wavebreakmedia/Shutterstock

Bootstrap Methods and Permutation Tests*

Introduction

The continuing revolution in computing is having a dramatic influence on statistics. The exploratory analysis of data is becoming easier as more graphs and calculations are automated. The statistical study of very large and very complex data sets is now feasible. Another impact of this fast and inexpensive computing is less obvious: new methods apply previously unthinkable amounts of computation to produce confidence intervals and tests of significance in settings that don't meet the conditions for safe application of the usual methods of inference.

Consider the commonly used t procedures for inference about means (Chapter 7) and for relationships between quantitative variables (Chapter 10). All these methods rest on the use of Normal distributions for data. While no data are exactly Normal, the t procedures are useful in practice because they

are *robust*. Nonetheless, we cannot use t confidence intervals and tests if the data are strongly skewed, unless our samples are quite large.

Other procedures cannot be used on non-Normal data even when the samples are large. For example, inference about spread based on Normal distributions is *not robust* and, therefore, is of little use in practice.

Finally, what should we do if we are interested in, say, a *ratio* of means, such as the ratio of average men's salary to average women's salary? There is no simple traditional inference method for this setting.

The methods of this chapter—bootstrap confidence intervals and permutation tests—apply the power of the computer to relax some of the conditions needed for traditional inference and to do inference in new settings. The big ideas of statistical inference remain the same. The fundamental reasoning is still based on asking, "What would happen if we applied this method many times?" Answers to this question are still given by confidence levels and P-values based on the sampling distributions of statistics.

The most important requirement for trustworthy conclusions about a population is still that *our data can be regarded as random samples from the population*—not even the computer can rescue voluntary response samples or confounded experiments. But the new methods set us free from the need for Normal data or large samples. They work the same way for many different statistics in many different settings. They can, with sufficient computing power, give results that are more accurate than those from traditional methods.

Bootstrap intervals and permutation tests are conceptually simple because they appeal directly to the basis of all inference: the sampling distribution that shows what would happen if we took very many samples under the same conditions. The new methods do have limitations, some of which we will illustrate. But their effectiveness and range of use are so great that they are now widely used in a variety of settings.

Software

Bootstrapping and permutation tests are feasible in practice only with software that automates the heavy computation that these methods require. If you are sufficiently expert, you can program at least the basic methods yourself. It is easier to use software that offers bootstrap intervals and permutation tests preprogrammed, just as most software offers the various t intervals and tests. You can expect the new methods to become more common in standard statistical software.

This chapter primarily uses R, the software choice of many statisticians doing research on resampling methods.[1] There are several packages of functions for resampling in R. We will focus on the boot package, which offers the most capabilities. Unlike software such as Minitab and SPSS, R is not menu driven and requires command line requests to load data and access various functions. All commmands used in this chapter are available on the text website.

JMP, SPSS, and SAS also offer preprogrammed bootstrap and permutation methods. JMP offers single-click bootstrapping capabilities to many of their tables of results. SPSS has an auxiliary bootstrap module that contains most of the methods described in this chapter. In SAS, the SURVEYSELECT procedure can be used to do the necessary resampling. The bootstrap macro contains most of the confidence interval methods offered by R. You can find links for downloading these modules or macros on the text website.

16.1 The Bootstrap Idea

When you complete this section, you will be able to:	• Randomly select bootstrap resamples from a small sample using software or a table of random digits. • Find the bootstrap standard error from a collection of resamples. • Use computer output to describe the results of a bootstrap analysis of the mean.

Here is the example we will use to introduce these methods.

EXAMPLE 16.1

FACE4

Average time looking at a Facebook profile. In Example 12.17 (page 670), we compared the amount of time a Facebook user spends reading different types of profiles. Here, let's focus on just the average time for the fourth profile (negative male). Figure 16.1(a) gives a histogram, and Figure 16.1(b) gives the Normal quantile plot of the 21 observations. The data are skewed to the right. Given the relatively small sample size, we have some concerns about using the t procedures for these data.
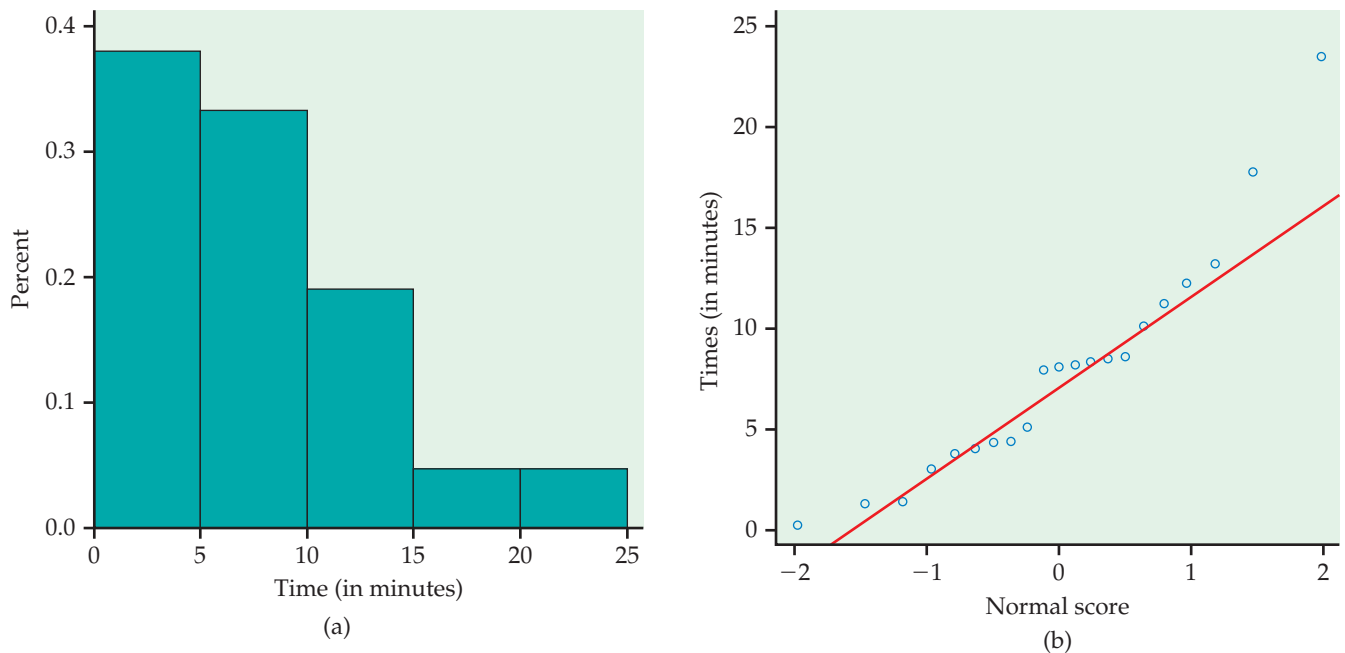

FIGURE 16.1 (a) The distribution of times (minutes) looking at a negative male Facebook profile page. (b) Normal quantile plot of the times, Example 16.1. The distribution is right-skewed.

