UNIT 5

Sampling Distributions

Chapter 7

Sampling Distributions

Stefano Paterna/Alamy

INTRODUCTION

In this chapter, we will return to a key idea about statistical inference from Chapter 4—making conclusions about a population based on data from a sample. Here are a few examples of statistical inference in practice:

- 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.
- To estimate how much gasoline prices vary in a large city, a reporter records the price per gallon of regular unleaded gasoline at a random sample of 10 gas stations in the city. 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?
- A battery manufacturer wants to make sure that the AA batteries it produces each hour meet certain standards. Quality control inspectors collect data from a random sample of 50 AA batteries produced during one hour and use the sample mean lifetime $\bar{x}$ to estimate the unknown population mean lifetime μ for all batteries produced that hour.

Let's look at the battery example a little more closely. To make an inference about the batteries produced in the given hour, we need to know how close the sample mean $\bar{x}$ is likely to be to the population mean μ. 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.

The following activity will help you get a feel for the distribution of two very common statistics, the sample mean $\bar{x}$ and the sample proportion $\hat{p}$.

ACTIVITY A penny for your thoughts?

peterspiro/Getty Images

In this activity, your class will investigate how the mean year $\bar{x}$ and the proportion of pennies from the 2000s $\hat{p}$ vary from sample to sample, using a large population of pennies of various ages.[1]

1. Each member of the class should randomly select 1 penny from the population and record the year of the penny with an "X" on the dotplot provided by your teacher. Return the penny to the population. Repeat this process until at least 100 pennies have been selected and recorded. This graph gives you an idea of what the population distribution of penny years looks like.
2. Each member of the class should then select an SRS of 5 pennies from the population and note the year on each penny.
   - Record the average year of these 5 pennies (rounded to the nearest year) with an "$\bar{x}$" on a new class dotplot. Make sure this dotplot is on the same scale as the dotplot in Step 1.
   - Record the proportion of pennies from the 2000s with a "$\hat{p}$" on a different dotplot provided by your teacher.

Return the pennies to the population. Repeat this process until there are at least 100 $\bar{x}$'s and 100 $\hat{p}$'s.

3. Repeat Step 2 with SRSs of size $n = 20$. Make sure these dotplots are on the same scale as the corresponding dotplots from Step 2.
4. Compare the distribution of X (year of penny) with the two distributions of $\bar{x}$ (mean year). How are the distributions similar? How are they different? What effect does sample size seem to have on the shape, center, and variability of the distribution of $\bar{x}$?
5. Compare the two distributions of $\hat{p}$. How are the distributions similar? How are they different? What effect does sample size seem to have on the shape, center, and variability of the distribution of $\hat{p}$?

Sampling distributions are the foundation of 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 many times?" If our data come from random sampling, the laws of probability help us answer this question. These laws also allow us to determine how far our estimates typically vary from the truth and what values of a statistic should be considered unusual.

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 involving proportions. Section 7.3 investigates sampling distributions involving means.

SECTION 7.1 What Is a Sampling Distribution?

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

- Distinguish between a parameter and a statistic.
- Create a sampling distribution using all possible samples from a small population.
- 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 if 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 U.S. residents with a college degree? Each March, the government's Current Population Survey (CPS) asks detailed questions about income. The random sample of about 70,000 U.S. college grads contacted in March 2016 had a mean "total money income" of \$73,750 in 2015.[2] That \$73,750 describes the sample, but we use it to estimate the mean income of all college grads in the United States.

Because of some very large incomes, the mean total income (\$73,750) was much larger than the median total income (\$55,071).

Parameters and Statistics

As we begin to use sample data to draw conclusions about a larger population, we must be clear about whether a number describes a sample or a population. For the sample of college graduates contacted by the CPS, the mean income was $\bar{x}$ = \$73,750. The number \$73,750 is a **statistic** because it describes this one CPS sample. The population that the poll wants to draw conclusions about is the nearly 100 million U.S. residents with a college degree. In this case, the **parameter** of interest is the mean income μ of all these college graduates. We don't know the value of this parameter, but we can estimate it using data from the sample.

A sample statistic is sometimes called a *point estimator* of the corresponding population parameter because the estimate—\$73,750 in this case—is a single point on the number line.

DEFINITION **Statistic, Parameter**

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

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

Recall our hint from Chapter 1 about **s** and **p**: **s**tatistics come from **s**amples, and **p**arameters come from **p**opulations. As long as we were doing data analysis, the distinction between population and sample rarely came up. Now that we are focusing on statistical inference, however, it is essential. The notation we use should reflect this distinction. The table shows three commonly used statistics and their corresponding parameters.

It is common practice to use Greek letters for parameters and Roman letters for statistics. In that case, the population proportion would be π (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® exam, however.

Sample statistic		Population parameter
$\bar{x}$ (the sample mean)	estimates	μ (the population mean)
$\hat{p}$ (the sample proportion)	estimates	p (the population proportion)
s_x (the sample SD)	estimates	σ (the population SD)

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 515 randomly selected U.S. adults if they believe in ghosts. Of the respondents, 160 said "Yes."[3]

(b) During the winter months, the temperatures outside the Starneses' cabin in Colorado can stay well below freezing 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.

RyersonClark/Getty Images

SOLUTION:

(a) *Population: all U.S. adults. Parameter: p = the proportion of all U.S. adults who believe in ghosts. Sample: the 515 people who were interviewed in this Gallup Poll. Statistic: $\hat{p}$ = the proportion in the sample who say they believe in ghosts = 160/515 = 0.31.*

(b) *Population: all times during the day in question. Parameter: the true minimum temperature in the cabin at all times that day. Sample: the 20 randomly selected times. Statistic: the sample minimum temperature, 38°F.*

Not all parameters and statistics have their own symbols. To distinguish parameters and statistics in these cases, use descriptors like "true" and "sample."

FOR PRACTICE, TRY EXERCISE 1

AP® EXAM TIP

Many students lose credit on the AP® Statistics exam when defining parameters because their description refers to the sample instead of the population or because the description isn't clear about which group of individuals the parameter is describing. When defining a parameter, we suggest including the word *all* or the word *true* in your description to make it clear that you aren't referring to a sample statistic.

The Idea of a Sampling Distribution

The students in Mrs. Gallas's class did the "Penny for your thoughts" activity at the beginning of the chapter. Figure 7.1 shows their "dotplot" of the sample mean year for 50 samples of size $n = 5$.

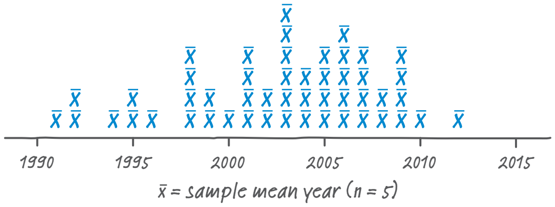


FIGURE 7.1 Distribution of the sample mean year of penny for 50 samples of size $n = 5$ from Mrs. Gallas's population of pennies.

It shouldn't be surprising that the statistic $\bar{x}$ is a random variable. After all, different samples of $n = 5$ pennies will produce different means. As you learned in Section 4.3, this basic fact is called **sampling variability**.

DEFINITION Sampling variability

Sampling variability refers to the fact that different random samples of the same size from the same population produce different values for a statistic.

Knowing how statistics vary from sample to sample is essential when making an inference about a population. Understanding sampling variability reminds us that the value of a statistic is unlikely to be exactly equal to the value of the parameter it is trying to estimate. It also lets us say how much we expect an estimate to vary from its corresponding parameter.

Mrs. Gallas's class took only 50 random samples of 5 pennies. However, there are many, many possible random samples of size 5 from Mrs. Gallas's large population of pennies. If the students took every one of those possible samples, calculated the value of $\bar{x}$ for each, and graphed all those $\bar{x}$ values, then we'd have a **sampling distribution.**

Remember that a distribution describes the possible values of a variable and how often these values occur. Thus, a sampling distribution shows the possible values of a *statistic* and how often these values occur.

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.

For large populations, it is too difficult to take all possible samples of size n to obtain the exact sampling distribution of a statistic. Instead, we can approximate a sampling distribution by taking many samples, calculating the value of the statistic for each of these samples, and graphing the results. Because the students in Mrs. Gallas's class didn't take all possible samples of 5 pennies, their dotplot of $\bar{x}$'s in Figure 7.1 is called an *approximate sampling distribution.*

The following example demonstrates how to construct a complete sampling distribution using a small population.

EXAMPLE Sampling heights

Creating a sampling distribution

Brian Miller

PROBLEM: John and Carol have four grown sons. Their heights (in inches) are 71, 75, 72, and 68.

(a) List all 6 possible samples of size 2.

(b) Calculate the mean of each sample and display the sampling distribution of the sample mean using a dotplot.

(c) Calculate the range of each sample and display the sampling distribution of the sample range using a dotplot.

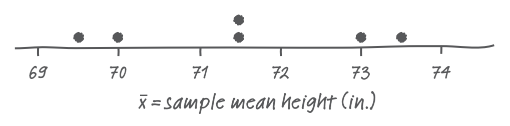
SOLUTION:

(a)

Sample
71, 75
71, 72
71, 68
75, 72
75, 68
72, 68

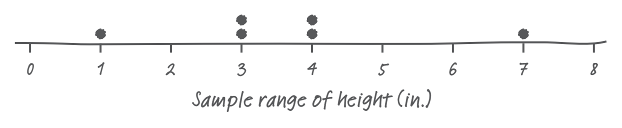
(b)

Sample	Sample mean
71, 75	73.0
71, 72	71.5
71, 68	69.5
75, 72	73.5
75, 68	71.5
72, 68	70.0



(c)

Sample	Sample range
71, 75	4
71, 72	1
71, 68	3
75, 72	3
75, 68	7
72, 68	4



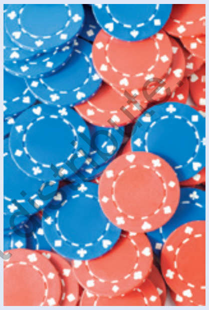
FOR PRACTICE, TRY EXERCISE 7

Being able to construct (or approximate) the sampling distribution of a statistic allows us to determine the values of the statistic that are likely to occur by chance alone—and the values that should be considered unusual. The following example shows how we can use a sampling distribution to evaluate a claim.

EXAMPLE Reaching for chips

Using a sampling distribution to evaluate a claim

PROBLEM: To determine how much homework time students will get in class, Mrs. Lin has a student select an SRS of 20 chips from a large bag. The number of red chips in the SRS determines the number of minutes in class students get to work on homework. Mrs. Lin claims that there are 200 chips in the bag and that 100 of them are red. When Jenna selected a random sample of 20 chips from the bag (without looking), she got 7 red chips. Does this provide convincing evidence that less than half of the chips in the bag are red?

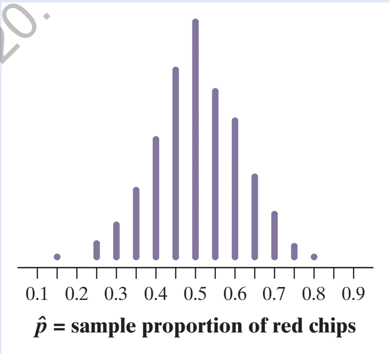


woodygraphs/Getty Images

(a) What is the evidence that less than half of the chips in the bag are red?

(b) Provide two explanations for the evidence described in part (a).

We used technology to simulate choosing 500 SRSs of size $n = 20$ from a population of 200 chips, 100 red and 100 blue. The dotplot shows $\hat{p}$ = the sample proportion of red chips for each of the 500 samples.



(c) There is one dot on the graph at 0.80. Explain what this value represents.

(d) Would it be surprising to get a sample proportion of $\hat{p} = 7/20 = 0.35$ or smaller in an SRS of size 20 when $p = 0.5$? Justify your answer.

(e) Based on your previous answers, is there convincing evidence that less than half of the chips in the large bag are red? Explain your reasoning.

SOLUTION:

(a) *Jenna's sample proportion was $\hat{p} = 7/20 = 0.35$, which is less than 0.50.*

(b) *It is possible that Mrs. Lin is telling the truth and Jenna got a $\hat{p}$ less than 0.50 because of sampling variability. It is also possible that Mrs. Lin is lying and less than half of the chips in the bag are red.*

(c) *In one simulated SRS of 20 chips, there were 16 red chips. So $\hat{p} = 16/20 = 0.80$ for this sample.*

(d) *No; there were many simulated samples that had $\hat{p}$ values less than or equal to 0.35.*

(e) *Because it isn't surprising to get a $\hat{p}$ less than or equal to 0.35 by chance alone when $p = 0.50$, there isn't convincing evidence that less than half of the chips in the bag are red.*

FOR PRACTICE, TRY EXERCISE 13

When we simulate a sampling distribution using assumed values for the parameters, like in the chips example, the resulting distribution is sometimes called a *randomization distribution.*

Suppose that Jenna's sample included only 3 red chips, giving $\hat{p} = 3/20 = 0.15$. Would this provide convincing evidence that less than half of the chips in the bag are red? Yes. According to the simulated sampling distribution in the example, it would be very unusual to get a $\hat{p}$ value this small when $p = 0.50$. Therefore, sampling variability would not be a plausible explanation for the outcome of Jenna's sample. The only plausible explanation for a $\hat{p}$ value of 0.15 is that less than half of the chips in the bag are red.

Figure 7.2 (on the next page) illustrates the process of choosing many random samples of 20 chips from a population of 100 red chips and 100 blue chips and finding the sample proportion of red chips $\hat{p}$ for each sample. Follow the flow of the figure from the population distribution on the left, to choosing an SRS, graphing the distribution of sample data, and finding the $\hat{p}$ for that particular 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 is the *approximate sampling distribution* of the statistic $\hat{p}$.

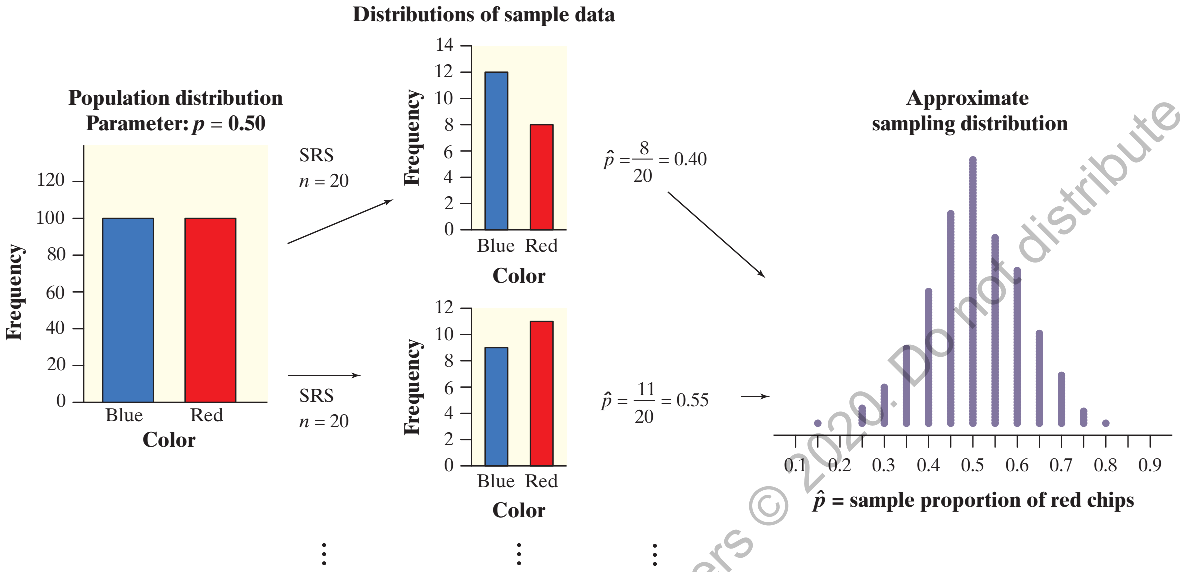


FIGURE 7.2 The idea of a sampling distribution is to take many samples from the same population, collect the value of the statistic from all the samples, and display the distribution of the statistic. The dotplot shows the approximate sampling distribution of $\hat{p} =$ the sample proportion of red chips.

AP® EXAM TIP

Terminology matters. Never just say "the distribution." Always say "the distribution of [blank]," being careful to distinguish the distribution of the population, the distribution of sample data, and the sampling distribution of a statistic. Likewise, don't use ambiguous terms like "sample distribution," which could refer to the distribution of sample data or to the sampling distribution of a statistic. You will lose credit on free response questions for misusing statistical terms.

As Figure 7.2 shows, there are three distinct distributions involved when we sample repeatedly and calculate the value of a statistic.

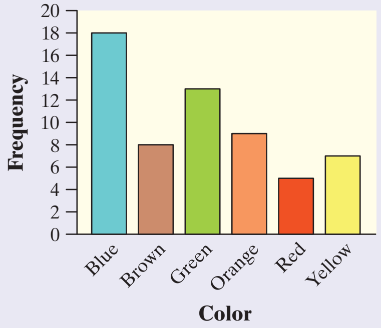
- The *population distribution* gives the values of the variable for all individuals in the population. In this case, the individuals are the 200 chips and the variable we're recording is color. Our parameter of interest is the proportion of red chips in the population, $p = 0.50$.
- The *distribution of sample data* shows the values of the variable for the individuals in a sample. In this case, the distribution of sample data shows the values of the variable color for the 20 chips in the sample. For each sample, we record a value for the statistic $\hat{p}$, the sample proportion of red chips.
- The *sampling distribution of the sample proportion* displays the values of $\hat{p}$ from all possible samples of the same size.

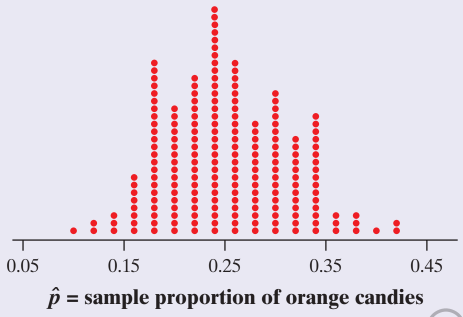
Remember that a sampling distribution describes how a *statistic* (e.g., $\hat{p}$) varies in many samples from the population. However, the population distribution and the distribution of sample data describe how *individuals* (e.g., chips) vary.

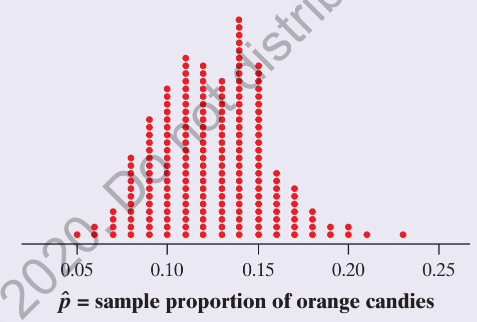
CHECK YOUR UNDERSTANDING

Mars,® Inc. says that the mix of colors in its M&M'S® Milk Chocolate Candies from its Hackettstown, NJ, factory is 25% blue, 25% orange, 12.5% green, 12.5% yellow, 12.5% red, and 12.5% brown. Assume that the company's claim is true and that you will randomly select 50 candies to estimate the proportion that are orange.

1. Identify the population, the parameter, the sample, and the statistic in this setting.
2. Graph the population distribution.
3. Imagine taking a random sample of 50 M&M'S® Milk Chocolate Candies. Make a graph showing a possible distribution of the sample data. Give the value of the statistic for this sample.
4. Which of these three graphs 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 will consider the shape, center, and variability of the sampling distribution. For reasons that will be clear later, we'll save shape for Sections 7.2 and 7.3.

Here is an activity that gets you thinking about the center and variability of a sampling distribution.

ACTIVITY The craft stick problem

In this activity, you will create a statistic for estimating the total number of craft sticks in a bag (N). The sticks are numbered 1, 2, 3, . . . , N. Near the end of the activity, your teacher will select a random sample of $n = 7$ sticks and read the number on each stick to the class. The team that has the best estimate for the total number of sticks will win a prize.

1. Form teams of three or four students. Each team will be given a statistic to begin their investigation.

Stephen Orsillo/Alamy

2. For now, assume that there are $N = 100$ sticks in the bag and that you will be selecting a sample of size $n = 7$. To investigate the quality of the statistic you were given, generate a simulated sampling distribution:
   (a) Using your TI-84 calculator, select an SRS of size 7 using the command RandIntNoRep(lower:1, upper:100, n:7). Calculate and record the value of your statistic for this simulated sample. [With a TI-83 or older TI-84 OS, use the command RandInt(lower:1, upper:100, n:7) and verify that there are no repeated numbers.]

(b) Repeat the previous step at least 9 more times, recording the value of your statistic each time.

(c) Display the simulated sampling distribution of your statistic on the dotplot provided by your teacher.

3. As a class, discuss the quality of each of these statistics. Do any of them consistently overestimate or consistently underestimate the truth? Are some of the statistics more variable than others?

4. In your group, spend about 10–15 minutes creating a few additional statistics that could be used to estimate the total number of sticks. Create a simulated sampling distribution for each one (as in Step 2) to determine which statistic you will use for the competition.

5. It is time for the final competition. Your teacher will select a random sample of 7 sticks from the bag and read out the stick numbers. On a sheet of paper, write the names of your group, the statistic you think is best (a formula), and the value of the statistic calculated from the sample provided by the teacher. The closest estimate wins!

CENTER: BIASED AND UNBIASED ESTIMATORS Figure 7.3 shows the simulated sampling distribution of $\hat{p}$ = proportion of red chips when selecting samples of size $n = 20$ from a population where $p = 0.5$.

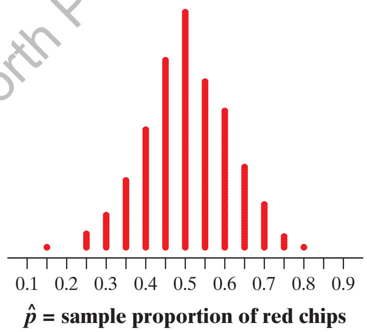


FIGURE 7.3 The distribution of $\hat{p}$ = the sample proportion of red chips in 500 SRSs of size $n = 20$ from a population where $p = 0.5$.

