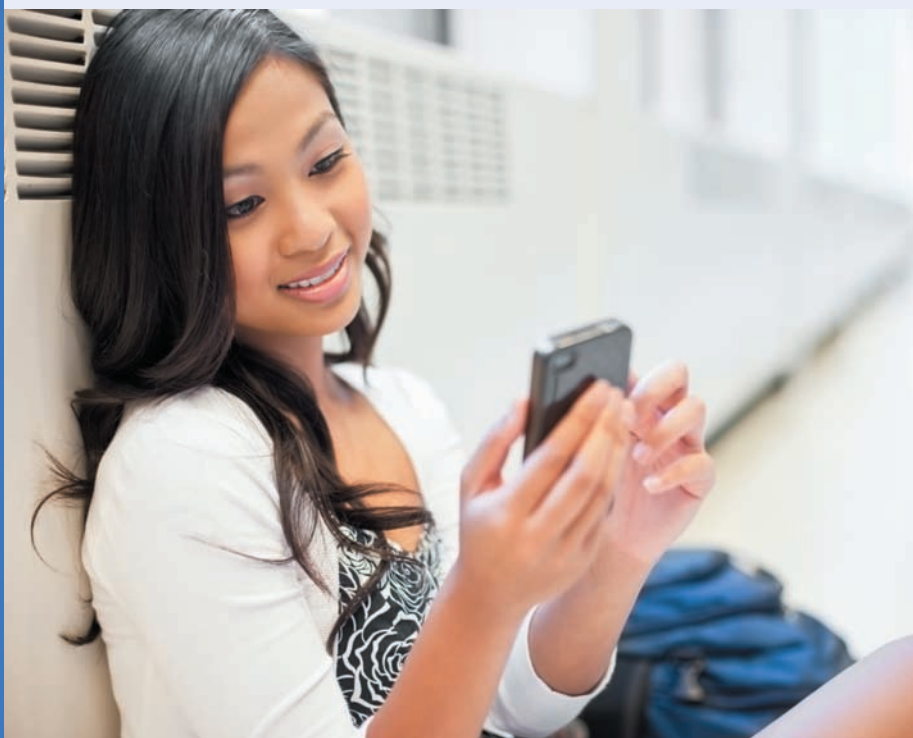


# Chapter

# 7

<b>Introduction</b>	422
<b>Section 7.1</b>	424
What Is a Sampling Distribution?	
<b>Section 7.2</b>	440
Sample Proportions	
<b>Section 7.3</b>	450
Sample Means	
<b>Free Response AP® Problem, Yay!</b>	464
<b>Chapter 7 Review</b>	465
<b>Chapter 7 Review Exercises</b>	466
<b>Chapter 7 AP® Statistics Practice Test</b>	468
<b>Cumulative AP® Practice Test 2</b>	470





# Sampling Distributions



## case study

### Building Better Batteries

Everyone wants to have the latest technological gadget. That's why iPods, digital cameras, smartphones, Game Boys, and the Wii have sold millions of units. These devices require lots of power and can drain batteries quickly. Battery manufacturers are constantly searching for ways to build longer-lasting batteries.

A particular manufacturer produces AA batteries that are designed to last an average of 17 hours with a standard deviation of 0.8 hours. Quality control inspectors select a random sample of 50 batteries during each hour of production, and they then drain them under conditions that mimic normal use. Here are the lifetimes (in hours) of the batteries from one such sample:

16.73	15.60	16.31	17.57	16.14	17.28	16.67	17.28	17.27	17.50	15.46	16.50	16.19
15.59	17.54	16.46	15.63	16.82	17.16	16.62	16.71	16.69	17.98	16.36	17.80	16.61
15.99	15.64	17.20	17.24	16.68	16.55	17.48	15.58	17.61	15.98	16.99	16.93	16.01
17.54	17.41	16.91	16.60	16.78	15.75	17.31	16.50	16.72	17.55	16.46		

Do these data suggest that the production process is working properly? Is it safe for plant managers to send out all the batteries produced in this hour for sale? In this chapter, you will develop the tools you need to help answer questions like this.

## Introduction

The battery manufacturer in the Case Study could find the true mean lifetime  $\mu$  of all the batteries produced in an hour. Quality control inspectors would simply measure the lifetime of each battery (by draining it) and then calculate the average. With this method, the company would know the truth about the population mean  $\mu$ , but it would have no batteries left to sell!

Instead of taking a census, the manufacturer collects data from a random sample of 50 batteries produced that hour. The company's goal is to use the sample mean lifetime  $\bar{x}$  to estimate the unknown population mean  $\mu$ . This is an example of *statistical inference*: we use information from a sample to draw conclusions about a larger population.

To make such an inference, we need to know how close the sample mean  $\bar{x}$  is likely to be to the population mean  $\mu$ . After all, different random samples of 50 batteries from the same hour of production would yield different values of  $\bar{x}$ . How can we describe this *sampling distribution* of possible  $\bar{x}$ -values? We can think of  $\bar{x}$  as a random variable because it takes numerical values that describe the outcomes of the random sampling process. As a result, we can examine its probability distribution using what we learned in Chapter 6.

This same reasoning applies to other types of inference settings. Here are a few examples.

- Each month, the Current Population Survey (CPS) interviews a random sample of individuals in about 60,000 U.S. households. The CPS uses the proportion of unemployed people in the sample  $\hat{p}$  to estimate the national unemployment rate  $p$ .
- Tom is cooking a large turkey breast for a holiday meal. He wants to be sure that the turkey is safe to eat, which requires a minimum internal temperature of 165°F. Tom uses a thermometer to measure the temperature of the turkey meat at four randomly chosen points. If the minimum reading in the sample is 170°F, can Tom safely serve the turkey?
- How much do gasoline prices vary in a large city? To find out, a reporter records the price per gallon of regular unleaded gasoline at a random sample of 10 gas stations in the city on the same day. The range (maximum – minimum) of the prices in the sample is 25 cents. What can the reporter say about the range of gas prices at all the city's stations?

The following Activity gives you a chance to estimate an unknown population value based on data from a random sample.

### ACTIVITY | The German Tank Problem

#### MATERIALS:

Tags or pieces of cardstock numbered 1 to  $N$ ; small brown paper bag; index card and prelabeled graph grid for each team; prizes for the winners

During World War II, the Allies captured several German tanks. Each tank had a serial number on it. Allied commanders wanted to know how many tanks the Germans had so that they could allocate their forces appropriately. They sent the serial numbers of the captured tanks to a group of mathematicians in Washington, D.C., and asked for an estimate of the total number of German tanks  $N$ . In this Activity, you and your teammates will play the role of the mathematicians.



More recently, people used the mathematicians' method from the German tank problem to estimate the number of iPhones produced.

1. Your teacher will create tags numbered 1, 2, 3, . . . ,  $N$  to represent the German tanks and place them in a bag. The class will be divided into teams of three or four students.
2. The teacher will mix the tags well and ask four students to draw one tag each from the bag. Each selected tag represents the serial number of a captured German tank. All four numbers should be written on the board for everyone to see. Return the tags to the bag.
3. Each team will have 15 minutes to come up with a statistical formula for estimating the total number of tanks  $N$  in the bag. You should have time to try several ideas. When you are satisfied with your method, calculate your estimate of  $N$ . Write your team members' names, your formula, and your estimate on the index card provided.
4. When time is called, each team must give its index card to your teacher. The teacher will make a chart on the board showing the formulas and estimates. Each team will have one minute to explain why it chose the formula it did.
5. The teacher will reveal the actual number of tanks. Which team came closest to the correct answer?
6. What if the Allies had captured four other German tanks? Which team's formula would consistently produce the best estimate? Students should help choose nine more simple random samples of four tanks from the bag. After each sample is taken, the four serial numbers chosen should be written on the board, and the tags should be returned to the bag and mixed thoroughly.
7. Each team should use its formula to estimate the total number of tanks  $N$  for each of the nine new samples. The team should then make a dotplot of its 10 estimates.
8. Compare the teams' dotplots. As a class, decide which team used the best method for estimating the number of tanks.

Sampling distributions are the key to inference when data are produced by random sampling. Because the results of random samples include an element of chance, we can't guarantee that our inferences are correct. What we can guarantee is that our methods usually give correct answers. The reasoning of statistical inference rests on asking, "How often would this method give a correct answer if I used it very many times?" If our data come from random sampling, the laws of probability answer the question "What would happen if we did this many times?"

Section 7.1 presents the basic ideas of sampling distributions. The most common applications of statistical inference involve proportions and means. Section 7.2 focuses on sampling distributions of sample proportions. Section 7.3 investigates sampling distributions of sample means.

## 7.1 What Is a Sampling Distribution?

### WHAT YOU WILL LEARN

By the end of the section, you should be able to:

- Distinguish between a parameter and a statistic.
- Use the sampling distribution of a statistic to evaluate a claim about a parameter.
- Distinguish among the distribution of a population, the distribution of a sample, and the sampling distribution of a statistic.
- Determine whether or not a statistic is an unbiased estimator of a population parameter.
- Describe the relationship between sample size and the variability of a statistic.

What is the average income of American households? Each March, the government's Current Population Survey (CPS) asks detailed questions about income. The random sample of about 60,000 households contacted in March 2012 had a mean "total money income" of \$69,677 in 2011.<sup>1</sup> (The median income was lower, of course, at \$50,054.) That \$69,677 describes the sample, but we use it to estimate the mean income of all households.

### Parameters and Statistics

As we begin to use sample data to draw conclusions about a wider population, we must be clear about whether a number describes a sample or a population. For the sample of households contacted by the CPS, the mean income was  $\bar{x} = \$69,677$ . The number \$69,677 is a **statistic** because it describes this one CPS sample. The population that the poll wants to draw conclusions about is all 121 million U.S. households. In this case, the **parameter** of interest is the mean income  $\mu$  of all these households. We don't know the value of this parameter.

#### DEFINITION: Parameter, statistic

A **parameter** is a number that describes some characteristic of the population.

A **statistic** is a number that describes some characteristic of a sample.

The value of a parameter is usually not known because we cannot examine the entire population. The value of a statistic can be computed directly from the sample data. We often use a statistic to estimate an unknown parameter.

Remember **s** and **p**: statistics come from samples, and parameters come from populations. As long as we were doing data analysis, the distinction between population and sample rarely came up. Now, however, it is essential. The notation we use should reflect this distinction. For instance, we write  $\mu$  (the Greek letter mu) for the population mean and  $\bar{x}$  for the sample mean. We use  $p$  to represent a population proportion. The sample proportion  $\hat{p}$  is used to estimate the unknown parameter  $p$ .

It is common practice to use Greek letters for parameters and Roman letters for statistics. In that case, the population proportion would be  $\pi$  (pi, the Greek letter for "p") and the sample proportion would be  $p$ . We'll stick with the notation that's used on the AP<sup>®</sup> exam, however.



## EXAMPLE

## From Ghosts to Cold Cabins

### Parameters and statistics

**PROBLEM:** Identify the population, the parameter, the sample, and the statistic in each of the following settings.

(a) The Gallup Poll asked a random sample of 515 U.S. adults whether or not they believe in ghosts. Of the respondents, 160 said “Yes.”<sup>2</sup>

(b) During the winter months, the temperatures outside the Starneses’ cabin in Colorado can stay well below freezing (32°F, or 0°C) for weeks at a time. To prevent the pipes from freezing, Mrs. Starnes sets the thermostat at 50°F. She wants to know how low the temperature actually gets in the cabin. A digital thermometer records the indoor temperature at 20 randomly chosen times during a given day. The minimum reading is 38°F.

### SOLUTION:

(a) The population is all U.S. adults, and the parameter of interest is  $p$ , the proportion of all U.S. adults who believe in ghosts. The sample is the 515 people who were interviewed in this Gallup Poll.

The statistic is  $\hat{p} = \frac{160}{515} = 0.31$ , the proportion of the sample who say they believe in ghosts.

(b) The population is all times during the day in question; the parameter of interest is the true minimum temperature in the cabin that day. The sample consists of the 20 temperature readings at randomly selected times. The statistic is the sample minimum, 38°F.

**For Practice** Try Exercise **1**



## CHECK YOUR UNDERSTANDING

Each boldface number in Questions 1 and 2 is the value of either a **parameter** or a **statistic**. In each case, state which it is and use appropriate notation to describe the number.

- On Tuesday, the bottles of Arizona Iced Tea filled in a plant were supposed to contain an average of **20** ounces of iced tea. Quality control inspectors sampled 50 bottles at random from the day’s production. These bottles contained an average of **19.6** ounces of iced tea.
- On a New York-to-Denver flight, 8% of the 125 passengers were selected for random security screening before boarding. According to the Transportation Security Administration, **10%** of passengers at this airport are chosen for random screening.

## Sampling Variability

How can  $\bar{x}$ , based on a sample of only a few thousand of the 121 million American households, be an accurate estimate of  $\mu$ ? After all, a second random sample taken at the same time would choose different households and likely produce a different value of  $\bar{x}$ . This basic fact is called **sampling variability**: the value of a statistic varies in repeated random sampling.





To make sense of sampling variability, we ask, “What would happen if we took many samples?” Here’s how to answer that question:

- Take a large number of samples from the same population.
- Calculate the statistic (like the sample mean  $\bar{x}$  or sample proportion  $\hat{p}$ ) for each sample.
- Make a graph of the values of the statistic.
- Examine the distribution displayed in the graph for shape, center, and spread, as well as outliers or other unusual features.

The following Activity gives you a chance to see sampling variability in action.

## ACTIVITY | Reaching for Chips

### MATERIALS:

200 colored chips, including 100 of the same color; large bag or other container



Before class, your teacher will prepare a population of 200 colored chips, with 100 having the same color (say, red). The parameter is the actual proportion  $p$  of red chips in the population:  $p = 0.50$ . In this Activity, you will investigate sampling variability by taking repeated random samples of size 20 from the population.

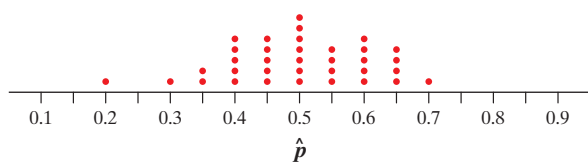
1. After your teacher has mixed the chips thoroughly, each student in the class should take a sample of 20 chips and note the sample proportion  $\hat{p}$  of red chips. When finished, the student should return all the chips to the bag, stir them up, and pass the bag to the next student.

*Note:* If your class has fewer than 25 students, have some students take two samples.

2. Each student should record the  $\hat{p}$ -value in a chart on the board and plot this value on a class dotplot. Label the graph scale from 0.10 to 0.90 with tick marks spaced 0.05 units apart.

3. Describe what you see: shape, center, spread, and any outliers or other unusual features.

When Mr. Caldwell’s class did the “Reaching for Chips” Activity, his 35 students produced the graph shown in Figure 7.1. Here’s what the class said about its distribution of  $\hat{p}$ -values.



**FIGURE 7.1** Dotplot of sample proportions obtained by the 35 students in Mr. Caldwell’s class.

**Shape:** The graph is roughly symmetric with a single peak at 0.5.

**Center:** The mean of our sample proportions is 0.499. This is the balance point of the distribution.

**Spread:** The standard deviation of our sample proportions is 0.112. The values of  $\hat{p}$  are typically about 0.112 away from the mean.

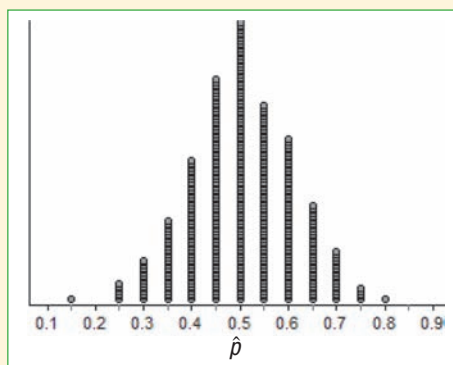
**Outliers:** There are no obvious outliers or other unusual features.

Of course, the class only took 35 different simple random samples of 20 chips. There are many, many possible SRSs of size 20 from a population of size 200 (about  $1.6 \cdot 10^{27}$ , actually). If we took every one of those possible samples, calculated the value of  $\hat{p}$  for each, and graphed all those  $\hat{p}$ -values, then we’d have a **sampling distribution**.

**DEFINITION: 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.

It's usually too difficult to take all possible samples of size  $n$  to obtain the sampling distribution of a statistic. Instead, we can use simulation to imitate the process of taking many, many samples and create an approximate sampling distribution.

**EXAMPLE****Reaching for Chips***Simulating a sampling distribution*

**FIGURE 7.2** Dotplot of the sample proportion  $\hat{p}$  of red chips in 500 simulated SRSs, created by Fathom software.

We used Fathom software to simulate choosing 500 SRSs of size  $n = 20$  from a population of 200 chips, 100 red and 100 blue. Figure 7.2 is a dotplot of the values of  $\hat{p}$ , the sample proportion of red chips, from these 500 samples.

**PROBLEM:**

- There is one dot on the graph at 0.15. Explain what this value represents.
- Describe the distribution. Are there any obvious outliers?
- Would it be surprising to get a sample proportion of 0.85 or higher in an SRS of size 20 when  $p = 0.5$ ? Justify your answer.
- Suppose your teacher prepares a bag with 200 chips and claims that half of them are red. A classmate takes an SRS of 20 chips; 17 of them are red. What would you conclude about your teacher's claim? Explain.

**SOLUTION:**

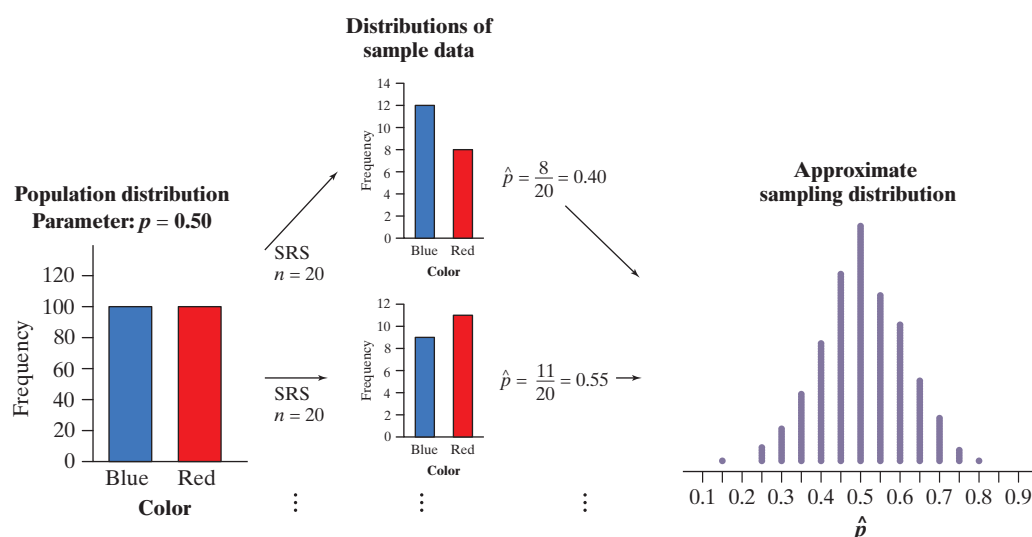
- In one SRS of 20 chips, there were 3 red chips. So  $\hat{p} = 3/20 = 0.15$  for this sample.
- Shape:* Symmetric, unimodal, and somewhat bell-shaped. *Center:* Around 0.5. *Spread:* The values of  $\hat{p}$  fall mostly between 0.25 and 0.75. *Outliers:* One sample with  $\hat{p} = 0.15$  stands out.
- It is very unlikely to obtain an SRS of 20 chips in which  $\hat{p} = 0.85$  from a population in which  $p = 0.5$ . A value of  $\hat{p}$  this large or larger never occurred in 500 simulated samples.
- This student's result gives strong evidence against the teacher's claim. As noted in part (c), it is very unlikely to get a sample proportion of 0.85 or higher when  $p = 0.5$ .

**For Practice** Try Exercise **9**

Strictly speaking, the sampling distribution is the ideal pattern that would emerge if we looked at *all* possible samples of size 20 from our population of chips. A distribution obtained from simulating a smaller number of random samples, like the 500 values of  $\hat{p}$  in Figure 7.2, is only an approximation to the sampling distribution. One of the uses of probability theory in statistics is to obtain sampling distributions without simulation. We'll get to the theory later. The interpretation of a sampling distribution is the same, however, whether we obtain it by simulation or by the mathematics of probability.



Figure 7.3 illustrates the process of choosing many random samples of 20 chips and finding the sample proportion of red chips  $\hat{p}$  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 first sample has  $\hat{p} = 0.40$ . The second sample is a different group of chips, with  $\hat{p} = 0.55$ , and so on. The dotplot at the right of the figure shows the distribution of the values of  $\hat{p}$  from 500 separate SRSs of size 20. This dotplot displays the approximate sampling distribution of the statistic  $\hat{p}$ .



**FIGURE 7.3** The idea of a sampling distribution: take many samples from the same population, collect the  $\hat{p}$ 's from all the samples, and display the distribution of the  $\hat{p}$ 's. The dotplot shows the results of 500 samples.

As Figure 7.3 shows, there are three distinct distributions involved when we sample repeatedly and measure a variable of interest. The **population distribution** gives the values of the variable for all individuals in the population. In this case, the individuals are the 200 chips and the variable we're recording is color. Our parameter of interest is the proportion of red chips in the population,  $p = 0.50$ . Each random sample that we take consists of 20 chips.

The **distribution of sample data** shows the values of the variable “color” for the individuals in the sample. For each sample, we record a value for the statistic  $\hat{p}$ , the sample proportion of red chips. Finally, we collect the values of  $\hat{p}$  from all possible samples of the same size and display them in the *sampling distribution*.

*Be careful: The population distribution and the distribution of sample data describe individuals. A sampling distribution describes how a statistic varies in many samples from the population.*



**AP® EXAM TIP** Terminology matters. Don't say “sample distribution” when you mean sampling distribution. You will lose credit on free response questions for misusing statistical terms.