The big idea: Resampling and the bootstrap distribution

LOOK BACK

sampling distribution, p. 286

Statistical inference is based on the sampling distributions of sample statistics. A sampling distribution is based on many random samples from the population. The bootstrap is a way of finding the sampling distribution, at least approximately, from just one sample. Here is the procedure:
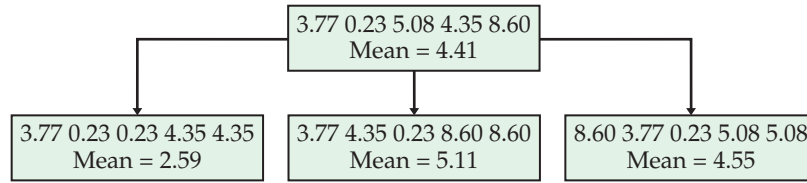
| 3.77 0.23 5.08 4.35 8.60 |
| Mean = 4.41 |

| 3.77 0.23 0.23 4.35 4.35 | 3.77 4.35 0.23 8.60 8.60 | 8.60 3.77 0.23 5.08 5.08 |
| Mean = 2.59 | Mean = 5.11 | Mean = 4.55 |

FIGURE 16.2 The resampling idea. The top box is a sample of size $n = 5$ from the Facebook profile viewing time data. The three lower boxes are three resamples from this original sample. Some values from the original sample are repeated in the resamples because each resample is formed by sampling with replacement. We calculate the statistic of interest, the sample mean in this example, for the original sample and each resample.

LOOK BACK

resample, p. 424

Step 1: Resampling. In Example 16.1, we have just one random sample of 21 observations. In place of many samples from the population, create many resamples by repeatedly *sampling with replacement* from this one random sample. Each resample is the same size as the original random sample.

sampling with replacement

Sampling with replacement means that after we randomly draw an observation from the original sample, we put it back before drawing the next observation. Think of drawing a number from a hat and then putting it back before drawing again. As a result, any number can be drawn more than once. If we sampled *without* replacement, we'd get the same set of numbers we started with, though in a different order. Figure 16.2 illustrates three resamples from a sample of five observations. In practice, we draw hundreds or thousands of resamples, not just three.

Step 2: Bootstrap distribution. The sampling distribution of a statistic describes the values taken by the statistic in all possible samples of the population of the same size. The **bootstrap distribution** of a statistic summarizes the values taken by the statistic in all possible resamples of the same size. The bootstrap distribution gives information (that is, shape and spread) about the sampling distribution.

bootstrap distribution

THE BOOTSTRAP IDEA

The original sample is representative of the population from which it was drawn. Thus, resamples from this original sample represent what we would get if we took many samples from the population. The **bootstrap distribution** of a statistic, based on the resamples, represents the sampling distribution of the statistic.

EXAMPLE 16.2

FACE4

Bootstrap distribution of the mean time looking at a Facebook profile. In Example 16.1, we want to estimate the average time viewing a negative male Facebook profile, μ, so the statistic is the sample mean $\bar{x}$. For our one sample of subjects, $\bar{x} = 7.87$ minutes. When we resample, we get different values of $\bar{x}$, just as we would if we randomly sampled a new group of subjects to survey.

We randomly generated 3000 resamples for these data. The mean for the resamples is 7.89 minutes, and the standard deviation is 1.22 minutes. Figure 16.3(a) gives a histogram of the bootstrap distribution of the means of 3000 resamples from the viewing time data. The Normal density curve with
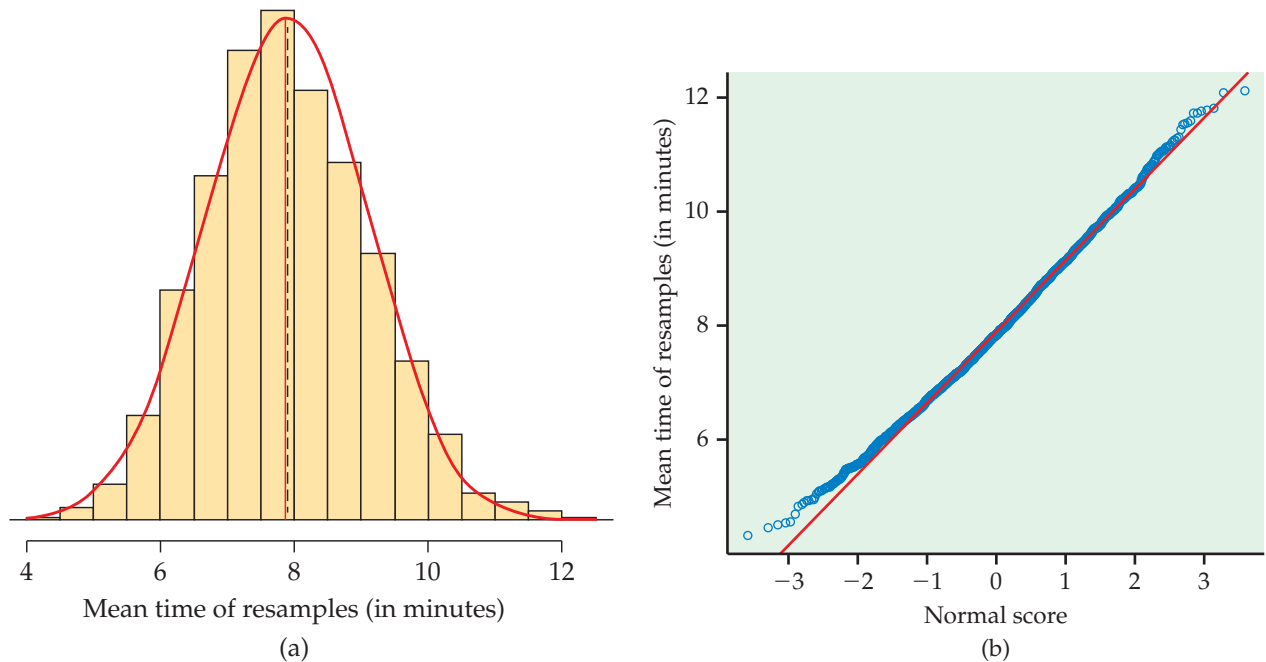
(a)

(b)

FIGURE 16.3 (a) The bootstrap distribution of 3000 resample means from the sample of Facebook profile viewing time data. The smooth curve is the Normal density function for the distribution which matches the mean and standard deviation of the distribution of the resample means. (b) The Normal quantile plot confirms that the bootstrap distribution is slightly skewed to the right but fits the Normal distribution quite well.

the mean 7.89, and standard deviation 1.22 is superimposed on the histogram. A Normal quantile plot is given in Figure 16.3(b). The Normal curve fits the data well, but some skewness is still evident.

According to the bootstrap idea, the bootstrap distribution represents the sampling distribution. Let's compare the bootstrap distribution with what we know about the sampling distribution.

LOOK BACK

central limit theorem, p. 298

mean and standard deviation of $\bar{x}$, p. 297

Shape: We see that the bootstrap distribution is nearly Normal. The central limit theorem says that the sampling distribution of the sample mean $\bar{x}$ is approximately Normal if n is large. So the bootstrap distribution shape is close to the shape we expect the sampling distribution to have.