Notice 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.

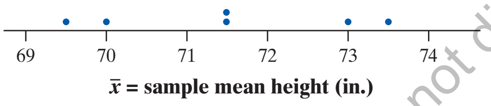
DEFINITION Unbiased estimator

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.

In a particular sample, the value of an unbiased estimator might be greater than the value of the parameter or it might be less than the value of the parameter. However, because the sampling distribution of the statistic is centered at the true value, the statistic will not consistently overestimate or consistently underestimate the parameter. This fits with our definition of bias from Chapter 4. The design of a statistical study shows bias if it is very likely to underestimate or very likely to overestimate the value we want to know.

We will confirm mathematically 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?

Let's revisit the "Sampling heights" example with John and Carol's four sons to investigate one of these statistics. Recall that the heights of the four sons are 71, 75, 72, and 68 inches. Here again is the sampling distribution of the sample mean $\bar{x}$ for samples of size 2:



To determine if the sample mean is an unbiased estimator of the population mean, we need to compare the mean of the sampling distribution to the value we are trying to estimate—the mean of the population μ.

The mean of the sampling distribution of $\bar{x}$ is

$$\mu_{\bar{x}} = \frac{69.5 + 70 + 71.5 + 71.5 + 73 + 73.5}{6} = 71.5$$

The mean of the population is

$$\mu = \frac{71 + 75 + 72 + 68}{4} = 71.5$$

Because these values are equal, this example suggests that the sample mean $\bar{x}$ is an unbiased estimator of the population mean μ. We will confirm this fact in Section 7.3.

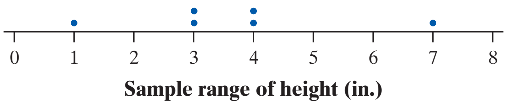
EXAMPLE

Estimating the range

Biased and unbiased estimators

Brian Miller

PROBLEM: In the "Sampling heights" example, we created the sampling distribution of the sample range for samples of size $n = 2$ from the population of John and Carol's four sons with heights of 71, 75, 72, and 68 inches tall. Is the sample range an unbiased estimator of the population range? Explain your answer.



SOLUTION:

The mean of the sampling distribution of the sample range is

$$\frac{1 + 3 + 3 + 4 + 4 + 7}{6} = 3.67$$

The range of the population is

$$\text{Population range} = 75 - 68 = 7$$

Because the mean of the sampling distribution of the sample range (3.67) is not equal to the value it is trying to estimate (7), the sample range is not an unbiased estimator of the population range.

FOR PRACTICE, TRY EXERCISE 19

Because the sample range is consistently smaller than the population range, the sample range is a *biased estimator* of the population range.

Think About It

WHY DO WE DIVIDE BY $n - 1$ WHEN CALCULATING THE SAMPLE STANDARD DEVIATION? Now that we know about sampling distributions and unbiased estimators, we can finally answer this question. In Chapter 1, you learned that the formula for the sample standard deviation is $s_x = \sqrt{\dfrac{\sum(x_i - \bar{x})^2}{n-1}}$. You also learned that the value obtained before taking the square root in the standard deviation calculation is known as the variance. That is, the sample variance is $s_x^2 = \dfrac{\sum(x_i - \bar{x})^2}{n-1}$.

In an inference setting involving a quantitative variable, we might be interested in estimating the variance σ^2 of the population distribution. The most logical choice for our estimator is the sample variance s_x^2. We used technology to select 500 SRSs of size $n = 4$ from a population where the population variance is $\sigma^2 = 9$. For each sample, we recorded the value of two statistics:

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

Figure 7.4 shows the approximate sampling distributions of these two statistics. The blue vertical lines mark the means of these two distributions.

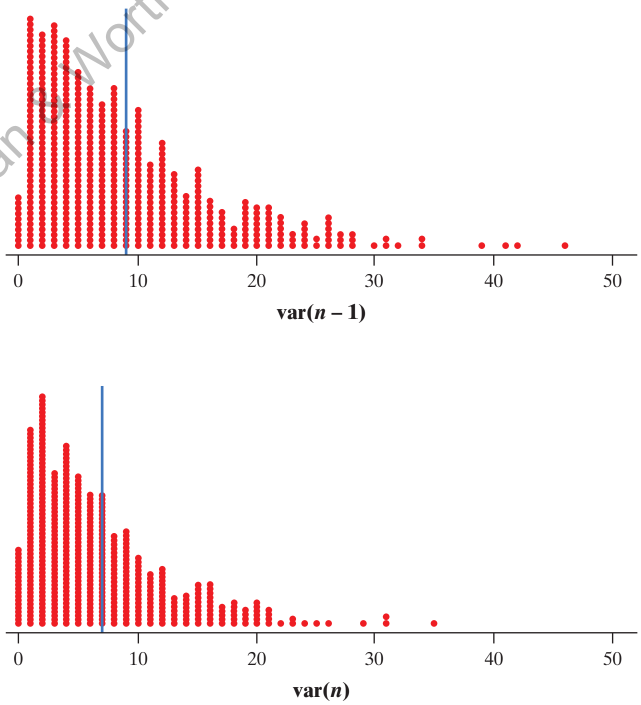


FIGURE 7.4 Results from a simulation of 500 SRSs of size $n = 4$ from a population with variance $\sigma^2 = 9$. The sample variance s_x^2 (labeled "var($n - 1$)" in the figure) is an unbiased estimator of the population variance. The "var(n)" statistic is a biased estimator of the population variance.

We can see that "var(n)" is a *biased* estimator of the population variance. The mean of its approximate sampling distribution (marked with a blue line segment) is clearly less than the value of the population parameter, $\sigma^2 = 9$. However, the statistic "var($n - 1$)" (otherwise known as the sample variance s_x^2) is an unbiased estimator. Its values are centered at $\sigma^2 = 9$. That's why we divide by $n - 1$ when calculating the sample variance—and when calculating the sample standard deviation.

VARIABILITY: SMALLER 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 consistently 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.

To investigate the variability of a statistic, let's consider the proportion of people in a random sample who have ever watched the show *Survivor*. According to Nielsen ratings, *Survivor* was one of the most-watched television shows in the United States every week that it aired. Suppose that the true proportion of U.S. adults who have ever watched *Survivor* is $p = 0.37$.

The top dotplot in Figure 7.5 shows the results of drawing 400 SRSs of size $n = 100$ from a large population with $p = 0.37$ and recording the value of $\hat{p} =$ the sample proportion who have ever watched *Survivor*. We see that a sample of 100 people often gave a $\hat{p}$ quite far from the population parameter, $p = 0.37$.

Let's repeat our simulation, this time taking 400 SRSs of size $n = 1000$ from a large population with proportion $p = 0.37$ who have watched *Survivor*. The bottom dotplot in Figure 7.5 displays the distribution of the 400 values of $\hat{p}$ from these larger samples. Both graphs are drawn on the same horizontal scale to make comparison easy.

We can see that the variability shown in the top dotplot in Figure 7.5 is much greater than the variability shown in the bottom dotplot. With samples of size 100, the standard deviation of these $\hat{p}$ values is about 0.047. Using SRSs of size 1000, the standard deviation of these $\hat{p}$ values is about 0.016. This confirms what we learned in Section 4.3: larger random samples tend to produce estimates that are closer to the true population value.

One important and surprising fact is that the variability of a statistic does *not* depend very much on the size of the population, as long as the sample size is less than 10% of the population size. Suppose that in a small town of 25,000 people, 37% of the population have watched *Survivor*. That is, $p = 0.37$. Let's simulate taking 400 SRSs of size 1000 from this small town and compute $\hat{p}$, the sample proportion who have watched *Survivor*. The results are shown in Figure 7.6.

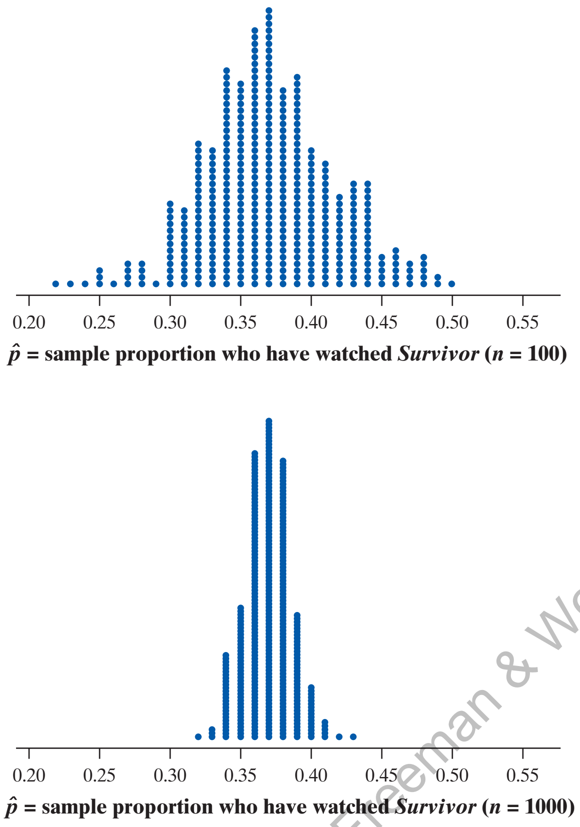


FIGURE 7.5 The approximate sampling distribution of the sample proportion $\hat{p}$ from SRSs of size $n = 100$ and $n = 1000$ chosen from a large population with proportion $p = 0.37$. Both dotplots show the results of 400 SRSs.

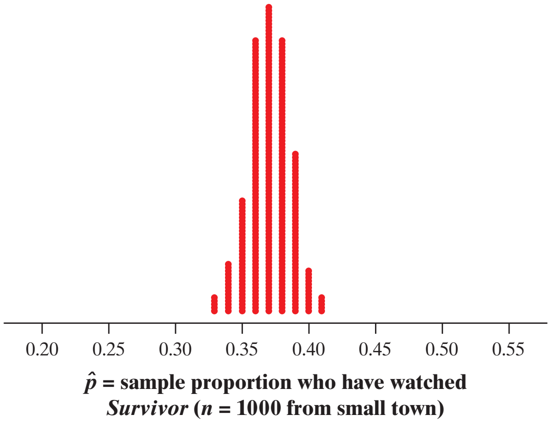


FIGURE 7.6 The approximate sampling distribution of the sample proportion $\hat{p}$ from 400 SRSs of size $n = 1000$ chosen from a population of 25,000 individuals with proportion $p = 0.37$.

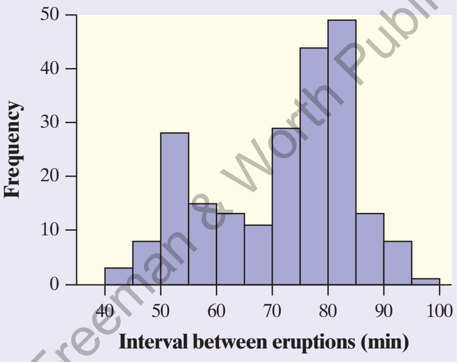
Sergey Skleznev/Shutterstock

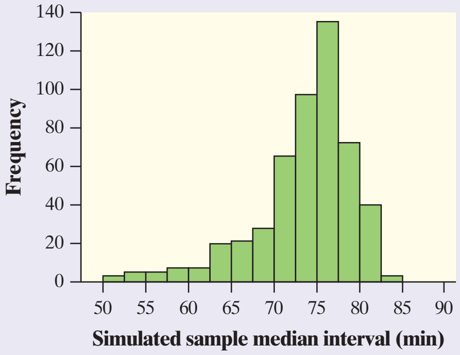
Notice that the distribution of $\hat{p}$ looks nearly the same when sampling from a small town of 25,000 residents as when sampling from the entire United States. In fact, the standard deviation for each sampling distribution is approximately 0.016.

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 if 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 mostly on the size of the scoop.

CHECK YOUR UNDERSTANDING

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





1. Does the simulation provide evidence that the sample median is a biased estimator of the population median? Justify your answer.
2. Suppose we had taken samples of size 20 instead of size 10. Would the variability of the sampling distribution of the sample median be larger, smaller, or about the same? Justify your answer.
3. Describe the shape of the sampling distribution of the sample median.

CHOOSING AN ESTIMATOR In many cases, it is obvious which statistic should be used as an estimator of a population parameter. If we want to estimate the population mean μ, use the sample mean $\bar{x}$. If we want to estimate a population proportion p, use the sample proportion $\hat{p}$. However, in other cases, there isn't an obvious best choice. When trying to estimate the population maximum in the "Craft stick" activity, there were many different estimators that could be used.

To decide which estimator to use when there are several choices, consider both bias and variability. Imagine the true value of the population parameter as the bull's-eye on a target and 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 statistics do not center on the population parameter. In other words, our estimates are not *accurate*. High variability means that repeated shots are widely scattered on the target. Repeated samples do not give very similar results. In other words, our estimates are not very *precise*. Figure 7.7 shows this target illustration.

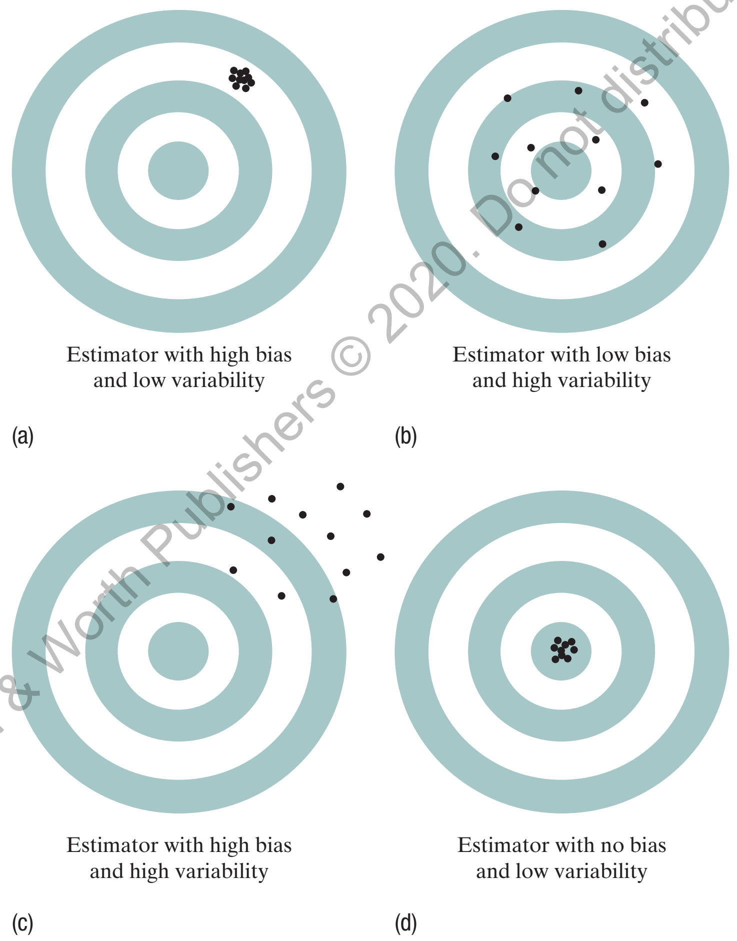


FIGURE 7.7 Bias and variability of four different estimators. (a) High bias, low variability. (b) Low bias, high variability. (c) High bias, high variability. (d) The ideal estimator: no bias, low variability.

Notice that an estimator with low variability can also have high bias, as in Figure 7.7(a). And an estimator with low or no bias can be quite variable, as in Figure 7.7(b). Ideally, we'd like to use an estimator that is unbiased with low variability, as in Figure 7.7(d).

AP® EXAM TIP

Make sure to understand the difference between accuracy and precision when writing responses on the AP® Statistics Exam. Many students use "accurate" when they really mean "precise." For example, a response that says "increasing the sample size will make an estimate more accurate" is incorrect. It should say that increasing the sample size will make an estimate more precise. If you can't remember which term to use, don't use either of them. Instead, explain what you mean without using statistical vocabulary.

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 an unbiased estimator if the center (mean) of its sampling distribution is equal to the true value of the parameter.
- The **variability** of a statistic is described by the spread of its sampling distribution. Larger samples result in sampling distributions with less variability.
- When trying to estimate a parameter, choose a statistic with low or no bias and minimum variability.

Section 7.1 Exercises

For Exercises 1–6, identify the population, the parameter, the sample, and the statistic in each setting.

pg 470

1. **Healthy living** From a large group of people who signed a card saying they intended to quit smoking, 1000 people were selected at random. It turned out that 210 (21%) of these individuals had not smoked over the past 6 months.

2. **Unemployment** Each month, the Current Population Survey interviews about 60,000 randomly selected U.S. adults. One of their goals is to estimate the national unemployment rate. In October 2016, 4.9% of those interviewed were unemployed.

3. **Fillings** How much do prices vary for filling a cavity? To find out, an insurance company randomly selects 10 dental practices in California and asks for the cash (non-insurance) price for this procedure at each practice. The interquartile range is $74.

4. **Warm turkey** 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 is 170°F.

5. **Iced tea** 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 selected 50 bottles at random from the day's production. These bottles contained an average of 19.6 ounces of iced tea.

6. **Bearings** A production run of ball bearings is supposed to have a mean diameter of 2.5000 centimeters (cm). An inspector chooses 100 bearings at random from the run. These bearings have mean diameter 2.5009 cm.

Exercises 7–10 refer to the small population of 5 students in the table.

Name	Gender	Quiz score
Abigail	Female	10
Bobby	Male	5
Carlos	Male	10
DeAnna	Female	7
Emily	Female	9

pg 472

7. **Sample means** List all 10 possible SRSs of size $n = 2$, calculate the mean quiz score for each sample, and display the sampling distribution of the sample mean on a dotplot.

8. **Sample minimums** List all 10 possible SRSs of size $n = 3$, calculate the minimum quiz score for each sample, and display the sampling distribution of the sample minimum on a dotplot.

9. **Sample proportions** List all 10 possible SRSs of size $n = 2$, calculate the proportion of females for each sample, and display the sampling distribution of the sample proportion on a dotplot.

10. **Sample medians** List all 10 possible SRSs of size $n = 3$, calculate the median quiz score for each sample, and display the sampling distribution of the sample median on a dotplot.

11. **Doing homework** A school newspaper article claims that 60% of the students at a large high school completed their assigned homework last week. Assume that this claim is true for the 2000 students at the school.

(a) Make a bar graph of the population distribution.

(b) Imagine one possible SRS of size 100 from this population. Sketch a bar graph of the distribution of sample data.

12. **Tall girls** According to the National Center for Health Statistics, the distribution of height 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. Assume this claim is true for the three hundred 16-year-old females at a large high school.

(a) Make a graph of the population distribution.

(b) Imagine one possible SRS of size 20 from this population. Sketch a dotplot of the distribution of sample data.

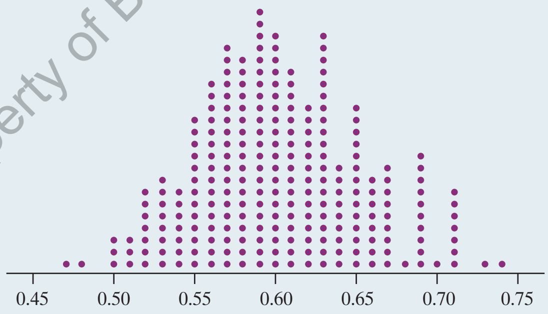
pg 473

13. **More homework** Some skeptical AP® Statistics students want to investigate the newspaper's claim in Exercise 11, so they choose an SRS of 100 students from the school to interview. In their sample, 45 students completed their homework last week. Does this provide convincing evidence that less than 60% of all students at the school completed their assigned homework last week?

(a) What is the evidence that less than 60% of all students completed their assigned homework last week?

(b) Provide two explanations for the evidence described in part (a).

We used technology to simulate choosing 250 SRSs of size $n = 100$ from a population of 2000 students where 60% completed their assigned homework last week. The dotplot shows $\hat{p}$ = the sample proportion of students who completed their assigned homework last week for each of the 250 simulated samples.



$\hat{p}$ = **simulated sample proportion who completed their assigned homework last week**

(c) There is one dot on the graph at 0.73. Explain what this value represents.

(d) Would it be surprising to get a sample proportion of $\hat{p} = 0.45$ or smaller in an SRS of size 100 when $p = 0.60$? Justify your answer.

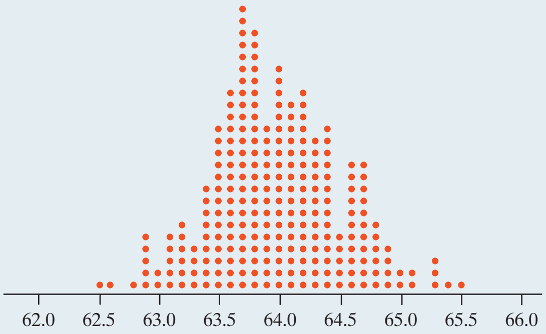
(e) Based on your previous answers, is there convincing evidence that less than 60% of all students at the school completed their assigned homework last week? Explain your reasoning.

14. **Tall girls?** To see if the claim made in Exercise 12 is true at their high school, an AP® Statistics class chooses an SRS of twenty 16-year-old females at the school and measures their heights. In their sample, the mean height is 64.7 inches. Does this provide convincing evidence that 16-year-old females at this school are taller than 64 inches, on average?

(a) What is the evidence that the average height of all 16-year-old females at this school is greater than 64 inches, on average?

(b) Provide two explanations for the evidence described in part (a).

We used technology to simulate choosing 250 SRSs of size $n = 20$ from a population of three hundred 16-year-old females whose heights follow a Normal distribution with mean $\mu = 64$ inches and standard deviation $\sigma = 2.5$ inches. The dotplot shows $\bar{x}$ = the sample mean height for each of the 250 simulated samples.



$\bar{x}$ = **simulated sample mean height (in.)**

(c) There is one dot on the graph at 62.5. Explain what this value represents.

