. Census Bureau says that the 10 most common names in the United States are (in order) Smith, Johnson, Williams, Brown, Jones, Miller, Davis, Garcia, Rodriguez, and Wilson.[25] These names account for 4.9% of all U.S. residents. Out of curiosity, you look at the authors of the textbooks for your current courses. There are 12 authors in all. Would you be surprised if none of the names of these authors were among the 10 most common? Give a probability to support your answer and explain the reasoning behind your calculation.

5.89 Benford's law. It is a striking fact that the first digits of numbers in legitimate records often follow a distribution known as Benford's law. Here it is:

First digit	1	2	3	4	5	6	7	8	9
Proportion	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

Fake records usually have fewer first digits 1, 2, and 3. What is the approximate probability, if Benford's law holds, that among 1000 randomly chosen invoices there are 575 or fewer in amounts with first digit 1, 2, or 3?

5.90 Genetics of peas. According to genetic theory, the blossom color in the second generation of a certain cross of sweet peas should be red or white in a 3:1 ratio. That is, each plant has probability 3/4 of having red blossoms, and the blossom colors of separate plants are independent.

(a) What is the probability that exactly 8 out of 10 of these plants have red blossoms?

(b) What is the mean number of red-blossomed plants when 130 plants of this type are grown from seeds?

(c) What is the probability of obtaining at least 90 red-blossomed plants when 130 plants are grown from seeds?

5.91 Leaking gas tanks. Leakage from underground gasoline tanks at service stations can damage the environment. It is estimated that 25% of these tanks leak. You examine 15 tanks chosen at random, independently of each other.

(a) What is the mean number of leaking tanks in such samples of 15?

(b) What is the probability that 10 or more of the 15 tanks leak?

(c) Now you do a larger study, examining a random sample of 2000 tanks nationally. What is the probability that at least 540 of these tanks are leaking?

5.92 A roulette payoff. A $1 bet on a single number on a casino's roulette wheel pays $35 if the ball ends up in the number slot you choose. Here is the distribution of the payoff X:

Payoff X	$0	$35
Probability	0.974	0.026

Each spin of the roulette wheel is independent of other spins.

(a) What are the mean and standard deviation of X?

(b) Sam comes to the casino weekly and bets on 10 spins of the roulette wheel. What does the law of large numbers say about the average payoff Sam receives from his bets each visit?

(c) What does the central limit theorem say about the distribution of Sam's average payoff after betting on 520 spins in a year?

(d) Sam comes out ahead for the year if his average payoff is greater than \$1 (the amount he bet on each spin). What is the probability that Sam ends the year ahead? The true probability is 0.396. Does using the central limit theorem provide a reasonable approximation?

5.93 A roulette payoff revisited. Refer to the previous exercise. In part (d), the central limit theorem was used to approximate the probability that Sam ends the year ahead. The estimate was about 0.10 too large. Let's see if we can get closer using the Normal approximation to the binomial with the continuity correction.

(a) If Sam plans to bet on 520 roulette spins, he needs to win at least \$520 to break even. If each win gives him \$35, what is the minimum number of wins m he must have?

(b) Given $p = 1/38 = 0.026$, what are the mean and standard deviation of X, the number of wins in 520 roulette spins?

(c) Use the information in the previous two parts to compute $P(X \geq m)$ with the continuity correction. Does your answer get closer to the exact probability 0.396?

5.94 Learning a foreign language. Does delaying oral practice hinder learning a foreign language? Researchers randomly assigned 25 beginning students of Russian to begin speaking practice immediately and another 25 to delay speaking for four weeks. At the end of the semester both groups took a standard test of comprehension of spoken Russian. Suppose that in the population of all beginning students, the test scores for early speaking vary according to the $N(32, 6)$ distribution and scores for delayed speaking have the $N(29, 5)$ distribution.

(a) What is the sampling distribution of the mean score $\bar{x}$ in the early-speaking group in many repetitions of the experiment? What is the sampling distribution of the mean score $\bar{y}$ in the delayed-speaking group?

(b) If the experiment were repeated many times, what would be the sampling distribution of the difference $\bar{y} - \bar{x}$ between the mean scores in the two groups?

(c) What is the probability that the experiment will find (misleadingly) that the mean score for delayed speaking is at least as large as that for early speaking?

5.95 Summer employment of college students. Suppose (as is roughly true) that 88% of college men and 82% of college women were employed last summer. A sample survey interviews SRSs of 400 college men and 400 college women. The two samples are of course independent.