Center: The bootstrap distribution is centered close to the mean of the original sample, 7.89 minutes versus 7.87 minutes for the original sample. Therefore, the mean of the bootstrap distribution has little bias as an estimator of the mean of the original sample. We know that the sampling distribution of $\bar{x}$ is centered at the population mean μ, that is, that $\bar{x}$ is an unbiased estimate of μ. So the resampling distribution behaves (starting from the original sample) as we expect the sampling distribution to behave (starting from the population).

Spread: The histogram and density curve in Figure 16.3(a) picture the variation among the resample means. We can get a numerical measure by calculating their standard deviation. Because this is the standard deviation of the 3000 values of $\bar{x}$ that make up the bootstrap distribution, we call it the bootstrap standard error

bootstrap standard error of $\bar{x}$. The numerical value is 1.22. In fact, we know that the standard deviation of $\bar{x}$ is $\sigma/\sqrt{n}$, where σ is the standard deviation of

individual observations in the population. Our usual estimate of this quantity is the standard error of $\bar{x}$, $s/\sqrt{n}$, where s is the standard deviation of our one random sample. For these data, $s = 5.65$ and

$$\frac{s}{\sqrt{n}} = \frac{5.65}{\sqrt{21}} = 1.23$$

The bootstrap standard error 1.22 is very close to the theory-based estimate 1.23.

In discussing Example 16.2, we took advantage of the fact that statistical theory tells us a great deal about the sampling distribution of the sample mean $\bar{x}$. We found that the bootstrap distribution created by resampling matches the properties of this sampling distribution. The heavy computation needed to produce the bootstrap distribution replaces the heavy theory (central limit theorem, mean, and standard deviation of $\bar{x}$) that tells us about the sampling distribution.

The great advantage of the resampling idea is that it often works even when theory does not apply. Of course, theory also has its advantages: we know exactly when it works. We don't know exactly when resampling works, so that "When can I safely bootstrap?" is a somewhat subtle issue.

Figure 16.4 illustrates the bootstrap idea by comparing three distributions. Figure 16.4(a) shows the idea of the sampling distribution of the sample mean $\bar{x}$: take many simple random samples (SRS) from the population, calculate the mean $\bar{x}$ for each sample, and collect these $\bar{x}$-values into a distribution.

Figure 16.4(b) shows how traditional inference works: statistical theory tells us that if the population has a Normal distribution, then the sampling distribution of $\bar{x}$ is also Normal. If the population is not Normal but our sample is large, we can use the central limit theorem. If μ and σ are the mean and standard deviation of the population, the sampling distribution of $\bar{x}$ has mean μ and standard deviation $\sigma/\sqrt{n}$. When it is available, theory is wonderful: we know the sampling distribution without the impractical task of actually taking many samples from the population.

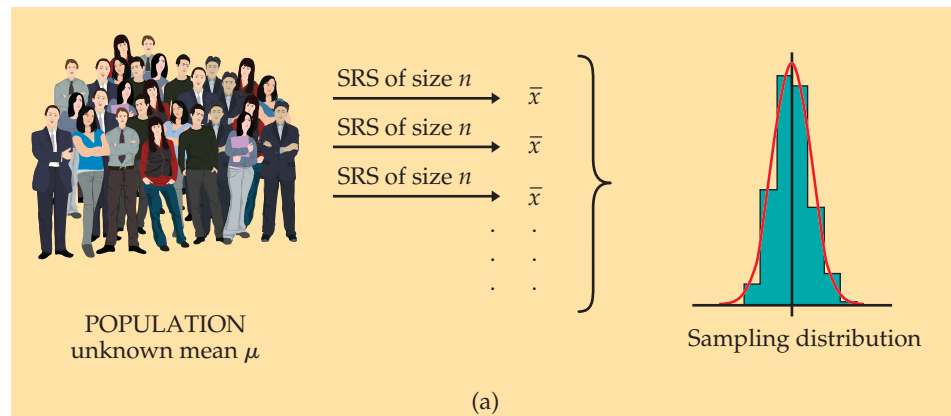**LOOK BACK**

central limit theorem, p. 298

FIGURE 16.4 (a) The idea of the sampling distribution of the sample mean $\bar{x}$: take very many samples, collect the $\bar{x}$ values from each, and look at the distribution of these values. (b) The theory shortcut: if we know that the population values follow a Normal distribution, theory tells us that the sampling distribution of $\bar{x}$ is also Normal. (c) The bootstrap idea: when theory fails and we can afford only one sample, that sample stands in for the population, and the distribution of $\bar{x}$ in many resamples stands in for the sampling distribution.
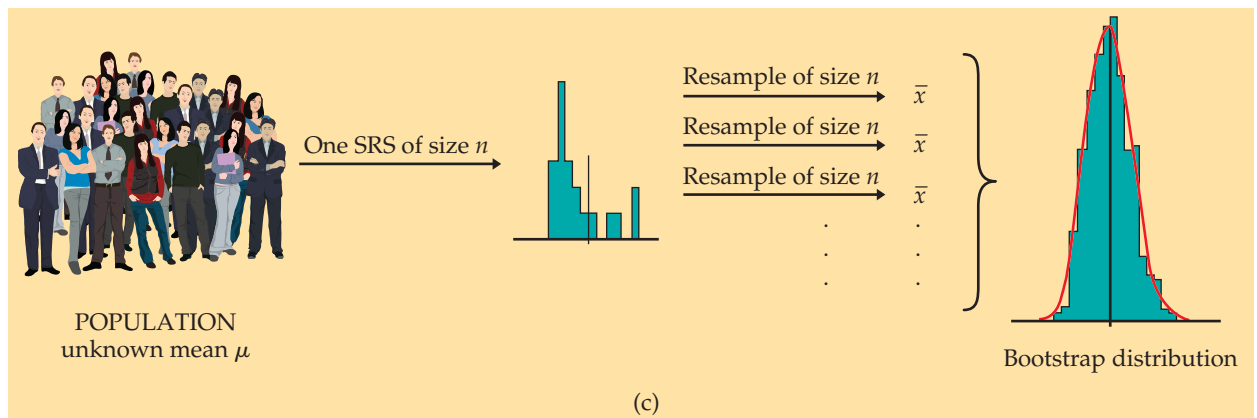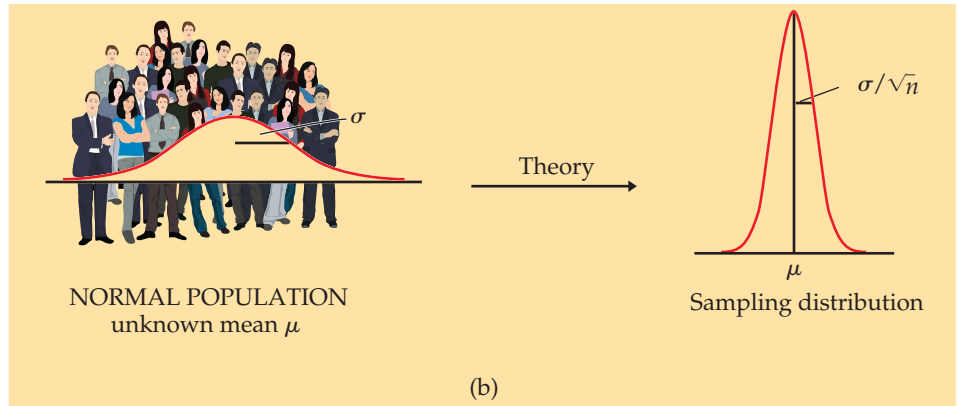
(b)

(c)

FIGURE 16.4 (*Continued*)

Figure 16.4(c) shows the bootstrap idea: we avoid the task of taking many samples from the population by instead taking many resamples from a single sample. The values of $\bar{x}$ from these resamples form the bootstrap distribution. We use the bootstrap distribution rather than theory to learn about the sampling distribution.