(d) Would it be surprising to get a sample mean of $\bar{x} = 64.7$ or larger in an SRS of size 20 when $\mu = 64$ inches and $\sigma = 2.5$ inches? Justify your answer.

(e) Based on your previous answers, is there convincing evidence that the average height of all 16-year-old females at this school is greater than 64 inches? Explain your reasoning.

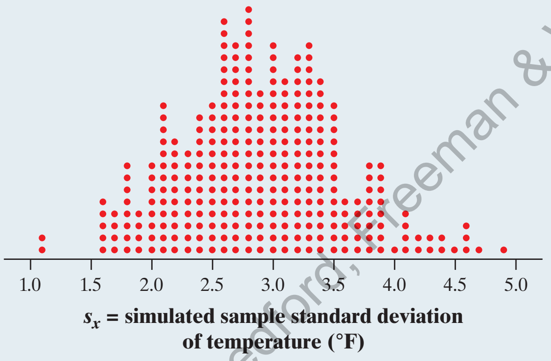
15. **Even more homework** Refer to Exercises 11 and 13. Suppose that the sample proportion of students who did all their assigned homework last week is $\hat{p} = 57/100 = 0.57$. Would this sample proportion provide convincing evidence that less than 60% of all

students at the school completed all their assigned homework last week? Explain your reasoning.

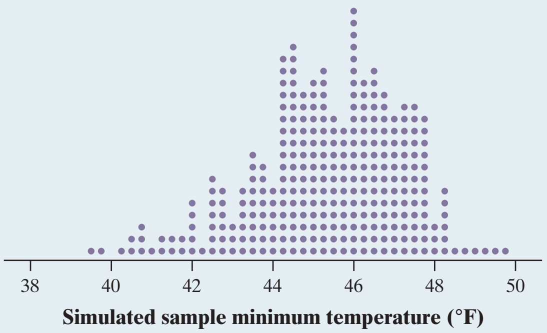
16. **Even more tall girls** Refer to Exercises 12 and 14. Suppose that the sample mean height of the twenty 16-year-old females is $\bar{x} = 65.8$ inches. Would this sample mean provide convincing evidence that the average height of all 16-year-old females at this school is greater than 64 inches? Explain your reasoning.

Exercises 17 and 18 refer to the following setting. During the winter months, outside temperatures at 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. The manufacturer claims that the thermostat allows variation in home temperature that follows a Normal distribution with $\sigma = 3$°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 $\mu = 50$°F and standard deviation $\sigma = 3$°F.

17. **Cold cabin?** The dotplot shows the results of taking 300 SRSs of 10 temperature readings from a Normal population with $\mu = 50$ and $\sigma = 3$ and recording the sample standard deviation s_x each time. Suppose that the standard deviation from an actual sample is $s_x = 5$°F. What would you conclude about the thermostat manufacturer's claim? Explain your reasoning.



18. **Really cold cabin** The dotplot shows the results of taking 300 SRSs of 10 temperature readings from a Normal population with $\mu = 50$ and $\sigma = 3$ and recording the sample minimum each time. Suppose that the minimum of an actual sample is 40°F. What would you conclude about the thermostat manufacturer's claim? Explain your reasoning.



Exercises 19–22 refer to the small population of 4 cars listed in the table.

Color	Age (years)
Red	1
White	5
Silver	8
Red	20

19. **Sample proportions** List all 6 possible SRSs of size $n = 2$, calculate the proportion of red cars in each sample, and display the sampling distribution of the sample proportion on a dotplot. Is the sample proportion an unbiased estimator of the population proportion? Explain your answer.

pg 477

20. **Sample minimums** List all 6 possible SRSs of size $n = 2$, calculate the minimum age for each sample, and display the sampling distribution of the sample minimum on a dotplot. Is the sample minimum an unbiased estimator of the population minimum? Explain your answer.

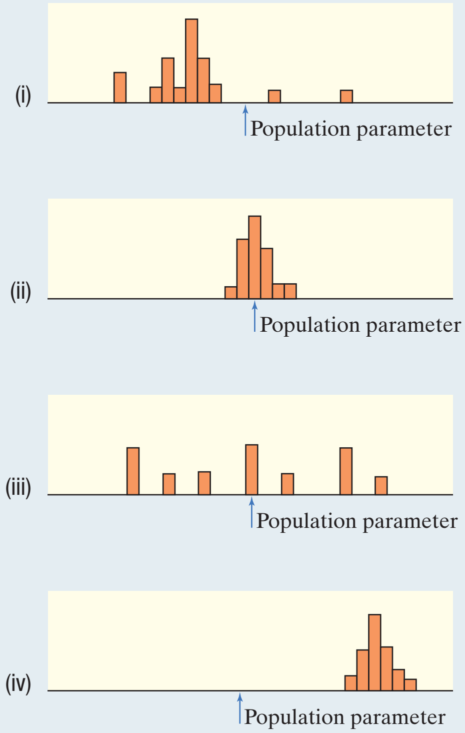
21. **More sample proportions** List all 4 possible SRSs of size $n = 3$, calculate the proportion of red cars in the sample, and display the sampling distribution of the sample proportion on a dotplot with the same scale as the dotplot in Exercise 19. How does the variability of this sampling distribution compare with the variability of the sampling distribution from Exercise 19? What does this indicate about increasing the sample size?

22. **More sample minimums** List all 4 possible SRSs of size $n = 3$, calculate the minimum age for each sample, and display the sampling distribution of the sample minimum on a dotplot with the same scale as the dotplot in Exercise 20. How does the variability of this sampling distribution compare with the variability of the sampling distribution from Exercise 20? What does this indicate about increasing the sample size?

23. **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 μ in this population. Explain to someone who knows little about statistics what it means to say that $\bar{x}$ is an unbiased estimator of μ.

24. **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.

25. **Bias and variability** The figure shows approximate sampling distributions of 4 different statistics intended to estimate the same parameter.



(a) Which statistics are unbiased estimators? Justify your answer.

(b) Which statistic does the best job of estimating the parameter? Explain your answer.

Multiple Choice: *Select the best answer for Exercises 26–30.*

26. 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?

(a) 78% is a population and 72% is a sample.

(b) 72% is a population and 78% is a sample.

(c) 78% is a parameter and 72% is a statistic.

(d) 72% is a parameter and 78% is a statistic.

(e) 72% is a parameter and 100 is a statistic.

27. A statistic is an unbiased estimator of a parameter when

(a) the statistic is calculated from a random sample.

(b) in a single sample, the value of the statistic is equal to the value of the parameter.

(c) in many samples, the values of the statistic are very close to the value of the parameter.

(d) in many samples, the values of the statistic are centered at the value of the parameter.

(e) in many samples, the distribution of the statistic has a shape that is approximately Normal.

28. In a residential neighborhood, the distribution of house values is unimodal and skewed to the right, with a median of \$200,000 and an *IQR* of \$100,000. For which of the following sample sizes is the sample median most likely to be above \$250,000?

(a) $n = 10$

(b) $n = 50$

(c) $n = 100$

(d) $n = 1000$

(e) Impossible to determine without more information.

29. Increasing the sample size of an opinion poll will reduce the

(a) bias of the estimates made from the data collected in the poll.

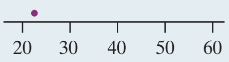
(b) variability of the estimates made from the data collected in the poll.

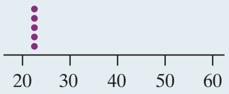
(c) effect of nonresponse on the poll.

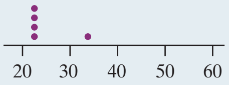
(d) variability of opinions in the sample.

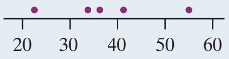
(e) variability of opinions in the population.

30. The math department at a small school has 5 teachers. The ages of these teachers are 23, 34, 37, 42, and 58. Suppose you select a random sample of 4 teachers and calculate the sample minimum age. Which of the following shows the sampling distribution of the sample minimum age?

(a)


(b)


(c)


(d)


(e) None of these.

Recycle and Review

31. **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

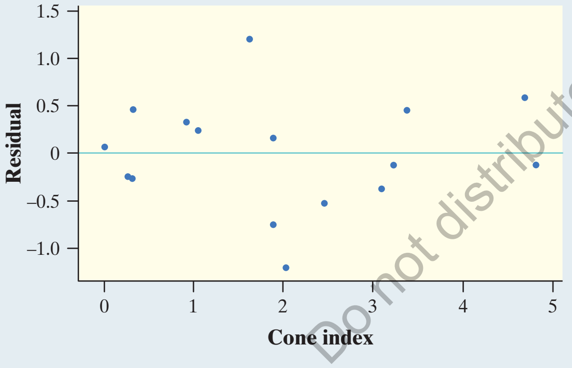
for osteoporosis is a BMD score that is 2.5 standard deviations below the mean for young adults. BMD measurements in a population of people similar in age and gender roughly follow a Normal distribution.

(a) What percent of healthy young adults have osteoporosis by the WHO criterion?

(b) 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?

32. **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.[4] Computer output from a least-squares regression on these data and a residual plot are shown here.

```
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%
```



(a) Is a linear model appropriate for these data? Explain.

(b) Give the equation for the least-squares regression line. Define any variables you use.

(c) Interpret the values of r^2 and s in context.

SECTION 7.2 Sample Proportions

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

- Calculate the mean and standard deviation of the sampling distribution of a sample proportion $\hat{p}$ and interpret the standard deviation.
- Determine if the sampling distribution of $\hat{p}$ is approximately Normal.
- Calculate the mean and the standard deviation of the sampling distribution of a difference in sample proportions $\hat{p}_1 - \hat{p}_2$ and interpret the standard deviation.
- Determine if the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately Normal.
- If appropriate, use a Normal distribution to calculate probabilities involving $\hat{p}$ or $\hat{p}_1 - \hat{p}_2$.

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.[5] 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 all U.S. teenagers know that Columbus discovered America in 1492. In this section, we’ll use sampling distributions to clarify what “about” means.

I DON'T GET IT.
THERE ISN'T A "CLEVELAND DAY"... THERE ISN'T AN "AKRON DAY"...
AS FAR AS I KNOW THERE ISN'T A "CINCINNATI DAY"...
SO WHY IS THERE A "COLUMBUS DAY"?
THE WORLD HAS MANY MYSTERIES.
www.foxtrot.com
10-10
©2005 Bill Amend / Distributed by Universal Press Syndicate
AMEND

FOXTROT©2005 Bill Amend. Andrews McMeel Syndicate.

The Sampling Distribution of $\hat{p}$

When Mrs. Gallas's class did the "Penny for your thoughts" activity at the beginning of the chapter, her students produced the "dotplot" in Figure 7.8. This graph approximates the **sampling distribution of the sample proportion** of pennies from the 2000s for samples of size $n = 20$ from Mrs. Gallas's population of pennies.

DEFINITION **Sampling distribution of the sample proportion**

The **sampling distribution of the sample proportion** $\hat{p}$ describes the distribution of values taken by the sample proportion $\hat{p}$ in all possible samples of the same size from the same population.

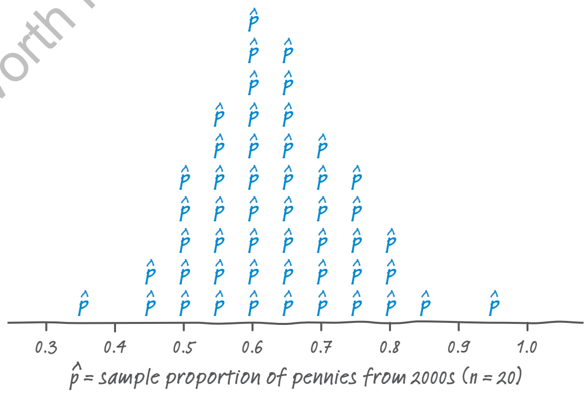


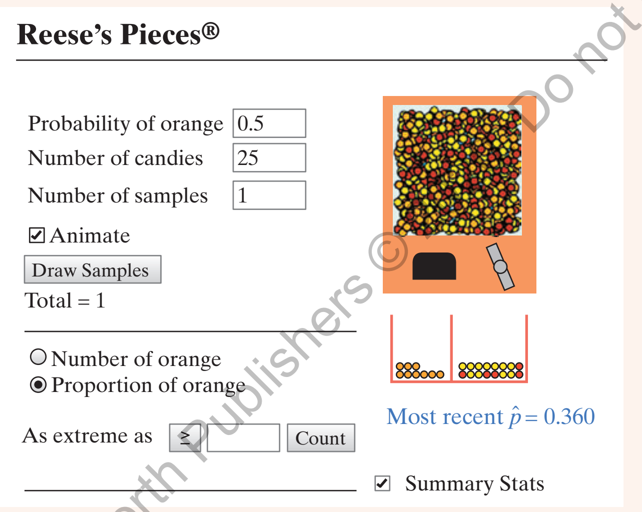
FIGURE 7.8 Approximate sampling distribution of the sample proportion of pennies from the 2000s in 50 samples of size $n = 20$ from a population of pennies.

This distribution is roughly symmetric with a mean of about 0.65 and a standard deviation of about 0.10. By the end of this section, you should be able to anticipate the shape, center, and variability of distributions like this one without getting your hands dirty in a jar of pennies.

ACTIVITY The candy machine

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 shape, center, and variability of the sampling distribution of the sample proportion $\hat{p}$.

1. Launch the *Reese's Pieces*® applet at www.rossmanchance.com/applets. Make sure the Probability of orange = 0.5, Number of candies (sample size) = 25, and Number of samples = 1. Choose "Proportion of orange" as the statistic to be calculated and check the box for Summary stats to be calculated, as shown in the screen shot for Step 2.
2. Click on the "Draw Samples" button. An animated simple random sample of $n = 25$ candies should be dispensed. The following screen shot shows the results of one such sample where the sample proportion of orange candies was $\hat{p} = 0.360$. How far was your sample proportion of orange candies from the actual population proportion, $p = 0.50$?



Reese's Pieces®

Probability of orange 0.5
Number of candies 25
Number of samples 1
☑ Animate
Draw Samples
Total = 1

○ Number of orange
● Proportion of orange

As extreme as ≥ Count

Most recent $\hat{p} = 0.360$

☑ Summary Stats

3. Click "Draw Samples" 9 more times so that you have a total of 10 sample proportions. 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 990 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 1000 samples of 25 candies from the machine. Describe the shape, center, and variability of the approximate sampling distribution of $\hat{p}$ shown in the dotplot.
5. How does the sampling distribution of $\hat{p}$ change if the proportion of orange candies in the machine is different from $p = 0.5$? Use the applet to investigate this question. Then write a brief summary of what you learned.
6. How does the sampling distribution of $\hat{p}$ change if the machine dispenses a different number of candies (n)? Use the applet to investigate this question. 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 this question. Then write a brief summary of what you learned.

The graphs in Figure 7.9 show approximate sampling distributions of $\hat{p}$ for different combinations of p (population proportion) and n (sample size).

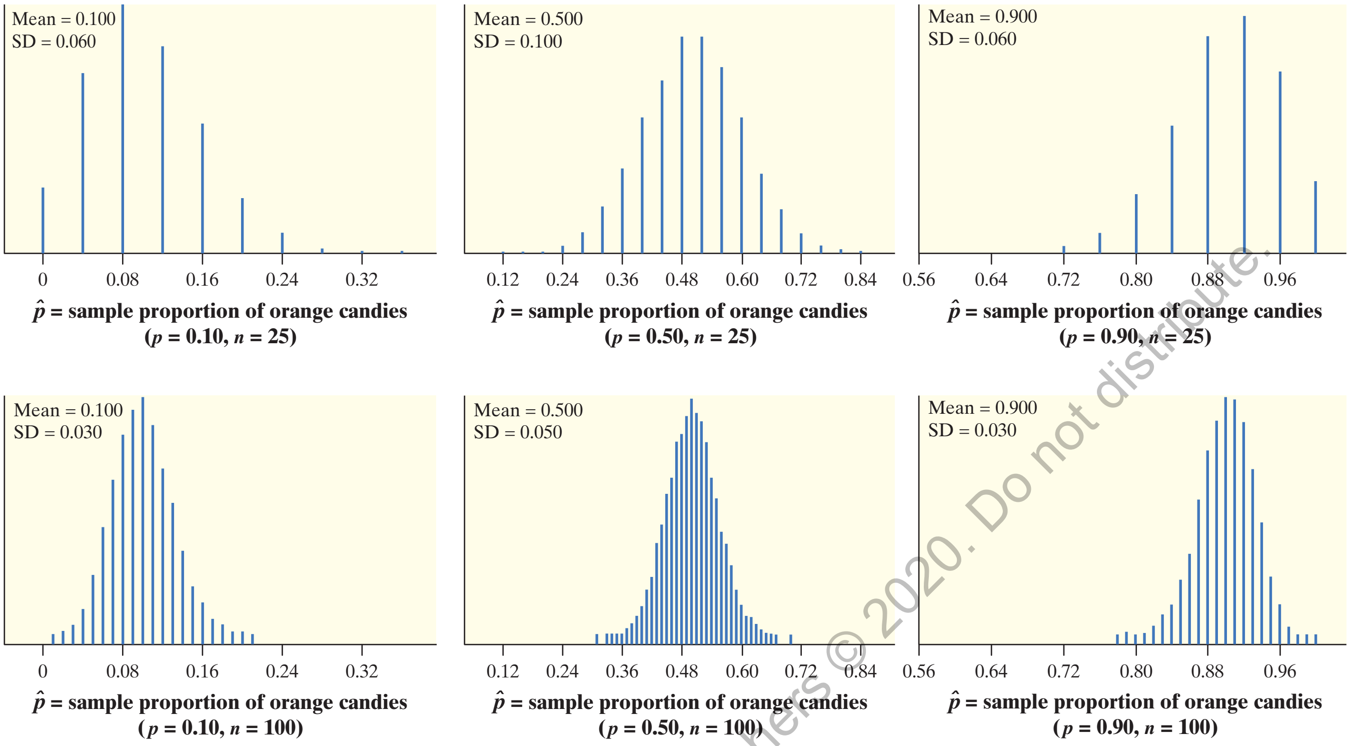


FIGURE 7.9 Approximate sampling distributions of $\hat{p}$ = sample proportion of orange candies for different combinations of p (population proportion) and n (sample size).

What do these graphs teach us about the sampling distribution of $\hat{p}$?

Shape: When n is small and p is close to 0, the sampling distribution of $\hat{p}$ is skewed to the right. When n is small and p is close to 1, the sampling distribution of $\hat{p}$ is skewed to the left. Finally, the sampling distribution of $\hat{p}$ becomes more Normal when p is closer to 0.5 or n is larger (or both).

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

Variability: The value of $\sigma_{\hat{p}}$ depends on both n and p. For a specific sample size, the standard deviation $\sigma_{\hat{p}}$ is larger for values of p close to 0.5 and smaller for values of p close to 0 or 1. For a specific value of p, the standard deviation $\sigma_{\hat{p}}$ gets smaller as n gets larger. *Specifically, multiplying the sample size by 4 cuts the standard deviation in half.*

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

SAMPLING DISTRIBUTION OF A SAMPLE PROPORTION $\hat{p}$

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 approximately

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

as long as the *10% condition* is satisfied: $n < 0.10N$. The value $\sigma_{\hat{p}}$ measures the typical distance between a sample proportion $\hat{p}$ and the population proportion p.

- The sampling distribution of $\hat{p}$ is **approximately Normal** as long as the *Large Counts condition* is satisfied: $np \geq 10$ and $n(1-p) \geq 10$.

We call it the "Large Counts" condition because *np* is the expected *count* of successes in the sample and $n(1-p)$ is the expected *count* of failures in the sample.

The two conditions mentioned in the preceding box are very important.

- *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.
- *10% condition:* When we sample *with* replacement, the standard deviation of the sampling distribution of $\hat{p}$ is exactly $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$. When we sample *without* replacement, the observations are not independent, and the actual standard deviation of the sampling distribution of $\hat{p}$ is smaller than the value given by the formula. If the sample size is less than 10% of the population size, however, the value given by the formula is nearly correct.

Because larger random samples give better information, it sometimes makes sense to sample more than 10% of a population. In such a case, there is an adjustment we can make to the formula for $\sigma_{\hat{p}}$ that correctly reduces the standard deviation. The adjustment is called a *finite population correction (FPC)*. We'll avoid situations that require the FPC in this text.

EXAMPLE

Backing the pack
The sampling distribution of $\hat{p}$

bo1982/Getty Images

PROBLEM: Suppose that 84% of students at a large high school regularly use a backpack to carry their books from class to class. Imagine taking an SRS of 100 students and calculating $\hat{p}$ = the proportion of students in the sample who regularly use a backpack.

(a) Identify the mean of the sampling distribution of $\hat{p}$.

(b) Calculate and interpret the standard deviation of the sampling distribution of $\hat{p}$. Verify that the 10% condition is met.

(c) Describe the shape of the sampling distribution of $\hat{p}$. Justify your answer.

SOLUTION:

(a) $\mu_{\hat{p}} = 0.84$

(b) Assuming that $n = 100$ students is less than 10% of students in a large high school, the standard deviation is approximately

$$\sigma_{\hat{p}} = \sqrt{\frac{0.84(1-0.84)}{100}} = 0.0367$$

In SRSs of size 100, the sample proportion of students who regularly use a backpack will typically vary by about 0.0367 from the true proportion of $p = 0.84$.

(c) Because $100(0.84) = 84 \geq 10$ and $100(1-0.84) = 16 \geq 10$, the sampling distribution of $\hat{p}$ is approximately Normal.

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

When $n < 0.10N$,

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

When $np \geq 10$ and $n(1-p) \geq 10$, the sampling distribution of $\hat{p}$ is approximately Normal.