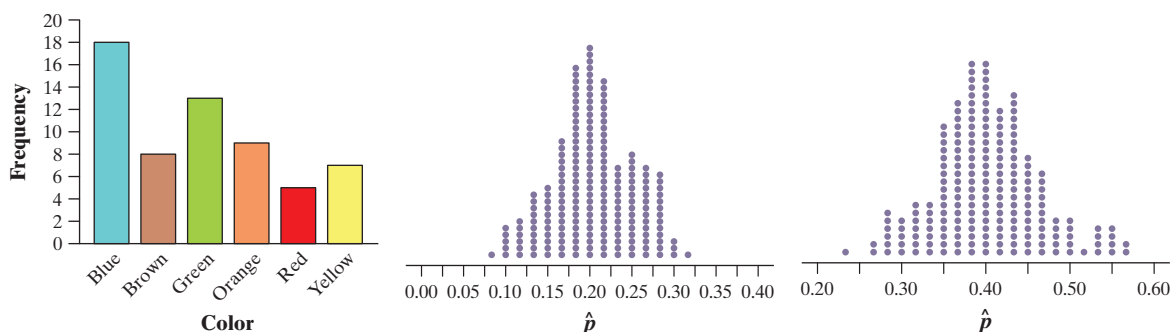
## CHECK YOUR UNDERSTANDING

Mars, Incorporated, says that the mix of colors in its M&M'S® Milk Chocolate Candies is 24% blue, 20% orange, 16% green, 14% yellow, 13% red, and 13% brown. Assume that the company's claim is true. We want to examine the proportion of orange M&M'S in repeated random samples of 50 candies.

1. Graph the population distribution. Identify the individuals, the variable, and the parameter of interest.

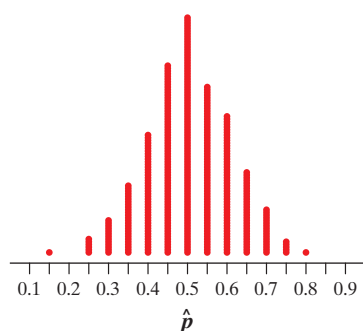


- Imagine taking an SRS of 50 M&M'S. Make a graph showing a possible distribution of the sample data. Give the value of the appropriate statistic for this sample.
- Which of the graphs that follow could be the approximate sampling distribution of the statistic? Explain your choice.



## Describing Sampling Distributions

The fact that statistics from random samples have definite sampling distributions allows us to answer the question “How trustworthy is a statistic as an estimate of a parameter?” To get a complete answer, we consider the shape, center, and spread of the sampling distribution. For reasons that will be clear later, we’ll save shape for last.



**Center: Biased and unbiased estimators** Let’s return to the familiar chips example. How well does the sample proportion of red chips estimate the population proportion of red chips,  $p = 0.5$ ? The dotplot in the margin shows the approximate sampling distribution of  $\hat{p}$  once again. We noted earlier that the center of this distribution is very close to 0.5, the parameter value. In fact, if we took all possible samples of 20 chips from the population, calculated  $\hat{p}$  for each sample, and then found the mean of all those  $\hat{p}$ -values, we’d get *exactly* 0.5. For this reason, we say that  $\hat{p}$  is an **unbiased estimator** of  $p$ .

### DEFINITION: Unbiased estimator

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

If we take many samples, the value of an unbiased estimator will sometimes exceed the value of the parameter and sometimes be less. However, because the sampling distribution of the statistic is centered at the true value, we will not consistently overestimate or underestimate the parameter. This is consistent with our definition of bias from Chapter 4.

We will confirm in Section 7.2 that the sample proportion  $\hat{p}$  is an unbiased estimator of the population proportion  $p$ . This is a very helpful result if we’re dealing with a categorical variable (like color). With quantitative variables, we might be interested in estimating the population mean, median, minimum, maximum,  $Q_1$ ,  $Q_3$ , variance, standard deviation,  $IQR$ , or range. Which (if any) of these are unbiased estimators? The following Activity should shed some light on this question.

## ACTIVITY | Sampling heights

### MATERIALS:

Small piece of cardstock for each student; bag



In this Activity, you will use a population of quantitative data to investigate whether a given statistic is an unbiased estimator of its corresponding population parameter.

1. Each student should write his or her height (in inches) neatly on a small piece of cardstock and then place it in the bag.
2. After your teacher has mixed the cards thoroughly, each student in the class should take a sample of four cards and record the heights of the four chosen students. When finished, the student should return the cards to the bag, mix them up, and pass the bag to the next student.

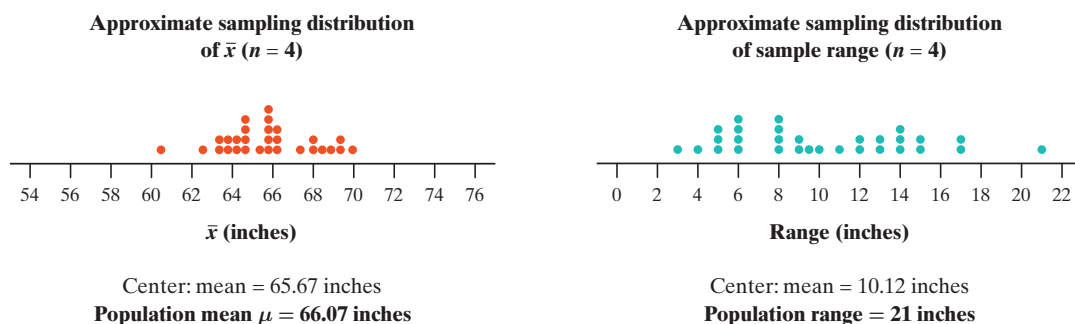
*Note:* If your class has fewer than 25 students, have some students take two samples.

3. For your SRS of four students from the class, calculate the sample mean  $\bar{x}$  and the sample range (maximum – minimum) of the heights. Then go to the board and record the heights of the four students in your sample, the sample mean  $\bar{x}$ , and the sample range in a chart like the one below.

Height (in.)	Sample mean ( $\bar{x}$ )	Sample range (max – min)
62, 75, 68, 63	67	75 – 62 = 13

4. Plot the values of your sample mean and sample range on the two class dotplots drawn by your teacher.
5. Once everyone has finished, find the population mean  $\mu$  and the population range.
6. Based on your approximate sampling distributions of  $\bar{x}$  and the sample range, which statistic appears to be an unbiased estimator? Which appears to be a *biased estimator*?

When Mrs. Washington's class did the "Sampling Heights" Activity, they produced the graphs shown in Figure 7.4. Her students concluded that the sample mean  $\bar{x}$  is probably an unbiased estimator of the population mean  $\mu$ . Their reason: the center of the approximate sampling distribution of  $\bar{x}$ , 65.67 inches, is close to the population mean of 66.07 inches. On the other hand, Mrs. Washington's

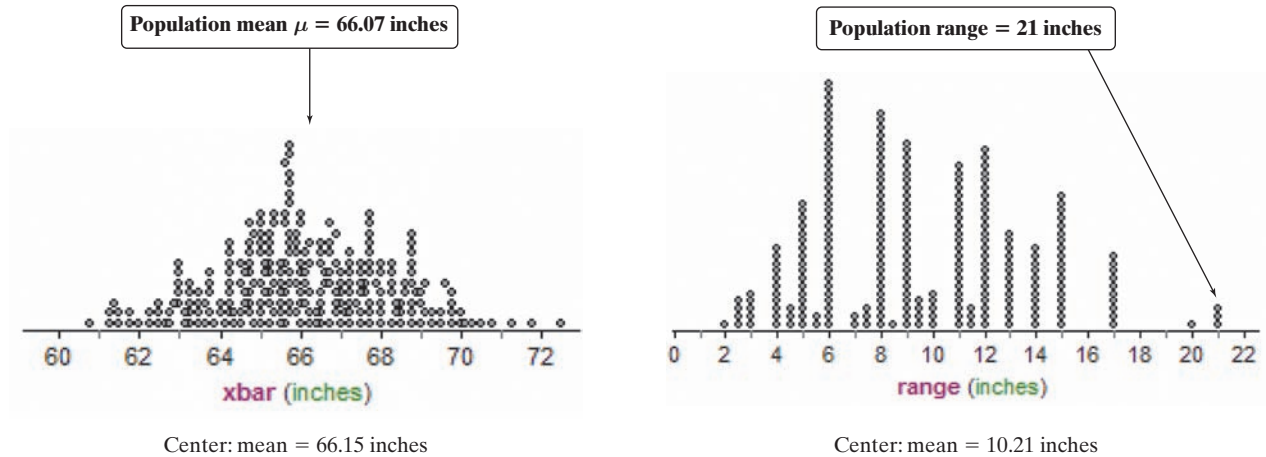


**FIGURE 7.4** Results from Mrs. Washington's class. The sample mean appears to be an unbiased estimator. The sample range appears to be a biased estimator.



students decided that the sample range is a **biased estimator** of the population range. Why? Because the center of the sampling distribution for this statistic was 10.12 inches, much less than the corresponding parameter value of 21 inches.

To confirm the class's conclusions, we used Fathom software to simulate taking 250 SRSs of  $n = 4$  students. For each sample, we plotted the mean height  $\bar{x}$  and the range of the heights. Figure 7.5 shows the approximate sampling distributions for these two statistics. It looks like the class was right:  $\bar{x}$  is an unbiased estimator of  $\mu$ , but the sample range is clearly a biased estimator. The range of the sample heights tends to be much lower, on average, than the population range.



**FIGURE 7.5** Results from a Fathom simulation of 250 SRSs of size  $n = 4$  from the students in Mrs. Washington's class. The sample mean is an unbiased estimator. The sample range is a biased estimator.

### THINK ABOUT IT

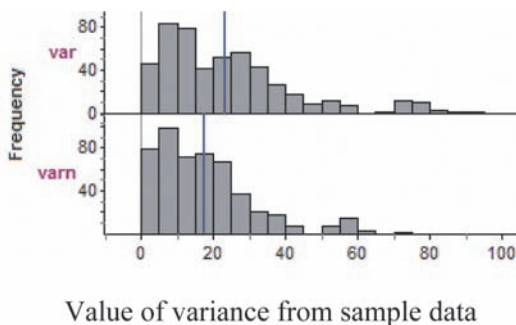
**Why do we divide by  $n - 1$  when calculating the sample variance?** In Chapter 1, we introduced the sample variance  $s_x^2$  as a measure of spread for a set of quantitative data. The idea of  $s_x^2$  is simple: it's a number that describes the "average" squared deviation of the values in the sample from their mean  $\bar{x}$ . It probably surprised you when we computed this average by dividing by  $n - 1$

instead of  $n$ . Now we're ready to tell you why we defined  $s_x^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$ .

In an inference setting involving a quantitative variable, we might be interested in estimating the variance  $\sigma^2$  of the population distribution. The most logical choice for our estimator is the sample variance  $s_x^2$ . We used Fathom software to take 500 SRSs of size  $n = 4$  from the population distribution of heights in Mrs. Washington's class. Note that the population variance is  $\sigma^2 = 22.19$ . For each sample, we recorded the value of two statistics:

$$\text{var} = s_x^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 \text{ (the sample variance)}$$

$$\text{varn} = \frac{1}{n} \sum (x_i - \bar{x})^2$$



**FIGURE 7.6** Results from a Fathom simulation of 500 SRSs of size  $n = 4$  from the population distribution of heights in Mrs. Washington's class. The sample variance  $s_x^2$  (labeled "var" in the figure) is an unbiased estimator. The "varn" statistic (dividing by  $n$  instead of  $n - 1$ ) is a biased estimator.

Figure 7.6 shows the approximate sampling distributions of these two statistics. We used histograms to show the overall pattern more clearly. The vertical lines mark the means of these two distributions.

We can see that “var” is a *biased* estimator of the population variance. The mean of its sampling distribution (marked with a blue line segment) is clearly less than the value of the population parameter, 22.19. However, the statistic “var” (otherwise known as the sample variance  $s_x^2$ ) is an unbiased estimator. Its values are centered at 22.19. That’s why we divide by  $n - 1$  and not  $n$  when calculating the sample variance: to get an unbiased estimator of the population variance.

**Spread: Low variability is better!** To get a trustworthy estimate of an unknown population parameter, start by using a statistic that’s an unbiased estimator. This ensures that you won’t consistently overestimate or underestimate the parameter. Unfortunately, using an unbiased estimator doesn’t guarantee that the value of your statistic will be close to the actual parameter value. The following example illustrates what we mean.

## EXAMPLE

### Who Watches *Survivor*?

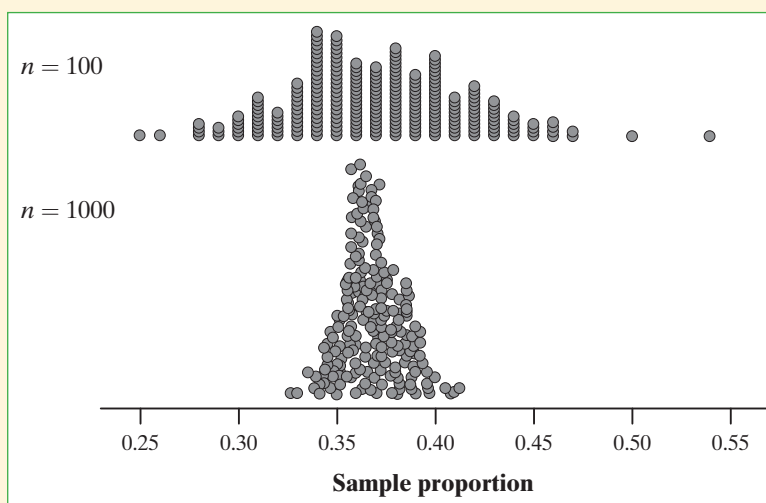
#### Why sample size matters

Television executives and companies who advertise on TV are interested in how many viewers watch particular shows. According to Nielsen ratings, *Survivor* was one of the most-watched television shows in the United States during every week that it aired. Suppose that the true proportion of U.S. adults who have watched *Survivor* is  $p = 0.37$ .

The top dotplot in Figure 7.7 shows the results of drawing 400 SRSs of size  $n = 100$  from a population with  $p = 0.37$ . We see that a sample of 100 people often gave a  $\hat{p}$  quite far from the population parameter. That is why a Gallup Poll asked not 100, but 1000 people whether they had watched *Survivor*. Let’s repeat our simulation, this time taking 400 SRSs of size  $n = 1000$  from a population with proportion  $p = 0.37$  who have watched *Survivor*. The bottom dotplot in Figure 7.7 displays the distribution of the 400 values of  $\hat{p}$  from these new samples. Both graphs are drawn on the same horizontal scale to make comparison easy.



**FIGURE 7.7** The approximate sampling distribution of the sample proportion  $\hat{p}$  from SRSs of size  $n = 100$  and  $n = 1000$  drawn from a population with proportion  $p = 0.37$  who have watched *Survivor*. Both dotplots show the results of 400 SRSs.





We can see that the spread of the top dotplot in Figure 7.7 is much greater than the spread of the bottom dotplot. With samples of size 100, the values of  $\hat{p}$  vary from 0.25 to 0.54. The standard deviation of these  $\hat{p}$ -values is about 0.05. Using SRSs of size 1000, the values of  $\hat{p}$  only vary from 0.328 to 0.412. The standard deviation of these  $\hat{p}$ -values is about 0.015, so most random samples of 1000 people give a  $\hat{p}$  that is within 0.03 of the actual population parameter,  $p = 0.37$ .

The sample proportion  $\hat{p}$  from a random sample of any size is an unbiased estimator of the parameter  $p$ . As we can see from the previous example, though, larger random samples have a clear advantage. They are much more likely to produce an estimate close to the true value of the parameter. Said another way, larger random samples give us more *precise* estimates than smaller random samples. That's because a large random sample gives us more information about the underlying population than a smaller sample does.

*Taking a larger sample doesn't fix bias. Remember that even a very large voluntary response sample or convenience sample is worthless because of bias.*



There are general rules for describing how the spread of the sampling distribution of a statistic decreases as the sample size increases. In Sections 7.2 and 7.3, we'll reveal these rules for the sampling distributions of  $\hat{p}$  and  $\bar{x}$ . One important and surprising fact is that the variability of a statistic in repeated sampling does *not* depend very much on the size of the population.

### VARIABILITY OF A STATISTIC

The **variability of a statistic** is described by the spread of its sampling distribution. This spread is determined mainly by the size of the random sample. Larger samples give smaller spreads. The spread of the sampling distribution does not depend much on the size of the population, as long as the population is at least 10 times larger than the sample.

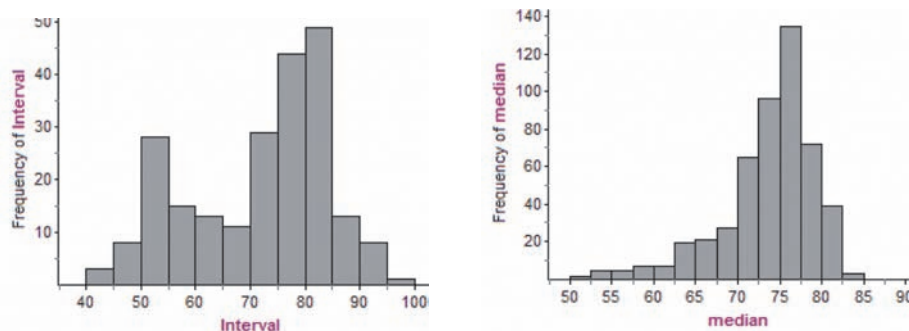
Why does the size of the population have little influence on the behavior of statistics from random samples? Imagine sampling harvested corn by thrusting a scoop into a large sack of corn kernels. The scoop doesn't know whether it is surrounded by a bag of corn or by an entire truckload. As long as the 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 a statistic is controlled by the size of the sample has important consequences for designing samples. Suppose a researcher wants to estimate the proportion of all U.S. adults who use Twitter regularly. A random sample of 1000 or 1500 people will give a fairly precise estimate of the parameter because the sample size is large. Now consider another researcher who wants to estimate the proportion of all Ohio State University students who use Twitter regularly. It can take just as large an SRS to estimate the proportion of Ohio State University students who use Twitter regularly as to estimate with equal precision the proportion of all U.S. adults who use Twitter regularly. We can't expect to need a smaller SRS at Ohio State just because there are about 60,000 Ohio State students and about 235 million adults in the United States.





## CHECK YOUR UNDERSTANDING



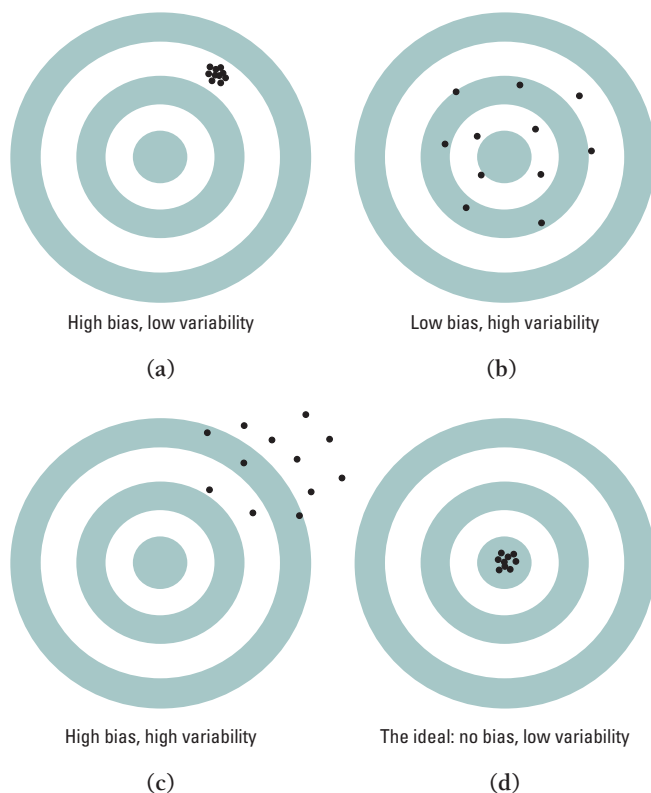
The histogram above left shows the intervals (in minutes) between eruptions of Old Faithful geyser for all 222 recorded eruptions during a particular month. For this population, the median is 75 minutes. We used Fathom software to take 500 SRSs of size 10 from the population. The 500 values of the sample median are displayed in the histogram above right. The mean of these 500 values is 73.5.

1. Is the sample median an unbiased estimator of the population median? Justify your answer.
2. Suppose we had taken samples of size 20 instead of size 10. Would the spread of the sampling distribution be larger, smaller, or about the same? Justify your answer.
3. Describe the shape of the sampling distribution.

**Bias, variability, and shape** 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 target. Both bias and variability describe what happens when we take many shots at the target. *Bias* means that our aim is off and we consistently miss the bull's-eye in the same direction. Our sample values do not center on the population value. *High variability* means that repeated shots are widely scattered on the target. Repeated samples do not give very similar results. Figure 7.8 shows this target illustration of the two types of error.

Notice that low variability (shots are close together) can accompany high bias (shots are consistently away from the bull's-eye in one direction). And low or no bias (shots center on the bull's-eye) can accompany high variability (shots are widely scattered). Ideally, we'd like our estimates to be *accurate* (unbiased) and *precise* (have low variability). See Figure 7.8(d).

The following example attempts to tie these ideas together in a familiar setting.



**FIGURE 7.8** Bias and variability. (a) High bias, low variability. (b) Low bias, high variability. (c) High bias, high variability. (d) The ideal: no bias, low variability.

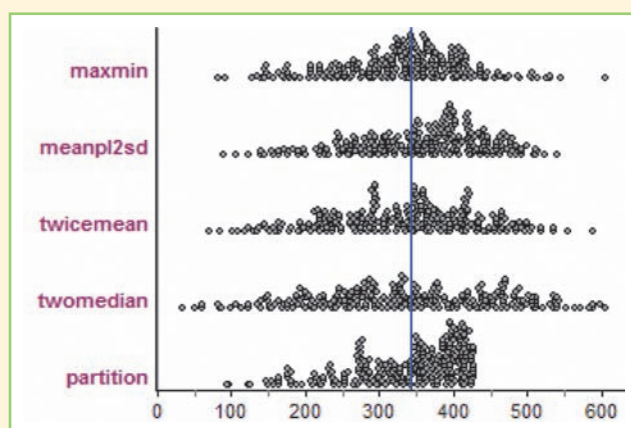


## EXAMPLE

## The German Tank Problem

### Evaluating estimators: Shape, center, spread

Refer to the Activity on page 422. Mrs. Friedman's student teams came up with four different methods for estimating the number of tanks in the bag: (1) "maxmin" = maximum + minimum, (2) "meanpl2sd" =  $\bar{x} + 2s_x$ , (3) "twicemean" =  $2\bar{x}$ , and (4) "twomedian" =  $2(\text{median})$ . She added one more method, called "partition." Figure 7.9 shows the results of taking 250 SRSs of 4 tanks and recording the value of the five statistics for each sample. The vertical line marks the actual value of the population parameter  $N$ : there were 342 tanks in the bag.



**FIGURE 7.9** Results from a Fathom simulation of 250 SRSs of 4 tanks. The approximate sampling distributions of five different statistics are shown.