USE YOUR KNOWLEDGE

16.1 A small bootstrap example. To illustrate the bootstrap procedure, let's bootstrap a small random subset of the Facebook profile data:

FACE46

4.02   3.03   4.35   8.33   1.40   5.08

(a) Sample *with replacement* from this initial simple random sample (SRS) by rolling a die. Rolling a 1 means select the first member of the SRS, a 2 means select the second member, and so on. (You can also use Table B of random digits, responding only to digits 1 to 6.) Create 20 resamples of size $n = 6$.

(b) Calculate the sample mean for each of the resamples.

(c) Make a stemplot of the means of the 20 resamples. This is the bootstrap distribution.

(d) Calculate the bootstrap standard error.

> **16.2 Standard deviation versus standard error.** Explain the difference between the standard deviation of a sample and the standard error of a statistic such as the sample mean.

Thinking about the bootstrap idea

It might appear that resampling creates new data out of nothing. This seems suspicious. Even the name "bootstrap" comes from the impossible image of "pulling yourself up by your own bootstraps."[2] But the resampled observations are not used as if they were new data. The bootstrap distribution of the resample means is used only to estimate how the sample mean of one actual sample of size 21 would vary because of random sampling.

Using the same data for two purposes—to estimate a parameter and also to estimate the variability of the estimate—is perfectly legitimate. We do exactly this when we calculate $\bar{x}$ to estimate μ and then calculate $s/\sqrt{n}$ from the same data to estimate the variability of $\bar{x}$.

What is new? First of all, we don't rely on the formula $s/\sqrt{n}$ to estimate the standard deviation of $\bar{x}$. Instead, we use the ordinary standard deviation of the many $\bar{x}$-values from our many resamples.[3] Suppose that we take B resamples and call the means of these resamples $\bar{x}^*$ to distinguish them from the mean $\bar{x}$ of the original sample. We would then find the mean and standard deviation of the $\bar{x}^*$'s in the usual way.

To make clear that these are the mean and standard deviation of the means of the B resamples rather than the mean $\bar{x}$ and standard deviation s of the original sample, we use a distinct notation:

$$\text{mean}_{\text{boot}} = \frac{1}{B}\sum \bar{x}^*$$

$$\text{SE}_{\text{boot}} = \sqrt{\frac{1}{B-1}\sum (\bar{x}^* - \text{mean}_{\text{boot}})^2}$$

LOOK BACK

describing distributions with numbers, p. 27

These formulas go all the way back to Chapter 1. Once we have the values $\bar{x}^*$, we can just ask our software for their mean and standard deviation.

Because we will often apply the bootstrap to statistics other than the sample mean, here is the general definition for the bootstrap standard error.

BOOTSTRAP STANDARD ERROR

The **bootstrap standard error, SE$_{\text{boot}}$,** of a statistic is the standard deviation of the bootstrap distribution of that statistic.

Second, we don't appeal to the central limit theorem or other theory to tell us that a sampling distribution is roughly Normal. We look at the bootstrap distribution to see if it is roughly Normal (or not). In most cases, the bootstrap distribution has approximately the same shape and spread as the sampling distribution, but it is centered at the original sample statistic value rather than the parameter value.

In summary, the bootstrap allows us to calculate standard errors for statistics for which we don't have formulas and to check Normality of the sampling distribution of statistics that theory doesn't easily handle. To apply the bootstrap idea, we must start with a statistic that estimates the parameter we

are interested in. We come up with a suitable statistic by appealing to another principle that we have often applied without thinking about it.

> **THE PLUG-IN PRINCIPLE**
>
> To estimate a parameter, a quantity that describes the population, use the statistic that is the corresponding quantity for the sample.

The plug-in principle tells us to estimate a population mean μ by the sample mean $\bar{x}$ and a population standard deviation σ by the sample standard deviation s. Estimate a population median by the sample median and a population regression line by the least-squares line calculated from a sample. The bootstrap idea itself is a form of the plug-in principle: substitute the data for the population and then draw samples (resamples) to mimic the process of building a sampling distribution.

Using software

Software is essential for bootstrapping in practice. Here is an outline of the program you would write if your software can choose random samples from a set of data but does not have bootstrap functions:

```
Repeat B times {
    Draw a resample with replacement from the data.
    Calculate the resample statistic.
    Save the resample statistic into a variable.
}
Make a histogram and Normal quantile plot of the B resample statistics. Calculate
the standard deviation of the B statistics.
```

EXAMPLE 16.3

FACE4

Using software. R has packages that contain various bootstrap functions, so we do not have to write them ourselves. If the 21 viewing times are saved as a variable, we can use functions to resample from the data, calculate the means of the resamples, and request both graphs and printed output. We can also ask that the bootstrap results be saved for later access.

The function plot.boot will generate graphs similar to those in Figure 16.3 so you can assess Normality. Figure 16.5 contains the default output from a call of the function boot. The variable Time contains the 21 viewing times, the function theta is specified to be the mean, and we request 3000 resamples.

FIGURE 16.5 R output for the Facebook profile viewing time bootstrap, Example 16.3.


```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = Time, statistic = theta, R = 3000)


Bootstrap Statistics :
     original      bias     std. error
t1*  7.870476   0.02295317   1.216978
```

The **original** entry gives the mean $\bar{x} = 7.87$ of the original sample. **Bias** is the difference between the mean of the resample means and the original mean. If we add the entries for **bias** and **original**, we get the mean of the resample means, mean$_{\text{boot}}$:

$$7.87 + 0.02 = 7.89$$

The bootstrap standard error is displayed under **std. error**. All these values except **original** will differ a bit if you take another 3000 resamples because resamples are drawn at random.

SECTION 16.1 SUMMARY

• To bootstrap a statistic such as the sample mean, draw hundreds of **resamples with replacement** from a single original sample, calculate the statistic for each resample, and inspect the **bootstrap distribution** of the resample statistics.

• A bootstrap distribution approximates the sampling distribution of the statistic. This is an example of the **plug-in principle**: use a quantity based on the sample to approximate a similar quantity from the population.

• A bootstrap distribution usually has approximately the same shape and spread as the sampling distribution. It is centered at the statistic (from the original sample) while the sampling distribution is centered at the parameter (of the population).