FOR PRACTICE, TRY EXERCISE 33

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

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

The sample proportion of successes is closely related to X:

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

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 a random variable by a constant multiplies both the mean and the standard deviation by that constant. We have

$$\mu_{\hat{p}} = \frac{1}{n}\mu_X = \frac{1}{n}(np) = p$$

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

What about shape? Multiplying a random variable by a positive 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 remember the subsection in Chapter 6 about the Normal approximation to a binomial distribution, you already know that a Normal distribution can be used to approximate the sampling distribution of $\hat{p}$ whenever both np and $n(1-p)$ are at least 10.

CHECK YOUR UNDERSTANDING

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

1. Identify the mean of the sampling distribution of $\hat{p}$.
2. Calculate and interpret the standard deviation of the sampling distribution of $\hat{p}$. Check that the 10% condition is met.

continues on next page

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 estimate 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 $\hat{p}$

Ivan Vukovic/Shutterstock

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

SOLUTION:

Let $\hat{p}$ = sample proportion of all first-year college students who attend college within 50 miles of home.

$\mu_{\hat{p}} = 0.35$

Assuming that 1500 < 10% of all first-year college students,

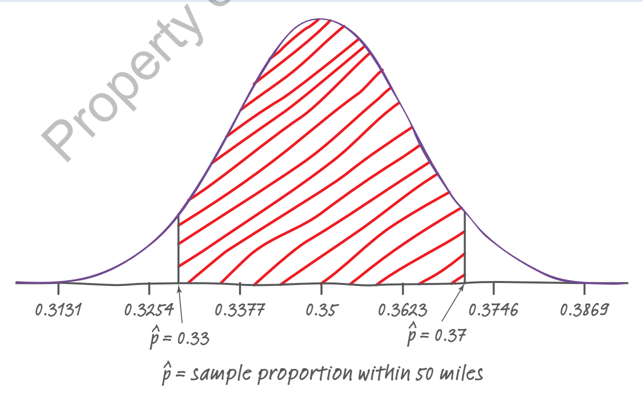
$$\sigma_{\hat{p}} = \sqrt{\frac{(0.35)(0.65)}{1500}} = 0.0123$$

Calculate the mean and standard deviation of the sampling distribution of $\hat{p}$.

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

Because $np = 1500(0.35) = 525 \geq 10$ and $n(1-p) = 1500(0.65) = 975 \geq 10$, the distribution of $\hat{p}$ is approximately Normal.

Justify that the distribution of $\hat{p}$ is approximately Normal using the Large Counts condition.



1. Draw a Normal distribution.

(i) $z = \dfrac{0.33 - 0.35}{0.0123} = -1.63$ and $z = \dfrac{0.37 - 0.35}{0.0123} = 1.63$

Using Table A: $P(0.33 \leq \hat{p} \leq 0.37) = P(-1.63 \leq z \leq 1.63)$
$= 0.9484 - 0.0516 = 0.8968$

Using technology: normalcdf(lower: -1.63, upper: 1.63, mean: 0, SD: 1)
$= 0.8969$

(ii) normalcdf(lower: 0.33, upper: 0.37, mean: 0.35, SD: 0.0123) $= 0.8961$

2. Perform calculations.
(i) Standardize and use Table A or technology; or
(ii) Use technology without standardizing.
Be sure to answer the question that was asked.

FOR PRACTICE, TRY EXERCISE 43

In the preceding example, about 90% of all SRSs of size 1500 from this population will give a result within 2 percentage points of the truth about the population. This result also suggests that in about 90% of all SRSs of size 1500 from this population, the true proportion will be within 2 percentage points of the sample proportion. This fact will become very important in Chapter 8 when we use sample data to create an interval of plausible values for a population parameter.

The Sampling Distribution of a Difference Between Two Proportions

Are males or females more likely to use Twitter? Many statistical questions involve comparing the proportion of individuals with a certain characteristic in two populations. Let's call these parameters of interest p_1 and p_2. The preferred strategy is to take a separate random sample from each population and to compare the sample proportions $\hat{p}_1$ and $\hat{p}_2$ with that characteristic.

Which of two treatments is more successful for helping people quit smoking? A randomized experiment can be used to answer this question. This time, the parameters p_1 and p_2 that we want to compare are the true proportions of successful outcomes for each treatment. We use the proportions of successes in the two treatment groups, $\hat{p}_1$ and $\hat{p}_2$, to make the comparison. Here's a table that summarizes these two situations:

Population or treatment	Parameter	Statistic	Sample size
1	p_1	$\hat{p}_1$	n_1
2	p_2	$\hat{p}_2$	n_2

MachineHeadz/Getty Images

We compare the populations or treatments by doing inference about the difference $p_1 - p_2$ between the parameters. The statistic that estimates this difference is the difference between the two sample proportions, $\hat{p}_1 - \hat{p}_2$. To explore the sampling distribution of $\hat{p}_1 - \hat{p}_2$, let's start with two populations having a known proportion of successes.

Suppose there are two large high schools—each with more than 2000 students—in a certain town. At School 1, 70% of students did their homework last night ($p_1 = 0.70$). Only 50% of the students at School 2 did their homework last night ($p_2 = 0.50$). The counselor at School 1 selects an SRS of 100 students and records the proportion $\hat{p}_1$ that did the homework. School 2's counselor selects an SRS of 200 students and records the proportion $\hat{p}_2$ that did the homework. What can we say about the difference $\hat{p}_1 - \hat{p}_2$ in the sample proportions?

Earlier in this section, we saw that the sampling distribution of a sample proportion $\hat{p}$ has the following properties:

Shape: Approximately Normal if $np \geq 10$ and $n(1-p) \geq 10$

Center: $\mu_{\hat{p}} = p$

Variability: $\sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$ if $n < 0.10N$

For the sampling distributions of $\hat{p}_1$ and $\hat{p}_2$ in this case:

	Sampling distribution of $\hat{p}_1$	Sampling distribution of $\hat{p}_2$
Shape	Approximately Normal; $n_1p_1 = 100(0.70) = 70 \geq 10$ and $n_1(1-p_1) = 100(0.30) = 30 \geq 10$	Approximately Normal; $n_2p_2 = 200(0.50) = 100 \geq 10$ and $n_2(1-p_2) = 200(0.50) = 100 \geq 10$
Center	$\mu_{\hat{p}_1} = p_1 = 0.70$	$\mu_{\hat{p}_2} = p_2 = 0.50$
Variability	$\sigma_{\hat{p}_1} = \sqrt{\dfrac{p_1(1-p_1)}{n_1}} = \sqrt{\dfrac{0.7(0.3)}{100}}$ $= 0.0458$ because 100 < 10% of all students at School 1.	$\sigma_{\hat{p}_2} = \sqrt{\dfrac{p_2(1-p_2)}{n_2}} = \sqrt{\dfrac{0.5(0.5)}{200}}$ $= 0.0354$ because 200 < 10% of all students at School 2.

What about the sampling distribution of $\hat{p}_1 - \hat{p}_2$? We used software to randomly select 100 students from School 1 and 200 students from School 2. Our first set of samples gave $\hat{p}_1 = 0.72$ and $\hat{p}_2 = 0.47$, resulting in a difference of $\hat{p}_1 - \hat{p}_2 = 0.72 - 0.47 = 0.25$. A red dot for this value appears in Figure 7.10. The dotplot shows the results of repeating this process 500 times.

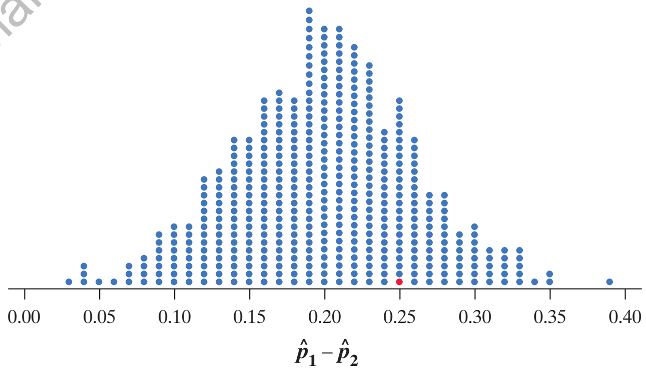


FIGURE 7.10 Simulated sampling distribution of the difference in sample proportions $\hat{p}_1 - \hat{p}_2$ in 500 SRSs of size $n_1 = 100$ from a population with $p_1 = 0.70$ and 500 SRSs of size $n_2 = 200$ from a population with $p_2 = 0.50$.

The figure suggests that the sampling distribution of $\hat{p}_1 - \hat{p}_2$ has an approximately Normal shape. This makes sense from what you learned in Section 6.2 because we are subtracting two independent random variables, $\hat{p}_1$ and $\hat{p}_2$, that have approximately Normal distributions.

The mean of the sampling distribution is 0.20. The true proportion of students who did last night's homework at School 1 is $p_1 = 0.70$ and at School 2 is $p_2 = 0.50$. We expect the difference $\hat{p}_1 - \hat{p}_2$ to center on the actual difference in the population proportions, $p_1 - p_2 = 0.70 - 0.50 = 0.20$. The standard deviation of the sampling distribution is 0.058. It can be found using the formula

$$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} = \sqrt{\frac{0.7(0.3)}{100} + \frac{0.5(0.5)}{200}} = 0.058$$

That is, the difference (School 1 – School 2) in the sample proportions of students at the two schools who did their homework last night typically varies by about 0.058 from the true difference in proportions of 0.20.

THE SAMPLING DISTRIBUTION OF $\hat{p}_1 - \hat{p}_2$

Choose an SRS of size n_1 from Population 1 with proportion of successes p_1 and an independent SRS of size n_2 from Population 2 with proportion of successes p_2. Then:

- The sampling distribution of $\hat{p}_1 - \hat{p}_2$ is **approximately Normal** if the *Large Counts condition* is met for both samples: $n_1p_1 \geq 10$, $n_1(1-p_1) \geq 10$, $n_2p_2 \geq 10$, and $n_2(1-p_2) \geq 10$.
- The **mean** of the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is $\mu_{\hat{p}_1-\hat{p}_2} = p_1 - p_2$.
- The **standard deviation** of the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately

$$\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

as long as the *10% condition* is met for both samples: $n_1 < 0.10N_1$ and $n_2 < 0.10N_2$.

Note that the formula for the standard deviation is exactly correct only when we have two types of independence:

- Independent samples, so that we can add the variances of $\hat{p}_1$ and $\hat{p}_2$.
- Independent observations within each sample. When sampling without replacement, the actual value of the standard deviation is smaller than the formula suggests. However, if the 10% condition is met for both samples, the difference is negligible.

The standard deviation of the sampling distribution tells us how much the difference in sample proportions will typically vary from the difference in the population proportions if we repeat the random sampling process many times.

Think About It

WHERE DO THE FORMULAS FOR THE MEAN AND STANDARD DEVIATION OF THE SAMPLING DISTRIBUTION OF $\hat{p}_1 - \hat{p}_2$ COME FROM? Both $\hat{p}_1$ and $\hat{p}_2$ are random variables. That is, their values would vary in repeated independent SRSs of size n_1 and n_2. Independent random samples yield independent random variables $\hat{p}_1$ and $\hat{p}_2$. The statistic $\hat{p}_1 - \hat{p}_2$ is the difference of these two independent random variables.

In Chapter 6, we learned that for any two random variables X and Y,

$$\mu_{X-Y} = \mu_X - \mu_Y$$

For the random variables $\hat{p}_1$ and $\hat{p}_2$, we have

$$\mu_{\hat{p}_1-\hat{p}_2} = \mu_{\hat{p}_1} - \mu_{\hat{p}_2} = p_1 - p_2$$

We also learned in Chapter 6 that for *independent* random variables X and Y,

$$\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y$$

For the random variables $\hat{p}_1$ and $\hat{p}_2$, we have

$$\sigma^2_{\hat{p}_1-\hat{p}_2} = \sigma^2_{\hat{p}_1} + \sigma^2_{\hat{p}_2} = \left(\sqrt{\frac{p_1(1-p_1)}{n_1}}\right)^2 + \left(\sqrt{\frac{p_2(1-p_2)}{n_2}}\right)^2$$

$$= \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

So $\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}$.

When the conditions are met, we can use the Normal density curve shown in Figure 7.11 to model the sampling distribution of $\hat{p}_1 - \hat{p}_2$. Note that this would allow us to calculate probabilities involving $\hat{p}_1 - \hat{p}_2$ with a Normal distribution.

When we analyzed the results of randomized experiments in Section 4.3, we used simulation to create a *randomization distribution* by repeatedly reallocating individuals to treatment groups. Fortunately for us, randomization distributions of $\hat{p}_1 - \hat{p}_2$ roughly follow the same rules for shape, center, and variability as sampling distributions of $\hat{p}_1 - \hat{p}_2$.

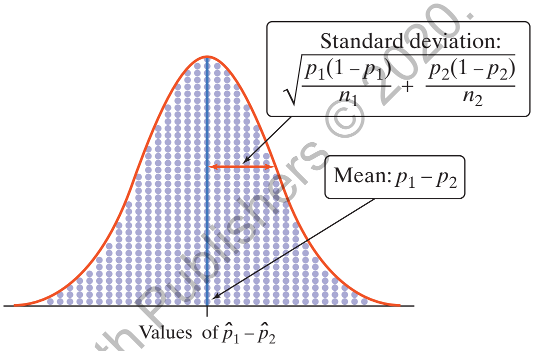


FIGURE 7.11 Select independent SRSs from two populations having proportions of successes p_1 and p_2. The proportions of successes in the two samples are $\hat{p}_1$ and $\hat{p}_2$. When the conditions are met, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately Normal with mean $p_1 - p_2$ and standard deviation $\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}$.

EXAMPLE

Yummy goldfish!

The sampling distribution of $\hat{p}_1 - \hat{p}_2$

PROBLEM: Your teacher brings two bags of colored goldfish crackers to class. Bag 1 has 25% red crackers and Bag 2 has 35% red crackers. Each bag contains more than 1000 crackers. Using a paper cup, your teacher takes an SRS of 50 crackers from Bag 1 and an independent SRS of 40 crackers from Bag 2. Let $\hat{p}_1 - \hat{p}_2$ be the difference in the sample proportions of red crackers.

(a) What is the shape of the sampling distribution of $\hat{p}_1 - \hat{p}_2$? Why?

(b) Find the mean of the sampling distribution.

(c) Calculate and interpret the standard deviation of the sampling distribution. Verify that the 10% condition is met.

(d) Estimate the probability that the proportion of red crackers in the sample from Bag 1 is greater than the proportion of red crackers in the sample from Bag 2.

Envision/Getty Images

SOLUTION:

(a) Approximately Normal, because $n_1 p_1 = 50(0.25) = 12.5$, $n_1(1 - p_1) = 50(0.75) = 37.5$, $n_2 p_2 = 40(0.35) = 14$, and $n_2(1 - p_2) = 40(0.65) = 26$ are all ≥ 10.

Note that these values are the expected numbers of successes and failures in the two samples.

(b) $\mu_{\hat{p}_1 - \hat{p}_2} = 0.25 - 0.35 = -0.10$

$\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$

(c) Because $50 < 10\%$ of all crackers in Bag 1 and $40 < 10\%$ of all crackers in Bag 2,

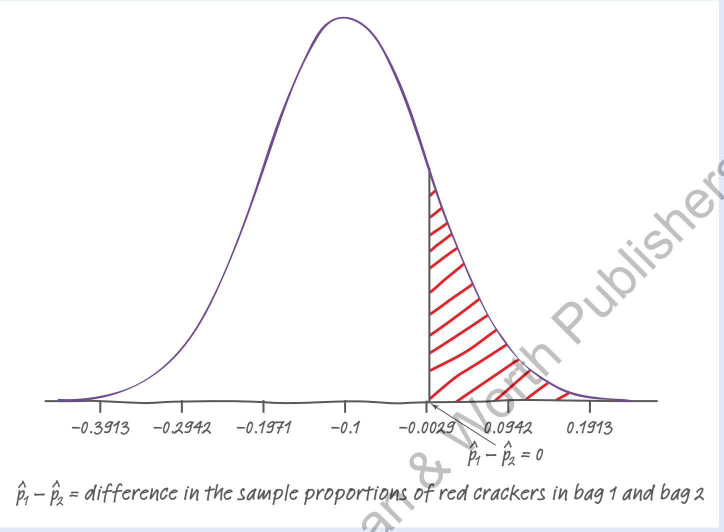
$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{0.25(0.75)}{50} + \frac{0.35(0.65)}{40}} = 0.0971$$

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}$$

The difference (Bag 1 − Bag 2) in the sample proportions of red goldfish crackers typically varies by about 0.097 from the true difference in proportions of −0.10.

(d) We want to find $P(\hat{p}_1 > \hat{p}_2)$, which is equivalent to $P(\hat{p}_1 - \hat{p}_2 > 0)$.

1. Draw a Normal distribution.



(i) $z = \dfrac{0 - (-0.10)}{0.0971} = 1.03$

Using Table A: $P(\hat{p}_1 - \hat{p}_2 > 0) = P(z > 1.03) = 1 - 0.8485 = 0.1515$

Using technology: normalcdf (lower:1.03, upper:1000, mean:0, SD:1) = 0.1515

(ii) normalcdf(lower:0, upper:1000, mean: −0.10, SD:0.0971) = 0.1515

2. Perform calculations.
(i) Standardize and use Table A or technology; or
(ii) Use technology without standardizing.
Be sure to answer the question that was asked.

FOR PRACTICE, TRY EXERCISE 49

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 the sample proportion** $\hat{p}$ describes how the statistic $\hat{p}$ varies in all possible samples of the same size from the population.
    - **Shape:** The sampling distribution of $\hat{p}$ is **approximately Normal** when both $np \geq 10$ and $n(1 - p) \geq 10$ (the *Large Counts condition*).

  - **Center:** The **mean** of the sampling distribution of $\hat{p}$ is $\mu_{\hat{p}} = p$. So $\hat{p}$ is an unbiased estimator of p.
  - **Variability:** The **standard deviation** of the sampling distribution of $\hat{p}$ is approximately $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$ for an SRS of size n. This formula can be used if the sample size is less than 10% of the population size (the *10% condition*).
- Choose independent SRSs of size n_1 from Population 1 with proportion of successes p_1 and of size n_2 from Population 2 with proportion of successes p_2. The sampling distribution of $\hat{p}_1 - \hat{p}_2$ has the following properties:
  - **Shape: Approximately Normal** if the Large Counts condition is met: n_1p_1, $n_1(1-p_1)$, n_2p_2, and $n_2(1-p_2)$ are all at least 10.
  - **Center:** The **mean** of the sampling distribution is $\mu_{\hat{p}_1-\hat{p}_2} = p_1 - p_2$.
  - **Variability:** The **standard deviation** of the sampling distribution is approximately $\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}$ as long as the 10% condition is met: $n_1 < 0.10N_1$ and $n_2 < 0.10N_2$.

Section 7.2 Exercises

pg 490

33. **Registered voters** In a congressional district, 55% of registered voters are Democrats. A polling organization selects a random sample of 500 registered voters from this district. Let $\hat{p}$ = the proportion of Democrats in the sample.

(a) Identify the mean of the sampling distribution of $\hat{p}$.

(b) Calculate and interpret the standard deviation of the sampling distribution of $\hat{p}$. Verify that the 10% condition is met.

(c) Describe the shape of the sampling distribution of $\hat{p}$. Justify your answer.

34. **Married with children** According to a recent U.S. Bureau of Labor Statistics report, the proportion of married couples with children in which both parents work outside the home is 59%.[6] You select an SRS of 50 married couples with children and let $\hat{p}$ = the sample proportion of couples in which both parents work outside the home.

(a) Identify the mean of the sampling distribution of $\hat{p}$.

(b) Calculate and interpret the standard deviation of the sampling distribution of $\hat{p}$. Verify that the 10% condition is met.

(c) Describe the shape of the sampling distribution of $\hat{p}$. Justify your answer.

35. **Orange Skittles®** The makers of Skittles claim that 20% of Skittles candies are orange. Suppose this claim is true. You select a random sample of 30 Skittles from a large bag. Let $\hat{p}$ = the proportion of orange Skittles in the sample.

(a) Identify the mean of the sampling distribution of $\hat{p}$.

(b) Calculate and interpret the standard deviation of the sampling distribution of $\hat{p}$. Verify that the 10% condition is met.

(c) Describe the shape of the sampling distribution of $\hat{p}$. Justify your answer.

36. **Male workers** A factory employs 3000 unionized workers, 90% of whom are male. A random sample of 15 workers is selected for a survey about worker satisfaction. Let $\hat{p}$ = the proportion of males in the sample.

(a) Identify the mean of the sampling distribution of $\hat{p}$.

(b) Calculate and interpret the standard deviation of the sampling distribution of $\hat{p}$. Verify that the 10% condition is met.

(c) Describe the shape of the sampling distribution of $\hat{p}$. Justify your answer.

37. **More Skittles®** 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. **More workers** 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. **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.

(a) Show that the 10% condition is not met in this case.

(b) What effect does violating the 10% condition have on the standard deviation of the sampling distribution of $\hat{p}$?

40. **Don't pick me!** Instead of collecting homework from all of her students, Mrs. Friedman randomly selects 5 of her 30 students and collects homework from only those students. Let $\hat{p}$ be the proportion of students in the sample that completed their homework.

(a) Show that the 10% condition is not met in this case.

(b) What effect does violating the 10% condition have on the standard deviation of the sampling distribution of $\hat{p}$?

41. **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 $\hat{p}$?

(b) Find the standard deviation of the sampling distribution of $\hat{p}$. Verify that the 10% condition is met.

(c) Verify that the sampling distribution of $\hat{p}$ is approximately Normal.

(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, assuming the milk industry spokesman's claim is true.

(e) Does this poll give convincing evidence against the spokesman's claim? Explain your reasoning.

42. **Do you go to church?** The Gallup Poll asked a random sample of 1785 adults if 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 $\hat{p}$?

(b) Find the standard deviation of the sampling distribution of $\hat{p}$. Verify that the 10% condition is met.

(c) Verify that the sampling distribution of $\hat{p}$ is approximately Normal.

(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, assuming the newspaper report's claim is true.