**PROBLEM:** Use the information in Figure 7.9 to help answer these questions.

- Which of the four statistics proposed by the student teams is the best estimator? Justify your answer.
- Why was the partition method, which uses the statistic  $(5/4) \cdot \text{maximum}$ , recommended by the mathematicians in Washington, D.C.?

**SOLUTION:**

- Meanpl2sd is a biased estimator: the center of its sampling distribution is too high. This statistic produces consistent overestimates of the number of tanks. The other three statistics proposed by the students appear to be unbiased estimators. All three sampling distributions have roughly symmetric shapes, so these statistics are about equally likely to underestimate or overestimate the number of tanks. Because maxmin has the smallest variability

among the three, it would generally produce estimates that are closer to the actual number of tanks. Among the students' proposed statistics, maxmin would be the best estimator.

- The partition method uses a statistic  $(5/4) \cdot \text{maximum}$  that is an unbiased estimator and that has much less variability than any of the student teams' statistics. Its sampling distribution is left-skewed, so the mean of the distribution is less than its median. Because more than half of the dots in the graph are to the right of the mean, the statistic is more likely to overestimate than underestimate the number of tanks. The mathematicians believed that it would be better to err on the side of caution and give the military commanders an estimate that is slightly too high.

**For Practice** Try Exercise 19

The lesson about center and spread is clear: given a choice of statistics to estimate an unknown parameter, choose one with no or low bias and minimum variability. Shape is a more complicated issue. We have seen sampling distributions that are left-skewed, right-skewed, roughly symmetric, and even approximately Normal. The same statistic can have sampling distributions with different shapes depending on the population distribution and the sample size. Our advice: be sure to consider the shape of the sampling distribution before doing inference.

## Section 7.1

## Summary


- A **parameter** is a number that describes a population. To estimate an unknown parameter, use a **statistic** calculated from a sample.
- The **population distribution** of a variable describes the values of the variable for all individuals in a population. The **sampling distribution** of a statistic describes the values of the statistic in all possible samples of the same size from the same population. Don't confuse the sampling distribution with a **distribution of sample data**, which gives the values of the variable for all individuals in a particular sample.
- A statistic can be an **unbiased estimator** or a **biased estimator** of a parameter. A statistic is a biased estimator if the center (mean) of its sampling distribution is not equal to the true value of the parameter.
- The **variability** of a statistic is described by the spread of its sampling distribution. Larger samples give smaller spread.
- When trying to estimate a parameter, choose a statistic with low or no bias and minimum variability.

## Section 7.1

## Exercises

For Exercises 1 and 2, identify the population, the parameter, the sample, and the statistic in each setting.

## 1. Healthy living

- pg 425  (a) A random sample of 1000 people who signed a card saying they intended to quit smoking were contacted 9 months later. It turned out that 210 (21%) of the sampled individuals had not smoked over the past 6 months.
- (b) Tom is cooking a large turkey breast for a holiday meal. He wants to be sure that the turkey is safe to eat, which requires a minimum internal temperature of 165°F. Tom uses a thermometer to measure the temperature of the turkey meat at four randomly chosen points. The minimum reading in the sample is 170°F.

## 2. The economy

- (a) Each month, the Current Population Survey interviews a random sample of individuals in about 60,000 U.S. households. One of their goals is to estimate the national unemployment rate. In October 2012, 7.9% of those interviewed were unemployed.
- (b) How much do gasoline prices vary in a large city? To find out, a reporter records the price per gallon of regular unleaded gasoline at a random sample of 10 gas stations in the city on the same day. The range (maximum – minimum) of the prices in the sample is 25 cents.

For each boldface number in Exercises 3 to 6, (1) state whether it is a parameter or a statistic and (2) use appropriate notation to describe each number; for example,  $p = 0.65$ .

3. **Get your bearings** A large container is full of ball bearings with mean diameter **2.5003** centimeters (cm). This is within the specifications for acceptance of the container by the purchaser. By chance, an inspector chooses 100 bearings from the container that have mean diameter **2.5009** cm. Because this is outside the specified limits, the container is mistakenly rejected.
4. **Voters** Voter registration records show that **41%** of voters in a state are registered as Democrats. To test a random digit dialing device, you use it to call 250 randomly chosen residential telephones in the state. Of the registered voters contacted, **33%** are registered Democrats.
5. **Unlisted numbers** A telemarketing firm in a large city uses a device that dials residential telephone numbers in that city at random. Of the first 100 numbers dialed, **48%** are unlisted. This is not surprising because **52%** of all residential phones in the city are unlisted.
6. **How tall?** A random sample of female college students has a mean height of **64.5** inches, which is greater than the **63**-inch mean height of all adult American women.



Exercises 7 and 8 refer to the small population  $\{2, 6, 8, 10, 10, 12\}$  with mean  $\mu = 8$  and range 10.

### 7. Sampling distribution

- List all 15 possible SRSs of size  $n = 2$  from the population. Find the value of  $\bar{x}$  for each sample.
- Make a graph of the sampling distribution of  $\bar{x}$ . Describe what you see.

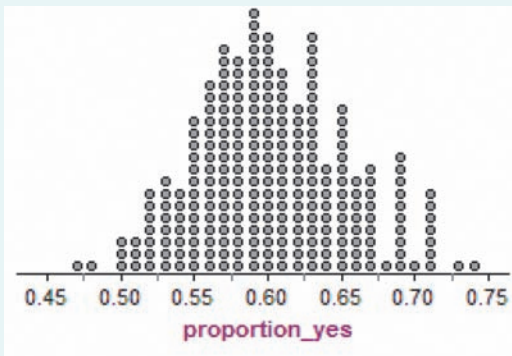
### 8. Sampling distribution

- List all 15 possible SRSs of size  $n = 2$  from the population. Find the value of the range for each sample.
- Make a graph of the sampling distribution of the sample range. Describe what you see.

pg 427



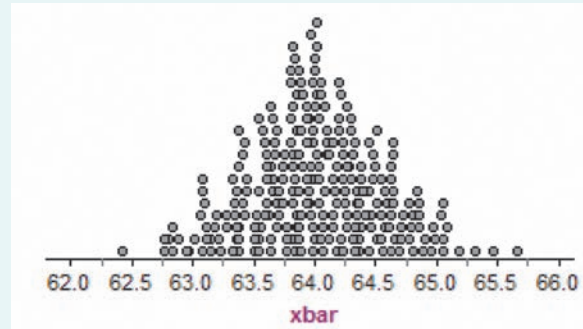
**9. Doing homework** A school newspaper article claims that 60% of the students at a large high school did all their assigned homework last week. Some skeptical AP<sup>®</sup> Statistics students want to investigate whether this claim is true, so they choose an SRS of 100 students from the school to interview. What values of the sample proportion  $\hat{p}$  would be consistent with the claim that the population proportion of students who completed all their homework is  $p = 0.60$ ? To find out, we used Fathom software to simulate choosing 250 SRSs of size  $n = 100$  students from a population in which  $p = 0.60$ . The figure below is a dotplot of the sample proportion  $\hat{p}$  of students who did all their homework.



- There is one dot on the graph at 0.73. Explain what this value represents.
- Describe the distribution. Are there any obvious outliers?
- Would it be surprising to get a sample proportion of 0.45 or lower in an SRS of size 100 when  $p = 0.6$ ? Justify your answer.
- Suppose that 45 of the 100 students in the actual sample say that they did all their homework last week. What would you conclude about the newspaper article's claim? Explain.

10. **Tall girls** According to the National Center for Health Statistics, the distribution of heights for 16-year-old females is modeled well by a Normal density curve with mean  $\mu = 64$  inches and standard deviation  $\sigma = 2.5$  inches. To see if this

distribution applies at their high school, an AP<sup>®</sup> Statistics class takes an SRS of 20 of the 300 16-year-old females at the school and measures their heights. What values of the sample mean  $\bar{x}$  would be consistent with the population distribution being  $N(64, 2.5)$ ? To find out, we used Fathom software to simulate choosing 250 SRSs of size  $n = 20$  students from a population that is  $N(64, 2.5)$ . The figure below is a dotplot of the sample mean height  $\bar{x}$  of the students in each sample.



- There is one dot on the graph at 62.4. Explain what this value represents.
- Describe the distribution. Are there any obvious outliers?
- Would it be surprising to get a sample mean of 64.7 or more in an SRS of size 20 when  $\mu = 64$ ? Justify your answer.
- Suppose that the average height of the 20 girls in the class's actual sample is  $\bar{x} = 64.7$ . What would you conclude about the population mean height  $\mu$  for the 16-year-old females at the school? Explain.

### 11. Doing homework

- Refer to Exercise 9. Make a bar graph of the population distribution given that the newspaper's claim is correct.
- Sketch a possible graph of the distribution of sample data for the SRS of size 100 taken by the AP<sup>®</sup> Statistics students.

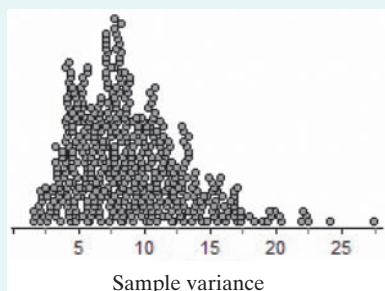
### 12. Tall girls

- Refer to Exercise 10. Make a graph of the population distribution.
- Sketch a possible dotplot of the distribution of sample data for the SRS of size 20 taken by the AP<sup>®</sup> Statistics class.

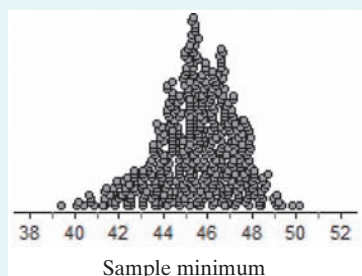
*Exercises 13 and 14 refer to the following setting.* During the winter months, outside temperatures at the Starneses' cabin in Colorado can stay well below freezing ( $32^{\circ}\text{F}$ , or  $0^{\circ}\text{C}$ ) for weeks at a time. To prevent the pipes from freezing, Mrs. Starnes sets the thermostat at  $50^{\circ}\text{F}$ . The manufacturer claims that the thermostat allows variation in home temperature that follows a Normal distribution with  $\sigma = 3^{\circ}\text{F}$ . To test this claim, Mrs. Starnes programs her digital thermometer to take an SRS of  $n = 10$  readings during a 24-hour period. Suppose the thermostat is

working properly and that the actual temperatures in the cabin vary according to a Normal distribution with mean  $\mu = 50^\circ\text{F}$  and standard deviation  $\sigma = 3^\circ\text{F}$ .

13. **Cold cabin?** The Fathom screen shot below shows the results of taking 500 SRSs of 10 temperature readings from a population distribution that is  $N(50, 3)$  and recording the sample variance  $s_x^2$  each time.



- (a) Describe the approximate sampling distribution.
- (b) Suppose that the variance from an actual sample is  $s_x^2 = 25$ . What would you conclude about the thermostat manufacturer's claim? Explain.
14. **Cold cabin?** The Fathom screen shot below shows the results of taking 500 SRSs of 10 temperature readings from a population distribution that is  $N(50, 3)$  and recording the sample minimum each time.

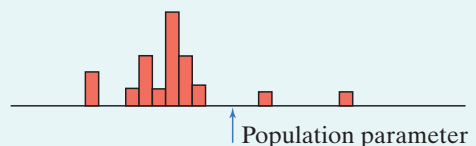


- (a) Describe the approximate sampling distribution.
- (b) Suppose that the minimum of an actual sample is  $40^\circ\text{F}$ . What would you conclude about the thermostat manufacturer's claim? Explain.
15. **A sample of teens** A study of the health of teenagers plans to measure the blood cholesterol levels 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  $\mu$  in this population. Explain to someone who knows little about statistics what it means to say that  $\bar{x}$  is an unbiased estimator of  $\mu$ .
16. **Predict the election** A polling organization plans to ask a random sample of likely voters who they plan to vote for in an upcoming election. The researchers will report the sample proportion  $\hat{p}$  that favors the incumbent as an estimate of the population proportion  $p$  that favors the incumbent. Explain to someone who knows little about statistics what it means to say that  $\hat{p}$  is an unbiased estimator of  $p$ .

17. **A sample of teens** Refer to Exercise 15. The sample mean  $\bar{x}$  is an unbiased estimator of the population mean  $\mu$  no matter what size SRS the study chooses. Explain to someone who knows nothing about statistics why a large random sample will give more trustworthy results than a small random sample.

18. **Predict the election** Refer to Exercise 16. The sample proportion  $\hat{p}$  is an unbiased estimator of the population proportion  $p$  no matter what size random sample the polling organization chooses. Explain to someone who knows nothing about statistics why a large random sample will give more trustworthy results than a small random sample.

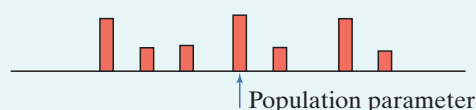
19. **Bias and variability** The figure below shows histograms of four sampling distributions of different statistics intended to estimate the same parameter.



(i)



(ii)



(iii)



(iv)

- (a) Which statistics are unbiased estimators? Justify your answer.
- (b) Which statistic does the best job of estimating the parameter? Explain.
20. **IRS audits** The Internal Revenue Service plans to examine an SRS of individual federal income tax returns. The parameter of interest is the proportion of all returns claiming itemized deductions. Which would be better for estimating this parameter: an SRS of 20,000 returns or an SRS of 2000 returns? Justify your answer.





**Multiple choice: Select the best answer for Exercises 21 to 24.**

21. At a particular college, 78% of all students are receiving some kind of financial aid. The school newspaper selects a random sample of 100 students and 72% of the respondents say they are receiving some sort of financial aid. Which of the following is true?
- 78% is a population and 72% is a sample.
  - 72% is a population and 78% is a sample.
  - 78% is a parameter and 72% is a statistic.
  - 72% is a parameter and 78% is a statistic.
  - 78% is a parameter and 100 is a statistic.
22. A statistic is an unbiased estimator of a parameter when
- the statistic is calculated from a random sample.
  - in a single sample, the value of the statistic is equal to the value of the parameter.
  - in many samples, the values of the statistic are very close to the value of the parameter.
  - in many samples, the values of the statistic are centered at the value of the parameter.
  - in many samples, the distribution of the statistic has a shape that is approximately Normal.
23. In a residential neighborhood, the median value of a house is \$200,000. For which of the following sample sizes is the sample median most likely to be above \$250,000?
- $n = 10$
  - $n = 50$
  - $n = 100$
  - $n = 1000$
  - Impossible to determine without more information.
24. Increasing the sample size of an opinion poll will reduce the
- bias of the estimates made from the data collected in the poll.
  - variability of the estimates made from the data collected in the poll.
  - effect of nonresponse on the poll.
  - variability of opinions in the sample.
  - variability of opinions in the population.

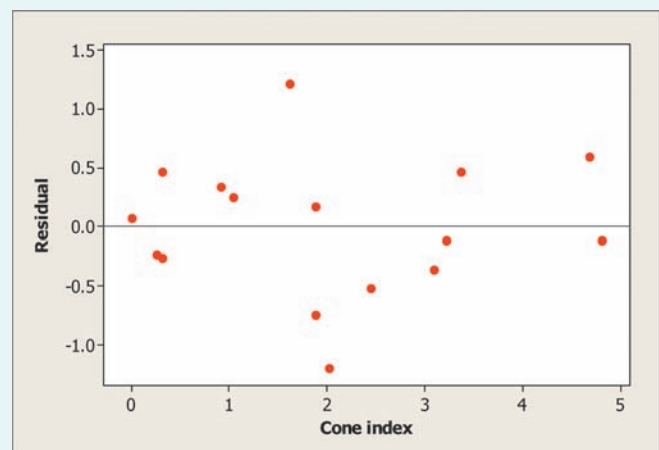
25. **Dem bones (2.2)** Osteoporosis is a condition in which the bones become brittle due to loss of minerals. To diagnose osteoporosis, an elaborate

apparatus measures bone mineral density (BMD). BMD is usually reported in standardized form. The standardization is based on a population of healthy young adults. The World Health Organization (WHO) criterion for osteoporosis is a BMD score that is 2.5 standard deviations below the mean for young adults. BMD measurements in a population of people similar in age and gender roughly follow a Normal distribution.

- What percent of healthy young adults have osteoporosis by the WHO criterion?
- Women aged 70 to 79 are, of course, not young adults. The mean BMD in this age group is about  $-2$  on the standard scale for young adults. Suppose that the standard deviation is the same as for young adults. What percent of this older population has osteoporosis?

26. **Squirrels and their food supply (3.2)** Animal species produce more offspring when their supply of food goes up. Some animals appear able to anticipate unusual food abundance. Red squirrels eat seeds from pinecones, a food source that sometimes has very large crops. Researchers collected data on an index of the abundance of pinecones and the average number of offspring per female over 16 years.<sup>3</sup> Computer output from a least-squares regression on these data and a residual plot are shown below.

Predictor	Coef	SE Coef	T	P
Constant	1.4146	0.2517	5.62	0.000
Cone index	0.4399	0.1016	4.33	0.001
S = 0.600309 R-Sq = 57.2% R-Sq(adj) = 54.2%				



- Give the equation for the least-squares regression line. Define any variables you use.
- Is a linear model appropriate for these data? Explain.
- Interpret the values of  $r^2$  and  $s$  in context.



## 7.2 Sample Proportions

### WHAT YOU WILL LEARN

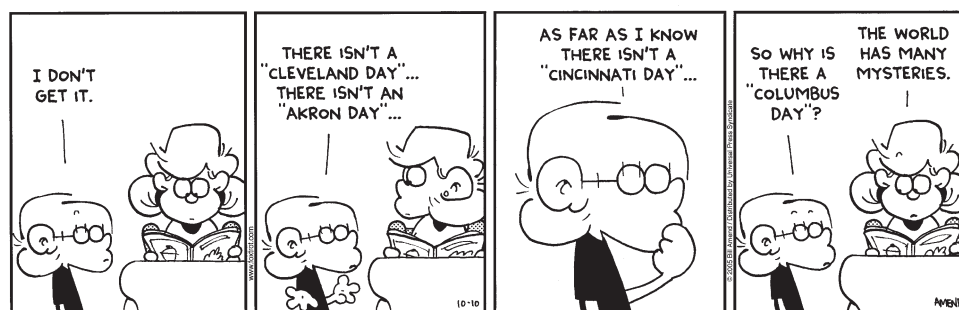
By the end of the section, you should be able to:

- Find the mean and standard deviation of the sampling distribution of a sample proportion  $\hat{p}$ . Check the 10% condition before calculating  $\sigma_{\hat{p}}$ .
- Determine if the sampling distribution of  $\hat{p}$  is approximately Normal.
- If appropriate, use a Normal distribution to calculate probabilities involving  $\hat{p}$ .

What proportion of U.S. teens know that 1492 was the year in which Columbus “discovered” America? A Gallup Poll found that 210 out of a random sample of 501 American teens aged 13 to 17 knew this historically important date.<sup>4</sup> The sample proportion

$$\hat{p} = \frac{210}{501} = 0.42$$

is the statistic that we use to gain information about the unknown population proportion  $p$ . Because another random sample of 501 teens would likely result in a different estimate, we can only say that “about” 42% of U.S. teenagers know that Columbus discovered America in 1492. In this section, we’ll use sampling distributions to clarify what “about” means.



### The Sampling Distribution of $\hat{p}$

How good is the statistic  $\hat{p}$  as an estimate of the parameter  $p$ ? To find out, we ask, “What would happen if we took many samples?” The **sampling distribution of  $\hat{p}$**  answers this question. How do we determine the shape, center, and spread of the sampling distribution of  $\hat{p}$ ? Let’s start with a simulation.

### ACTIVITY | The Candy Machine

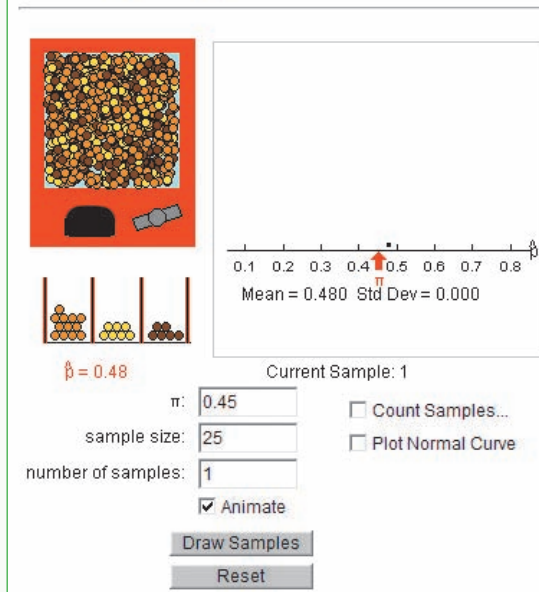
#### MATERIALS:

Computer with Internet access—one for the class or one per pair of students

Imagine a very large candy machine filled with orange, brown, and yellow candies. When you insert money, the machine dispenses a sample of candies. In this Activity, you will use an applet to investigate the sample-to-sample variability in the proportion of orange candies dispensed by the machine.



## Sampling Reese's Pieces

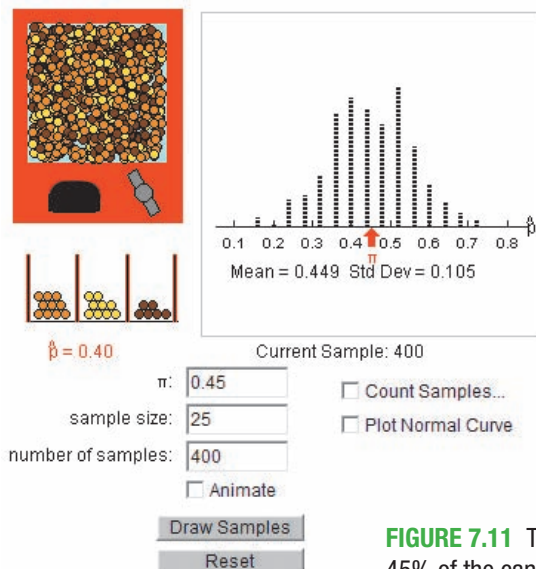


**FIGURE 7.10** The result of taking one SRS of 25 candies from a large candy machine in which 45% of the candies are orange.