• Use graphs and numerical summaries to determine whether the bootstrap distribution is approximately Normal and centered at the original statistic and to assess its spread. The **bootstrap standard error** is the standard deviation of the bootstrap distribution.

• The bootstrap does not replace or add to the original data. We use the bootstrap distribution as a way to estimate the sampling distribution of a statistic based on the original data.

SECTION 16.1 EXERCISES

For Exercises 16.1 and 16.2, see pages 16-7–16-8.

16.3 Gosset's data on double stout sales. William Sealy Gosset worked at the Guinness Brewery in Dublin and made substantial contributions to the practice of statistics. In Exercise 1.65 (page 48), we examined Gosset's data on the change in the double stout market before and after World War I (1914–1918). For various regions in England and Scotland, he calculated the ratio of sales in 1925, after the war, as a percent of sales in 1913, before the war. Here are the data for a sample of six of the regions in the original data: [▥ STOUT6]

Bristol	94	Glasgow	66
English P	46	Liverpool	140
English Agents	78	Scottish	24

(a) Do you think that these data appear to be from a Normal distribution? Give reasons for your answer.

(b) Select five resamples from this set of data.

(c) Compute the mean for each resample.

16.4 Find the bootstrap standard error. Refer to your work in the previous exercise. [▥ STOUT6]

(a) Would you expect the bootstrap standard error to be larger, smaller, or approximately equal to the standard deviation of the original sample of six regions? Explain your answer.

(b) Find the bootstrap standard error.

16.5 Read the output. Figure 16.6 gives a histogram and a Normal quantile plot for 3000 resample means from R. Interpret these plots.
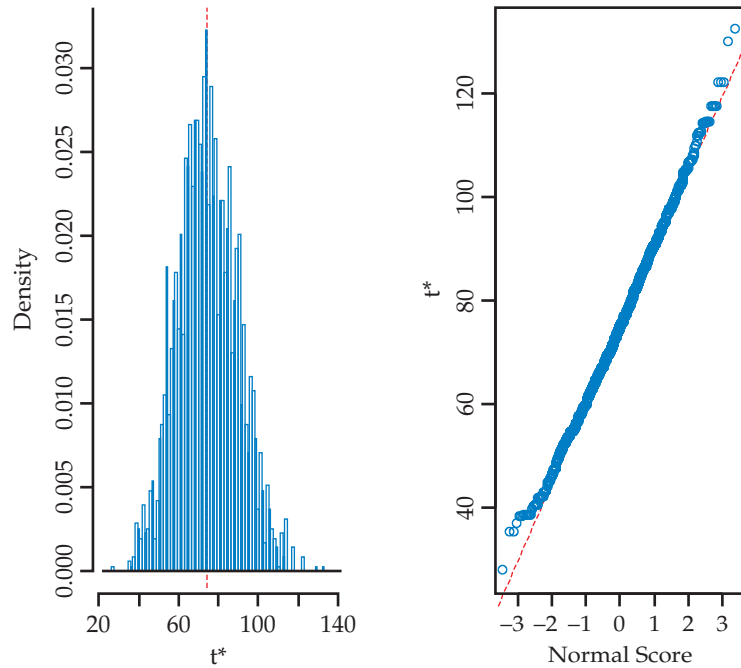
FIGURE 16.6 R output for the change in double stout sales bootstrap, Exercise 16.5.


```
R Console                                    _  □  X

  ORDINARY NONPARAMETRIC BOOTSTRAP

  Call:
  boot(data = stout, statistic = theta, R = 3000)

  Bootstrap Statistics :
        original      bias      std.error
  t1*    74.66667   -0.2038889   14.90047
```

FIGURE 16.7 R output for the change in double stout sales bootstrap, Exercise 16.6.

16.6 Read the output. Figure 16.7 gives output from R for the sample of regions in Exercise 16.3. Summarize the results of the analysis using this output.

16.7 What's wrong? Explain what is wrong with each of the following statements.

(a) The standard deviation of the bootstrap distribution will be approximately the same as the standard deviation of the original sample.

(b) The bootstrap distribution is created by resampling without replacement from the original sample.

(c) When generating the resamples, it is best to use a sample size smaller than the size of the original sample.

(d) The bootstrap distribution is created by resampling with replacement from the population.

Inspecting the bootstrap distribution of a statistic helps us judge whether the sampling distribution of the statistic is close to Normal. Bootstrap the sample mean $\bar{x}$ for each of the data sets in Exercises 16.8 through 16.12 using 2000 resamples. Construct a histogram and a Normal quantile plot to assess Normality of the bootstrap distribution. On the basis of your work, do you expect the sampling distribution of $\bar{x}$ to be close to Normal? Save your bootstrap results for later analysis.

16.8 Bootstrap distribution of average IQ score. The distribution of the 60 IQ test scores in Table 1.1 (page 14) is roughly Normal (see Figure 1.7) and the sample size is large enough that we expect a Normal sampling distribution. IQ

16.9 Bootstrap distribution of StubHub! prices. We examined the distribution of the 518 tickets for the National Collegiate Athletic Association (NCAA) Women's Final Four Basketball Championship posted for sale on StubHub! on June 28, 2014, in Example 1.48 (page 69). The distribution is clearly not Normal; it has three peaks possibly corresponding to three types of seats. We view these data as coming from a process that gives seat prices for an event such as this. STUB1

16.10 Bootstrap distribution of time spent watching traditional television. The hours per week spent watching

traditional television in a random sample of eight full-time U.S. college students (Example 7.1, page 411) are

3.0   16.5   10.5   40.5   5.5   33.5   0.0   6.5

The distribution has no outliers, but we cannot comfortably assess Normality from such a small sample. [📊] **TVTIME**

16.11 Bootstrap distribution of *Titanic* passenger ages. In Example 1.36 (page 52), we examined the distribution of the ages of the passengers on the *Titanic*. There is a single mode around 25, a short left tail, and a long right tail. We view these data as coming from a process that would generate similar data. [📊] **TITANIC**

16.12 Bootstrap distribution of average audio file length. The lengths (in seconds) of audio files found on an iPod (Table 7.5, page 470) are skewed. We previously transformed the data prior to using t procedures. [📊] **SONGS**

16.13 Standard error versus the bootstrap standard error. We have two ways to estimate the standard deviation of a sample mean $\bar{x}$: use the formula $s/\sqrt{n}$ for the standard error, or use the bootstrap standard error.

(a) Find the sample standard deviation s for the 60 IQ test scores in Exercise 16.8 and use it to find the standard error $s/\sqrt{n}$ of the sample mean. How closely does your result agree with the bootstrap standard error from your resampling in Exercise 16.8?

(b) Find the sample standard deviation s for the StubHub! ticket price data in Exercise 16.9 and use it to find the standard error $s/\sqrt{n}$ of the sample mean. How closely does your result agree with the bootstrap standard error from your resampling in Exercise 16.9?