(e) Does this poll give convincing evidence against the newspaper's claim? Explain your reasoning.

pg 492

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

44. **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?

45. **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?

(b) Based on your answer to part (a), is there convincing evidence that less than 90% of all orders from this company are shipped within three working days? Explain your reasoning.

46. **Wait times** A hospital claims that 75% of people who come to its emergency room are seen by a doctor within 30 minutes of checking in. To verify this claim, an auditor inspects the medical records of 55 randomly selected patients who checked into the emergency room during the last year. Only 32 (58.2%) of these patients were seen by a doctor within 30 minutes of checking in.

(a) If the wait time is less than 30 minutes for 75% of all patients in the emergency room, what is the probability that the proportion of patients who wait less than 30 minutes is 0.582 or less in a random sample of 55 patients?

(b) Based on your answer to part (a), is there convincing evidence that less than 75% of all patients in the emergency room wait less than 30 minutes? Explain your reasoning.

47. **AP® enrollment** Suppose that 30% of female students and 25% of male students at a large high school are enrolled in an AP® class. Independent random samples of 20 females and 20 males are selected and are asked if they are enrolled in an AP® class. Let $\hat{p}_F$ represent the sample proportion of females enrolled in an AP® class and let $\hat{p}_M$ represent the sample proportion of males enrolled in an AP® class.

(a) Find the mean of the sampling distribution of $\hat{p}_F - \hat{p}_M$.

(b) Calculate and interpret the standard deviation of the sampling distribution. Verify that the 10% condition is met.

(c) Is the shape of the sampling distribution approximately Normal? Justify your answer.

48. **Athletic participation** Suppose that 20% of students at high school A and 18% of students at high school B participate on a school athletic team. Independent random samples of 30 students from each school are selected and are asked if they participate on a school athletic team. Let $\hat{p}_A$ represent the sample proportion of students at school A who participate on a school athletic team and let $\hat{p}_B$ represent the sample proportion of students at school B who participate on a school athletic team.

(a) Find the mean of the sampling distribution of $\hat{p}_A - \hat{p}_B$.

(b) Calculate and interpret the standard deviation of the sampling distribution. Verify that the 10% condition is met.

(c) Is the shape of the sampling distribution approximately Normal? Justify your answer.

49. **I want red!** A candy maker offers Child and Adult bags of jelly beans with different color mixes. The company claims that the Child mix has 30% red jelly beans, while the Adult mix contains 15% red jelly beans. Assume that the candy maker's claim is true. Suppose we take a random sample of 50 jelly beans from the Child mix and an independent random sample of 100 jelly beans from the Adult mix. Let $\hat{p}_C$ and $\hat{p}_A$ be the sample proportions of red jelly beans from the Child and Adult mixes, respectively.

pg 496

(a) What is the shape of the sampling distribution of $\hat{p}_C - \hat{p}_A$? Why?

(b) Find the mean of the sampling distribution.

(c) Calculate and interpret the standard deviation of the sampling distribution. Verify that the 10% condition is met.

(d) Find the probability that the proportion of red jelly beans in the Child sample is less than or equal to the proportion of red jelly beans in the Adult sample, assuming that the company's claim is true.

50. **Literacy** A researcher reports that 80% of high school graduates, but only 40% of high school dropouts, would pass a basic literacy test.[7] Assume that the researcher's claim is true. Suppose we give a basic literacy test to a random sample of 60 high school graduates and an independent random sample of 75 high school dropouts. Let $\hat{p}_G$ and $\hat{p}_D$ be the sample proportions of graduates and dropouts, respectively, who pass the test.

(a) What is the shape of the sampling distribution of $\hat{p}_G - \hat{p}_D$? Why?

(b) Find the mean of the sampling distribution.

(c) Calculate and interpret the standard deviation of the sampling distribution. Verify that the 10% condition is met.

(d) Find the probability that the proportion of graduates who pass the test is at most 0.20 higher than the proportion of dropouts who pass, assuming that the researcher's report is correct.

51. **I want red!** Refer to Exercise 49. Suppose that the Child and Adult samples contain an equal proportion of red jelly beans. Based on your result in part (d) of Exercise 49, would this give you reason to doubt the company's claim? Explain your reasoning.

52. **Literacy** Refer to Exercise 50. Suppose that the difference (Graduate – Dropout) in the sample proportions who pass the test is exactly 0.20. Based on your result in part (d) in Exercise 50, would this give you reason to doubt the researcher's claim? Explain your reasoning.

Multiple Choice *Select the best answer for Exercises 53–56.*

Exercises 53–55 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.

53. 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

54. Decreasing the sample size from 750 to 375 would multiply the standard deviation by

(a) 2.

(b) $\sqrt{2}$

(c) 1/2.

(d) $1/\sqrt{2}$.

(e) none of these.

55. 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.

56. 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) $P\left(z < \dfrac{0.50 - 0.55}{100}\right)$

(b) $P\left(z < \dfrac{0.50 - 0.55}{\sqrt{\dfrac{0.55(0.45)}{100}}}\right)$

(c) $P\left(z < \dfrac{0.55 - 0.50}{\sqrt{\dfrac{0.55(0.45)}{100}}}\right)$

(d) $P\left(z < \dfrac{0.50 - 0.55}{\sqrt{100(0.55)(0.45)}}\right)$

(e) $P\left(z < \dfrac{0.55 - 0.50}{\sqrt{100(0.55)(0.45)}}\right)$

Recycle and Review

57. **Sharing music online** (5.2, 5.3) 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.[8]

(a) Make a two-way table that displays this information.

(b) What percent of Internet users neither download nor share music files?

(c) Given that an Internet user downloads music files online, what is the probability that this person also shares music files?

58. **Whole grains** (4.2) A series of observational studies revealed that people who typically consume 3 servings of whole grain per day have about a 20% lower risk of dying from heart disease and about a 15% lower risk of dying from stroke or cancer than those who consume no whole grains.[9]

(a) Explain how confounding makes it difficult to establish a cause-and-effect relationship between whole grain consumption and risk of dying from heart disease, stroke, or cancer, based on these studies.

(b) Explain how researchers could establish a cause-and-effect relationship in this context.

SECTION 7.3 Sample Means

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

- Calculate the mean and standard deviation of the sampling distribution of a sample mean $\bar{x}$ and interpret the standard deviation.
- Explain how the shape of the sampling distribution of $\bar{x}$ is affected by the shape of the population distribution and the sample size.
- Calculate the mean and the standard deviation of the sampling distribution of a difference in sample means $\bar{x}_1 - \bar{x}_2$ and interpret the standard deviation.
- Determine if the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is approximately Normal.
- If appropriate, use a Normal distribution to calculate probabilities involving $\bar{x}$ or $\bar{x}_1 - \bar{x}_2$.

Sample proportions arise most often when we are interested in categorical variables. We then ask questions like "What proportion of U.S. adults has 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.

The Sampling Distribution of $\bar{x}$

When Mrs. Gallas's class did the "Penny for your thoughts" activity at the beginning of the chapter, her students produced the "dotplots" in Figure 7.12. These graphs approximate the **sampling distribution of the sample mean** year of penny for samples of size $n = 5$ and for samples of size $n = 20$ from Mrs. Gallas's population of pennies.

DEFINITION **Sampling distribution of the sample mean**

The **sampling distribution of the sample mean $\bar{x}$** describes the distribution of values taken by the sample mean $\bar{x}$ in all possible samples of the same size from the same population.

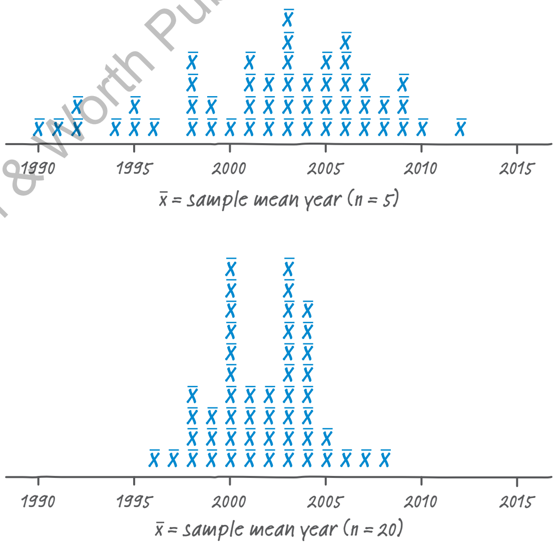


FIGURE 7.12 Approximate sampling distribution of the sample mean year of pennies in 50 samples of size $n = 5$ and 50 samples of size $n = 20$ from a population of pennies.

How do these approximate sampling distributions compare?

- *Shape*: The distribution of $\bar{x}$ is slightly skewed to the left when using samples of size $n = 5$ but roughly symmetric when using samples of size $n = 20$.
- *Center*: The distribution of $\bar{x}$ is centered at around 2002 for both sample sizes ($\mu_{\bar{x}} \approx 2002$).
- *Variability*: The distribution of $\bar{x}$ is about half as variable when using samples of size $n = 20$ ($\sigma_{\bar{x}} \approx 2.6$) than with samples of size $n = 5$ ($\sigma_{\bar{x}} \approx 5.2$).

Like the sampling distribution of $\hat{p}$, there are some simple rules that describe the mean and standard deviation of the sampling distribution of $\bar{x}$. Describing the shape of the sampling distribution of $\bar{x}$ is more complicated, so we'll save that for later.

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

Suppose that $\bar{x}$ is the mean of an SRS of size n drawn from a large population with mean μ and standard deviation σ. 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 approximately

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

as long as the *10% condition* is satisfied: $n < 0.10N$. The value $\sigma_{\bar{x}}$ measures the typical distance between a sample mean $\bar{x}$ and the population mean μ.

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 μ.
- The variability of $\bar{x}$ depends on both the variability in the population σ and the sample size n. Values of $\bar{x}$ will be more variable for populations that have more variability. Values of $\bar{x}$ will be less variable for larger samples. *Specifically, multiplying the sample size by 4 cuts the standard deviation in half.*
- When we sample *with* replacement, the standard deviation of the sampling distribution of $\bar{x}$ is exactly $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. When we sample *without* replacement, the observations are not independent and the actual standard deviation of the sampling distribution of $\bar{x}$ is smaller than the value given by the formula. If the sample size is less than 10% of the population size, however, the value given by the formula is nearly correct. This doesn't mean you should aim for a small sample size! Bigger samples provide more information than smaller samples. If the sample size is more than 10% of the population size, we need to use the finite population correction to calculate the standard deviation of the sampling distribution of $\bar{x}$. We'll avoid these situations in this text.

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

AP® EXAM TIP

Notation matters. The symbols $\hat{p}$, $\bar{x}$, n, p, μ, σ, $\mu_{\hat{p}}$, $\sigma_{\hat{p}}$, $\mu_{\bar{x}}$, 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.

EXAMPLE Been to the movies recently?

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

PROBLEM: The number of movies viewed in the last year by students at a large high school has a mean of 19.3 movies with a standard deviation of 15.8 movies. Suppose we take an SRS of 100 students from this school and calculate $\bar{x}$ = the mean number of movies viewed by the members of the sample.

(a) Identify the mean of the sampling distribution of $\bar{x}$.

(b) Calculate and interpret the standard deviation of the sampling distribution of $\bar{x}$. Verify that the 10% condition is met.

PatriciaPix/Getty Images

SOLUTION:

(a) $\mu_{\bar{x}} = 19.3$ movies

(b) Assuming that $n = 100$ is less than 10% of students at the large high school, the standard deviation is approximately $\sigma_{\bar{x}} = \dfrac{15.8}{\sqrt{100}} = 1.58$ movies.

In SRSs of size 100, the sample mean number of movies viewed will typically vary by about 1.58 movies from the true mean of 19.3 movies.

$\mu_{\bar{x}} = \mu$

When $n < 0.10N$, $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$

FOR PRACTICE, TRY EXERCISE 59

Think About It

WHERE DO THE FORMULAS FOR THE MEAN AND STANDARD DEVIATION OF $\bar{x}$ COME FROM? Choose an SRS of size n from a population, and measure a quantitative variable X on each individual in the sample. Call the individual measurements $X_1, X_2, \ldots, X_n$. If the population is large relative to the sample, we can think of these X_i's as independent random variables, each with mean μ and standard deviation σ. 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

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

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

$$\mu_{\bar{x}} = \frac{1}{n}\mu_T = \frac{1}{n}(n\mu) = \mu$$

$$\sigma_{\bar{x}} = \frac{1}{n}\sigma_T = \frac{1}{n}(\sigma\sqrt{n}) = \sigma\sqrt{\frac{n}{n^2}} = \sigma\sqrt{\frac{1}{n}} = \sigma\frac{1}{\sqrt{n}} = \frac{\sigma}{\sqrt{n}}$$

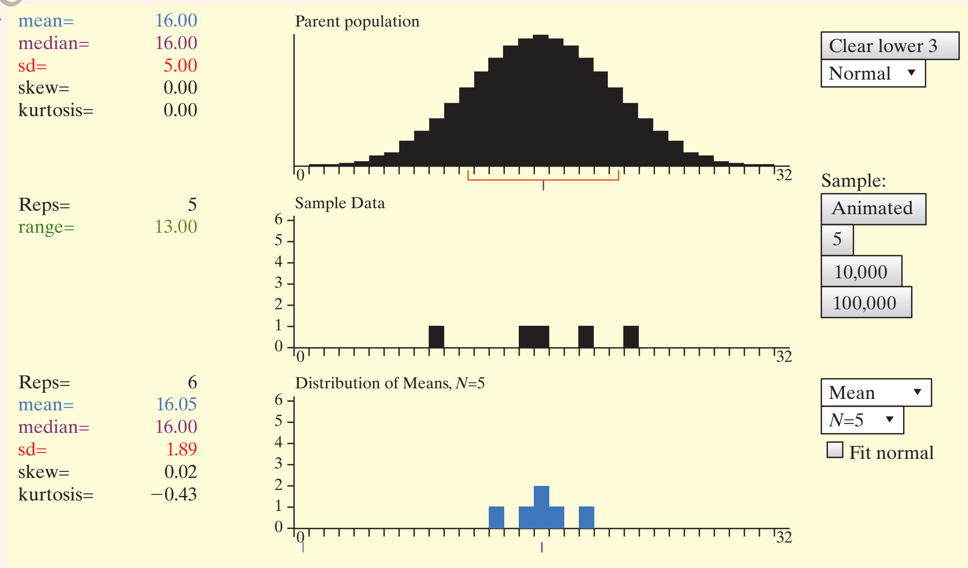
Sampling from a Normal Population

We have described the mean and standard deviation of the sampling distribution of a sample mean $\overline{x}$ but not its shape. That's because the shape of the sampling distribution of $\overline{x}$ 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 $\overline{x}$ for a Normal population

Professor David Lane of Rice University has developed a wonderful applet for investigating the sampling distribution of $\overline{x}$. 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. Go to http://onlinestatbook.com/stat_sim/sampling_dist/ or search for "online statbook sampling distributions applet" and go to the website. 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 screen shot.
2. There are choices for the population distribution: Normal, uniform, skewed, and custom. Keep the default option: 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?


mean= 16.00
median= 16.00
sd= 5.00
skew= 0.00
kurtosis= 0.00

Reps= 5
range= 13.00

Reps= 6
mean= 16.05
median= 16.00
sd= 1.89
skew= 0.02
kurtosis= −0.43

3. Click on "Clear lower 3" to start clean. Then click on the "100,000" button under "Sample:" to simulate taking 100,000 SRSs of size $n = 5$ from the population. Answer these questions:
   - Does the simulated sampling distribution of $\bar{x}$ (blue bars) have a recognizable shape? Click the box next to "Fit normal."
   - To the left of each distribution is a set of summary statistics. Compare the mean of the simulated sampling distribution with the mean of the population.
   - How is the standard deviation of the simulated 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 100,000 times. How do the two distributions of $\bar{x}$ compare: shape, center, and variability?
5. What have you learned about the shape of the sampling distribution of $\bar{x}$ when the population has a Normal shape?

As the preceding 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 THE SAMPLE MEAN $\bar{x}$ WHEN SAMPLING FROM A NORMAL POPULATION

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

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

Sampling from a Normal population

Alamy Images

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

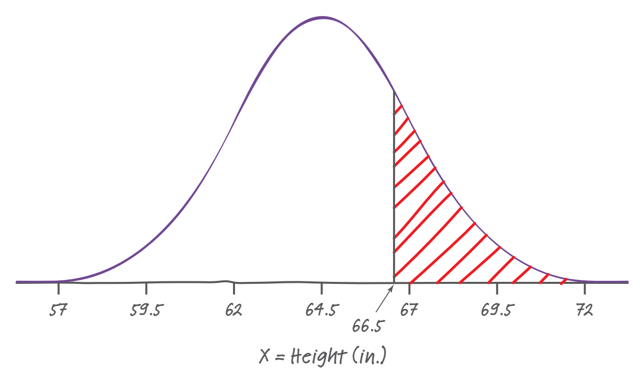
(a) Find the probability that a randomly selected young woman is taller than 66.5 inches.

(b) Find the probability that the mean height of an SRS of 10 young women exceeds 66.5 inches.

SOLUTION:

(a) Let $X =$ height of a randomly selected young woman.

1. Draw a Normal distribution.



(i) $z = \dfrac{66.5 - 64.5}{2.5} = 0.80$

Using Table A: $P(X > 66.5) = P(z > 0.80) = 1 - 0.7881$
$= 0.2119$

Using technology: normalcdf(lower:0.80, upper:1000, mean:0, SD:1) = 0.2119

(ii) normalcdf(lower:66.5, upper:1000, mean:64.5, SD:2.5) = 0.2119

2. Perform calculations.
(i) Standardize and use Table A or technology
or
(ii) Use technology without standardizing.
Be sure to answer the question that was asked.

(b) Let $\bar{x} =$ mean height of 10 randomly selected young women.

$\mu_{\bar{x}} = 64.5$

Because $10 < 10\%$ of all young women,

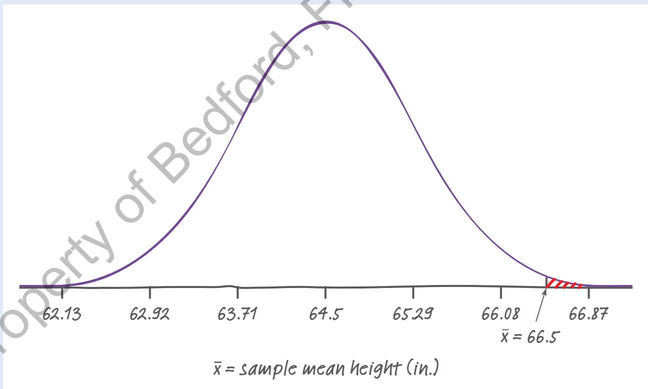
$\sigma_{\bar{x}} = \dfrac{2.5}{\sqrt{10}} = 0.79$

Calculate the mean and standard deviation of the sampling distribution of $\bar{x}$.

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

Because the population of heights is Normal, the distribution of $\bar{x}$ is also Normal.

Justify that the distribution of $\bar{x}$ is Normal.



(i) $z = \dfrac{66.5 - 64.5}{0.79} = 2.53$

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

Using technology: normalcdf(lower:2.53, upper:1000, mean:0, SD:1) = 0.0057

(ii) normalcdf(lower:66.5, upper:1000, mean:64.5, SD:0.79) = 0.0057

FOR PRACTICE, TRY EXERCISE 65

Figure 7.13 compares the population distribution and the sampling distribution of $\bar{x}$ for the example about young women's heights. 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.

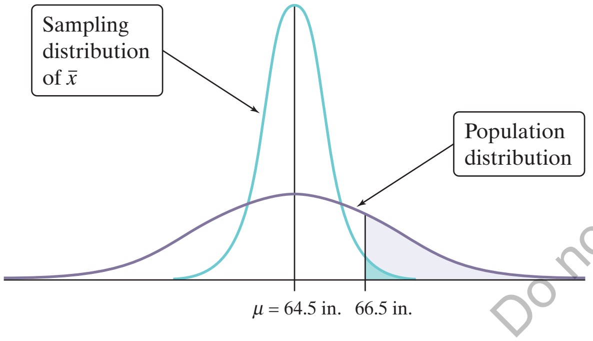
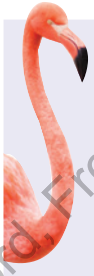


FIGURE 7.13 The sampling distribution of the mean height $\bar{x}$ for SRSs of 10 young women compared with the population distribution of young women's heights.

AP® EXAM TIP

Many students lose credit on probability calculations involving $\bar{x}$ because they forget to divide the population standard deviation by $\sqrt{n}$. Remember that averages are less variable than individual observations!

The fact that averages of several observations are less variable than individual observations is important in many settings. For example, it is common practice in science and medicine to repeat a measurement several times and report the average of the results.



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.

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 $\bar{x}$?
3. Calculate and interpret the standard deviation of the sampling distribution of $\bar{x}$. Verify that the 10% condition is met.
4. Find the probability that the mean pregnancy length for the women in the sample exceeds 270 days.