1. Launch the *Reese's Pieces*® applet at [www.rossmanchance.com](http://www.rossmanchance.com). Change the population proportion of orange candies to  $p = 0.45$  (the applet calls this value  $\pi$  instead of  $p$ ).
2. Click on the “Draw Samples” button. An animated simple random sample of  $n = 25$  candies should be dispensed. Figure 7.10 shows the results of one such sample. Was your sample proportion of orange candies close to the actual population proportion,  $p = 0.45$ ? Look at the value of  $\hat{p}$  in the applet window.
3. Click “Draw Samples” 9 more times, so that you have a total of 10 sample results. Look at the dotplot of your  $\hat{p}$ -values. What is the mean of your 10 sample proportions? What is their standard deviation?
4. To take many more samples quickly, enter 390 in the “number of samples” box. Click on the Animate box to turn the animation off. Then click “Draw Samples.” You have now taken a total of 400 samples of 25 candies from the machine. Describe the shape, center, and spread of the approximate sampling distribution of  $\hat{p}$  shown in the dotplot.

5. How would the sampling distribution of the sample proportion  $\hat{p}$  change if the machine dispensed  $n = 50$  candies each time instead of 25? “Reset” the applet. Take 400 samples of 50 candies. Describe the shape, center, and spread of the approximate sampling distribution.
6. How would the sampling distribution of  $\hat{p}$  change if the proportion of orange candies in the machine was  $p = 0.15$  instead of  $p = 0.45$ ? Does your answer depend on whether  $n = 25$  or  $n = 50$ ? Use the applet to investigate these questions. Then write a brief summary of what you learned.
7. For what combinations of  $n$  and  $p$  is the sampling distribution of  $\hat{p}$  approximately Normal? Use the applet to investigate.

Figure 7.11 shows one set of possible results from Step 4 of “The Candy Machine” Activity. Let’s describe what we see.



**FIGURE 7.11** The result of taking 400 SRSs of 25 candies from a large candy machine in which 45% of the candies are orange. The dotplot shows the approximate sampling distribution of  $\hat{p}$ .

**Shape:** Roughly symmetric and somewhat bell-shaped. It looks as though a Normal curve would approximate this distribution fairly well.

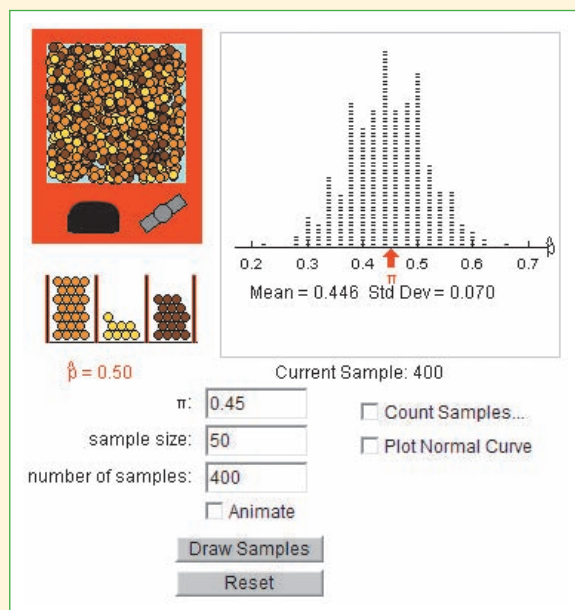
**Center:** The mean of the 400 sample proportions is 0.449. This is quite close to the actual population proportion,  $p = 0.45$ .

**Spread:** The standard deviation of the 400 values of  $\hat{p}$  from these samples is 0.105.

The dotplot in Figure 7.11 is the approximate sampling distribution of  $\hat{p}$ . If we took all possible SRSs of  $n = 25$  candies from the machine and graphed the value of  $\hat{p}$  for each sample, then we’d have the sampling distribution of  $\hat{p}$ . We can get an idea of its shape, center, and spread from Figure 7.11.

## EXAMPLE

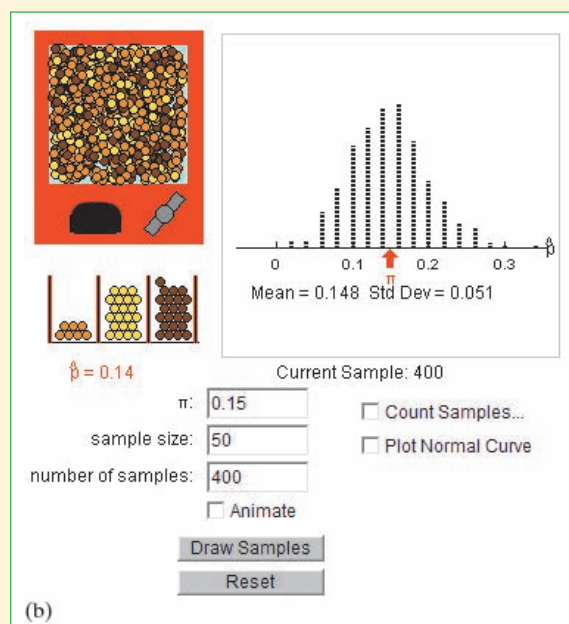
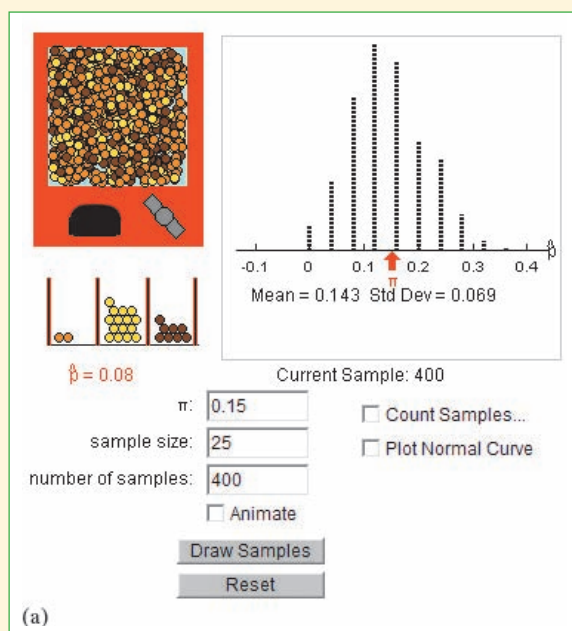
## Sampling Candies

*Effect of  $n$  and  $p$  on shape, center, and spread*

**FIGURE 7.12** The approximate sampling distribution of  $\hat{p}$  for 400 SRSs of 50 candies from a population in which  $p = 0.45$  of the candies are orange.

In a similar way, we can explore the sampling distribution of  $\hat{p}$  when  $n = 50$  (Step 5 of the Activity). As Figure 7.12 shows, the dotplot is once again roughly symmetric and somewhat bell-shaped. This graph is also centered at about 0.45. With samples of size 50, however, there is less spread in the values of  $\hat{p}$ . The standard deviation in Figure 7.12 is 0.070. For the samples of size 25 in Figure 7.11, it is 0.105. To repeat what we said earlier, larger samples give the sampling distribution a smaller spread.

What if the actual proportion of orange candies in the machine were  $p = 0.15$ ? Figure 7.13(a) shows the approximate sampling distribution of  $\hat{p}$  when  $n = 25$ . Notice that the dotplot is slightly right-skewed. The graph is centered close to the population parameter,  $p = 0.15$ . As for the spread, it's similar to the standard deviation in Figure 7.12, where  $n = 50$  and  $p = 0.45$ . If we increase the sample size to  $n = 50$ , the sampling distribution of  $\hat{p}$  should show less variability. The standard deviation in Figure 7.13(b) confirms this. Note that we can't just visually compare the graphs because the horizontal scales are different. The dotplot is more symmetrical than the graph in Figure 7.13(a) and is once again centered at a value that is close to  $p = 0.15$ .



**FIGURE 7.13** The result of taking 400 SRSs of (a) size  $n = 25$  and (b) size  $n = 50$  candies from a large candy machine in which 15% of the candies are orange. The dotplots show the approximate sampling distribution of  $\hat{p}$  in each case.



What have we learned so far about the sampling distribution of  $\hat{p}$ ?

**Shape:** In some cases, the sampling distribution of  $\hat{p}$  can be approximated by a Normal curve. This seems to depend on both the sample size  $n$  and the population proportion  $p$ .

**Center:** The mean of the distribution is  $\mu_{\hat{p}} = p$ . This makes sense because the sample proportion  $\hat{p}$  is an *unbiased estimator* of  $p$ .

**Spread:** For a specific value of  $p$ , the standard deviation  $\sigma_{\hat{p}}$  gets smaller as  $n$  gets larger. The value of  $\sigma_{\hat{p}}$  depends on both  $n$  and  $p$ .

To sort out the details of shape and spread, we need to make an important connection between the sample proportion  $\hat{p}$  and the number of “successes”  $X$  in the sample.

In the candy machine example, we started by taking repeated SRSs of  $n = 25$  candies from a population with proportion  $p = 0.45$  of orange candies. For any such sample, we can think of each candy that comes out of the machine as a trial of this chance process. A “success” occurs when we get an orange candy. Let  $X$  = the number of orange candies obtained. As long as the number of candies in the machine is very large,  $X$  will have close to a binomial distribution with  $n = 25$  and  $p = 0.45$ . (Refer to the 10% condition on page 401.) The sample proportion of successes is closely related to  $X$ :

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

**THINK  
ABOUT IT**

**How is the sampling distribution of  $\hat{p}$  related to the binomial count  $X$ ?** From Chapter 6, we know that the mean and standard deviation of a binomial random variable  $X$  are

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

Because  $\hat{p} = X/n = (1/n)X$ , we’re just multiplying the random variable  $X$  by a constant  $(1/n)$  to get the random variable  $\hat{p}$ . Recall from Chapter 6 that multiplying by a constant multiplies both the mean and the standard deviation of the new random variable by that constant. We have

$$\mu_{\hat{p}} = \frac{1}{n}(np) = p \quad (\text{confirming that } \hat{p} \text{ is an unbiased estimator of } p)$$

$$\sigma_{\hat{p}} = \frac{1}{n}\sqrt{np(1-p)} = \sqrt{\frac{np(1-p)}{n^2}} = \sqrt{\frac{p(1-p)}{n}}$$

(as sample size increases, spread decreases)

That takes care of center and spread. What about shape? Multiplying a random variable by a constant doesn’t change the shape of the probability distribution. So the sampling distribution of  $\hat{p}$  will have the same shape as the distribution of the binomial random variable  $X$ .

If you studied the optional material in Chapter 6 about the Normal approximation to a binomial distribution, then you already know the punch line. Whenever  $np$  and  $n(1-p)$  are at least 10, a Normal distribution can be used to approximate the sampling distribution of  $\hat{p}$ .

Here's a summary of the important facts about the sampling distribution of  $\hat{p}$ .

### SAMPLING DISTRIBUTION OF A SAMPLE PROPORTION

Choose an SRS of size  $n$  from a population of size  $N$  with proportion  $p$  of successes. Let  $\hat{p}$  be the sample proportion of successes. Then:

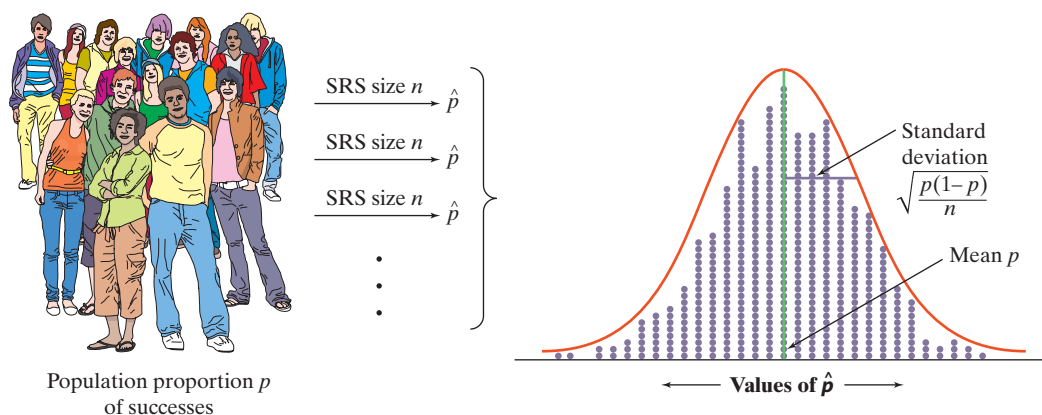
- The **mean** of the sampling distribution of  $\hat{p}$  is  $\mu_{\hat{p}} = p$ .
- The **standard deviation** of the sampling distribution of  $\hat{p}$  is

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

as long as the *10% condition* is satisfied:  $n \leq \frac{1}{10}N$ .

- As  $n$  increases, the sampling distribution of  $\hat{p}$  becomes **approximately Normal**. Before you perform Normal calculations, check that the *Large Counts condition* is satisfied:  $np \geq 10$  and  $n(1-p) \geq 10$ .

Figure 7.14 displays the facts in a form that helps you recall the big idea of a sampling distribution. The mean of the sampling distribution of  $\hat{p}$  is the true value of the population proportion  $p$ . The standard deviation of  $\hat{p}$  gets smaller as the sample size  $n$  increases. In fact, because the sample size  $n$  is under the square root sign, we'd have to take a sample four times as large to cut the standard deviation in half.



**FIGURE 7.14** Select a large SRS from a population in which proportion  $p$  are successes. The sampling distribution of the proportion  $\hat{p}$  of successes in the sample is approximately Normal. The mean is  $p$  and the standard deviation is  $\sqrt{p(1-p)/n}$ .

The two conditions in the preceding box are very important. (1) *Large Counts condition*: If we assume that the sampling distribution of  $\hat{p}$  is approximately Normal when it isn't, any calculations we make using a Normal distribution will be flawed. (2) *10% condition*: When we're sampling without replacement from a (finite) population, the observations are not independent, because knowing the outcome of one trial helps us predict the outcome of future trials. But the standard deviation formula assumes that the observations are independent. If we sample too large a fraction of the population, our calculated value of  $\sigma_{\hat{p}}$  will be inaccurate.





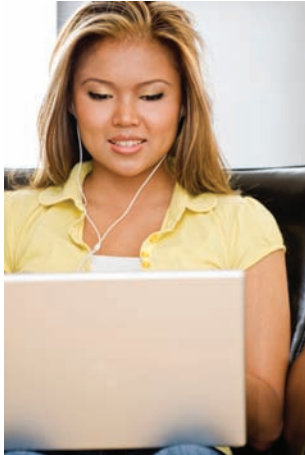
Because larger random samples give better information, it sometimes makes sense to sample more than 10% of a population. In such a case, there's a more accurate formula for calculating the standard deviation  $\sigma_{\hat{p}}$ . It uses something called a *finite population correction* (FPC). We'll avoid situations that require the FPC in this text.



### CHECK YOUR UNDERSTANDING

About 75% of young adult Internet users (ages 18 to 29) watch online videos. Suppose that a sample survey contacts an SRS of 1000 young adult Internet users and calculates the proportion  $\hat{p}$  in this sample who watch online videos.

1. What is the mean of the sampling distribution of  $\hat{p}$ ? Explain.
2. Find the standard deviation of the sampling distribution of  $\hat{p}$ . Check that the 10% condition is met.
3. Is the sampling distribution of  $\hat{p}$  approximately Normal? Check that the Large Counts condition is met.
4. If the sample size were 9000 rather than 1000, how would this change the sampling distribution of  $\hat{p}$ ?



## Using the Normal Approximation for $\hat{p}$

Inference about a population proportion  $p$  is based on the sampling distribution of  $\hat{p}$ . When the sample size is large enough for  $np$  and  $n(1 - p)$  to both be at least 10 (the *Large Counts condition*), the sampling distribution of  $\hat{p}$  is approximately Normal. In that case, we can use a Normal distribution to calculate the probability of obtaining an SRS in which  $\hat{p}$  lies in a specified interval of values. Here is an example.

### EXAMPLE

## Going to College

### Normal calculations involving $\hat{p}$

A polling organization asks an SRS of 1500 first-year college students how far away their home is. Suppose that 35% of all first-year students attend college within 50 miles of home.

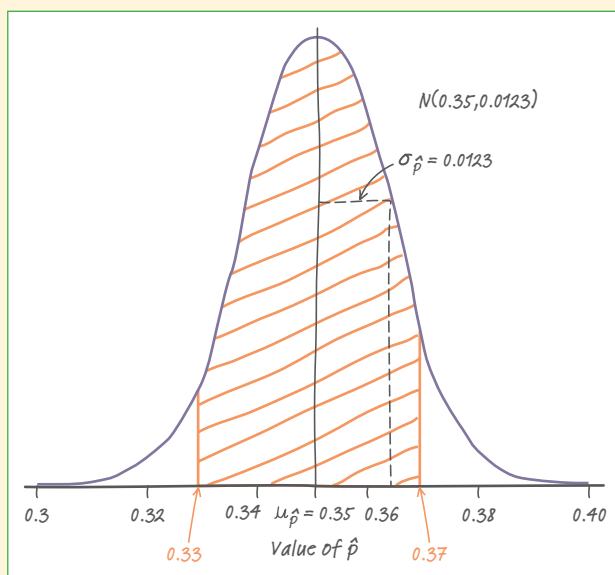
**PROBLEM:** Find the probability that the random sample of 1500 students will give a result within 2 percentage points of this true value. Show your work.

**SOLUTION:**

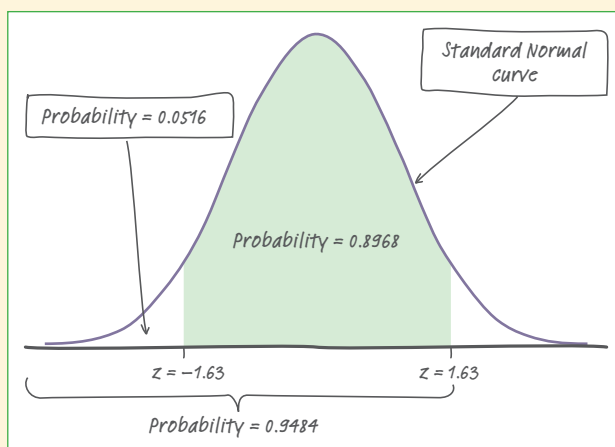
**Step 1: State the distribution and the values of interest.** We want to find the probability that  $\hat{p}$  falls between 0.33 and 0.37 (within 2 percentage points, or 0.02, of 0.35). In symbols, that's  $P(0.33 \leq \hat{p} \leq 0.37)$ . We have an SRS of size  $n = 1500$  drawn from a population in which the proportion  $p = 0.35$  attend college within 50 miles of home. What do we know about the sampling distribution of  $\hat{p}$ ?

- Its mean is  $\mu_{\hat{p}} = p = 0.35$ .





**FIGURE 7.15** The Normal approximation to the sampling distribution of  $\hat{p}$ .



**FIGURE 7.16** Probabilities as areas under the standard Normal curve.

- What about the standard deviation? We need to check the 10% condition. To use the standard deviation formula we derived, the population must contain at least  $10(1500) = 15,000$  people. There are over 1.7 million first-year college students, so

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(0.35)(0.65)}{1500}} = 0.0123$$

- Can we use a Normal distribution to approximate the sampling distribution of  $\hat{p}$ ? Check the Large Counts condition:  $np = 1500(0.35) = 525$  and  $n(1-p) = 1500(0.65) = 975$ . Both are much larger than 10, so the Normal approximation will be quite accurate.

Figure 7.15 shows the Normal distribution that we'll use with the area of interest shaded and the mean, standard deviation, and boundary values labeled.

**Step 2: Perform calculations—show your work!** The standardized scores for the two boundary values are

$$z = \frac{0.33 - 0.35}{0.0123} = -1.63 \quad \text{and} \quad z = \frac{0.37 - 0.35}{0.0123} = 1.63$$

Figure 7.16 shows the area under the standard Normal curve corresponding to these standardized values. Using Table A, the desired probability is

$$\begin{aligned} P(0.33 \leq \hat{p} \leq 0.37) &= P(-1.63 \leq Z \leq 1.63) \\ &= 0.9484 - 0.0516 = 0.8968 \end{aligned}$$

*Using technology:* The command `normalcdf(lower:0.33, upper:0.37,  $\mu$ :0.35,  $\sigma$ :0.0123)` gives an area of 0.8961.

**Step 3: Answer the question.** About 90% of all SRSs of size 1500 will give a result within 2 percentage points of the truth about the population.

**For Practice** Try Exercise 39

## Section 7.2

## Summary

- When we want information about the population proportion  $p$  of successes, we often take an SRS and use the sample proportion  $\hat{p}$  to estimate the unknown parameter  $p$ . The **sampling distribution** of  $\hat{p}$  describes how the sample proportion varies in all possible samples from the population.
- The **mean** of the sampling distribution of  $\hat{p}$  is equal to the population proportion  $p$ . That is,  $\hat{p}$  is an unbiased estimator of  $p$ .



- The **standard deviation** of the sampling distribution of  $\hat{p}$  is  $\sqrt{p(1-p)/n}$  for an SRS of size  $n$ . This formula can be used if the population is at least 10 times as large as the sample (the *10% condition*). The standard deviation of  $\hat{p}$  gets smaller as the sample size  $n$  gets larger. Because of the square root, a sample four times larger is needed to cut the standard deviation in half.
- When the sample size  $n$  is large, the sampling distribution of  $\hat{p}$  is close to a Normal distribution with mean  $p$  and standard deviation  $\sqrt{p(1-p)/n}$ . In practice, use this **Normal approximation** when both  $np \geq 10$  and  $n(1-p) \geq 10$  (the *Large Counts condition*).

## Section 7.2 Exercises

- 27. The candy machine** Suppose a large candy machine has 45% orange candies. Use Figures 7.11 and 7.12 (pages 441 and 442) to help answer the following questions.
- Would you be surprised if a sample of 25 candies from the machine contained 8 orange candies (that's 32% orange)? How about 5 orange candies (20% orange)? Explain.
  - Which is more surprising: getting a sample of 25 candies in which 32% are orange or getting a sample of 50 candies in which 32% are orange? Explain.
- 28. The candy machine** Suppose a large candy machine has 15% orange candies. Use Figure 7.13 (page 442) to help answer the following questions.
- Would you be surprised if a sample of 25 candies from the machine contained 8 orange candies (that's 32% orange)? How about 5 orange candies (20% orange)? Explain.
  - Which is more surprising: getting a sample of 25 candies in which 32% are orange or getting a sample of 50 candies in which 32% are orange? Explain.
- 29. The candy machine** Suppose a large candy machine has 45% orange candies. Imagine taking an SRS of 25 candies from the machine and observing the sample proportion  $\hat{p}$  of orange candies.
- What is the mean of the sampling distribution of  $\hat{p}$ ? Why?
  - Find the standard deviation of the sampling distribution of  $\hat{p}$ . Check to see if the 10% condition is met.
  - Is the sampling distribution of  $\hat{p}$  approximately Normal? Check to see if the Large Counts condition is met.
  - If the sample size were 100 rather than 25, how would this change the sampling distribution of  $\hat{p}$ ?
- 30. The candy machine** Suppose a large candy machine has 15% orange candies. Imagine taking an SRS of 25 candies from the machine and observing the sample proportion  $\hat{p}$  of orange candies.
- What is the mean of the sampling distribution of  $\hat{p}$ ? Why?
  - Find the standard deviation of the sampling distribution of  $\hat{p}$ . Check to see if the 10% condition is met.
  - Is the sampling distribution of  $\hat{p}$  approximately Normal? Check to see if the Large Counts condition is met.
  - If the sample size were 225 rather than 25, how would this change the sampling distribution of  $\hat{p}$ ?
- 31. Airport security** The Transportation Security Administration (TSA) is responsible for airport safety. On some flights, TSA officers randomly select passengers for an extra security check before boarding. One such flight had 76 passengers—12 in first class and 64 in coach class. TSA officers selected an SRS of 10 passengers for screening. Let  $\hat{p}$  be the proportion of first-class passengers in the sample.
- Is the 10% condition met in this case? Justify your answer.
  - Is the Large Counts condition met in this case? Justify your answer.