(c) Find the sample standard deviation s for the eight traditional television viewing times in Exercise 16.10 and use it to find the standard error $s/\sqrt{n}$ of the sample mean. How closely does your result agree with the bootstrap standard error from your resampling in Exercise 16.10?

16.14 Service center call lengths. Table 1.2 (page 17) gives the service center call lengths for a sample of 80 calls. See Example 1.15 (page 15) for more details about these data. [📊] **CALLS80**

(a) Make a histogram of the call lengths. The distribution is strongly skewed.

(b) The central limit theorem says that the sampling distribution of the sample mean $\bar{x}$ becomes Normal as the sample size increases. Is the sampling distribution roughly Normal for $n = 80$? To find out, bootstrap these data using 1000 resamples and inspect the bootstrap distribution of the mean. The central part of the distribution is close to Normal. In what way do the tails depart from Normality?

16.15 More on service center call lengths. Here is an SRS of 10 of the service center call lengths from Exercise 16.14: [📊] **CALLS10**

104   102   35   211   56   325   67   9   179   59

We expect the sampling distribution of $\bar{x}$ to be less close to Normal for samples of size 10 than for samples of size 80 from a skewed distribution.

(a) Create and inspect the bootstrap distribution of the sample mean for these data using 1000 resamples. Compared with your distribution from the previous exercise, is this distribution closer to or farther away from Normal?

(b) Compare the bootstrap standard errors for your two sets of resamples. Why is the standard error larger for the smaller SRS?

16.2 First Steps in Using the Bootstrap

When you complete this section, you will be able to:	• Determine when it is appropriate to use the bootstrap standard error and the t distribution to find a confidence interval.
	• Use the bootstrap standard error and the t distribution to find a confidence interval.

To introduce the key ideas of resampling and bootstrap distributions, we studied an example in which we knew quite a bit about the actual sampling distribution. We saw that the bootstrap distribution agrees with the sampling distribution in *shape* and *spread*.

The *center* of the bootstrap distribution is not the same as the center of the sampling distribution. The sampling distribution of a statistic used to estimate a parameter is centered at the actual value of the parameter in the population, plus any bias. The bootstrap distribution is centered at the value

of the statistic for the original sample, plus any bias. The key fact is that the two biases are similar even though the two centers may not be.

The bootstrap method is most useful in settings where we don't know the sampling distribution of the statistic. The principles are

• **Shape:** Because the shape of the bootstrap distribution approximates the shape of the sampling distribution, we can use the bootstrap distribution to check Normality of the sampling distribution.

• **Center:** A statistic is biased as an estimate of the parameter if its sampling distribution is not centered at the true value of the parameter. We can check bias by seeing whether the bootstrap distribution of the statistic is centered at the value of the statistic for the original sample.

More precisely, the bias of a statistic is the difference between the mean of its sampling distribution and the true value of the parameter.

bootstrap estimate of bias

The **bootstrap estimate of bias** is the difference between the mean of the bootstrap distribution and the value of the statistic in the original sample.

• **Spread:** The bootstrap standard error of a statistic is the standard deviation of its bootstrap distribution. The bootstrap standard error estimates the standard deviation of the sampling distribution of the statistic.

Bootstrap *t* confidence intervals

If the bootstrap distribution of a statistic shows a Normal shape and small bias, we can get a confidence interval for the parameter by using the bootstrap standard error and the familiar *t* distribution. An example will show how this works.

EXAMPLE 16.4

Digital Vision/Thinkstock

GPA

Grade point averages. A study of college students at a large university looked at grade point average (GPA) after three semesters of college as a measure of success. In Example 11.1 (page 608), we examined predictors of GPA. Let's take a look at the distribution of the GPA for the 150 students in this study.

A histogram is given in Figure 16.8(a). The Normal quantile plot is given in Figure 16.8(b). The distribution is strongly skewed to the left. The Normal quantile plot suggests that there are several students with perfect (4.0) GPAs and one at the lower end of the distribution (0.0). These data are not Normally distributed.

The first step is to abandon the mean as a measure of center in favor of a statistic that focuses on the central part of the distribution. We might choose the median, but in this case, we will use the 25% trimmed mean, the mean of the middle 50% of the observations. The median is the middle observation or the mean of the two middle observations. The trimmed mean often does a better job of representing the average of typical observations than does the median.

Our *parameter* is the 25% trimmed mean of the population of college student GPAs after three semesters at this large university. By the plug-in principle, the *statistic* that estimates this parameter is the 25% trimmed mean of the sample of 150 students. Because 25% of 150 is 37.5, we drop the 37 lowest and 37 highest GPAs and find the mean of the remaining 76 GPAs. The statistic is
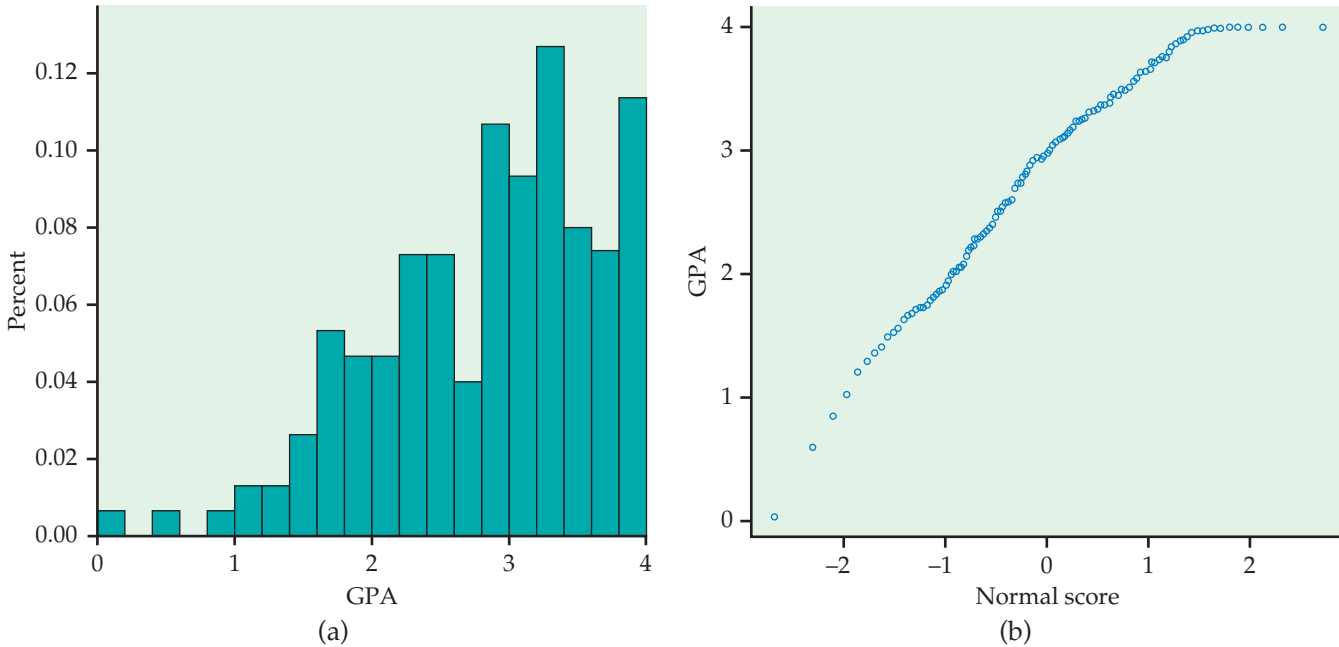
$$\bar{x}_{25\%} = 2.950$$

(a)   (b)

FIGURE 16.8 Histogram and Normal quantile plot for 150 grade point averages, Example 16.4. The distribution is strongly skewed.