(a) What is the approximate distribution of the proportion $\hat{p}_F$ of women who worked last summer? What is the approximate distribution of the proportion $\hat{p}_M$ of men who worked?

(b) The survey wants to compare men and women. What is the approximate distribution of the difference in the proportions who worked, $\hat{p}_M - \hat{p}_F$? Explain the reasoning behind your answer.

(c) What is the probability that in the sample a higher proportion of women than men worked last summer?

5.96 Income of working couples. A study of working couples measures the income X of the husband and the income Y of the wife in a large number of couples in which both partners are employed. Suppose that you knew the means μ_X and μ_Y and the variances σ_X^2 and σ_Y^2 of both variables in the population.

(a) Is it reasonable to take the mean of the total income $X + Y$ to be $\mu_X + \mu_Y$? Explain your answer.

(b) Is it reasonable to take the variance of the total income to be $\sigma_X^2 + \sigma_Y^2$? Explain your answer.

5.97 A random walk. A particle moves along the line in a random walk. That is, the particle starts at the origin (position 0) and moves either right or left in independent steps of length 1. If the particle moves to the right with probability 0.6, its movement at the ith step is a random variable X_i with distribution

$$P(X_i = 1) = 0.6$$
$$P(X_i = -1) = 0.4$$

The position of the particle after k steps is the sum of these random movements,

$$Y = X_1 + X_2 + \cdots + X_k$$

Use the central limit theorem to find the approximate probability that the position of the particle after 500 steps is at least 200 to the right.

5.98 A lottery payoff. A \$1 bet in a state lottery's Pick 3 game pays \$500 if the three-digit number you choose exactly matches the winning number, which is drawn at random. Here is the distribution of the payoff X:

Payoff X	$0	$500
Probability	0.999	0.001

Each day's drawing is independent of other drawings.

(a) Joe buys a Pick 3 ticket twice a week. The number of times he wins follows a B(104, 0.001) distribution. Using the Poisson approximation to the binomial, what is the probability that he wins at least once?

(b) The exact binomial probability is 0.0988. How accurate is the Poisson approximation here?

(c) If Joe pays $5 a ticket, he needs to win at least twice a year to come out ahead. Using the Poisson approximation, what is the probability that Joe comes out ahead?

5.99 Poisson distribution? Suppose you find in your spam folder an average of two spam emails every 10 minutes. Furthermore, you find that the rate of spam mail from midnight to 6 A.M. is twice the rate during other parts of the day. Explain whether or not the Poisson distribution is an appropriate model for the spam process.

5.100 Tossing a die. You are tossing a balanced die that has probability 1/6 of coming up 1 on each toss. Tosses are independent. We are interested in how long we must wait to get the first 1.

(a) The probability of a 1 on the first toss is 1/6. What is the probability that the first toss is not a 1 and the second toss is a 1?

(b) What is the probability that the first two tosses are not 1s and the third toss is a 1? This is the probability that the first 1 occurs on the third toss.

(c) Now you see the pattern. What is the probability that the first 1 occurs on the fourth toss? On the fifth toss?

5.101 The geometric distribution. Generalize your work in Exercise 5.100 (page 337). You have independent trials, each resulting in a success or a failure. The probability of a success is p on each trial. The binomial distribution describes the count of successes in a fixed number of trials. Now the number of trials is not fixed; instead, continue until you get a success. The random variable Y is the number of the trial on which the first success occurs. What are the possible values of Y? What is the probability $P(Y = k)$ for any of these values? (*Comment:* The distribution of the number of trials to the first success is called a **geometric distribution**.)

5.102 Wi-fi interruptions. Suppose that the number of wi-fi interruptions on your home network follows the Poisson distribution with an average of 0.9 wi-fi interruptions per day.

(a) Show that the probability of no interruptions on a given day is 0.4066.

(b) Treating each day as a trial in a binomial setting, use the binomial formula to compute the probability of no interruptions in a week.

(c) Now, instead of using the binomial model, let's use the Poisson distribution exclusively. What is the mean number of wi-fi interruptions during a week?

(d) Based on the Poisson mean of part (c), use the Poisson distribution to compute the probability of no interruptions in a week. Confirm that this probability is the same as found part (b). Explain in words why the two ways of computing no interruptions in a week give the same result.

(e) Explain why using the binomial distribution to compute the probability that only one day in the week will not be interruption free would not give the same probability had we used the Poisson distribution to compute that only one interruption occurs during the week.