**32. Scrabble** In the game of Scrabble, each player begins by drawing 7 tiles from a bag containing 100 tiles. There are 42 vowels, 56 consonants, and 2 blank tiles in the bag. Cait chooses an SRS of 7 tiles. Let  $\hat{p}$  be the proportion of vowels in her sample.

- (a) Is the 10% condition met in this case? Justify your answer.
- (b) Is the Large Counts condition met in this case? Justify your answer.

*In Exercises 33 and 34, explain why you cannot use the methods of this section to find the desired probability.*

**33. Hispanic workers** A factory employs 3000 unionized workers, of whom 30% are Hispanic. The 15-member union executive committee contains 3 Hispanics. What would be the probability of 3 or fewer Hispanics if the executive committee were chosen at random from all the workers?

**34. Studious athletes** A university is concerned about the academic standing of its intercollegiate athletes. A study committee chooses an SRS of 50 of the 316 athletes to interview in detail. Suppose that 40% of the athletes have been told by coaches to neglect their studies on at least one occasion. What is the probability that at least 15 in the sample are among this group?

**35. Do you drink the cereal milk?** A *USA Today* Poll asked a random sample of 1012 U.S. adults what they do with the milk in the bowl after they have eaten the cereal. Let  $\hat{p}$  be the proportion of people in the sample who drink the cereal milk. A spokesman for the dairy industry claims that 70% of all U.S. adults drink the cereal milk. Suppose this claim is true.

- (a) What is the mean of the sampling distribution of  $\hat{p}$ ? Why?
- (b) Find the standard deviation of the sampling distribution of  $\hat{p}$ . Check to see if the 10% condition is met.
- (c) Is the sampling distribution of  $\hat{p}$  approximately Normal? Check to see if the Large Counts condition is met.
- (d) Of the poll respondents, 67% said that they drink the cereal milk. Find the probability of obtaining a sample of 1012 adults in which 67% or fewer say they drink the cereal milk if the milk industry spokesman's claim is true. Does this poll give convincing evidence against the claim? Explain.

**36. Do you go to church?** The Gallup Poll asked a random sample of 1785 adults whether they

attended church during the past week. Let  $\hat{p}$  be the proportion of people in the sample who attended church. A newspaper report claims that 40% of all U.S. adults went to church last week. Suppose this claim is true.

- (a) What is the mean of the sampling distribution of  $\hat{p}$ ? Why?
- (b) Find the standard deviation of the sampling distribution of  $\hat{p}$ . Check to see if the 10% condition is met.
- (c) Is the sampling distribution of  $\hat{p}$  approximately Normal? Check to see if the Large Counts condition is met.
- (d) Of the poll respondents, 44% said they did attend church last week. Find the probability of obtaining a sample of 1785 adults in which 44% or more say they attended church last week if the newspaper report's claim is true. Does this poll give convincing evidence against the claim? Explain.

**37. Do you drink the cereal milk?** What sample size would be required to reduce the standard deviation of the sampling distribution to one-half the value you found in Exercise 35(b)? Justify your answer.

**38. Do you go to church?** What sample size would be required to reduce the standard deviation of the sampling distribution to one-third the value you found in Exercise 36(b)? Justify your answer.

**39. Students on diets** A sample survey interviews an SRS of 267 college women. Suppose that 70% of college women have been on a diet within the past 12 months. What is the probability that 75% or more of the women in the sample have been on a diet? Show your work.

pg 445



**40. Who owns a Harley?** Harley-Davidson motorcycles make up 14% of all the motorcycles registered in the United States. You plan to interview an SRS of 500 motorcycle owners. How likely is your sample to contain 20% or more who own Harleys? Show your work.

**41. On-time shipping** A mail-order company advertises that it ships 90% of its orders within three working days. You select an SRS of 100 of the 5000 orders received in the past week for an audit. The audit reveals that 86 of these orders were shipped on time.

- (a) If the company really ships 90% of its orders on time, what is the probability that the proportion in an SRS of 100 orders is 0.86 or less? Show your work.
- (b) A critic says, "Aha! You claim 90%, but in your sample the on-time percentage is lower than that."



So the 90% claim is wrong.” Explain in simple language why your probability calculation in (a) shows that the result of the sample does not refute the 90% claim.

- 42. Underage drinking** The Harvard College Alcohol Study finds that 67% of college students support efforts to “crack down on underage drinking.” Does this result hold at a large local college? To find out, college administrators survey an SRS of 100 students and find that 62 support a crackdown on underage drinking.
- (a) Suppose that the proportion of all students attending this college who support a crackdown is 67%, the same as the national proportion. What is the probability that the proportion in an SRS of 100 students is 0.62 or less? Show your work.
- (b) A writer in the college’s student paper says that “support for a crackdown is lower at our school than nationally.” Write a short letter to the editor explaining why the survey does not support this conclusion.

**Multiple choice: Select the best answer for Exercises 43 to 46.** Exercises 43 to 45 refer to the following setting. The magazine *Sports Illustrated* asked a random sample of 750 Division I college athletes, “Do you believe performance-enhancing drugs are a problem in college sports?” Suppose that 30% of all Division I athletes think that these drugs are a problem. Let  $\hat{p}$  be the sample proportion who say that these drugs are a problem.

- 43.** Which of the following are the mean and standard deviation of the sampling distribution of the sample proportion  $\hat{p}$ ?
- (a) Mean = 0.30, SD = 0.017  
 (b) Mean = 0.30, SD = 0.55  
 (c) Mean = 0.30, SD = 0.0003  
 (d) Mean = 225, SD = 12.5  
 (e) Mean = 225, SD = 157.5
- 44.** Decreasing the sample size from 750 to 375 would multiply the standard deviation by
- (a) 2.      (c) 1/2.      (e) none of these.  
 (b)  $\sqrt{2}$ .      (d)  $1/\sqrt{2}$ .
- 45.** The sampling distribution of  $\hat{p}$  is approximately Normal because
- (a) there are at least 7500 Division I college athletes.

- (b)  $np = 225$  and  $n(1 - p) = 525$  are both at least 10.  
 (c) a random sample was chosen.  
 (d) the athletes’ responses are quantitative.  
 (e) the sampling distribution of  $\hat{p}$  always has this shape.
- 46.** In a congressional district, 55% of the registered voters are Democrats. Which of the following is equivalent to the probability of getting less than 50% Democrats in a random sample of size 100?

- (a)  $P\left(Z < \frac{0.50 - 0.55}{100}\right)$   
 (b)  $P\left(Z < \frac{0.50 - 0.55}{\sqrt{\frac{0.55(0.45)}{100}}}\right)$   
 (c)  $P\left(Z < \frac{0.55 - 0.50}{\sqrt{\frac{0.55(0.45)}{100}}}\right)$   
 (d)  $P\left(Z < \frac{0.50 - 0.55}{\sqrt{100(0.55)(0.45)}}\right)$   
 (e)  $P\left(Z < \frac{0.55 - 0.50}{\sqrt{100(0.55)(0.45)}}\right)$

**47. Sharing music online (5.2)** A sample survey reports that 29% of Internet users download music files online, 21% share music files from their computers, and 12% both download and share music.<sup>5</sup> Make a Venn diagram that displays this information. What percent of Internet users neither download nor share music files?

**48. California’s endangered animals (4.1)** The California Department of Fish and Game publishes a list of the state’s endangered animals. The reptiles on the list are given below.

Desert tortoise	Southern rubber boa
Olive Ridley sea turtle	Loggerhead sea turtle
Island night lizard	Barefoot banded gecko
Flat-tailed horned lizard	Coachella Valley fringe-toed lizard
Green sea turtle	Blunt-nosed leopard lizard
Leatherback sea turtle	Giant garter snake
Alameda whip snake	San Francisco garter snake

- (a) Describe how you would use Table D at line 111 to choose an SRS of 3 of these reptiles to study.
- (b) Use your method from part (a) to select your sample. Identify the reptiles you chose.

## 7.3 Sample Means

### WHAT YOU WILL LEARN

By the end of the section, you should be able to:

- Find the mean and standard deviation of the sampling distribution of a sample mean  $\bar{x}$ . Check the 10% condition before calculating  $\sigma_{\bar{x}}$ .
- Explain how the shape of the sampling distribution of  $\bar{x}$  is affected by the shape of the population distribution and the sample size.
- If appropriate, use a Normal distribution to calculate probabilities involving  $\bar{x}$ .

Sample proportions arise most often when we are interested in categorical variables. We then ask questions like “What proportion of U.S. adults have watched *Survivor*?” or “What percent of the adult population attended church last week?” But when we record quantitative variables—household income, lifetime of car brake pads, blood pressure—we are interested in other statistics, such as the median or mean or standard deviation of the variable. The sample mean  $\bar{x}$  is the most common statistic computed from quantitative data. This section describes the sampling distribution of the sample mean. The following Activity and the subsequent example give you a sense of what lies ahead.

### ACTIVITY | Penny for Your Thoughts

#### MATERIALS:

Large container with several hundred pennies



Your teacher will assemble a large population of pennies of various ages.<sup>6</sup> In this Activity, your class will investigate the sampling distribution of the mean year  $\bar{x}$  in a sample of pennies for SRSs of several different sizes. Then, you will compare these distributions of the mean year with the population distribution.

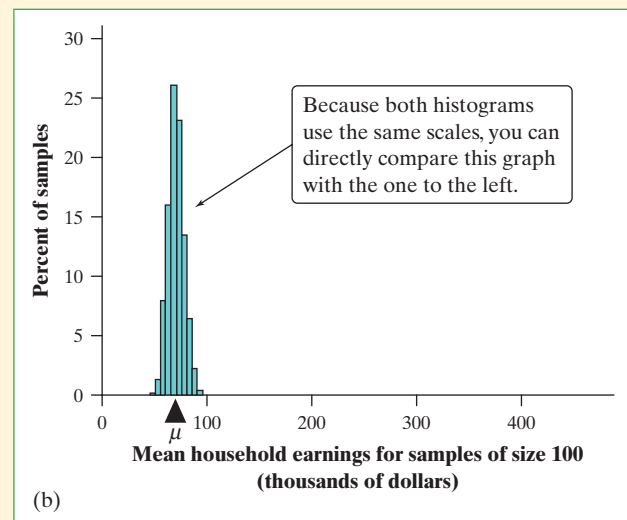
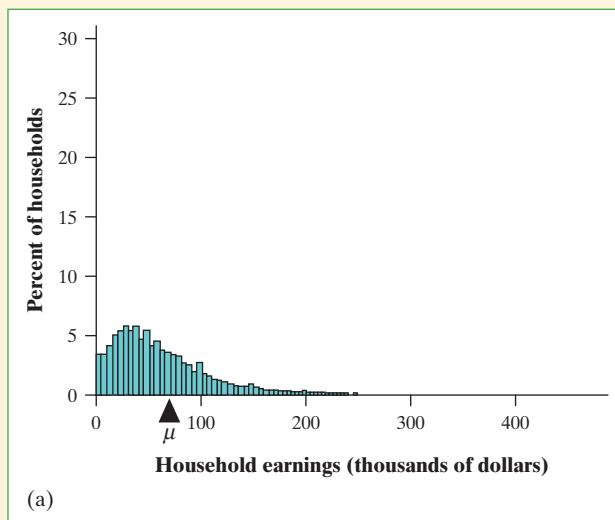
1. Your teacher will provide a dotplot of the population distribution of penny years.
2. Have each member of the class take an SRS of 5 pennies from the population and record the year on each penny. Be sure to replace these coins in the container before the next student takes a sample. If your class has fewer than 25 students, have each person take two samples.
3. Calculate the mean year  $\bar{x}$  of the 5 pennies in your sample.
4. Make a class dotplot of the sample mean years for SRSs of size 5 using the same scale as you did for the population distribution. Use  $\bar{x}$ 's instead of dots when making the graph.
5. Repeat the process in Steps 2 to 4 for samples of size 25. Use the same scale for your dotplot and place it beside the graph for samples of size 5.
6. Compare the population distribution with the two approximate sampling distributions of  $\bar{x}$ . What do you notice about shape, center, and spread as the sample size increases?

**EXAMPLE****Making Money***A first look at the sampling distribution of  $\bar{x}$* 

Figure 7.17(a) is a histogram of the earnings of a population of 61,742 households that had earned income greater than zero in a recent year.<sup>7</sup>

As we expect, the distribution of earned incomes is strongly skewed to the right and very spread out. The right tail of the distribution is even longer than the histogram shows because there are too few high incomes for their bars to be visible on this scale. We cut off the earnings scale at \$400,000 to save space. The mean earnings for these 61,742 households was  $\mu = \$69,750$ .

Take an SRS of 100 households. The mean earnings in this sample is  $\bar{x} = \$66,807$ . That's less than the mean of the population. Take another SRS of size 100. The mean for this sample is  $\bar{x} = \$70,820$ . That's higher than the mean of the population. What would happen if we did this many times? Figure 7.17(b) is a histogram of the mean earnings for 500 samples, each of size  $n = 100$ . The scales in Figures 7.17(a) and 7.17(b) are the same, for easy comparison. Although the distribution of individual earnings is skewed and very spread out, the distribution of sample means is roughly symmetric and much less spread out. Both distributions are centered at  $\mu = \$69,750$ .



**FIGURE 7.17** (a) The distribution of earned income in a population of 61,472 households. (b) The distribution of the mean earnings  $\bar{x}$  for 500 SRSs of  $n = 100$  households from this population.

This example illustrates an important fact that we will make precise in this section: averages are less variable than individual observations.

## The Sampling Distribution of $\bar{x}$ : Mean and Standard Deviation

Figure 7.17 suggests that when we choose many SRSs from a population, the sampling distribution of the sample mean is centered at the population mean  $\mu$  and is less spread out than the population distribution. Here are the facts.



### MEAN AND STANDARD DEVIATION OF THE SAMPLING DISTRIBUTION OF $\bar{x}$

Suppose that  $\bar{x}$  is the mean of an SRS of size  $n$  drawn from a large population with mean  $\mu$  and standard deviation  $\sigma$ . Then:

- The **mean** of the sampling distribution of  $\bar{x}$  is  $\mu_{\bar{x}} = \mu$ .
- The **standard deviation** of the sampling distribution of  $\bar{x}$  is

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

as long as the *10% condition* is satisfied:  $n \leq \frac{1}{10} N$ .

**AP® EXAM TIP** Notation matters. The symbols  $\hat{p}$ ,  $\bar{x}$ ,  $\rho$ ,  $\mu$ ,  $\sigma$ ,  $\mu_{\hat{p}}$ ,  $\sigma_{\hat{p}}$ ,  $\mu_{\bar{x}}$ , and  $\sigma_{\bar{x}}$  all have specific and different meanings. Either use notation correctly—or don't use it at all. You can expect to lose credit if you use incorrect notation.

The behavior of  $\bar{x}$  in repeated samples is much like that of the sample proportion  $\hat{p}$ :

- The sample mean  $\bar{x}$  is an *unbiased estimator* of the population mean  $\mu$ .
- The values of  $\bar{x}$  are less spread out for larger samples. Their standard deviation decreases at the rate  $\sqrt{n}$ , so you must take a sample four times as large to cut the standard deviation of the distribution of  $\bar{x}$  in half.
- You should use the formula  $\sigma/\sqrt{n}$  for the standard deviation of  $\bar{x}$  only when the population is at least 10 times as large as the sample (the *10% condition*).

Notice that these facts about the mean and standard deviation of  $\bar{x}$  are true *no matter what shape the population distribution has*.

## EXAMPLE

### This Wine Stinks

#### Mean and standard deviation of $\bar{x}$

Sulfur compounds such as dimethyl sulfide (DMS) are sometimes present in wine. DMS causes “off-odors” in wine, so winemakers want to know the odor threshold, the lowest concentration of DMS that the human nose can detect. Extensive studies have found that the DMS odor threshold of adults follows a distribution with mean  $\mu = 25$  micrograms per liter and standard deviation  $\sigma = 7$  micrograms per liter. Suppose we take an SRS of 10 adults and determine the mean odor threshold  $\bar{x}$  for the individuals in the sample.

#### PROBLEM:

- What is the mean of the sampling distribution of  $\bar{x}$ ? Explain.
- What is the standard deviation of the sampling distribution of  $\bar{x}$ ? Check that the 10% condition is met.

#### SOLUTION:

- Because  $\bar{x}$  is an unbiased estimator of  $\mu$ ,  $\mu_{\bar{x}} = \mu = 25$  micrograms per liter.
- The standard deviation is  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{7}{\sqrt{10}} = 2.214$  because there are at least  $10(10) = 100$  adults in the population.


**THINK  
ABOUT IT**

**Can we confirm the formulas for the mean and standard deviation of  $\bar{x}$ ?** Choose an SRS of size  $n$  from a population, and measure a variable  $X$  on each individual in the sample. Call the individual measurements  $X_1, X_2, \dots, X_n$ . If the population is large relative to the sample, we can think of these  $X_i$ 's as independent random variables, each with mean  $\mu$  and standard deviation  $\sigma$ . Because

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

we can use the rules for random variables from Chapter 6 to find the mean and standard deviation of  $\bar{x}$ . If we let  $T = X_1 + X_2 + \cdots + X_n$ , then  $\bar{x} = \frac{1}{n}T$ .

Using the addition rules for means and variances, we get

$$\begin{aligned}\mu_T &= \mu_{X_1} + \mu_{X_2} + \cdots + \mu_{X_n} = \mu + \mu + \cdots + \mu = n\mu \\ \sigma_T^2 &= \sigma_{X_1}^2 + \sigma_{X_2}^2 + \cdots + \sigma_{X_n}^2 = \sigma^2 + \sigma^2 + \cdots + \sigma^2 = n\sigma^2 \\ \Rightarrow \sigma_T &= \sqrt{n\sigma^2} = \sigma\sqrt{n}\end{aligned}$$

Because  $\bar{x}$  is just a constant multiple of the random variable  $T$ ,

$$\begin{aligned}\mu_{\bar{x}} &= \frac{1}{n}\mu_T = \frac{1}{n}(n\mu) = \mu \\ \sigma_{\bar{x}} &= \frac{1}{n}\sigma_T = \frac{1}{n}(\sigma\sqrt{n}) = \sigma\sqrt{\frac{n}{n^2}} = \sigma\sqrt{\frac{1}{n}} = \sigma\frac{1}{\sqrt{n}} = \frac{\sigma}{\sqrt{n}}\end{aligned}$$

## Sampling from a Normal Population

We have described the mean and standard deviation of the sampling distribution of a sample mean  $\bar{x}$  but not its shape. That's because the shape of the distribution of  $\bar{x}$  depends on the shape of the population distribution. In one important case, there is a simple relationship between the two distributions. The following Activity shows what we mean.

### ACTIVITY

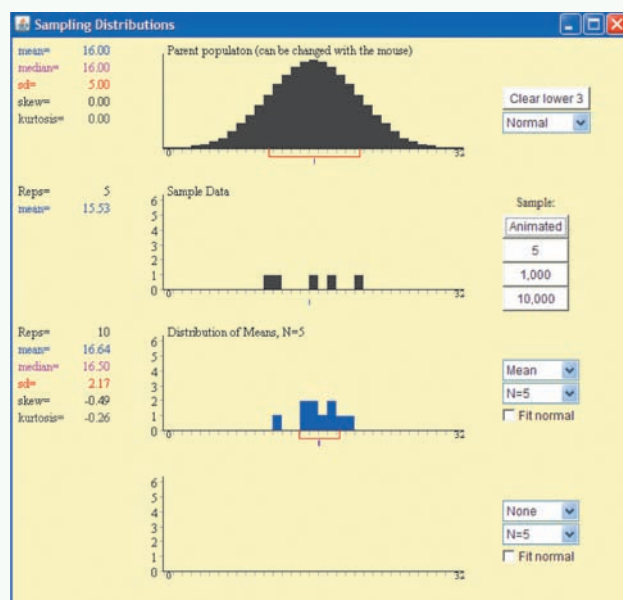
### Exploring the Sampling Distribution of $\bar{x}$ for a Normal Population

#### MATERIALS:

Computer with Internet access—one for the class or one per pair of students

Professor David Lane of Rice University has developed a wonderful applet for investigating the sampling distribution of  $\bar{x}$ . It's dynamic, and it's fun to play with. In this Activity, you'll use Professor Lane's applet to explore the shape of the sampling distribution when the population is Normally distributed.

1. Search for "online statbook sampling distributions applet" and go to the Web site. When the BEGIN button appears on the left side of the screen, click on it. You will then see a yellow page entitled "Sampling Distributions" like the one in the following figure.



2. There are choices for the population distribution: Normal, uniform, skewed, and custom. The default is Normal. Click the “Animated” button. What happens? Click the button several more times. What do the black boxes represent? What is the blue square that drops down onto the plot below? What does the red horizontal band under the population histogram tell us?

Look at the left panel. Important numbers are displayed there. Did you notice that the colors of the numbers match up with the objects to the right? As you make things happen, the numbers change accordingly, like an automatic scorekeeper.

3. Click on “Clear lower 3” to start clean. Then click on the “1,000” button under “Sample:” repeatedly until you have simulated taking 10,000 SRSs of size  $n = 5$  from the population (look for “Reps = 10000” on the left panel in black letters). Answer these questions:

- Does the approximate sampling distribution (blue bars) have a recognizable shape? Click the box next to “Fit normal.”
  - Compare the mean of the approximate sampling distribution with the mean of the population.
  - How is the standard deviation of the approximate sampling distribution related to the standard deviation of the population?
4. Click “Clear lower 3.” Use the drop-down menus to set up the bottom graph to display the mean for samples of size  $n = 20$ . Then sample 10,000 times. How do the two distributions of  $\bar{x}$  compare: shape, center, and spread?
5. What have you learned about the shape of the sampling distribution of  $\bar{x}$  when the population has a Normal shape?