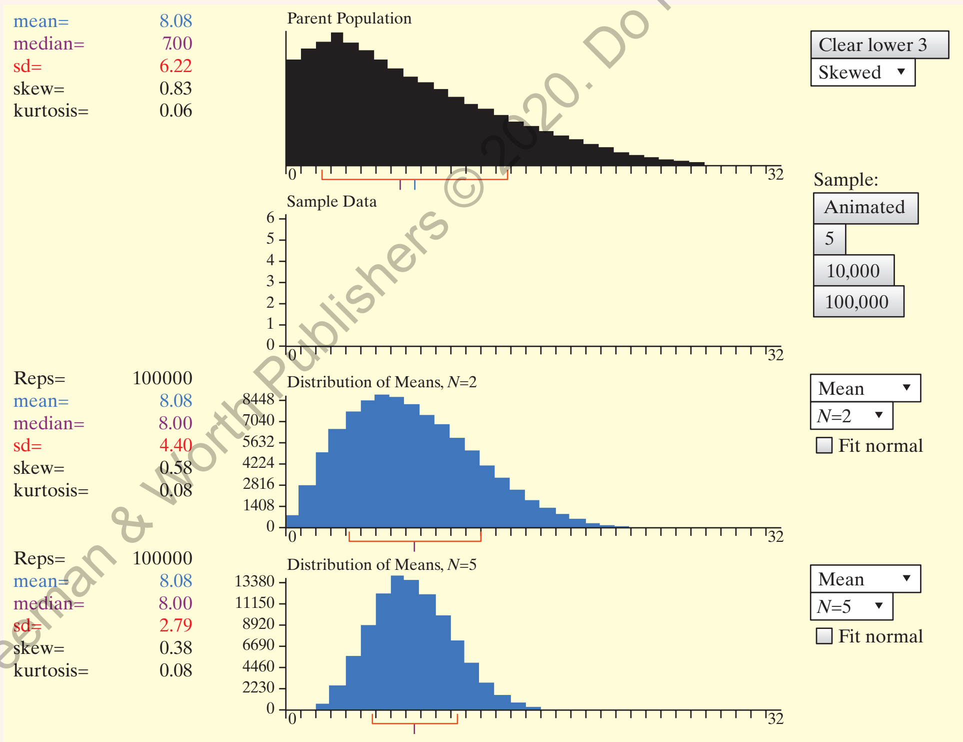
The Central Limit Theorem

Most population distributions are not Normal. What is the shape of the sampling distribution of $\bar{x}$ when sampling from a non-Normal population? The following activity sheds some light on this question.

ACTIVITY Exploring the sampling distribution of $\bar{x}$ for non-Normal populations

Let's use the sampling distributions applet from the preceding activity to investigate what happens when we start with a non-Normal population distribution.

1. Go to onlinestatbook.com/stat_sim/sampling_dist/ 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 100,000 SRSs. Describe what you see.



mean= 8.08
median= 7.00
sd= 6.22
skew= 0.83
kurtosis= 0.06

Reps= 100000
mean= 8.08
median= 8.00
sd= 4.40
skew= 0.58
kurtosis= 0.08

Reps= 100000
mean= 8.08
median= 8.00
sd= 2.79
skew= 0.38
kurtosis= 0.08

Skewed
Mean, N=2
Mean, N=5

2. Change the sample sizes to $n = 10$ and $n = 16$ and take 100,000 samples. What do you notice?

3. Now change the sample sizes to $n = 20$ and $n = 25$ and take 100,000 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 $\bar{x}$.

The screen shots in Figure 7.14 show the approximate sampling distributions of $\bar{x}$ for samples of size $n = 2$ and samples of size $n = 25$ from three different populations.

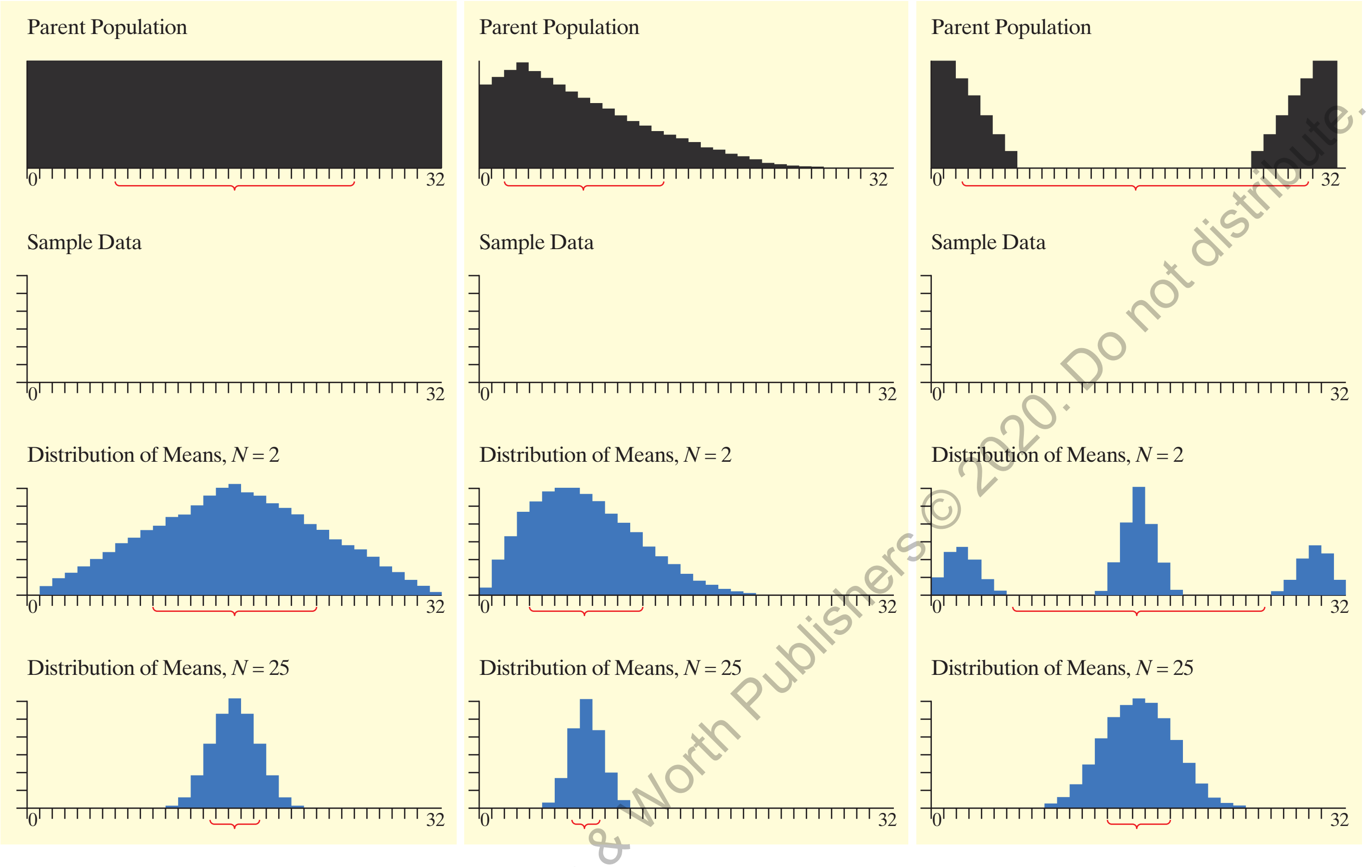


FIGURE 7.14 Approximate sampling distributions of $\bar{x}$ for different population shapes and sample sizes.

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 mean μ and standard deviation σ, and that the observations in the sample are independent. This important 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 μ and standard deviation σ. The **central limit theorem (CLT)** says that when n is sufficiently 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.

As Figure 7.14 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.*

SHAPE OF THE SAMPLING DISTRIBUTION OF THE SAMPLE MEAN $\bar{x}$

- If the population distribution is Normal, the sampling distribution of $\bar{x}$ will also be Normal, no matter what the sample size n is.
- If the population distribution is not Normal, the sampling distribution of $\bar{x}$ will be approximately Normal when the sample size is sufficiently large ($n \geq 30$ in most cases). If the sample size is small and the population distribution is not Normal, the sampling distribution of $\bar{x}$ will retain some characteristics of the population distribution (e.g., skewness).

EXAMPLE

Free oil changes
Calculations using the CLT

Joe Belanger/Alamy

PROBLEM: Keith is the manager of an auto-care center. Based on service records from the past year, the time (in hours) that a technician requires to complete a standard oil change and inspection follows a right-skewed distribution with $\mu = 30$ minutes and $\sigma = 20$ minutes. For a promotion, Keith randomly selects 40 current customers and offers them a free oil change and inspection if they redeem the offer during the next month. Keith budgets an average of 35 minutes per customer for a technician to complete the work. Will this be enough?

(a) Calculate the probability that the average time it takes to complete the work exceeds 35 minutes.

(b) How much average time per customer should Keith budget if he wants to be 99% certain that he doesn't go "over budget"?

SOLUTION:

(a) Let $\bar{x}$ = sample mean time to complete work (in minutes).

$\mu_{\bar{x}} = 30$

Assuming $40 < 10\%$ of all current customers,

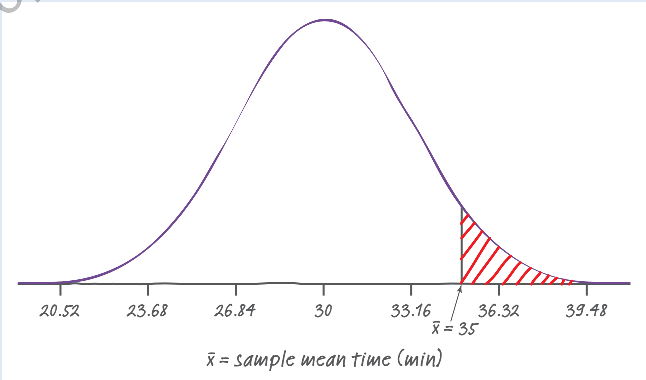
$$\sigma_{\bar{x}} = \frac{20}{\sqrt{40}} = 3.16$$

Because the sample size is large ($40 \geq 30$), the distribution of $\bar{x}$ is approximately Normal.

Calculate the mean and standard deviation of the sampling distribution of $\bar{x}$.

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

Justify that the distribution of $\bar{x}$ is approximately Normal.



1. Draw a Normal distribution.

(i) $z = \dfrac{35-30}{3.16} = 1.58$

Using Table A: $P(\bar{x} > 35) = P(z > 1.58) = 1 - 0.9429 = 0.0571$

Using technology:
normalcdf(lower:1.58, upper:1000, mean:0, SD:1) = 0.0571

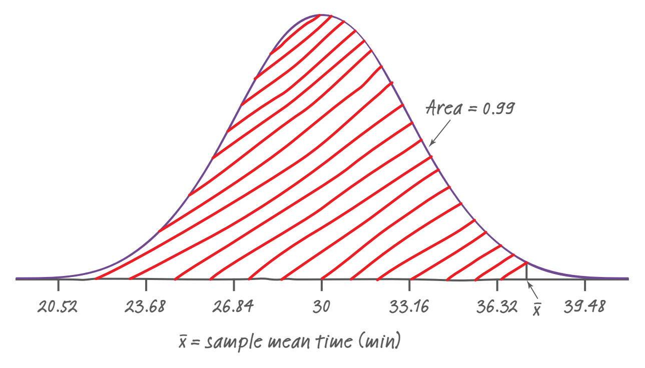
(ii) normalcdf(lower:35, upper:1000, mean:30, SD:3.16) = 0.0568

There is only a 5.68% probability that Keith hasn't budgeted enough time to complete the work.

2. Perform calculations.

(i) Standardize and use Table A or technology; or
(ii) Use technology without standardizing.
Be sure to answer the question that was asked.

(b)



1. Draw a Normal distribution.

(i) *Using Table A:* 0.99 area to left → $z = 2.33$

Using technology: invnorm(area:0.99, mean:0, SD:1) = 2.33

$$2.33 = \frac{\bar{x}-30}{3.16} \rightarrow \bar{x} = 37.4 \text{ minutes}$$

(ii) invnorm(area:0.99, mean:30, SD:3.16) = 37.4 minutes

To be 99% sure he has budgeted enough time, Keith should plan for an average of 37.4 minutes per customer.

2. Perform calculations.

(i) Use Table A or technology to find the value of z with the indicated area under the standard Normal curve, then "unstandardize" to transform back to the original distribution; or
(ii) Use technology to find the desired value without standardizing.
Be sure to answer the question that was asked.

FOR PRACTICE, TRY EXERCISE 73

What if Keith decided to give away only 10 free oil changes? Because the population distribution is skewed to the right and the sample size is small $(10 < 30)$, we can't use a Normal distribution to do probability calculations. The sampling distribution of $\bar{x}$ is likely to be skewed to the right—although not as strongly as the population distribution itself.

The Sampling Distribution of a Difference Between Two Means

In the preceding section, we developed methods for comparing two proportions. What if we want to compare the mean of some quantitative variable for the individuals in Population 1 and Population 2, such as the mean income for high school graduates and non-high school graduates? Our parameters of interest are the population means μ_1 and μ_2. Once again, the best approach is to take independent random samples from each population and to compare the sample means $\bar{x}_1$ and $\bar{x}_2$.

Suppose we want to compare the average effectiveness of two headache medications in a randomized experiment. In this case, the parameters μ_1 and μ_2 are the true mean responses for Treatment 1 and Treatment 2, respectively. We use the mean response in the two groups, $\bar{x}_1$ and $\bar{x}_2$, to make the comparison. Here's a table that summarizes these two situations:

Population or treatment	Parameter	Statistic	Sample size
1	μ_1	$\bar{x}_1$	n_1
2	μ_2	$\bar{x}_2$	n_2

We compare the populations or treatments by doing inference about the difference $\mu_1 - \mu_2$ between the parameters. The statistic that estimates this difference is the difference between the two sample means, $\bar{x}_1 - \bar{x}_2$.

To explore the sampling distribution of $\bar{x}_1 - \bar{x}_2$, let's start with two Normally distributed populations having known means and standard deviations. Based on information from the U.S. National Health and Nutrition Examination Survey (NHANES), the heights of 10-year-old girls can be modeled by a Normal distribution with mean $\mu_G = 56.4$ inches and standard deviation $\sigma_G = 2.7$ inches. The heights of 10-year-old boys can be modeled by a Normal distribution with mean $\mu_B = 55.7$ inches and standard deviation $\sigma_B = 3.8$ inches.[10] The table summarizes this information.

Monkey Business Images/Shutterstock

Population	Shape	Mean	Standard deviation
10-year-old girls	Approximately Normal	$\mu_G = 56.4$ in	$\sigma_G = 2.7$ in
10-year-old boys	Approximately Normal	$\mu_B = 55.7$ in	$\sigma_B = 3.8$ in

Suppose we take independent SRSs of 12 girls and 8 boys of this age and measure their heights. What can we say about the difference $\bar{x}_G - \bar{x}_B$ in the average heights of the sample of girls and the sample of boys?

Earlier in this section, we saw that the sampling distribution of a sample mean $\bar{x}$ has the following properties:

Shape: (1) If the population distribution is Normal, then so is the sampling distribution of $\bar{x}$; (2) If the population distribution isn't Normal, the sampling distribution of $\bar{x}$ will be approximately Normal if the sample size is large enough (say, $n \geq 30$) by the central limit theorem (CLT).

Center: $\mu_{\bar{x}} = \mu$

Variability: $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$ if $n < 0.10N$

For the sampling distributions of $\bar{x}_G$ and $\bar{x}_B$ in this case:

	Sampling distribution of $\bar{x}_G$	Sampling distribution of $\bar{x}_B$
Shape	Approximately Normal, because the population distribution is approximately Normal	Approximately Normal, because the population distribution is approximately Normal
Center	$\mu_{\bar{x}_G} = \mu_G = 56.4$ inches	$\mu_{\bar{x}_B} = \mu_B = 55.7$ inches
Variability	$\sigma_{\bar{x}_G} = \dfrac{\sigma_G}{\sqrt{n_G}} = \dfrac{2.7}{\sqrt{12}} = 0.78$ inch because $12 < 10\%$ of all 10-year-old girls in the United States.	$\sigma_{\bar{x}_B} = \dfrac{\sigma_B}{\sqrt{n_B}} = \dfrac{3.8}{\sqrt{8}} = 1.34$ inches because $8 < 10\%$ of all 10-year-old boys in the United States.

What about the sampling distribution of $\bar{x}_G - \bar{x}_B$? We used software to take independent SRSs of 12 ten-year-old girls and 8 ten-year-old boys. Our first

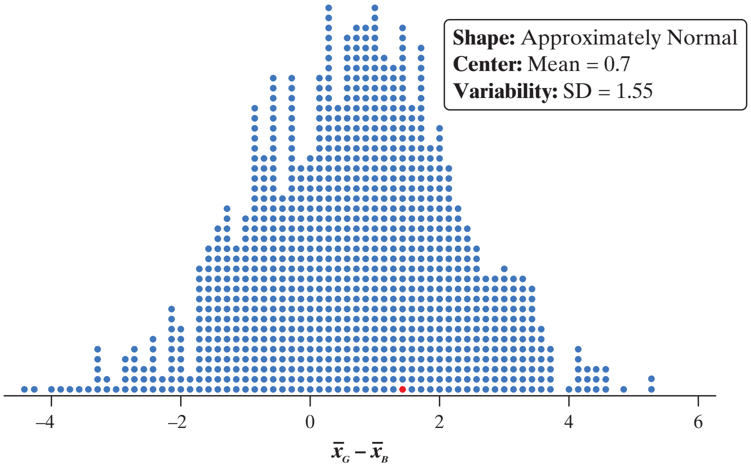


FIGURE 7.15 Simulated sampling distribution of the difference in sample means $\bar{x}_G - \bar{x}_B$ in 1000 SRSs of size $n_G = 12$ from an approximately Normally distributed population with $\mu_G = 56.4$ inches and $\sigma_G = 2.7$ inches and 1000 SRSs of size $n_B = 8$ from an approximately Normally distributed population with $\mu_B = 55.7$ inches and $\sigma_B = 3.8$ inches.

set of samples gave $\bar{x}_G = 56.09$ inches and $\bar{x}_B = 54.68$ inches, resulting in a difference of $\bar{x}_G - \bar{x}_B = 56.09 - 54.68 = 1.41$ inches. A red dot for this value appears in Figure 7.15. The dotplot shows the results of repeating this process 1000 times.

The figure suggests that the sampling distribution of $\bar{x}_G - \bar{x}_B$ has an approximately Normal shape. This makes sense from what you learned in Section 6.2 because we are subtracting two independent random variables, $\bar{x}_G$ and $\bar{x}_B$, that have approximately Normal distributions.

The mean of the sampling distribution is 0.7. The true mean height of all 10-year-old girls is $\mu_G = 56.4$ inches and the true mean height of all 10-year-old boys is $\mu_B = 55.7$ inches. We expect the difference $\bar{x}_G - \bar{x}_B$ to center on the actual difference in the population means, $\mu_G - \mu_B = 56.4 - 55.7 = 0.7$ inch. The standard deviation of the sampling distribution is 1.55 inches. It can be found using the formula

$$\sqrt{\frac{\sigma_G^2}{n_G} + \frac{\sigma_B^2}{n_B}} = \sqrt{\frac{2.7^2}{12} + \frac{3.8^2}{8}} = 1.55$$

That is, the difference (Girls – Boys) in the sample mean heights typically varies by about 1.55 inches from the true difference in mean heights of 0.7 inch.

THE SAMPLING DISTRIBUTION OF $\bar{x}_1 - \bar{x}_2$

Choose an SRS of size n_1 from Population 1 with mean μ_1 and standard deviation σ_1 and an independent SRS of size n_2 from Population 2 with mean μ_2 and standard deviation σ_2. Then:

- The sampling distribution of $\bar{x}_1 - \bar{x}_2$ is **Normal** if both population distributions are Normal. It is **approximately Normal** if both sample sizes are large $(n_1 \geq 30$ and $n_2 \geq 30)$ or if one population is Normally distributed and the other sample size is large.
- The **mean** of the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$.
- The **standard deviation** of the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is approximately

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

  as long as the *10% condition* is met for both samples: $n_1 < 0.10N_1$ and $n_2 < 0.10N_2$.

Note that the formula for the standard deviation is exactly correct only when we have two types of independence:

- Independent samples, so that we can add the variances of $\bar{x}_1$ and $\bar{x}_2$.
- Independent observations within each sample. When sampling without replacement, the actual value of the standard deviation is smaller than the

formula suggests. However, if the 10% condition is met for both samples, the difference is negligible.

The standard deviation of the sampling distribution tells us how much the difference in sample means will typically vary from the difference in the population means if we repeat the random sampling process many times.

Think About It

WHERE DO THE FORMULAS FOR THE MEAN AND STANDARD DEVIATION OF THE SAMPLING DISTRIBUTION OF $\bar{x}_1 - \bar{x}_2$ COME FROM? Both $\bar{x}_1$ and $\bar{x}_2$ are random variables. That is, their values would vary in repeated independent SRSs of size n_1 and n_2. Independent random samples yield independent random variables $\bar{x}_1$ and $\bar{x}_2$. The statistic $\bar{x}_1 - \bar{x}_2$ is the difference of these two independent random variables.

In Chapter 6, we learned that for any two random variables X and Y,

$$\mu_{X-Y} = \mu_X - \mu_Y$$

For the random variables $\bar{x}_1$ and $\bar{x}_2$, we have

$$\mu_{\bar{x}_1-\bar{x}_2} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2} = \mu_1 - \mu_2$$

We also learned in Chapter 6 that for *independent* random variables X and Y,

$$\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y$$

For the random variables $\bar{x}_1$ and $\bar{x}_2$, we have

$$\sigma^2_{\bar{x}_1-\bar{x}_2} = \sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2} = \left(\frac{\sigma_1}{\sqrt{n_1}}\right)^2 + \left(\frac{\sigma_2}{\sqrt{n_2}}\right)^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

So $\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$.

When the conditions are met, we can use the Normal density curve shown in Figure 7.16 to model the sampling distribution of $\bar{x}_1 - \bar{x}_2$. Note that this would allow us to calculate probabilities involving $\bar{x}_1 - \bar{x}_2$ with a Normal distribution.

When we analyzed the results of randomized experiments in Section 4.3, we used simulation to create a *randomization distribution* by repeatedly reallocating individuals to treatment groups. Fortunately for us, randomization distributions of $\bar{x}_1 - \bar{x}_2$ roughly follow the same rules for shape, center, and variability as sampling distributions of $\bar{x}_1 - \bar{x}_2$.