Given the relatively large sample size from this strongly skewed distribution, we can use the central limit theorem to argue that the sampling distribution would be approximately Normal with mean near 2.950. Estimating its standard deviation, however, is a more difficult task. We can't simply use the standard error of the sample mean based on the remaining 76 observations, as that will underestimate the true variability.

Fortunately, we don't need any distribution facts to use the bootstrap. We bootstrap the 25% trimmed mean just as we bootstrapped the sample mean: draw 3000 resamples of size 150 from the 150 GPAs, calculate the 25% trimmed mean for each resample, and form the bootstrap distribution from these 3000 values.

Figure 16.9 shows the bootstrap distribution of the 25% trimmed mean. Here is the summary output from R:

ORDINARY NONPARAMETRIC BOOTSTRAP

Call:
boot(data = GPA, statistic = theta, R = 3000)

Bootstrap Statistics :
         original   bias   std. error
t1*   2.949605   −0.002912   0.0778597

What do we see?

Shape: The bootstrap distribution is close to Normal. This suggests that the sampling distribution of the trimmed mean is also close to Normal.

Center: The bootstrap estimate of bias is −0.003, which is small relative to the value 2.950 of the statistic. So the statistic (the trimmed mean of the sample) has small bias as an estimate of the parameter (the trimmed mean of the population).
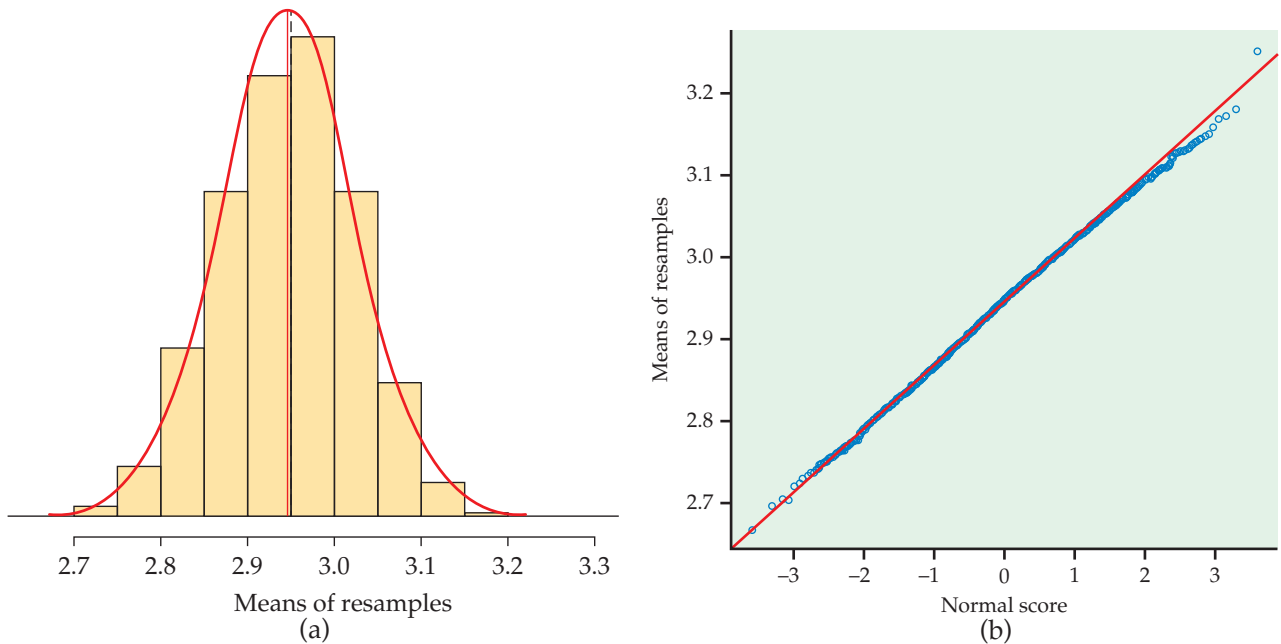
FIGURE 16.9 The bootstrap distribution of the 25% trimmed means for 3000 resamples from the GPA data in Example 16.4. The bootstrap distribution is approximately Normal.

Spread: The bootstrap standard error of the statistic is

$$SE_{boot} = 0.078$$

This is an estimate of the standard deviation of the sampling distribution of the trimmed mean.

Recall the familiar one-sample t confidence interval (page 411) for the mean of a Normal population:

$$\bar{x} \pm t^*SE = \bar{x} \pm t^*\frac{s}{\sqrt{n}}$$

This interval is based on the Normal sampling distribution of the sample mean $\bar{x}$ and the formula $SE = s/\sqrt{n}$ for the standard error of $\bar{x}$. When a bootstrap distribution is approximately Normal and has small bias, we can essentially use the same idea with the bootstrap standard error to get a confidence interval for any parameter.

BOOTSTRAP t CONFIDENCE INTERVAL

Suppose that the bootstrap distribution of a statistic from an SRS of size n is approximately Normal and that the bootstrap estimate of bias is small. An approximate **level C confidence interval** for the parameter that corresponds to this statistic by the plug-in principle is

$$\text{statistic} \pm t^*SE_{boot}$$

where SE_{boot} is the bootstrap standard error for this statistic and t^* is the critical value of the $t(n-1)$ distribution with area C between $-t^*$ and t^*.

EXAMPLE 16.5

GPA

Bootstrap distribution of the trimmed mean. We want to estimate the 25% trimmed mean of the population of all college student GPAs after three semesters at this large university. We have an SRS of size $n = 150$ but use only the middle 76 observations for our estimate. The preceding software output shows that the trimmed mean of this sample is $\bar{x}_{25\%} = 2.950$ and that the bootstrap standard error of this statistic is $SE_{boot} = 0.078$. A 95% confidence interval for the population trimmed mean is, therefore,

$$\bar{x}_{25\%} \pm t^*SE_{boot} = 2.950 \pm (2.000)(0.078)$$
$$= 2.950 \pm 0.156$$
$$= (2.794, 3.106)$$

Because Table D does not have entries for $[n - 2(37)] - 1 = 75$ degrees of freedom, we used $t^* = 2.000$, the entry for 60 degrees of freedom.

We are 95% confident that the 25% trimmed mean (the mean of the middle 50%) for the population of college student GPAs after three semesters at this large university is between 2.794 and 3.106.

USE YOUR KNOWLEDGE

16.16 Bootstrap t confidence interval. Recall Example 16.1 (page 16-3). Suppose a bootstrap distribution was created using 3000 resamples and that the mean and standard deviation of the resample means were 7.91 and 1.26, respectively.