As the previous Activity demonstrates, if the population distribution is Normal, then so is the sampling distribution of  $\bar{x}$ . *This is true no matter what the sample size is.*

### SAMPLING DISTRIBUTION OF A SAMPLE MEAN FROM A NORMAL POPULATION

Suppose that a population is Normally distributed with mean  $\mu$  and standard deviation  $\sigma$ . Then the sampling distribution of  $\bar{x}$  has the Normal distribution with mean  $\mu$  and standard deviation (provided the 10% condition is met)  $\sigma/\sqrt{n}$ .

We already knew the mean and standard deviation of the sampling distribution. All we have added is the Normal shape. Now we have enough information to calculate probabilities involving  $\bar{x}$  when the population distribution is Normal.



## EXAMPLE

### Young Women's Heights

#### Finding probabilities involving the sample mean

**PROBLEM:** The height of young women follows a Normal distribution with mean  $\mu = 64.5$  inches and standard deviation  $\sigma = 2.5$  inches.

- (a) Find the probability that a randomly selected young woman is taller than 66.5 inches. Show your work.
- (b) Find the probability that the mean height of an SRS of 10 young women exceeds 66.5 inches. Show your work.

#### SOLUTION:

(a) **Step 1: State the distribution and the values of interest.** Let  $X$  be the height of a randomly selected young woman. The random variable  $X$  follows a Normal distribution with  $\mu = 64.5$  inches and  $\sigma = 2.5$  inches. We want to find  $P(X > 66.5)$ . Figure 7.18 shows the distribution (purple curve) with the area of interest shaded and the mean, standard deviation, and boundary value labeled.

**Step 2: Perform calculations—show your work!** The standardized score for the boundary value is  $z = \frac{66.5 - 64.5}{2.5} = 0.80$ . Using Table A,  $P(X > 66.5) = P(Z > 0.80) = 1 - 0.7881 = 0.2119$ .

*Using technology:* The command `normalcdf(lower:66.5, upper:10000,  $\mu$ :64.5,  $\sigma$ :2.5)` gives an area of 0.2119.

**Step 3: Answer the question.** The probability of choosing a young woman at random whose height exceeds 66.5 inches is about 0.21.

(b) **Step 1: State the distribution and the values of interest.** For an SRS of 10 young women, the sampling distribution of their sample mean height  $\bar{x}$  will have mean  $\mu_{\bar{x}} = \mu = 64.5$  inches. The 10% condition is met because there are at least  $10(10) = 100$  young women in the population. So the standard deviation is  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.5}{\sqrt{10}} = 0.79$ . Because the population distribution is Normal, the values of  $\bar{x}$  will follow an  $N(64.5, 0.79)$  distribution. We want to find  $P(\bar{x} > 66.5)$  inches. Figure 7.18 shows the distribution (blue curve) with the area of interest shaded and the mean, standard deviation, and boundary value labeled.

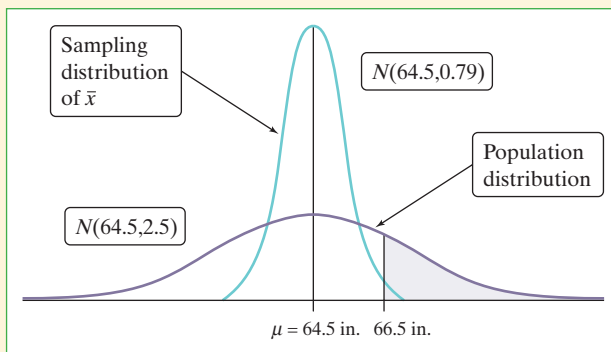
**Step 2: Perform calculations—show your work!** The standardized score for the boundary value is

$$z = \frac{66.5 - 64.5}{0.79} = 2.53.$$

Using Table A,  $P(\bar{x} > 66.5) = P(Z > 2.53) = 1 - 0.9943 = 0.0057$ .

*Using technology:* The command `normalcdf(lower:66.5, upper:10000,  $\mu$ :64.5,  $\sigma$ :0.79)` gives an area of 0.0057.

**Step 3: Answer the question.** It is very unlikely (less than a 1% chance) that we would choose an SRS of 10 young women whose average height exceeds 66.5 inches.



**FIGURE 7.18** The sampling distribution of the mean height  $\bar{x}$  for SRSs of 10 young women compared with the population distribution of young women's heights.

Figure 7.18 compares the population distribution and the sampling distribution of  $\bar{x}$ . It also shows the areas corresponding to the probabilities that we computed. You can see that it is much less likely for the average height of 10 randomly selected young women to exceed 66.5 inches than it is for the height of one randomly selected young woman to exceed 66.5 inches.

The fact that averages of several observations are less variable than individual observations is important in many settings. For example, it is common practice to repeat a measurement several times and report the average of the results. Think of the results of  $n$  repeated measurements as an SRS from the population of outcomes we would get if we repeated the measurement forever. The average of the  $n$  results (the sample mean  $\bar{x}$ ) is less variable than a single measurement.



### CHECK YOUR UNDERSTANDING

The length of human pregnancies from conception to birth varies according to a distribution that is approximately Normal with mean 266 days and standard deviation 16 days.

1. Find the probability that a randomly chosen pregnant woman has a pregnancy that lasts for more than 270 days. Show your work.

Suppose we choose an SRS of 6 pregnant women. Let  $\bar{x}$  = the mean pregnancy length for the sample.

2. What is the mean of the sampling distribution of  $\bar{x}$ ? Explain.
3. Compute the standard deviation of the sampling distribution of  $\bar{x}$ . Check that the 10% condition is met.
4. Find the probability that the mean pregnancy length for the women in the sample exceeds 270 days. Show your work.

## The Central Limit Theorem

Most population distributions are not Normal. The household incomes in Figure 7.17(a) on page 451, for example, are strongly skewed. Yet Figure 7.17(b) suggests that the distribution of means for samples of size 100 is approximately Normal. What is the shape of the sampling distribution of  $\bar{x}$  when the population distribution isn't Normal? The following Activity sheds some light on this question.

### ACTIVITY

### Exploring the Sampling Distribution of $\bar{x}$ for a Non-Normal Population

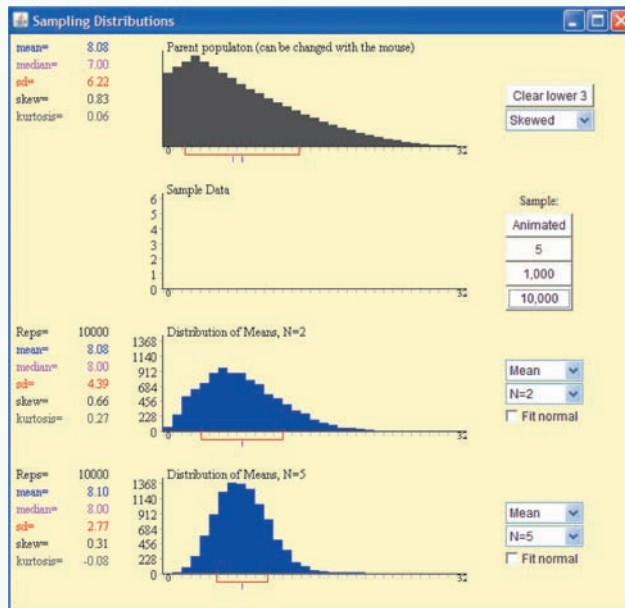
#### MATERIALS:

Computer with Internet access—one for the class or one per pair of students

Let's use the sampling distributions applet from the previous Activity (page 453) to investigate what happens when we start with a non-Normal population distribution.

1. Go to the Web site and launch the applet. Select "Skewed" population. Set the bottom two graphs to display the mean—one for samples of size 2 and the other for samples of size 5. Click the Animated button a few times to be sure you see what's happening. Then "Clear lower 3" and take 10,000 SRSs. Describe what you see.
2. Change the sample sizes to  $n = 10$  and  $n = 16$  and repeat Step 1. What do you notice?





3. Now change the sample sizes to  $n = 20$  and  $n = 25$  and take 10,000 more samples. Did this confirm what you saw in Step 2?

4. Clear the page, and select “Custom” distribution. Click on a point on the population graph to insert a bar of that height. Or click on a point on the horizontal axis, and drag up to define a bar. Make a distribution that looks as strange as you can. (Note: You can shorten a bar or get rid of it completely by clicking on the top of the bar and dragging down to the axis.) Then repeat Steps 1 to 3 for your custom distribution. Cool, huh?

5. Summarize what you learned about the shape of the sampling distribution of  $\bar{x}$ .

It is a remarkable fact that as the sample size increases, the sampling distribution of  $\bar{x}$  changes shape: it looks less like that of the population and more like a Normal distribution. When the sample size is large enough, the sampling distribution of  $\bar{x}$  is very close to Normal. This is true no matter what shape the population distribution has, as long as the population has a finite standard deviation  $\sigma$ . This famous fact of probability theory is called the **central limit theorem** (sometimes abbreviated as CLT).

#### DEFINITION: Central limit theorem (CLT)

Draw an SRS of size  $n$  from any population with mean  $\mu$  and finite standard deviation  $\sigma$ . The **central limit theorem (CLT)** says that when  $n$  is large, the sampling distribution of the sample mean  $\bar{x}$  is approximately Normal.

How large a sample size  $n$  is needed for the sampling distribution of  $\bar{x}$  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. In that case, the sampling distribution of  $\bar{x}$  will also be very non-Normal if the sample size is small. *Be sure you understand what the CLT does—and doesn't—say.*

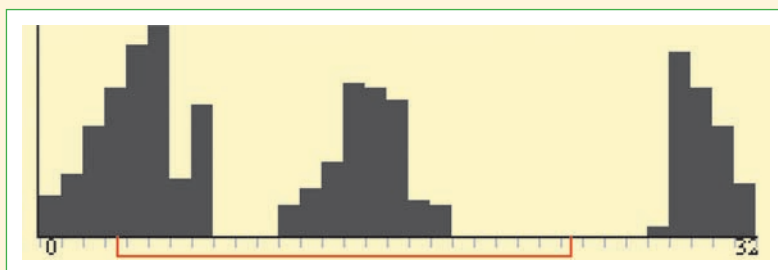


## EXAMPLE

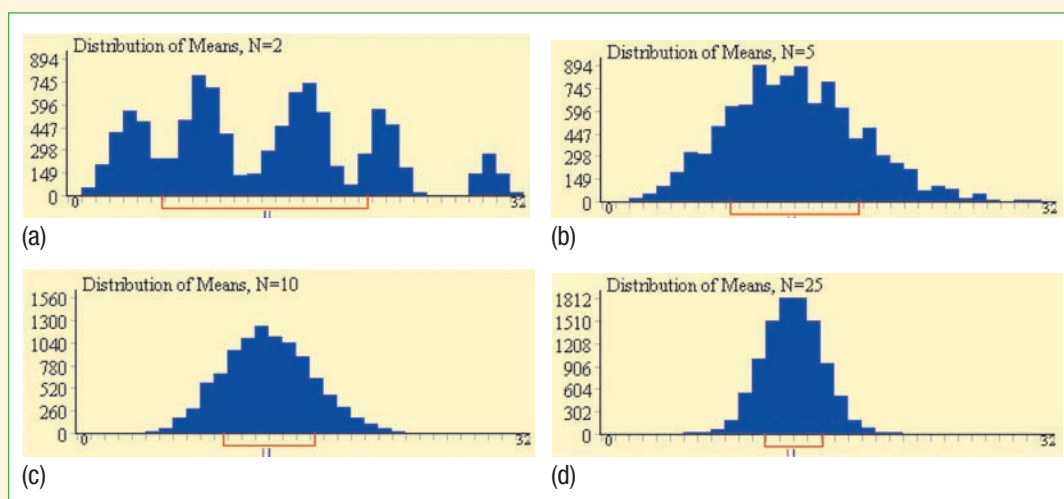
### A Strange Population Distribution

#### The CLT in action

We used the sampling distribution applet to create a population distribution with a very strange shape. See the graph at the top of the next page.



$n = 25$ , the sampling distribution is even more Normal. The contrast between the shapes of the population distribution and the distribution of the mean when  $n = 10$  or  $25$  is striking.



**FIGURE 7.19** The central limit theorem in action: the distribution of sample means  $\bar{x}$  from a strongly non-Normal population becomes more Normal as the sample size increases. (a) The distribution of  $\bar{x}$  for samples of size 2. (b) The distribution of  $\bar{x}$  for samples of size 5. (c) The distribution of  $\bar{x}$  for samples of size 10. (d) The distribution of  $\bar{x}$  for samples of size 25.

As the previous example illustrates, even when the population distribution is very non-Normal, the sampling distribution of  $\bar{x}$  often looks approximately Normal with sample sizes as small as  $n = 25$ . To be safe, we'll require that  $n$  be at least 30 to invoke the CLT. With that issue settled, we can now state the *Normal/Large Sample condition* for sample means.

### NORMAL/LARGE SAMPLE CONDITION FOR SAMPLE MEANS

- If the population distribution is Normal, then so is the sampling distribution of  $\bar{x}$ . This is true no matter what the sample size  $n$  is.
- If the population distribution is not Normal, the central limit theorem tells us that the sampling distribution of  $\bar{x}$  will be approximately Normal in most cases if  $n \geq 30$ .

The central limit theorem allows us to use Normal probability calculations to answer questions about sample means from many observations even when the population distribution is not Normal.



## EXAMPLE

### Servicing Air Conditioners

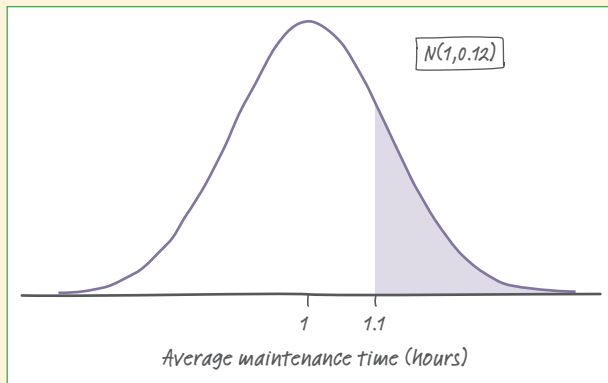
#### Calculations using the CLT

Your company has a contract to perform preventive maintenance on thousands of air-conditioning units in a large city. Based on service records from the past year, the time (in hours) that a technician requires to complete the work follows a strongly right-skewed distribution with  $\mu = 1$  hour and  $\sigma = 1$  hour. In the coming week, your company will service an SRS of 70 air-conditioning units in the city. You plan to budget an average of 1.1 hours per unit for a technician to complete the work. Will this be enough?

**PROBLEM:** What is the probability that the average maintenance time  $\bar{x}$  for 70 units exceeds 1.1 hours? Show your work.

**SOLUTION:**

**Step 1: State the distribution and the values of interest.** The sampling distribution of the sample mean time  $\bar{x}$  spent working on 70 units has



**FIGURE 7.20** The Normal approximation from the central limit theorem for the average time needed to maintain an air conditioner.

- mean  $\mu_{\bar{x}} = \mu = 1$  hour

- standard deviation  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{70}} = \frac{1}{\sqrt{70}} = 0.12$  because the 10% condition is met (there are more than  $10(70) = 700$  air-conditioning units in the population)

- an approximately Normal shape because the Normal/Large Sample condition is met:  $n = 70 \geq 30$

The distribution of  $\bar{x}$  is therefore approximately  $N(1, 0.12)$ . We want to find  $P(\bar{x} > 1.1)$ . Figure 7.20 shows the Normal curve with the area of interest shaded and the mean, standard deviation, and boundary value labeled.

**Step 2: Perform calculations—show your work!** The standardized score for the boundary value is

$$z = \frac{1.1 - 1}{0.12} = 0.83$$

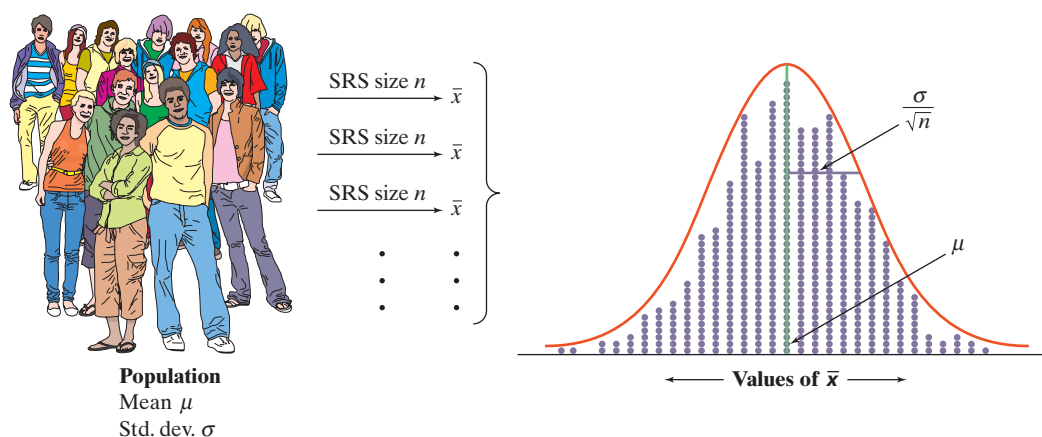
Using Table A,  $P(\bar{x} > 1.1) = P(Z > 0.83) = 1 - 0.7967 = 0.2033$ .

*Using technology:* The command `normalcdf(lower:1.1, upper:10000,  $\mu$ :1,  $\sigma$ :0.12)` gives an area of 0.2023.

**Step 3: Answer the question.** If you budget 1.1 hours per unit, there is about a 20% chance that the technicians will not complete the work within the budgeted time. You will have to decide if this risk is worth taking or if you should schedule more time for the work.

**For Practice** Try Exercise **63**

Figure 7.21 on the next page summarizes the facts about the sampling distribution of  $\bar{x}$ . It reminds us of the big idea of a sampling distribution. Keep taking random samples of size  $n$  from a population with mean  $\mu$ . Find the sample mean  $\bar{x}$  for each sample. Collect all the  $\bar{x}$ 's and display their distribution: the sampling distribution of  $\bar{x}$ . Sampling distributions are the key to understanding statistical inference. Keep this figure in mind for future reference.



**FIGURE 7.21** The sampling distribution of a sample mean  $\bar{x}$  has mean  $\mu$  and standard deviation  $\sigma/\sqrt{n}$ . It has a Normal shape if the population distribution is Normal. If the population distribution isn't Normal, the sampling distribution of  $\bar{x}$  is approximately Normal if the sample size is large enough.

**case closed**

## Building Better Batteries



Refer to the chapter-opening Case Study on page 421. Assuming the process is working properly, the population distribution of battery lifetimes has mean  $\mu = 17$  hours and standard deviation  $\sigma = 0.8$ . We don't know the shape of the population distribution.

1. Make an appropriate graph to display the sample data. Describe what you see.
2. Assume that the battery production process is working properly. Describe the shape, center, and spread of the sampling distribution of  $\bar{x}$  for random samples of 50 batteries. Justify your answers.

For the random sample of 50 batteries, the average lifetime was  $\bar{x} = 16.718$  hours.

3. Find the probability of obtaining a random sample of 50 batteries with a mean lifetime of 16.718 hours or less if the production process is working properly. Show your work. Based on your answer, do you believe that the process is working properly? Why or why not?

The plant manager also wants to know what proportion  $p$  of all the batteries produced that day lasted less than 16.5 hours, which he has declared “unsuitable.” From past experience, about 27% of batteries made at the plant are unsuitable. If the manager does not find convincing evidence that the proportion of unsuitable batteries  $p$  produced that day is greater than 0.27, the whole batch of batteries will be shipped to customers.



4. Assume that the actual proportion of unsuitable batteries produced that day is  $p = 0.27$ . Describe the shape, center, and spread of the sampling distribution of  $\hat{p}$  for random samples of 50 batteries. Justify your answers.

For the random sample of 50 batteries, the sample proportion with lifetimes less than 16.5 hours was  $\hat{p} = 0.32$ .

5. Find the probability of obtaining a random sample of 50 batteries in which 32% or more of the batteries are unsuitable if  $p = 0.27$ . Show your work. Based on your answer, should the entire batch of batteries be shipped to customers? Why or why not?


## Section 7.3

## Summary

- When we want information about the population mean  $\mu$  for some variable, we often take an SRS and use the sample mean  $\bar{x}$  to estimate the unknown parameter  $\mu$ . The **sampling distribution** of  $\bar{x}$  describes how the statistic  $\bar{x}$  varies in *all* possible samples of the same size from the population.
- The **mean** of the sampling distribution is  $\mu$ , so  $\bar{x}$  is an unbiased estimator of  $\mu$ .
- The **standard deviation** of the sampling distribution of  $\bar{x}$  is  $\sigma/\sqrt{n}$  for an SRS of size  $n$  if the population has standard deviation  $\sigma$ . That is, averages are less variable than individual observations. This formula can be used if the population is at least 10 times as large as the sample (*10% condition*).
- Choose an SRS of size  $n$  from a population with mean  $\mu$  and standard deviation  $\sigma$ . If the population distribution is Normal, then so is the sampling distribution of the sample mean  $\bar{x}$ . If the population distribution is not Normal, the **central limit theorem (CLT)** states that when  $n$  is large, the sampling distribution of  $\bar{x}$  is approximately Normal.
- We can use a Normal distribution to calculate approximate probabilities for events involving  $\bar{x}$  whenever the *Normal/Large Sample condition* is met:
  - If the population distribution is Normal, so is the sampling distribution of  $\bar{x}$ .
  - If  $n \geq 30$ , the CLT tells us that the sampling distribution of  $\bar{x}$  will be approximately Normal in most cases.

## Section 7.3

## Exercises

- pg 452  49. **Songs on an iPod** David's iPod has about 10,000 songs. The distribution of the play times for these songs is heavily skewed to the right with a mean of 225 seconds and a standard deviation of 60 seconds.

Suppose we choose an SRS of 10 songs from this population and calculate the mean play time  $\bar{x}$  of these songs. What are the mean and the standard deviation of the sampling distribution of  $\bar{x}$ ? Explain.