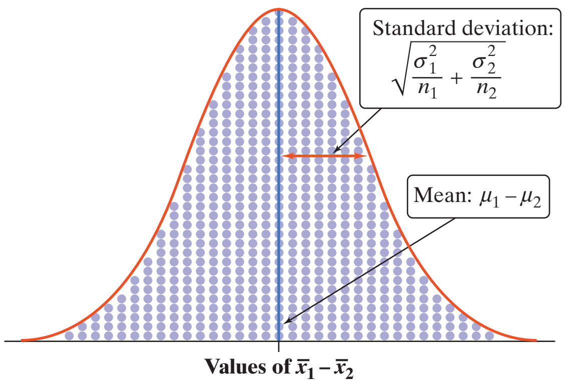


FIGURE 7.16 Select independent SRSs from two populations having means μ_1 and μ_2 and standard deviations σ_1 and σ_2. The two sample means are $\bar{x}_1$ and $\bar{x}_2$. When the conditions are met, the sampling distribution of the difference $\bar{x}_1 - \bar{x}_2$ is approximately Normal with mean $\mu_1 - \mu_2$ and standard deviation $\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$.

EXAMPLE

Medium or large drink?

Describing the sampling distribution of $\bar{x}_1 - \bar{x}_2$

RichLegg/Getty Images

PROBLEM: A fast-food restaurant uses an automated filling machine to pour its soft drinks. The machine has different settings for small, medium, and large drink cups. According to the machine's manufacturer, when the large setting is chosen, the amount of liquid L dispensed by the machine follows a Normal distribution with mean 27 ounces and standard deviation 0.8 ounce. When the medium setting is chosen, the amount of liquid M dispensed follows a Normal distribution with mean 17 ounces and standard deviation 0.5 ounce. To test this claim, the manager selects independent random samples of 20 cups filled using the large setting and 25 cups filled using the medium setting during one week. Let $\bar{x}_L - \bar{x}_M$ be the difference in the sample mean amount of liquid under the two settings.

(a) What is the shape of the sampling distribution of $\bar{x}_L - \bar{x}_M$? Why?

(b) Find the mean of the sampling distribution.

(c) Calculate and interpret the standard deviation of the sampling distribution. Verify that the 10% condition is met.

(d) Estimate the probability that the difference in sample means is within 0.25 ounce of the true difference in means.

SOLUTION:

(a) Normal, because both population distributions are Normal.

(b) $\mu_{\bar{x}_L - \bar{x}_M} = 27 - 17 = 10$ ounces

$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$

(c) Because $20 < 10\%$ of all large cups of soft drinks and $25 < 10\%$ of all medium cups of soft drinks that week,

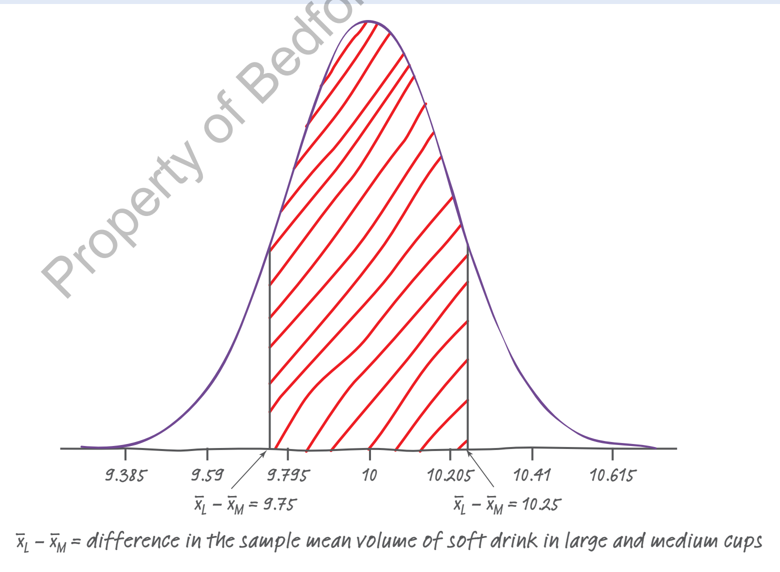
$$\sigma_{\bar{x}_L - \bar{x}_M} = \sqrt{\frac{0.80^2}{20} + \frac{0.50^2}{25}} = 0.205 \text{ ounce}$$

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

The difference (Large cup − Medium cup) in the sample mean amounts of liquid typically varies by about 0.2 ounce from the true difference in means of 10 ounces.

(d) We want to find $P(9.75 \le \bar{x}_L - \bar{x}_M \le 10.25)$.

1. Draw a Normal distribution.


9.385
9.59
9.795
10
10.205
10.41
10.615
$\bar{x}_L - \bar{x}_M = 9.75$
$\bar{x}_L - \bar{x}_M = 10.25$
$\bar{x}_L - \bar{x}_M$ = difference in the sample mean volume of soft drink in large and medium cups

(i) $z = \frac{9.75 - 10}{0.205} = -1.22$ and $z = \frac{10.25 - 10}{0.205} = 1.22$

Using Table A: $P(9.75 \leq \bar{x}_L - \bar{x}_M \leq 10.25) = P(-1.22 \leq z \leq 1.22) = 0.8888 - 0.1112 = 0.7776$.

Using technology: normalcdf(lower:−1.22, upper:1.22, mean:0, SD:1) = 0.7775.

(ii) normalcdf(lower:9.75, upper:10.25, mean:10, SD:0.205) = 0.7774.

2. Perform calculations.
(i) Standardize and use Table A or technology; or
(ii) Use technology without standardizing.
Be sure to answer the question that was asked.

FOR PRACTICE, TRY EXERCISE 83

Section 7.3 Summary

- When we want information about the population mean μ for some quantitative variable, we often take an SRS and use the sample mean $\bar{x}$ to estimate the unknown parameter μ. The **sampling distribution of the sample mean** $\bar{x}$ describes how the statistic $\bar{x}$ varies in all possible samples of the same size from the population.
  - **Shape:** 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 sufficiently large, the sampling distribution of $\bar{x}$ is approximately Normal. For most non-Normal populations, it is safe to use a Normal distribution to calculate probabilities involving $\bar{x}$ when $n \geq 30$.
  - **Center:** The **mean** of the sampling distribution of $\bar{x}$ is $\mu_{\bar{x}} = \mu$, so $\bar{x}$ is an unbiased estimator of μ.
  - **Variability:** The **standard deviation** of the sampling distribution of $\bar{x}$ is approximately $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ for an SRS of size n if the population has standard deviation σ. This formula can be used if the sample size is less than 10% of the population size (*10% condition*).
- Choose independent SRSs of size n_1 from Population 1 with mean μ_1 and standard deviation σ_1 and of size n_2 from Population 2 with mean μ_2 and standard deviation σ_2. The sampling distribution of $\bar{x}_1 - \bar{x}_2$ has the following properties:
  - **Shape:** Normal if both population distributions are Normal; approximately Normal if both sample sizes are large ($n_1 \geq 30$ and $n_2 \geq 30$) or if one population is Normally distributed and the other sample size is large.
  - **Center:** The **mean** of the sampling distribution is $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$.
  - **Variability:** The **standard deviation** of the sampling distribution is approximately $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ as long as the 10% condition is met for both samples: $n_1 < 0.10N_1$ and $n_2 < 0.10N_2$.

Section 7.3 Exercises

59. **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.

pg 504

(a) Identify the mean of the sampling distribution of $\bar{x}$.

(b) Calculate and interpret the standard deviation of the sampling distribution of $\bar{x}$. Verify that the 10% condition is met.

60. **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 $\bar{x}$. Assume the machine is working properly.

(a) Identify the mean of the sampling distribution of $\bar{x}$.

(b) Calculate and interpret the standard deviation of the sampling distribution of $\bar{x}$. Verify that the 10% condition is met.

61. **Songs on an iPod** Refer to Exercise 59. How many songs would you need to sample if you wanted the standard deviation of the sampling distribution of $\bar{x}$ to be 10 seconds? Justify your answer.

62. **Making auto parts** Refer to Exercise 60. 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.

63. **Screen time** Administrators at a small school with 200 students want to estimate the average amount of time students spend looking at a screen (phone, computer, television, and so on) per day. The administrators select a random sample of 50 students from the school to ask.

(a) Show that the 10% condition is not met in this case.

(b) What effect does violating the 10% condition have on the standard deviation of the sampling distribution of $\bar{x}$?

64. **Beautiful trees** As part of a school beautification project, students planted 35 trees along a road next to their school. Each month, students in the environmental science class randomly select 5 trees to estimate the average height of all the trees planted for the project.

(a) Show that the 10% condition is not met in this case.

(b) What effect does violating the 10% condition have on the standard deviation of the sampling distribution of $\bar{x}$?

65. **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.

pg 506

(a) What is the probability that a randomly selected bottle contains less than 295 ml?

(b) What is the probability that the mean contents of six randomly selected bottles is less than 295 ml?

66. **Cereal** A company's cereal boxes advertise that each box contains 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 $\sigma = 0.03$ ounce.

(a) What is the probability that a randomly selected box of the cereal contains less than 9.65 ounces of cereal?

(b) Now take an SRS of 5 boxes. What is the probability that the mean amount of cereal in these boxes is less than 9.65 ounces?

67. **Cholesterol** 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.

(a) Choose an SRS of 100 men from this population. Describe the sampling distribution of $\bar{x}$.

(b) Find the probability that $\bar{x}$ estimates μ within ± 3 mg/dl. (This is the probability that $\bar{x}$ takes a value between 185 and 191 mg/dl.)

(c) Choose an SRS of 1000 men from this population. Now what is the probability that $\bar{x}$ falls within ± 3 mg/dl of μ? In what sense is the larger sample "better"?

68. **Finch beaks** One dimension of bird beaks is "depth"—the height of the beak where it arises from the bird's head. During a research study on one island in the Galapagos archipelago, the beak depth of all Medium Ground Finches on the island was found to be Normally distributed with mean $\mu = 9.5$ millimeters (mm) and standard deviation $\sigma = 1.0$ mm.[11]

(a) Choose an SRS of 5 Medium Ground Finches from this population. Describe the sampling distribution of $\bar{x}$.

(b) Find the probability that $\bar{x}$ estimates μ within ± 0.5 mm. (This is the probability that $\bar{x}$ takes a value between 9 and 10 mm.)

(c) Choose an SRS of 50 Medium Ground Finches from this population. Now what is the probability that $\bar{x}$ falls within ± 0.5 mm of μ? In what sense is the larger sample "better"?

69. **Dead battery?** A car company claims 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. A consumer organization installs this type of battery in an SRS of 8 cars and calculates $\bar{x} = 42.2$ months.

(a) Find the probability that the sample mean lifetime is 42.2 months or less if the company's claim is true.

(b) Based on your answer to part (a), is there convincing evidence that the company is overstating the average lifetime of its batteries?

70. **Foiled again?** The manufacturer of a certain brand of aluminum foil claims that the amount of foil on each roll follows a Normal distribution with a mean of 250 square feet (ft^2) and a standard deviation of 2 ft^2. To test this claim, a restaurant randomly selects 10 rolls of this aluminum foil and carefully measures the mean area to be $\bar{x} = 249.6$ ft^2.

(a) Find the probability that the sample mean area is 249.6 ft^2 or less if the manufacturer's claim is true.

(b) Based on your answer to part (a), is there convincing evidence that the company is overstating the average area of its aluminum foil rolls?

71. **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.

(a) Describe the shape of the sampling distribution of $\bar{x}$ for SRSs of size $n = 5$ from the population of songs on David's iPod. Justify your answer.

(b) Describe the shape of the sampling distribution of $\bar{x}$ for SRSs of size $n = 100$ from the population of songs on David's iPod. Justify your answer.

72. **High school GPAs** The distribution of grade point average for students at a large high school is skewed to the left with a mean of 3.53 and a standard deviation of 1.02.

(a) Describe the shape of the sampling distribution of $\bar{x}$ for SRSs of size $n = 4$ from the population of students at this high school. Justify your answer.

(b) Describe the shape of the sampling distribution of $\bar{x}$ for SRSs of size $n = 50$ from the population of students at this high school. Justify your answer.

73. **More on insurance** An insurance company claims pg 511 that in the entire population of homeowners, the mean annual loss from fire is $\mu = \$250$ and the standard deviation of the loss is $\sigma = \$5000$. The distribution of losses is strongly right-skewed: many policies have \$0 loss, but a few have large losses. The company hopes to sell 1000 of these policies for \$300 each.

(a) Assuming that the company's claim is true, what is the probability that the mean loss from fire is greater than \$300 for an SRS of 1000 homeowners?

(b) If the company wants to be 90% certain that the mean loss from fire in an SRS of 1000 homeowners is less than the amount it charges for the policy, how much should the company charge?

74. **Cash grab** At a traveling carnival, a popular game is called the "Cash Grab." In this game, participants step into a sealed booth, a powerful fan turns on, and dollar bills are dropped from the ceiling. A customer has 30 seconds to grab as much cash as possible while the dollar bills swirl around. Over time, the operators of the game have determined that the mean amount grabbed is \$13 with a standard deviation of \$9. They charge \$15 to play the game and expect to have 40 customers at their next carnival.

(a) What is the probability that an SRS of 40 customers grab an average of \$15 or more?

(b) How much should the operators charge if they want to be 95% certain that the mean amount grabbed by an SRS of 40 customers is less than what they charge to play the game?

75. **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.

(a) Without doing any calculations, explain which event is more likely:

- randomly selecting 1 square yard of material and finding 2 or more flaws
- randomly selecting 50 square yards of material and finding an average of 2 or more flaws

(b) Explain why you cannot use a Normal distribution to calculate the probability of the first event in part (a).

(c) Calculate the probability of the second event in part (a).

76. **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 1.6 and standard deviation 0.75 in the population of all cars that enter at this interchange during rush hour.

(a) Without doing any calculations, explain which event is more likely:

- randomly selecting 1 car entering this interchange during rush hour and finding 2 or more people in the car
- randomly selecting 35 cars entering this interchange during rush hour and finding an average of 2 or more people in the cars

(b) Explain why you cannot use a Normal distribution to calculate the probability of the first event in part (a).

(c) Calculate the probability of the second event in part (a).

77. **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.

78. **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 variability of the sampling distribution of the sample mean decreases." Is the student right? Explain your answer.

79. **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. A commuter plane carries 30 passengers. Find the probability that the total weight of 30 randomly selected passengers exceeds 6000 pounds. (*Hint*: To calculate this probability, restate the problem in terms of the mean weight.)

80. **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 1-square-kilometer plots of land. Find the probability that the total number of lightning strikes in a random sample of 50 square-kilometer plots of land is less than 250. (*Hint*: To calculate this probability, restate the problem in terms of the mean number of strikes.)

81. **House prices** In the northern part of a large city, the distribution of home values is skewed to the right with a mean of \$410,000 and a standard deviation of \$250,000. In the southern part of the city, the distribution of home values is skewed to the right with a mean of \$375,000 and a standard deviation of \$240,000. Independent random samples of 10 houses in each part of the city are selected. Let $\bar{x}_N$ represent the sample mean value of homes in the northern part and let $\bar{x}_S$ represent the sample mean value of homes in the southern part.

(a) Find the mean of the sampling distribution of $\bar{x}_N - \bar{x}_S$.

(b) Calculate and interpret the standard deviation of the sampling distribution. Verify that the 10% condition is met.

(c) Is the shape of the sampling distribution approximately Normal? Justify your answer.

82. **Young players** In the National Football League (NFL), the distribution of age is skewed to the right with a mean of 26.2 years and a standard deviation of 3.24 years. In the National Basketball Association (NBA), the distribution of age is skewed to the right with a mean of 25.8 years and a standard deviation of 4.24 years. Independent random samples of 20 players in each league are selected. Let $\bar{x}_{NFL}$ represent the sample mean age of NFL players and let $\bar{x}_{NBA}$ represent the sample mean age of NBA players.

(a) Find the mean of the sampling distribution of $\bar{x}_{NFL} - \bar{x}_{NBA}$.

(b) Calculate and interpret the standard deviation of the sampling distribution. Verify that the 10% condition is met.

(c) Is the shape of the sampling distribution approximately Normal? Justify your answer.

83. **Cholesterol** The level of cholesterol in the blood for all men aged 20 to 34 follows a Normal distribution with mean $\mu_M = 188$ milligrams per deciliter (mg/dl) and standard deviation $\sigma_M = 41$ mg/dl. For 14-year-old boys, blood cholesterol levels follow a Normal distribution with mean $\mu_B = 170$ mg/dl and standard deviation $\sigma_B = 30$ mg/dl. Suppose we select independent SRSs of 25 men aged 20 to 34 and 36 boys aged 14 and calculate the sample mean cholesterol levels $\bar{x}_M$ and $\bar{x}_B$.

pg 516

(a) What is the shape of the sampling distribution of $\bar{x}_M - \bar{x}_B$? Why?

(b) Find the mean of the sampling distribution.

(c) Calculate and interpret the standard deviation of the sampling distribution. Verify that the 10% condition is met.

(d) Find the probability of getting a difference in sample means $\bar{x}_M - \bar{x}_B$ that's less than 0 mg/dl.

84. **How tall?** The heights of young men follow a Normal distribution with mean $\mu_M = 69.3$ inches and standard deviation $\sigma_M = 2.8$ inches. The heights of young women follow a Normal distribution with mean $\mu_W = 64.5$ inches and standard deviation $\sigma_W = 2.5$ inches. Suppose we select independent SRSs of 16 young men and 9 young women and calculate the sample mean heights $\bar{x}_M$ and $\bar{x}_W$.

(a) What is the shape of the sampling distribution of $\bar{x}_M - \bar{x}_W$? Why?

(b) Find the mean of the sampling distribution.

(c) Calculate and interpret the standard deviation of the sampling distribution. Verify that the 10% condition is met.

(d) Find the probability of getting a difference in sample means $\bar{x}_M - \bar{x}_W$ that's greater than 2 inches.

85. **Cholesterol** Refer to Exercise 83. Should we be surprised if the sample mean cholesterol level for the 14-year-old boys exceeds the sample mean cholesterol level for the men? Explain your answer.

86. **How tall?** Refer to Exercise 84. Should we be surprised if the sample mean height for the young men is at least 2 inches greater than the sample mean height for the young women? Explain your answer.

Multiple Choice: *Select the best answer for Exercises 87–90.*

87. The distribution of scores on the mathematics part of the SAT exam in a recent year was approximately Normal with mean 515 and standard deviation 114. Imagine choosing many SRSs of 100 students who took the exam and finding the average score for each sample. 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.

88. 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 μ

(e) To ensure that the observations in the sample are close to independent

89. 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 4 bottles are randomly selected and the number of ounces in each bottle is measured, then there is about a 95% probability 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

90. 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

(a) as we look at more and more bulbs, their average burnout time gets ever closer to the mean μ for all bulbs of this type.

(b) the average burnout time of a large number of bulbs has a sampling distribution with the same shape (strongly skewed) as the population distribution.

(c) the average burnout time of a large number of bulbs has a sampling distribution with a similar shape but not as extreme (skewed, but not as strongly) as the population distribution.

(d) the average burnout time of a large number of bulbs has a sampling distribution that is close to Normal.

(e) the average burnout time of a large number of bulbs has a sampling distribution that is exactly Normal.

Recycle and Review

Exercises 91 and 92 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) that is unemployed. Here are estimates 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

91. **Unemployment** (1.1) Find the unemployment rate for people with each level of education. Is there an association between unemployment rate and education? Explain your answer.

92. **Unemployment** (5.2, 5.3) Suppose that you randomly select one person 25 years of age or older.

(a) What is the probability that a randomly chosen person 25 years of age or older is in the labor force?

(b) 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?

(c) Are the events "in the labor force" and "college graduate" independent? Justify your answer.

Chapter 7 Wrap-Up

FRAPPY! FREE RESPONSE AP® PROBLEM, YAY!

The following problem is modeled after actual AP® 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.

(a) What is the shape of the sampling distribution of the sample mean? Explain.

(b) What are the mean and standard deviation of the sampling distribution of the sample mean?

(c) What is the probability that the mean number of absences in a random sample of 50 students is less than 1?

(d) 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 website (highschool.bfwpub.com/updatedtps6e). 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 7 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. The distribution of a statistic in all possible samples of the same size from the same population is called the sampling distribution of the statistic. Knowing the sampling distribution of a statistic tells us how far we can expect a statistic to vary from the parameter value and what values of the statistic should be considered unusual.

The third big idea is the distinction between the distribution of the population, the distribution of a sample, and the sampling distribution of a sample statistic. Reviewing the illustration on page 474 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 final big idea is how to describe a sampling distribution. To adequately describe a sampling distribution, you need to address shape, center, and variability. 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 underestimate or consistently overestimate the parameter in many samples. Ideally, the variability 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 less variability.

Section 7.2: Sample Proportions

In this section, you learned about the shape, center, and variability of the sampling distribution of a sample proportion $\hat{p}$. 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 sample size is less than 10% of the population size (the 10% condition), the standard deviation of the sampling distribution of the sample proportion is approximately $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$. The standard deviation measures how far the sample proportion $\hat{p}$ typically varies from the population proportion p.

In this section, you also learned about the sampling distribution of a difference in sample proportions $\hat{p}_1 - \hat{p}_2$. The shape of the sampling distribution of $\hat{p}_1 - \hat{p}_2$ will be approximately Normal when the Large Counts condition is met for both samples. The center of the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is $\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$. The standard deviation of the sampling distribution is approximately $\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}$ when the 10% condition is met for both samples.

Section 7.3: Sample Means

In this section, you learned about the shape, center, and variability of the sampling distribution of a sample mean $\bar{x}$. When the population is Normally distributed, the sampling distribution of $\bar{x}$ will also be Normal for any sample size. When the population distribution 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. You can use a Normal distribution to calculate probabilities involving the sampling distribution of $\bar{x}$ if the population is Normal or the sample size is at least 30.

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 μ. When the sample size is less than 10% of the population size (the 10% condition), the standard deviation of the sampling distribution of the sample mean is approximately $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. The standard deviation measures how far the sample mean $\bar{x}$ typically varies from the population mean μ.