50. **Making auto parts** A grinding machine in an auto parts plant prepares axles with a target diameter  $\mu = 40.125$  millimeters (mm). The machine has some variability, so the standard deviation of the diameters is  $\sigma = 0.002$  mm. The machine operator inspects a random sample of 4 axles each hour for quality control purposes and records the sample mean diameter  $\bar{x}$ . Assuming that the process is working properly, what are the mean and standard deviation of the sampling distribution of  $\bar{x}$ ? Explain.
51. **Songs on an iPod** Refer to Exercise 49. How many songs would you need to sample if you wanted the standard deviation of the sampling distribution of  $\bar{x}$  to be 30 seconds? Justify your answer.
52. **Making auto parts** Refer to Exercise 50. How many axles would you need to sample if you wanted the standard deviation of the sampling distribution of  $\bar{x}$  to be 0.0005 mm? Justify your answer.
53. **Larger sample** Suppose that the blood cholesterol level of all men aged 20 to 34 follows the Normal distribution with mean  $\mu = 188$  milligrams per deciliter (mg/dl) and standard deviation  $\sigma = 41$  mg/dl.
- Choose an SRS of 100 men from this population. Describe the sampling distribution of  $\bar{x}$ .
  - Find the probability that  $\bar{x}$  estimates  $\mu$  within  $\pm 3$  mg/dl. (This is the probability that  $\bar{x}$  takes a value between 185 and 191 mg/dl.) Show your work.
  - Choose an SRS of 1000 men from this population. Now what is the probability that  $\bar{x}$  falls within  $\pm 3$  mg/dl of  $\mu$ ? Show your work. In what sense is the larger sample “better”?
54. **Dead battery?** A car company has found that the lifetime of its batteries varies from car to car according to a Normal distribution with mean  $\mu = 48$  months and standard deviation  $\sigma = 8.2$  months. The company installs a new brand of battery on an SRS of 8 cars.
- If the new brand has the same lifetime distribution as the previous type of battery, describe the sampling distribution of the mean lifetime  $\bar{x}$ .
  - The average life of the batteries on these 8 cars turns out to be  $\bar{x} = 42.2$  months. Find the probability that the sample mean lifetime is 42.2 months or less if the lifetime distribution is unchanged. What conclusion would you draw?
55. **Bottling cola** A bottling company uses a filling machine to fill plastic bottles with cola. The bottles are supposed to contain 300 milliliters (ml). In fact, the contents vary according to a Normal distribution with mean  $\mu = 298$  ml and standard deviation  $\sigma = 3$  ml.
- What is the probability that a randomly selected bottle contains less than 295 ml? Show your work.
  - What is the probability that the mean contents of six randomly selected bottles are less than 295 ml? Show your work.
56. **Cereal** A company’s cereal boxes advertise 9.65 ounces of cereal. In fact, the amount of cereal in a randomly selected box follows a Normal distribution with mean  $\mu = 9.70$  ounces and standard deviation  $\sigma = 0.03$  ounces.
- What is the probability that a randomly selected box of the cereal contains less than 9.65 ounces of cereal? Show your work.
  - Now take an SRS of 5 boxes. What is the probability that the mean amount of cereal  $\bar{x}$  in these boxes is 9.65 ounces or less? Show your work.
57. **What does the CLT say?** Asked what the central limit theorem says, a student replies, “As you take larger and larger samples from a population, the histogram of the sample values looks more and more Normal.” Is the student right? Explain your answer.
58. **What does the CLT say?** Asked what the central limit theorem says, a student replies, “As you take larger and larger samples from a population, the spread of the sampling distribution of the sample mean decreases.” Is the student right? Explain your answer.
59. **Songs on an iPod** Refer to Exercise 49.
- Explain why you cannot safely calculate the probability that the mean play time  $\bar{x}$  is more than 4 minutes (240 seconds) for an SRS of 10 songs.
  - Suppose we take an SRS of 36 songs instead. Explain how the central limit theorem allows us to find the probability that the mean play time is more than 240 seconds. Then calculate this probability. Show your work.
60. **Lightning strikes** The number of lightning strikes on a square kilometer of open ground in a year has mean 6 and standard deviation 2.4. The National Lightning Detection Network (NLDN) uses automatic sensors to watch for lightning in a random sample of 10 one-square-kilometer plots of land.
- What are the mean and standard deviation of the sampling distribution of  $\bar{x}$ , the sample mean number of strikes per square kilometer?
  - Explain why you cannot safely calculate the probability that  $\bar{x} < 5$  based on a sample of size 10.
  - Suppose the NLDN takes a random sample of  $n = 50$  square kilometers instead. Explain how the central limit theorem allows us to find the probability that the mean number of lightning strikes per square kilometer is less than 5. Then calculate this probability. Show your work.





- 61. Airline passengers get heavier** In response to the increasing weight of airline passengers, the Federal Aviation Administration (FAA) told airlines to assume that passengers average 190 pounds in the summer, including clothes and carry-on baggage. But passengers vary, and the FAA 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 30 passengers.
- (a) Explain why you cannot calculate the probability that a randomly selected passenger weighs more than 200 pounds.
  - (b) Find the probability that the total weight of 30 randomly selected passengers exceeds 6000 pounds. Show your work. (*Hint:* To apply the central limit theorem, restate the problem in terms of the mean weight.)
- 62. How many people in a car?** A study of rush-hour traffic in San Francisco counts the number of people in each car entering a freeway at a suburban interchange. Suppose that this count has mean 1.5 and standard deviation 0.75 in the population of all cars that enter at this interchange during rush hours.
- (a) Could the exact distribution of the count be Normal? Why or why not?
  - (b) Traffic engineers estimate that the capacity of the interchange is 700 cars per hour. Find the probability that 700 randomly selected cars at this freeway entrance will carry more than 1075 people. Show your work. (*Hint:* Restate this event in terms of the mean number of people  $\bar{x}$  per car.)
- 63. More on insurance** An insurance company claims that in the entire population of homeowners, the mean annual loss from fire is  $\mu = \$250$  and the standard deviation of the loss is  $\sigma = \$1000$ . The distribution of losses is strongly right-skewed: many policies have \$0 loss, but a few have large losses. An auditor examines a random sample of 10,000 of the company's policies. If the company's claim is correct, what's the probability that the average loss from fire in the sample is no greater than \$275? Show your work.
- 64. Bad carpet** The number of flaws per square yard in a type of carpet material varies with mean 1.6 flaws per square yard and standard deviation 1.2 flaws per square yard. The population distribution cannot be Normal, because a count takes only whole-number values. An inspector studies a random sample of 200 square yards of the material, records the number of flaws found in each square yard, and calculates  $\bar{x}$ , the mean number of flaws per square yard inspected. Find the probability that the mean number of flaws exceeds 1.8 per square yard. Show your work.

**Multiple choice:** Select the best answer for Exercises 65 to 68.

- 65.** Scores on the mathematics part of the SAT exam in a recent year were roughly Normal with mean 515 and standard deviation 114. You choose an SRS of 100 students and average their SAT Math scores. Suppose that you do this many, many times. Which of the following are the mean and standard deviation of the sampling distribution of  $\bar{x}$ ?
- (a) Mean = 515, SD = 114
  - (b) Mean = 515, SD =  $114/\sqrt{100}$
  - (c) Mean =  $515/100$ , SD =  $114/100$
  - (d) Mean =  $515/100$ , SD =  $114/\sqrt{100}$
  - (e) Cannot be determined without knowing the 100 scores.
- 66.** Why is it important to check the 10% condition before calculating probabilities involving  $\bar{x}$ ?
- (a) To reduce the variability of the sampling distribution of  $\bar{x}$ .
  - (b) To ensure that the distribution of  $\bar{x}$  is approximately Normal.
  - (c) To ensure that we can generalize the results to a larger population.
  - (d) To ensure that  $\bar{x}$  will be an unbiased estimator of  $\mu$ .
  - (e) To ensure that the observations in the sample are close to independent.
- 67.** A newborn baby has extremely low birth weight (ELBW) if it weighs less than 1000 grams. A study of the health of such children in later years examined a random sample of 219 children. Their mean weight at birth was  $\bar{x} = 810$  grams. This sample mean is an *unbiased estimator* of the mean weight  $\mu$  in the population of all ELBW babies, which means that
- (a) in all possible samples of size 219 from this population, the mean of the values of  $\bar{x}$  will equal 810.
  - (b) in all possible samples of size 219 from this population, the mean of the values of  $\bar{x}$  will equal  $\mu$ .
  - (c) as we take larger and larger samples from this population,  $\bar{x}$  will get closer and closer to  $\mu$ .
  - (d) in all possible samples of size 219 from this population, the values of  $\bar{x}$  will have a distribution that is close to Normal.
  - (e) the person measuring the children's weights does so without any error.



68. The number of hours a lightbulb burns before failing varies from bulb to bulb. The population distribution of burnout times is strongly skewed to the right. The central limit theorem says that
- as we look at more and more bulbs, their average burnout time gets ever closer to the mean  $\mu$  for all bulbs of this type.
  - the average burnout time of a large number of bulbs has a sampling distribution with the same shape (strongly skewed) as the population distribution.
  - the average burnout time of a large number of bulbs has a sampling distribution with similar shape but not as extreme (skewed, but not as strongly) as the population distribution.
  - the average burnout time of a large number of bulbs has a sampling distribution that is close to Normal.
  - the average burnout time of a large number of bulbs has a sampling distribution that is exactly Normal.

*Exercises 69 to 72 refer to the following setting.* In the language of government statistics, you are “in the labor force” if you are available for work and either working or actively seeking work. The unemployment rate is the proportion of the labor force (not of the entire population) who are unemployed. Here are data from the Current Population Survey

for the civilian population aged 25 years and over in a recent year. The table entries are counts in thousands of people.

Highest education	Total population	In labor force	Employed
Didn't finish high school	27,669	12,470	11,408
High school but no college	59,860	37,834	35,857
Less than bachelor's degree	47,556	34,439	32,977
College graduate	51,582	40,390	39,293

69. **Unemployment (1.1)** Find the unemployment rate for people with each level of education. How does the unemployment rate change with education?
70. **Unemployment (5.1)** What is the probability that a randomly chosen person 25 years of age or older is in the labor force? Show your work.
71. **Unemployment (5.3)** If you know that a randomly chosen person 25 years of age or older is a college graduate, what is the probability that he or she is in the labor force? Show your work.
72. **Unemployment (5.3)** Are the events “in the labor force” and “college graduate” independent? Justify your answer.

## FRAPPY! Free Response AP<sup>®</sup> Problem, Yay!

The following problem is modeled after actual AP<sup>®</sup> Statistics exam free response questions. Your task is to generate a complete, concise response in 15 minutes.

*Directions: Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.*

The principal of a large high school is concerned about the number of absences for students at his school. To investigate, he prints a list showing the number of absences during the last month for each of the 2500 students at the school. For this population of students, the distribution of absences last month is skewed to the right with a mean of  $\mu = 1.1$  and a standard deviation of  $\sigma = 1.4$ .

Suppose that a random sample of 50 students is selected from the list printed by the principal and the sample mean number of absences is calculated.

- What is the shape of the sampling distribution of the sample mean? Explain.

- What are the mean and standard deviation of the sampling distribution of the sample mean?
- What is the probability that the mean number of absences in a random sample of 50 students is less than 1?
- Because the population distribution is skewed, the principal is considering using the median number of absences last month instead of the mean number of absences to summarize the distribution. Describe how the principal could use a simulation to estimate the standard deviation of the sampling distribution of the sample median for samples of size 50.

After you finish, you can view two example solutions on the book's Web site ([www.whfreeman.com/tps5e](http://www.whfreeman.com/tps5e)). Determine whether you think each solution is “complete,” “substantial,” “developing,” or “minimal.” If the solution is not complete, what improvements would you suggest to the student who wrote it? Finally, your teacher will provide you with a scoring rubric. Score your response and note what, if anything, you would do differently to improve your own score.

# Chapter Review



## Section 7.1: What Is a Sampling Distribution?

In this section, you learned the “big ideas” of sampling distributions. The first big idea is the difference between a statistic and a parameter. A parameter is a number that describes some characteristic of a population. A statistic estimates the value of a parameter using a sample from the population. Making the distinction between a statistic and a parameter will be crucial throughout the rest of the course.

The second big idea is that statistics vary. For example, the mean weight in a sample of high school students is a variable that will change from sample to sample. This means that statistics have distributions, but parameters do not. The distribution of a statistic in all possible samples of the same size is called the sampling distribution of the statistic.

The third big idea is the distinction between the distribution of the population, the distribution of the sample, and the sampling distribution of a sample statistic. Reviewing the illustration on page 428 will help you understand the difference between these three distributions. When you are writing your answers, be sure to indicate which distribution you are referring to. Don’t make ambiguous statements like “the distribution will become less variable.”

The fourth big idea is how to describe a sampling distribution. To adequately describe a sampling distribution, you need to address shape, center, and spread. If the center (mean) of the sampling distribution is the same as the value of the parameter being estimated, then the statistic is called an unbiased estimator. An estimator is unbiased if it doesn’t consistently under- or overestimate the parameter in many samples. Ideally, the spread of a sampling distribution will be very small, meaning that the statistic provides precise estimates of the parameter. Larger sample sizes result in sampling distributions with smaller spreads.

## Section 7.2: Sample Proportions

In this section, you learned about the shape, center, and spread of the sampling distribution of a sample proportion. When the Large Counts condition ( $np \geq 10$  and  $n(1-p) \geq 10$ ) is met, the sampling distribution of  $\hat{p}$  will

be approximately Normal. The mean of the sampling distribution of  $\hat{p}$  is  $\mu_{\hat{p}} = p$ , the population proportion. As a result, the sample proportion  $\hat{p}$  is an unbiased estimator of the population proportion  $p$ . When the 10% condition ( $n \leq \frac{1}{10}N$ ) is met, the standard deviation of the sampling distribution of the sample proportion is  $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$ . This formula tells us that the variability of the distribution of  $\hat{p}$  is smaller when the sample size is larger.

## Section 7.3: Sample Means

In this section, you learned about the shape, center, and spread of the sampling distribution of a sample mean. When the population is Normal, the sampling distribution of  $\bar{x}$  will also be Normal for any sample size. When the population is not Normal and the sample size is small, the sampling distribution of  $\bar{x}$  will resemble the population shape. However, the central limit theorem says that the sampling distribution of  $\bar{x}$  will become approximately Normal for larger sample sizes (typically when  $n \geq 30$ ), no matter what the population shape. When you are using a Normal distribution to calculate probabilities involving the sampling distribution of  $\bar{x}$ , make sure that the Normal/Large Sample condition is met.

The mean of the sampling distribution of  $\bar{x}$  is  $\mu_{\bar{x}} = \mu$ , the population mean. As a result, the sample mean  $\bar{x}$  is an unbiased estimator of the population mean  $\mu$ . When the 10% condition ( $n \leq \frac{1}{10}N$ ) is met, the standard deviation of the sampling distribution of the sample mean is  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ . This formula tells us that the variability of the distribution of  $\bar{x}$  is smaller when the sample size is larger.

Finally, when you are using a Normal distribution to calculate probabilities involving the sampling distribution of  $\hat{p}$  or  $\bar{x}$ , make sure that you (1) state the distribution and values of interest, (2) perform calculations—show your work, and (3) answer the question.



## What Did You Learn?

Learning Objective	Section	Related Example on Page(s)	Relevant Chapter Review Exercise(s)
Distinguish between a parameter and a statistic.	7.1	425	R7.1
Use the sampling distribution of a statistic to evaluate a claim about a parameter.	7.1	427	R7.5, R7.7
Distinguish among the distribution of a population, the distribution of a sample, and the sampling distribution of a statistic.	7.1	Discussion on 428	R7.2
Determine whether or not a statistic is an unbiased estimator of a population parameter.	7.1	Discussion on 430–431; 435	R7.3
Describe the relationship between sample size and the variability of a statistic.	7.1	432	R7.3
Find the mean and standard deviation of the sampling distribution of a sample proportion $\hat{p}$ . Check the 10% condition before calculating $\sigma_{\hat{p}}$ .	7.2	445	R7.4
Determine if the sampling distribution of $\hat{p}$ is approximately Normal.	7.2	445	R7.4
If appropriate, use a Normal distribution to calculate probabilities involving $\hat{p}$ .	7.2	445	R7.4, R7.5
Find the mean and standard deviation of the sampling distribution of a sample mean $\bar{x}$ . Check the 10% condition before calculating $\sigma_{\bar{x}}$ .	7.3	452	R7.6
Explain how the shape of the sampling distribution of $\bar{x}$ is affected by the shape of the population distribution and the sample size.	7.3	457	R7.6, R7.7
If appropriate, use a Normal distribution to calculate probabilities involving $\bar{x}$ .	7.3	455, 459	R7.6, R7.7

## Chapter 7 Chapter Review Exercises

*These exercises are designed to help you review the important ideas and methods of the chapter.*

**R7.1 Bad eggs** Sale of eggs that are contaminated with salmonella can cause food poisoning in consumers. A large egg producer takes an SRS of 200 eggs from all the eggs shipped in one day. The laboratory reports that 9 of these eggs had salmonella contami-

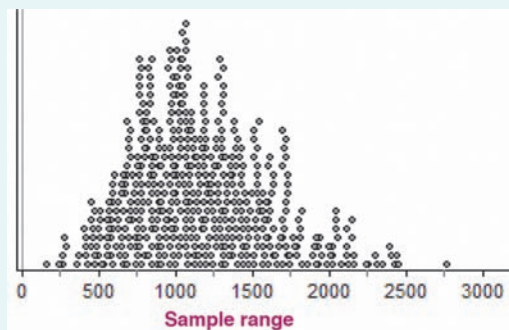
nation. Unknown to the producer, 3% of all eggs shipped had salmonella. Identify the population, the parameter, the sample, and the statistic.

*Exercises R7.2 and R7.3 refer to the following setting.* Researchers in Norway analyzed data on the birth weights of 400,000 newborns over a 6-year period. The distribution of birth weights is approximately Normal





with a mean of 3668 grams and a standard deviation of 511 grams.<sup>8</sup> In this population, the range (maximum – minimum) of birth weights is 3417 grams. We used Fathom software to take 500 SRSs of size  $n = 5$  and calculate the range (maximum – minimum) for each sample. The dotplot below shows the results.



### R7.2 Birth weights

- Sketch a graph that displays the distribution of birth weights for this population.
- Sketch a possible graph of the distribution of birth weights for an SRS of size 5.
- In the graph above, there is a dot at approximately 2750. Explain what this value represents.

### R7.3 Birth weights

- Is the sample range an unbiased estimator of the population range? Give evidence from the graph above to support your answer.
- Explain how we could decrease the variability of the sampling distribution of the sample range.

**R7.4. Do you jog?** The Gallup Poll once asked a random sample of 1540 adults, “Do you happen to jog?” Suppose that the true proportion of all adults who jog is  $p = 0.15$ .

- What is the mean of the sampling distribution of  $\hat{p}$ ? Justify your answer.
- Find the standard deviation of the sampling distribution of  $\hat{p}$ . Check that the 10% condition is met.
- Is the sampling distribution of  $\hat{p}$  approximately Normal? Justify your answer.
- Find the probability that between 13% and 17% of a random sample of 1540 adults are joggers.

**R7.5 Bag check** Thousands of travelers pass through the airport in Guadalajara, Mexico, each day. Before leaving the airport, each passenger must pass through the Customs inspection area. Customs agents want to be sure that passengers do not bring illegal items into the country. But they do not have time to search every traveler’s luggage. Instead, they require each person to press a button. Either a red

or a green bulb lights up. If the red light shows, the passenger will be searched by Customs agents. A green light means “go ahead.” Customs agents claim that the proportion of all travelers who will be stopped (red light) is 0.30, because the light has probability 0.30 of showing red on any push of the button. To test this claim, a concerned citizen watches a random sample of 100 travelers push the button. Only 20 get a red light.

- Assume that the Customs agents’ claim is true. Find the probability that the proportion of travelers who get a red light is as small as or smaller than the result in this sample. Show your work.
- Based on your results in (a), do you believe the Customs agents’ claim? Explain.

**R7.6 IQ tests** The Wechsler Adult Intelligence Scale (WAIS) is a common “IQ test” for adults. The distribution of WAIS scores for persons over 16 years of age is approximately Normal with mean 100 and standard deviation 15.

- What is the probability that a randomly chosen individual has a WAIS score of 105 or higher? Show your work.
- Find the mean and standard deviation of the sampling distribution of the average WAIS score  $\bar{x}$  for an SRS of 60 people.
- What is the probability that the average WAIS score of an SRS of 60 people is 105 or higher? Show your work.
- Would your answers to any of parts (a), (b), or (c) be affected if the distribution of WAIS scores in the adult population were distinctly non-Normal? Explain.

**R7.7 Detecting gypsy moths** The gypsy moth is a serious threat to oak and aspen trees. A state agriculture department places traps throughout the state to detect the moths. Each month, an SRS of 50 traps is inspected, the number of moths in each trap is recorded, and the mean number of moths is calculated. Based on years of data, the distribution of moth counts is discrete and strongly skewed, with a mean of 0.5 and a standard deviation of 0.7.

- Explain why it is reasonable to use a Normal distribution to approximate the sampling distribution of  $\bar{x}$  for SRSs of size 50.
- Estimate the probability that the mean number of moths in a sample of size 50 is greater than or equal to 0.6.
- In a recent month, the mean number of moths in an SRS of size 50 was 0.6. Based on this result, should the state agricultural department be worried that the moth population is getting larger in their state? Explain.