In this section, you also learned about the sampling distribution of a difference in sample means $\bar{x}_1 - \bar{x}_2$. The shape of the sampling distribution of $\bar{x}_1 - \bar{x}_2$ will be approximately Normal when both population distributions are Normal, both sample sizes are at least 30, or one population distribution is Normal and the sample size from the other population distribution is at least 30. The center of the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$. The standard deviation of the sampling distribution is approximately $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$ when the 10% condition is met for both samples.

Comparing sampling distributions

	Sampling distribution of $\hat{p}$	Sampling distribution of $\bar{x}$
Center	$\mu_{\hat{p}} = p$	$\mu_{\bar{x}} = \mu$
Variability	$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$ when the 10% condition is met	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ when the 10% condition is met
Shape	Approximately Normal when the Large Counts condition is met: $np \geq 10$ and $n(1-p) \geq 10$	Normal when the population distribution is Normal; approximately Normal if the population distribution is not Normal but the sample size is large ($n \geq 30$)
	Sampling distribution of $\hat{p}_1 - \hat{p}_2$	**Sampling distribution of $\bar{x}_1 - \bar{x}_2$**
Center	$\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$	$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$
Variability	$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$ when the 10% condition is met for both samples	$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ when the 10% condition is met for both samples
Shape	Approximately Normal when the Large Counts condition is met for both samples: $n_1p_1, n_1(1-p_1), n_2p_2, n_2(1-p_2)$ all ≥ 10	Normal when both population distributions are Normal; approximately Normal if the population distributions are not Normal but the sample sizes are both at least 30 or one population distribution is Normal and the sample size from the other population is at least 30.

What Did You Learn?

Learning Target	Section	Related Example on Page(s)	Relevant Chapter Review Exercise(s)
Distinguish between a parameter and a statistic.	7.1	470	R7.1
Create a sampling distribution using all possible samples from a small population.	7.1	472	R7.2
Use the sampling distribution of a statistic to evaluate a claim about a parameter.	7.1	473	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	Discussed on 474	R7.3
Determine if a statistic is an unbiased estimator of a population parameter.	7.1	477	R7.3
Describe the relationship between sample size and the variability of a statistic.	7.1	Discussed on 479	R7.2
Calculate the mean and standard deviation of the sampling distribution of a sample proportion $\hat{p}$ and interpret the standard deviation.	7.2	490	R7.4, R7.5
Determine if the sampling distribution of $\hat{p}$ is approximately Normal.	7.2	490	R7.4, R7.5
Calculate the mean and the standard deviation of the sampling distribution of a difference in sample proportions $\hat{p}_1 - \hat{p}_2$ and interpret the standard deviation.	7.2	496	R7.8
Determine if the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately Normal.	7.2	496	R7.8

Learning Target	Section	Related Example on Page(s)	Relevant Chapter Review Exercise(s)
If appropriate, use a Normal distribution to calculate probabilities involving $\hat{p}$ or $\hat{p}_1 - \hat{p}_2$.	7.2	492, 496	R7.4, R7.5, R7.8
Calculate the mean and standard deviation of the sampling distribution of a sample mean $\bar{x}$ and interpret the standard deviation.	7.3	504	R7.6, R7.7
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	Discussed on 510–511	R7.6, R7.7
Calculate the mean and the standard deviation of the sampling distribution of a difference in sample means $\bar{x}_1 - \bar{x}_2$ and interpret the standard deviation.	7.3	516	R7.9
Determine if the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is approximately Normal.	7.3	516	R7.9
If appropriate, use a Normal distribution to calculate probabilities involving $\bar{x}$ or $\bar{x}_1 - \bar{x}_2$.	7.3	506, 511, 516	R7.6, R7.7

Chapter 7 Review Exercises

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

R7.1 Bad eggs Selling eggs that are contaminated with salmonella can cause food poisoning in consumers. A large egg producer randomly selects 200 eggs from all the eggs shipped in one day. The laboratory reports that 9 of these eggs had salmonella contamination. Identify the population, the parameter, the sample, and the statistic.

R7.2 Five books An author has written 5 children's books. The numbers of pages in these books are 64, 66, 71, 73, and 76.

(a) List all 10 possible SRSs of size $n = 3$, calculate the median number of pages for each sample, and display the sampling distribution of the sample median on a dotplot.

(b) Describe how the variability of the sampling distribution of the sample median would change if the sample size was increased to $n = 4$.

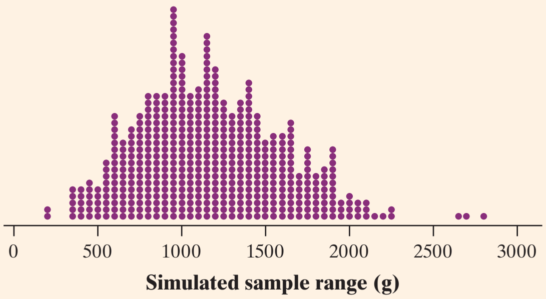
(c) Construct the sampling distribution of the sample median for samples of size $n = 4$. Does this sampling distribution support your answer to part (b)? Explain your reasoning.

R7.3 Birth weights 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.[12]

(a) Sketch a graph that displays the distribution of birth weights for this population.

(b) Sketch a possible graph of the distribution of birth weights for an SRS of size 5. Calculate the range for this sample.

In this population, the range (Maximum − Minimum) of birth weights is 3417 grams. We used technology to take 500 SRSs of size $n = 5$ and calculate the range (Maximum − Minimum) for each sample. The dotplot shows the results.



(c) In the graph provided, there is a dot at approximately 2800. Explain what this value represents.

(d) Is the sample range an unbiased estimator of the population range? Give evidence from the graph to support your answer.

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

(a) What is the mean of the sampling distribution of $\hat{p}$?

(b) Calculate and interpret the standard deviation of the sampling distribution of $\hat{p}$. Check that the 10% condition is met.

(c) Is the sampling distribution of $\hat{p}$ approximately Normal? Justify your answer.

(d) Find the probability that between 13% and 17% of people jog in a random sample of 1540 adults.

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 flashes, the passenger will be searched by customs agents. A green light means "Go ahead." Customs agents claim that 30% of all travelers will be stopped (red light), 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.

(a) Assume that the customs agents' claim is true. Find the probability that the proportion of travelers who get a red light in a random sample of 100 travelers is less than or equal to the result in this sample.

(b) Based on your results in part (a), is there convincing evidence that less than 30% of all travelers will be stopped? Explain your reasoning.

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.

(a) What is the probability that a randomly chosen individual has a WAIS score of 105 or greater?

(b) Find the mean and standard deviation of the sampling distribution of the average WAIS score $\bar{x}$ for an SRS of 60 people. Verify that the 10% condition is met. Interpret the standard deviation.

(c) What is the probability that the average WAIS score of an SRS of 60 people is 105 or greater?

(d) Would your answers to any of parts (a), (b), or (c) be affected if the distribution of WAIS scores in the adult population was distinctly non-Normal? Explain your reasoning.

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.

(a) Explain why it is reasonable to use a Normal distribution to approximate the sampling distribution of $\bar{x}$ for SRSs of size 50.

(b) Estimate the probability that the mean number of moths in a sample of size 50 is greater than or equal to 0.6.

(c) In a recent month, the mean number of moths in an SRS of size 50 was $\bar{x} = 0.6$. Based on this result, is there convincing evidence that the moth population is getting larger in this state? Explain your reasoning.

R7.8 American-made cars Nathan and Kyle both work for the Department of Motor Vehicles (DMV), but they live in different states. In Nathan's state, 80% of the registered cars are made by American manufacturers. In Kyle's state, only 60% of the registered cars are made by American manufacturers. Nathan selects a random sample of 100 cars in his state and Kyle selects a random sample of 70 cars in his state. Let $\hat{p}_N - \hat{p}_K$ be the difference (Nathan's state – Kyle's state) in the sample proportion of cars made by American manufacturers.

(a) What is the shape of the sampling distribution of $\hat{p}_N - \hat{p}_K$? Why?

(b) Find the mean of the sampling distribution.

(c) Calculate and interpret the standard deviation of the sampling distribution. Verify that the 10% condition is met.

(d) What is the probability that the proportion in the sample from Kyle's state exceeds the proportion in the sample from Nathan's state?

R7.9 Candles A company produces candles. Machine 1 makes candles with a mean length of 15 centimeters and a standard deviation of 0.15 centimeter. Machine 2 makes candles with a mean length of 15 centimeters and a standard deviation of 0.10 centimeter. A random sample of 49 candles is taken from each machine. Let $\bar{x}_1 - \bar{x}_2$ be the difference (Machine 1 – Machine 2) in the sample mean length of candles. Describe the shape, center, and variability of the sampling distribution of $\bar{x}_1 - \bar{x}_2$.

Chapter 7 AP® 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?

(a) 72% is a sample; 56% is a population.

(b) 72% and 56% are both statistics.

(c) 72% is a statistic and 56% is a parameter.

(d) 72% is a parameter and 56% is a statistic.

(e) 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

(a) reduce the bias of the estimate.

(b) increase the bias of the estimate.

(c) reduce the variability of the estimate.

(d) increase the variability of the estimate.

(e) 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 $\hat{p}$?

(a) 0.01

(b) 0.09

(c) 0.85

(d) 0.975

(e) 0.999

T7.4 The central limit theorem is important in statistics because it allows us to use a Normal distribution to find probabilities involving the sample mean if the

(a) sample size is sufficiently large (for any population).

(b) population is Normally distributed (for any sample size).

(c) population is Normally distributed and the sample size is reasonably large.

(d) population is Normally distributed and the population standard deviation is known (for any sample size).

(e) 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?

(a) We expect that the estimate from Johns Hopkins will be closer to the truth than the estimate from Ohio State because it comes from a smaller population.

(b) We expect that the estimate from Johns Hopkins will be closer to the truth than the estimate from Ohio State because it is based on a smaller sample size.

(c) We expect that the estimate from Ohio State will be closer to the truth than the estimate from Johns Hopkins because it comes from a larger population.

(d) We expect that the estimate from Ohio State will be closer to the truth than the estimate from Johns Hopkins because it is based on a larger sample size.

(e) We expect that the estimate from Johns Hopkins will be about the same distance from the truth as the estimate from Ohio State because both samples are 3% of their populations.

T7.6 A researcher initially plans to take an SRS of size 160 from a certain population and calculate the sample mean $\bar{x}$. Later, the researcher decides to increase the sample size so that the standard deviation of the sampling distribution of $\bar{x}$ will be half as big as when using a sample size of 160. What sample size should the researcher use?

(a) 40

(b) 80

(c) 320

(d) 640

(e) There is not enough information to determine the sample size.

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

(a) exactly Normal with mean 0.55 and standard deviation 0.03.

(b) approximately Normal with mean 0.55 and standard deviation 0.03.

(c) exactly Normal with mean 0.60 and standard deviation 0.03.

(d) approximately Normal with mean 0.60 and standard deviation 0.03.

(e) 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 measures how far the sample mean typically varies from the population mean.

(c) The sample mean is an unbiased estimator of the population mean.

(d) The sampling distribution shows how the sample mean is distributed around the population mean.

(e) The sampling distribution shows how the sample is distributed around the sample mean.

T7.9 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 μ 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 μ.

(c) as we take larger and larger samples from this population, $\bar{x}$ will get closer and closer to μ.

(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.

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.

T7.11 An SRS of size 100 is taken from Population A with proportion 0.8 of successes. An independent SRS of size 400 is taken from Population B with proportion 0.5 of successes. The sampling distribution of the difference (A − B) in sample proportions has what mean and standard deviation?

(a) mean = 0.3; standard deviation = 1.3

(b) mean = 0.3; standard deviation = 0.40

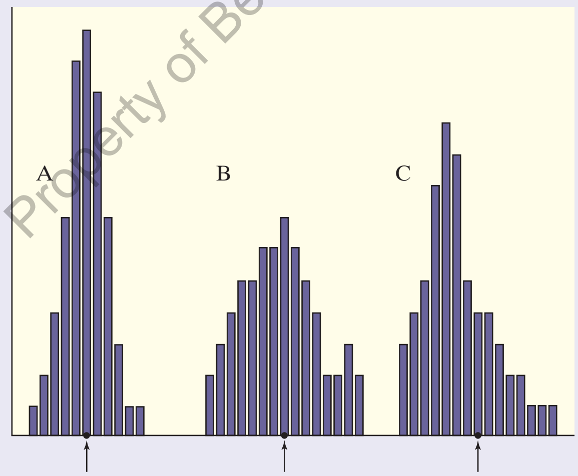
(c) mean = 0.3; standard deviation = 0.047

(d) mean = 0.3; standard deviation = 0.0022

(e) mean = 0.3; standard deviation = 0.0002

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.*

T7.12 Here 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.

T7.13 The amount that households pay service providers for access to the Internet varies quite a bit, but the mean monthly fee is \$50 and the standard deviation is \$20. The distribution is not Normal: many households pay a low rate as part of a bundle with phone or television service, but some pay much more for Internet only or for faster connections.[13] A sample survey asks an SRS of 50 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 \$55.

(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 \$55.

T7.14 According to government data, 22% of American children under the age of 6 live in households with incomes less than the official poverty level. A study of learning in early childhood chooses an SRS of 300 children from one state and finds that $\hat{p} = 0.29$.

(a) Find the probability that at least 29% of the sample are from poverty-level households, assuming that 22% of all children under the age of 6 in this state live in poverty-level households.

(b) Based on your answer to part (a), is there convincing evidence that the percentage of children under the age of 6 living in households with incomes less than the official poverty level in this state is greater than the national value of 22%? Explain your reasoning.

T7.15 In a children's book, the mean word length is 3.7 letters with a standard deviation of 2.1 letters. In a novel aimed at teenagers, the mean word length is 4.3 letters with a standard deviation of 2.5 letters. Both distributions of word length are unimodal and skewed to the right. Independent random samples of 35 words are selected from each book. Let $\bar{x}_C$ represent the sample mean word length in the children's book and let $\bar{x}_T$ represent the sample mean word length in the teen novel.

(a) Find the mean of the sampling distribution of $\bar{x}_C - \bar{x}_T$.

(b) Calculate and interpret the standard deviation of the sampling distribution. Verify that the 10% condition is met.

(c) Justify that the shape of the sampling distribution is approximately Normal.

(d) What is the probability that the sample mean word length is greater in the sample from the children's book than in the sample from the teen novel?

Cumulative AP® 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, $Q_1 = 18$, median = 20, $Q_3 = 40$, max = 75. If you wanted to construct a boxplot for the data set 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

Number of heads	0	1	2	3	4
Probability	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.

(b) 0.3125.

(c) 0.6875.

(d) 0.9375.

(e) 0.0625.

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. For SRSs of size 200, the distribution of sample mean IQ score 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. What would be a proper way to use a table of random digits to do this?

(a) One digit from the random digit table simulates one answer, with 5 = correct and all other digits = incorrect. Ten digits from the table simulate 10 answers.

(b) One digit from the random digit table simulates one answer, with 0 or 1 = correct and all other digits = incorrect. Ten digits from the table simulate 10 answers.

(c) One digit from the random digit table simulates one answer, with odd = correct and even = incorrect. Ten digits from the table simulate 10 answers.

(d) One digit from the random digit table simulates one answer, with 0 or 1 = correct and all other digits = incorrect, 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 = correct and 21 to 99 = incorrect. Ten pairs of digits from the table simulate 10 answers.

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

(a) $4\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^1$

(b) $\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^1$

(c) $4\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^3$

(d) $\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^3$

(e) $6\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^3$

AP2.6 On one episode of his show, a radio show host encouraged his listeners to visit his website and vote in a poll about proposed tax increases. Of the 4821 people who vote, 4277 are against the proposed increases. To which of the following populations should the results of this poll be generalized?

(a) All people who have ever listened to this show

(b) All people who listened to this episode of the show

(c) All people who visited the show host's website

(d) All people who voted in the poll

(e) All people who voted against the proposed increases

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 to reveal her attitude toward the product. You think there may be large differences in the responses of 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

(b) 0.8400

(c) 0.8600

(d) 0.8660

(e) 0.9900

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.

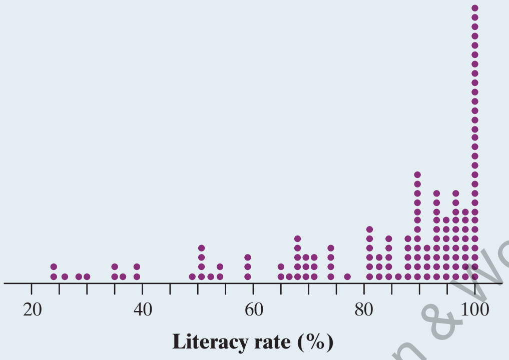
(b) For each 1% increase in the graduation rate, the percent of families living in poverty is predicted to decrease by approximately 0.802.

(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 decrease 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.

Questions AP2.11–AP2.13 refer to the following graph. 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%.



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) approximately 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 the distribution isn't symmetric.

(e) You can't say, because the distribution isn't Normal.

AP2.13 The country with a literacy rate of 49% is closest to which of the following percentiles?

(a) 6th

(b) 11th

(c) 28th

(d) 49th

(e) There is not enough information to calculate the percentile.

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 age.

(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 foot 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) variety of wheat, bushels per acre, a specific fertilizer.

(c) variety of wheat, 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 65 or fewer 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) $\binom{100}{65}(0.71)^{65}(0.29)^{35}$

(b) $\binom{100}{65}(0.29)^{65}(0.71)^{35}$

(c) $P\left(z \le \dfrac{0.65-0.71}{\sqrt{\dfrac{(0.71)(0.29)}{100}}}\right)$

(d) $P\left(z \le \dfrac{0.65-0.71}{\sqrt{\dfrac{(0.65)(0.35)}{100}}}\right)$

(e) $P\left(z \le \dfrac{0.65-0.71}{\dfrac{(0.71)(0.29)}{\sqrt{100}}}\right)$

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 dairy farm yielded the following information on the number of male and female calves born at various times of the day.

		Time of day			
		Day	Evening	Night	Total
Gender	Males	129	15	117	261
	Females	118	18	116	252
	Total	247	33	233	513

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

(a) $\dfrac{369}{513}$

(b) $\dfrac{485}{513}$

(c) $\dfrac{116}{513}$

(d) $\dfrac{116}{252}$

(e) $\dfrac{116}{233}$

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

(b) 2.60

(c) 3.02

(d) 3.40

(e) 9.10

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:

Amount won	$500	$100	$25	$0
Probability	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

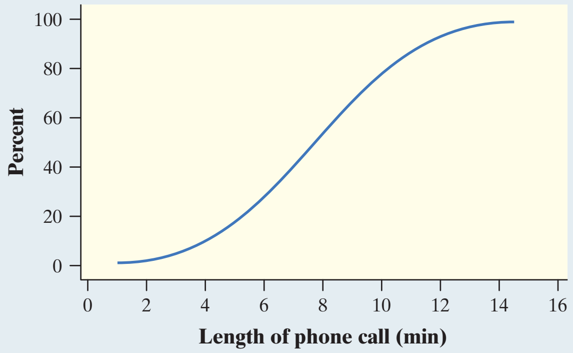
(b) $15.00, $53.85

(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. Here is a cumulative relative frequency graph 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.

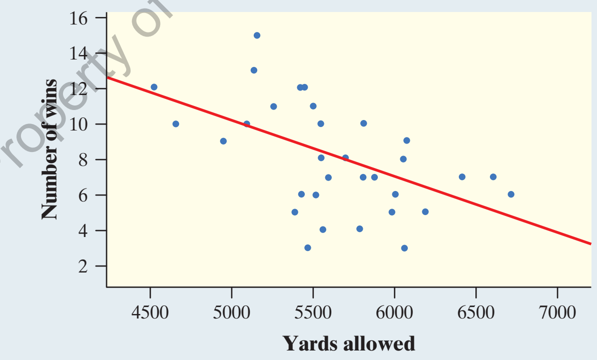
AP2.22 A health worker is interested in determining if omega-3 fish oil can help reduce cholesterol in adults. She obtains permission to examine the health records of 200 people in a large medical clinic and classifies them according to whether or not they take omega-3 fish oil. She also obtains their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

(a) Is this an observational study or an experiment? Justify your answer.

(b) 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.

(c) 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 The scatterplot shows the relationship between the number of yards allowed by teams in the National Football League and the number of wins for that team in a recent season, along with the least-squares regression line. Computer output is also provided.



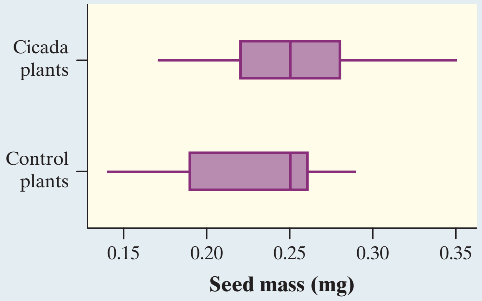
```
Term              Coef      SE Coef  T-Value  P-Value
Constant          25.66        5.37     4.78    0.000
Yards_allowed −0.003131  0.000948    −3.30    0.002
S = 2.65358     R-Sq = 26.65%      R-Sq(adj) = 24.21%
```

(a) State the equation of the least-squares regression line. Define any variables you use.

(b) Calculate and interpret the residual for the Seattle Seahawks, who allowed 4668 yards and won 10 games.

(c) The Carolina Panthers allowed 5167 yards and won 15 games. What effect does the point representing the Panthers have on the equation of the least-squares regression line? Explain.

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:



	n	Minimum	Q_1	Median	Q_3	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

(a) Write a few sentences comparing the distributions of seed mass for the two groups of plants.

(b) Based on the graphical displays, which distribution likely has the larger mean? Justify your answer.

(c) Explain the purpose of the random assignment in this study.

(d) 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. Furthermore, 30% of the novels have fewer than 400 pages. Suppose that you randomly select 50 novels from the library.

(a) What is the probability that the average number of pages in the sample is less than 500?

(b) What is the probability that at least 20 of the novels have fewer than 400 pages?