# Chapter 7 AP<sup>®</sup> Statistics Practice Test

## Section I: Multiple Choice *Select the best answer for each question.*

- T7.1** A study of voting chose 663 registered voters at random shortly after an election. Of these, 72% said they had voted in the election. Election records show that only 56% of registered voters voted in the election. Which of the following statements is true about the boldface numbers?
- 72% is a sample; 56% is a population.
  - 72% and 56% are both statistics.
  - 72% is a statistic and 56% is a parameter.
  - 72% is a parameter and 56% is a statistic.
  - 72% and 56% are both parameters.
- T7.2** The Gallup Poll has decided to increase the size of its random sample of voters from about 1500 people to about 4000 people right before an election. The poll is designed to estimate the proportion of voters who favor a new law banning smoking in public buildings. The effect of this increase is to
- reduce the bias of the estimate.
  - increase the bias of the estimate.
  - reduce the variability of the estimate.
  - increase the variability of the estimate.
  - reduce the bias and variability of the estimate.
- T7.3** Suppose we select an SRS of size  $n = 100$  from a large population having proportion  $p$  of successes. Let  $\hat{p}$  be the proportion of successes in the sample. For which value of  $p$  would it be safe to use the Normal approximation to the sampling distribution of  $\hat{p}$ ?
- 0.01
  - 0.09
  - 0.85
  - 0.975
  - 0.999
- T7.4** The central limit theorem is important in statistics because it allows us to use the Normal distribution to find probabilities involving the sample mean
- if the sample size is reasonably large (for any population).
  - if the population is Normally distributed and the sample size is reasonably large.
  - if the population is Normally distributed (for any sample size).
  - if the population is Normally distributed and the population standard deviation is known (for any sample size).
  - if the population size is reasonably large (whether the population distribution is known or not).
- T7.5** The number of undergraduates at Johns Hopkins University is approximately 2000, while the number at Ohio State University is approximately 60,000. At both schools, a simple random sample of about 3% of the undergraduates is taken. Each sample is used to estimate the proportion  $p$  of all students at that university who own an iPod. Suppose that, in fact,  $p = 0.80$  at both schools. Which of the following is the best conclusion?
- The estimate from Johns Hopkins has less sampling variability than that from Ohio State.
  - The estimate from Johns Hopkins has more sampling variability than that from Ohio State.
  - The two estimates have about the same amount of sampling variability.
  - It is impossible to make any statement about the sampling variability of the two estimates because the students surveyed were different.
  - None of the above.
- T7.6** A researcher initially plans to take an SRS of size  $n$  from a population that has mean 80 and standard deviation 20. If he were to double his sample size (to  $2n$ ), the standard deviation of the sampling distribution of the sample mean would be multiplied by
- $\sqrt{2}$ .
  - $1/\sqrt{2}$ .
  - 2.
  - $1/2$ .
  - $1/\sqrt{2n}$ .
- T7.7** The student newspaper at a large university asks an SRS of 250 undergraduates, “Do you favor eliminating the carnival from the term-end celebration?” All in all, 150 of the 250 are in favor. Suppose that (unknown to you) 55% of all undergraduates favor eliminating the carnival. If you took a very large number of SRSs of size  $n = 250$  from this population, the sampling distribution of the sample proportion  $\hat{p}$  would be
- exactly Normal with mean 0.55 and standard deviation 0.03.
  - approximately Normal with mean 0.55 and standard deviation 0.03.
  - exactly Normal with mean 0.60 and standard deviation 0.03.
  - approximately Normal with mean 0.60 and standard deviation 0.03.
  - heavily skewed with mean 0.55 and standard deviation 0.03.



**T7.8** Which of the following statements about the sampling distribution of the sample mean is *incorrect*?

- (a) The standard deviation of the sampling distribution will decrease as the sample size increases.
- (b) The standard deviation of the sampling distribution is a measure of the variability of the sample mean among repeated samples.
- (c) The sample mean is an unbiased estimator of the population mean.
- (d) The sampling distribution shows how the sample mean will vary in repeated samples.
- (e) The sampling distribution shows how the sample was distributed around the sample mean.

**T7.9** A machine is designed to fill 16-ounce bottles of shampoo. When the machine is working properly, the amount poured into the bottles follows a Normal distribution with mean 16.05 ounces and standard deviation 0.1 ounce. Assume that the machine is working properly. If four bottles are randomly selected and the number of ounces in each bottle is measured, then there is about

a 95% chance that the sample mean will fall in which of the following intervals?

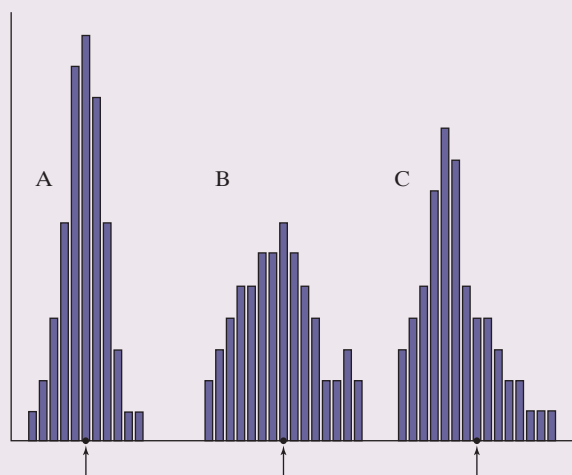
- (a) 16.05 to 16.15 ounces
- (b) 16.00 to 16.10 ounces
- (c) 15.95 to 16.15 ounces
- (d) 15.90 to 16.20 ounces
- (e) 15.85 to 16.25 ounces

**T7.10** Suppose that you are a student aide in the library and agree to be paid according to the “random pay” system. Each week, the librarian flips a coin. If the coin comes up heads, your pay for the week is \$80. If it comes up tails, your pay for the week is \$40. You work for the library for 100 weeks. Suppose we choose an SRS of 2 weeks and calculate your average earnings  $\bar{x}$ . The shape of the sampling distribution of  $\bar{x}$  will be

- (a) Normal.
- (b) approximately Normal.
- (c) right-skewed.
- (d) left-skewed.
- (e) symmetric but not Normal.

**Section II: Free Response** Show all your work. Indicate clearly the methods you use, because you will be graded on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

**T7.11** Below are histograms of the values taken by three sample statistics in several hundred samples from the same population. The true value of the population parameter is marked with an arrow on each histogram.



Which statistic would provide the best estimate of the parameter? Justify your answer.

**T7.12** The amount that households pay service providers for access to the Internet varies quite a bit, but the

mean monthly fee is \$38 and the standard deviation is \$10. The distribution is not Normal: many households pay a base rate for low-speed access, but some pay much more for faster connections. A sample survey asks an SRS of 500 households with Internet access how much they pay. Let  $\bar{x}$  be the mean amount paid.

- (a) Explain why you can't determine the probability that the amount a randomly selected household pays for access to the Internet exceeds \$39.
- (b) What are the mean and standard deviation of the sampling distribution of  $\bar{x}$ ?
- (c) What is the shape of the sampling distribution of  $\bar{x}$ ? Justify your answer.
- (d) Find the probability that the average fee paid by the sample of households exceeds \$39. Show your work.

**T7.13** According to government data, 22% of American children under the age of six live in households with incomes less than the official poverty level. A study of learning in early childhood chooses an SRS of 300 children. Find the probability that more than 20% of the sample are from poverty-level households. Be sure to check that you can use the Normal approximation.

# Cumulative AP<sup>®</sup> Practice Test 2

## Section I: Multiple Choice Choose the best answer for each question.

**AP2.1** The five-number summary for a data set is given by  $\min = 5$ ,  $Q_1 = 18$ ,  $\text{median} = 20$ ,  $Q_3 = 40$ ,  $\max = 75$ . If you wanted to construct a boxplot for the data set (that is, one that would show outliers, if any existed), what would be the maximum possible length of the right-side “whisker”?

- (a) 33 (b) 35 (c) 45 (d) 53 (e) 55

**AP2.2** The probability distribution for the number of heads in four tosses of a coin is given by

<b>Number of heads:</b>	0	1	2	3	4
<b>Probability:</b>	0.0625	0.2500	0.3750	0.2500	0.0625

The probability of getting at least one *tail* in four tosses of a coin is

- (a) 0.2500. (c) 0.6875. (e) 0.0625.  
(b) 0.3125. (d) 0.9375.

**AP2.3** In a certain large population of adults, the distribution of IQ scores is strongly left-skewed with a mean of 122 and a standard deviation of 5. Suppose 200 adults are randomly selected from this population for a market research study. The distribution of the sample mean of IQ scores is

- (a) left-skewed with mean 122 and standard deviation 0.35.  
(b) exactly Normal with mean 122 and standard deviation 5.  
(c) exactly Normal with mean 122 and standard deviation 0.35.  
(d) approximately Normal with mean 122 and standard deviation 5.  
(e) approximately Normal with mean 122 and standard deviation 0.35.

**AP2.4** A 10-question multiple-choice exam offers 5 choices for each question. Jason just guesses the answers, so he has probability  $1/5$  of getting any one answer correct. You want to perform a simulation to determine the number of correct answers that Jason gets. One correct way to use a table of random digits to do this is the following:

- (a) One digit from the random digit table simulates one answer, with 5 = right and all other digits = wrong. Ten digits from the table simulate 10 answers.  
(b) One digit from the random digit table simulates one answer, with 0 or 1 = right and all other digits = wrong. Ten digits from the table simulate 10 answers.  
(c) One digit from the random digit table simulates one answer, with odd = right and even = wrong. Ten digits from the table simulate 10 answers.

- (d) One digit from the random digit table simulates one answer, with 0 or 1 = right and all other digits = wrong, ignoring repeats. Ten digits from the table simulate 10 answers.  
(e) Two digits from the random digit table simulate one answer, with 00 to 20 = right and 21 to 99 = wrong. Ten pairs of digits from the table simulate 10 answers.

**AP2.5** Suppose we roll a fair die four times. The probability that a 6 occurs on exactly one of the rolls is

- (a)  $4\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^1$  (c)  $4\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^3$  (e)  $6\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^3$   
(b)  $\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^1$  (d)  $\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^3$

**AP2.6** You want to take an SRS of 50 of the 816 students who live in a dormitory on a college campus. You label the students 001 to 816 in alphabetical order. In the table of random digits, you read the entries

95592 94007 69769 33547 72450 16632 81194 14873

The first three students in your sample have labels

- (a) 955, 929, 400. (d) 929, 400, 769.  
(b) 400, 769, 769. (e) 400, 769, 335.  
(c) 559, 294, 007.

**AP2.7** The number of unbroken charcoal briquets in a 20-pound bag filled at the factory follows a Normal distribution with a mean of 450 briquets and a standard deviation of 20 briquets. The company expects that a certain number of the bags will be underfilled, so the company will replace for free the 5% of bags that have too few briquets. What is the minimum number of unbroken briquets the bag would have to contain for the company to avoid having to replace the bag for free?

- (a) 404 (b) 411 (c) 418 (d) 425 (e) 448

**AP2.8** You work for an advertising agency that is preparing a new television commercial to appeal to women. You have been asked to design an experiment to compare the effectiveness of three versions of the commercial. Each subject will be shown one of the three versions and then asked about her attitude toward the product. You think there may be large differences between women who are employed and those who are not. Because of these differences, you should use

- (a) a block design, but not a matched pairs design.  
(b) a completely randomized design.  
(c) a matched pairs design.



- (d) a simple random sample.
- (e) a stratified random sample.

**AP2.9** Suppose that you have torn a tendon and are facing surgery to repair it. The orthopedic surgeon explains the risks to you. Infection occurs in 3% of such operations, the repair fails in 14%, and both infection and failure occur together 1% of the time. What is the probability that the operation is successful for someone who has an operation that is free from infection?

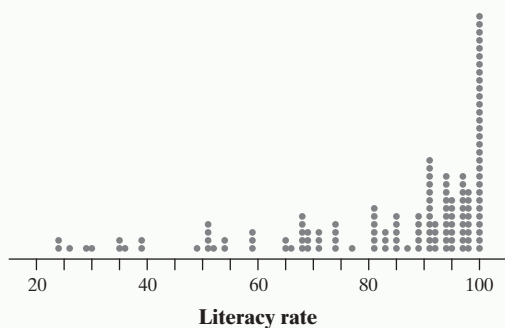
- (a) 0.8342                      (c) 0.8600                      (e) 0.9900
- (b) 0.8400                      (d) 0.8660

**AP2.10** Social scientists are interested in the association between high school graduation rate (HSGR, measured as a percent) and the percent of U.S. families living in poverty (POV). Data were collected from all 50 states and the District of Columbia, and a regression analysis was conducted.

The resulting least-squares regression line is given by  $\widehat{\text{POV}} = 59.2 - 0.620(\text{HSGR})$  with  $r^2 = 0.802$ . Based on the information, which of the following is the best interpretation for the slope of the least-squares regression line?

- (a) For each 1% increase in the graduation rate, the percent of families living in poverty is predicted to decrease by approximately 0.896.
- (b) For each 1% increase in the graduation rate, the percent of families living in poverty is predicted to decrease by approximately 0.802.
- (c) For each 1% increase in the graduation rate, the percent of families living in poverty is predicted to decrease by approximately 0.620.
- (d) For each 1% increase in the percent of families living in poverty, the graduation rate is predicted to increase by approximately 0.802.
- (e) For each 1% increase in the percent of families living in poverty, the graduation rate is predicted to decrease by approximately 0.620.

Here is a dotplot of the adult literacy rates in 177 countries in a recent year, according to the United Nations. For example, the lowest literacy rate was 23.6%, in the African country of Burkina Faso. Mali had the next lowest literacy rate at 24.0%. Use the graph to answer Questions AP2.11 to AP2.13.



**AP2.11** The overall shape of this distribution is

- (a) clearly skewed to the right.
- (b) clearly skewed to the left.
- (c) roughly symmetric.
- (d) uniform.
- (e) There is no clear shape.

**AP2.12** The mean of this distribution (*don't try to find it*) will be

- (a) very close to the median.
- (b) greater than the median.
- (c) less than the median.
- (d) You can't say, because distribution isn't symmetric.
- (e) You can't say, because the distribution isn't Normal.

**AP2.13** Based on the shape of this distribution, what measures of center and spread would be most appropriate to report?

- (a) The mean and standard deviation
- (b) The mean and the interquartile range
- (c) The median and the standard deviation
- (d) The median and the interquartile range
- (e) The mean and the range

**AP2.14** The correlation between the age and height of children under the age of 12 is found to be  $r = 0.60$ . Suppose we use the age  $x$  of a child to predict the height  $y$  of the child. What can we conclude?

- (a) The height is generally 60% of a child's weight.
- (b) About 60% of the time, age will accurately predict height.
- (c) Thirty-six percent of the variation in height is accounted for by the linear model relating height to age.
- (d) For every 1 year older a child is, the regression line predicts an increase of 0.6 feet in height.
- (e) Thirty-six percent of the time, the least-squares regression line accurately predicts height from age.

**AP2.15** An agronomist wants to test three different types of fertilizer (A, B, and C) on the yield of a new variety of wheat. The yield will be measured in bushels per acre. Six 1-acre plots of land were randomly assigned to each of the three fertilizers. The treatment, experimental unit, and response variable are, respectively,

- (a) a specific fertilizer, bushels per acre, a plot of land.
- (b) a plot of land, bushels per acre, a specific fertilizer.
- (c) random assignment, a plot of land, wheat yield.
- (d) a specific fertilizer, a plot of land, wheat yield.
- (e) a specific fertilizer, the agronomist, wheat yield.

**AP2.16** According to the U.S. Census, the proportion of adults in a certain county who owned their own home was 0.71. An SRS of 100 adults in a certain



section of the county found that 65 owned their home. Which one of the following represents the approximate probability of obtaining a sample of 100 adults in which fewer than 65 own their home, assuming that this section of the county has the same overall proportion of adults who own their home as does the entire county?

- (a)  $\binom{100}{65} (0.71)^{65} (0.29)^{35}$  (d)  $P\left(Z < \frac{0.65 - 0.71}{\sqrt{\frac{(0.65)(0.35)}{100}}}\right)$
- (b)  $\binom{100}{65} (0.29)^{65} (0.71)^{35}$  (e)  $P\left(Z < \frac{0.65 - 0.71}{\sqrt{\frac{(0.71)(0.29)}{100}}}\right)$
- (c)  $P\left(Z < \frac{0.65 - 0.71}{\sqrt{\frac{(0.71)(0.29)}{100}}}\right)$

**AP2.17** Which one of the following would be a correct interpretation if you have a  $z$ -score of +2.0 on an exam?

- (a) It means that you missed two questions on the exam.
- (b) It means that you got twice as many questions correct as the average student.
- (c) It means that your grade was 2 points higher than the mean grade on this exam.
- (d) It means that your grade was in the upper 2% of all grades on this exam.
- (e) It means that your grade is 2 standard deviations above the mean for this exam.

**AP2.18** Records from a random sample of dairy farms yielded the information below on the number of male and female calves born at various times of the day.

	Day	Evening	Night	Total
Males	129	15	117	261
Females	118	18	116	252
<b>Total</b>	<b>247</b>	<b>33</b>	<b>233</b>	<b>513</b>

What is the probability that a randomly selected calf was born in the night or was a female?

- (a)  $\frac{369}{513}$  (b)  $\frac{485}{513}$  (c)  $\frac{116}{513}$  (d)  $\frac{116}{252}$  (e)  $\frac{116}{233}$

**AP2.19** When people order books from a popular online source, they are shipped in standard-sized boxes. Suppose that the mean weight of the boxes is 1.5 pounds with a standard deviation of 0.3 pounds, the mean weight of the packing material is 0.5 pounds with a standard deviation of 0.1 pounds, and the mean weight of the books shipped is 12 pounds with a standard deviation of 3 pounds.

Assuming that the weights are independent, what is the standard deviation of the total weight of the boxes that are shipped from this source?

- (a) 1.84 (c) 3.02 (e) 9.10
- (b) 2.60 (d) 3.40

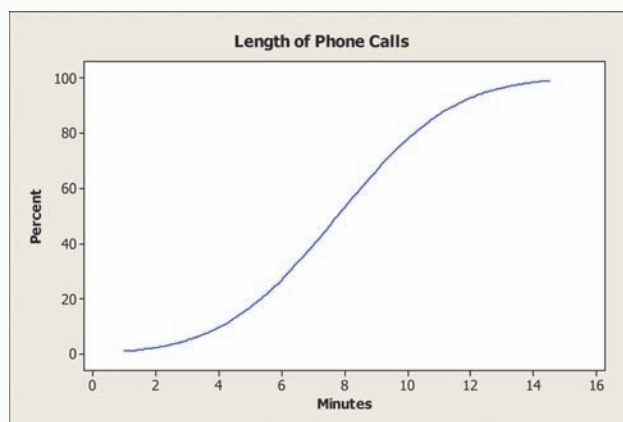
**AP2.20** A grocery chain runs a prize game by giving each customer a ticket that may win a prize when the box is scratched off. Printed on the ticket is a dollar value (\$500, \$100, \$25) or the statement "This ticket is not a winner." Monetary prizes can be redeemed for groceries at the store. Here is the probability distribution of the amount won on a randomly selected ticket:

<b>Amount won:</b>	\$500	\$100	\$25	\$0
<b>Probability:</b>	0.01	0.05	0.20	0.74

Which of the following are the mean and standard deviation, respectively, of the winnings?

- (a) \$15.00, \$2900.00
- (b) \$15.00, \$53.85
- (c) \$15.00, \$26.93
- (d) \$156.25, \$53.85
- (e) \$156.25, \$26.93

**AP2.21** A large company is interested in improving the efficiency of its customer service and decides to examine the length of the business phone calls made to clients by its sales staff. A cumulative relative frequency graph is shown below from data collected over the past year. According to the graph, the shortest 80% of calls will take how long to complete?



- (a) Less than 10 minutes
- (b) At least 10 minutes
- (c) Exactly 10 minutes
- (d) At least 5.5 minutes
- (e) Less than 5.5 minutes



**Section II: Free Response** Show all your work. Indicate clearly the methods you use, because you will be graded on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

**AP2.22** A health worker is interested in determining if omega-3 fish oil can help reduce cholesterol in adults. She obtains permission to examine the health records of 200 people in a large medical clinic and classifies them according to whether or not they take omega-3 fish oil. She also obtains their latest cholesterol readings and finds that the mean cholesterol reading for those who are taking omega-3 fish oil is 18 points lower than the mean for the group not taking omega-3 fish oil.

- Is this an observational study or an experiment? Justify your answer.
- Explain the concept of confounding in the context of this study and give one example of a variable that could be confounded with whether or not people take omega-3 fish oil.
- Researchers find that the 18-point difference in the mean cholesterol readings of the two groups is statistically significant. Can they conclude that omega-3 fish oil is the cause? Why or why not?

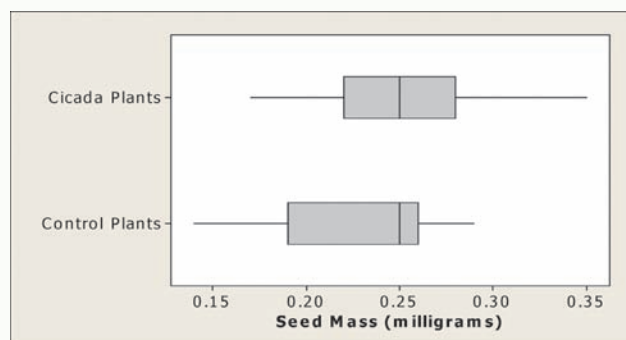
**AP2.23** There are four major blood types in humans: O, A, B, and AB. In a study conducted using blood specimens from the Blood Bank of Hawaii, individuals were classified according to blood type and ethnic group. The ethnic groups were Hawaiian, Hawaiian-White, Hawaiian-Chinese, and White. Suppose that a blood bank specimen is selected at random.

Blood type	Ethnic Group				Total
	Hawaiians	Hawaiian-White	Hawaiian-Chinese	White	
O	1903	4469	2206	53,759	<b>62,337</b>
A	2490	4671	2368	50,008	<b>59,537</b>
B	178	606	568	16,252	<b>17,604</b>
AB	99	236	243	5001	<b>5579</b>
<b>Total</b>	<b>4670</b>	<b>9982</b>	<b>5385</b>	<b>125,020</b>	<b>145,057</b>

- Find the probability that the specimen contains type O blood or comes from the Hawaiian-Chinese ethnic group. Show your work.
- What is the probability that the specimen contains type AB blood, given that it comes from the Hawaiian ethnic group? Show your work.
- Are the events “type B blood” and “Hawaiian ethnic group” independent? Give appropriate statistical evidence to support your answer.
- Now suppose that two blood bank specimens are selected at random. Find the probability that at

least one of the specimens contains type A blood from the White ethnic group.

**AP2.24** Every 17 years, swarms of cicadas emerge from the ground in the eastern United States, live for about six weeks, and then die. (There are several different “broods,” so we experience cicada eruptions more often than every 17 years.) There are so many cicadas that their dead bodies can serve as fertilizer and increase plant growth. In a study, a researcher added 10 dead cicadas under 39 randomly selected plants in a natural plot of American bellflowers on the forest floor, leaving other plants undisturbed. One of the response variables measured was the size of seeds produced by the plants. Here are the boxplots and summary statistics of seed mass (in milligrams) for 39 cicada plants and 33 undisturbed (control) plants:



Variable:	n	Minimum	Q <sub>1</sub>	Median	Q <sub>3</sub>	Maximum
Cicada plants:	39	0.17	0.22	0.25	0.28	0.35
Control plants:	33	0.14	0.19	0.25	0.26	0.29

- Write a few sentences comparing the distributions of seed mass for the two groups of plants.
- Based on the graphical displays, which distribution has the larger mean? Justify your answer.
- Explain the purpose of the random assignment in this study.
- Name one benefit and one drawback of only using American bellflowers in the study.

**AP2.25** In a city library, the mean number of pages in a novel is 525 with a standard deviation of 200. Approximately 30% of the novels have fewer than 400 pages. Suppose that you randomly select 50 novels from the library.

- What is the probability that the total number of pages is fewer than 25,000? Show your work.
- What is the probability that at least 20 of the novels have fewer than 400 pages? Show your work.