(a) What is the bootstrap estimate of the bias?

(b) What is the bootstrap standard error of $\bar{x}$?

(c) Assume that the bootstrap distribution is reasonably Normal. Because the bias is small relative to the observed $\bar{x}$, the bootstrap t confidence interval for the population mean μ is justified. Give the 95% bootstrap t confidence interval for μ.

16.17 Bootstrap t confidence interval for average audio file length. Return to or create the bootstrap distribution resamples on the sample mean for audio file length in Exercise 16.12 (page 16-12). In Example 7.25 (page 470), the t confidence interval was applied to the logarithm of the time measurements.

SONGS

(a) Inspect the bootstrap distribution. Is a bootstrap t confidence interval appropriate? Explain why or why not.

(b) Construct the 95% bootstrap t confidence interval.

(c) Compare the bootstrap results with the t confidence interval reported in Example 7.26 (page 471).

Bootstrapping to compare two groups

Two-sample problems are among the most common statistical settings. In a two-sample problem, we wish to compare two populations, such as male and female college students, based on separate samples from each population.

When both populations are roughly Normal, the two-sample t procedures compare the two population means. The bootstrap can also compare two populations, without the Normality condition and without the restriction to a comparison of means. The most important new idea is that bootstrap resampling must mimic the "separate samples" design that produced the original data.

LOOK BACK

two-sample
t confidence
interval,
p. 437

BOOTSTRAP FOR COMPARING TWO POPULATIONS

Given independent SRSs of sizes n and m from two populations:

1. Draw a resample of size n with replacement from the first sample and a separate resample of size m from the second sample. Compute a statistic that compares the two groups, such as the difference between the two sample means.

2. Repeat this resampling process thousands of times.

3. Construct the bootstrap distribution of the statistic. Inspect its shape, bias, and bootstrap standard error in the usual way.

EXAMPLE 16.6

GPA

Bootstrap comparison of GPAs. In Example 16.4, we looked at grade point average (GPA) after three semesters of college as a measure of success. How do GPAs compare between men and women? Figure 16.10 shows density curves and Normal quantile plots for the GPAs of 91 males and
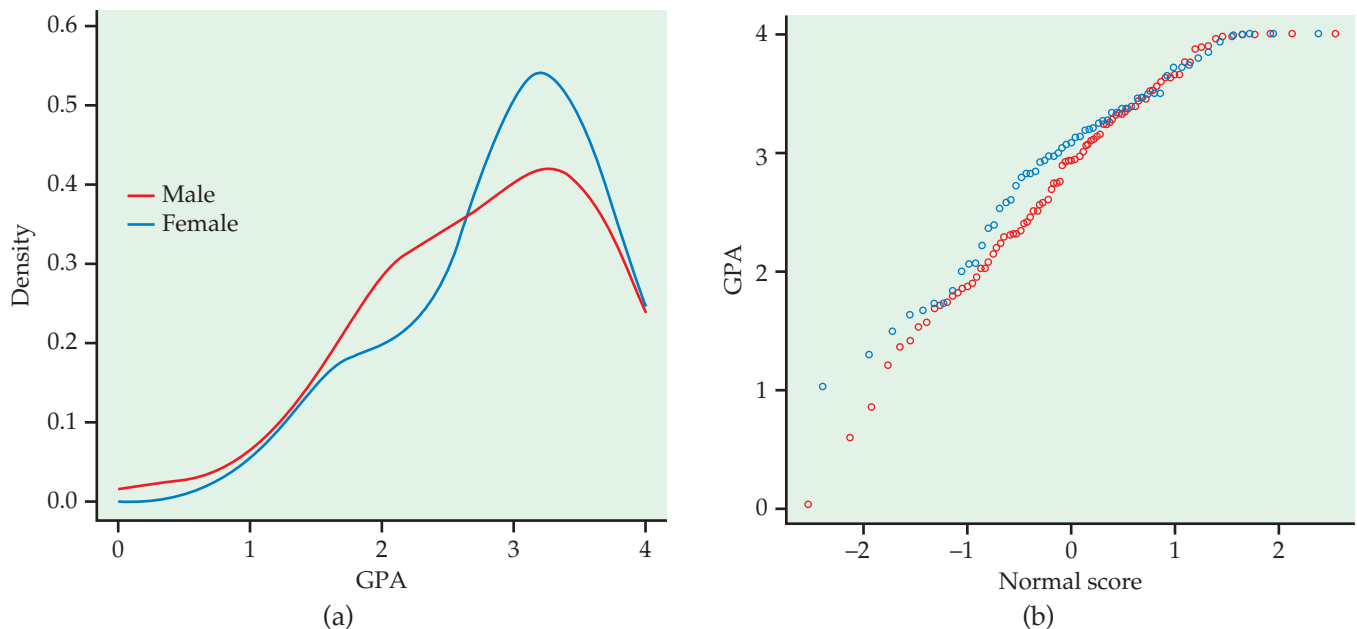

FIGURE 16.10 Density curves and Normal quantile plots of the distributions of GPA for males and females, Example 16.6.

59 females. The distributions are both far from Normal. Here are some summary statistics:

Sex	n	$\bar{x}$	s
Male	91	2.784	0.859
Female	59	2.933	0.748
Difference		−0.149	

The data suggest that GPAs tend to be slightly higher for females. The mean GPA for females is roughly 0.15 higher than the mean for males.

In the setting of Example 16.6, we want to estimate the difference between population means, $\mu_1 - \mu_2$. We might be somewhat reluctant to use the two-sample t confidence interval because both samples are very skewed. To compute the bootstrap standard error for the difference in sample means $\bar{x}_1 - \bar{x}_2$, resample separately from the two samples. Each of our 3000 resamples consists of two group resamples, one of size 91 drawn with replacement from the male data and one of size 59 drawn with replacement from the female data.

For each combined resample, compute the statistic $\bar{x}_1 - \bar{x}_2$. The 3000 differences form the bootstrap distribution. The bootstrap standard error is the standard deviation of the bootstrap distribution.

The boot function in R automates this bootstrap procedure. Here is the R output:

STRATIFIED BOOTSTRAP

Call:
boot(data = gpa, statistic = meanDiff, R = 3000, strata = sex)

Bootstrap Statistics:

	original	bias	std. error
t1*	-0.1490259	0.003989901	0.1327419

Figure 16.11 shows that the bootstrap distribution is close to Normal. We can trust the bootstrap t confidence interval for these data. A 95% confidence interval for the difference in mean GPAs (males versus females) is, therefore,

$$\bar{x}_{25\%} \pm t^*\text{SE}_{\text{boot}} = -0.149 \pm (2.009)(0.133)$$
$$= -0.149 \pm 0.267$$
$$= (-0.416, 0.118)$$

Because Table D does not have entries for $\min(n_1 - 1, n_2 - 1) = 58$ degrees of freedom, we used $t^* = 2.009$, the entry for 50 degrees of freedom.

We are 95% confident that the difference in the mean GPAs of males and females at this large university after three semesters is between −0.416 and 0.118. Because 0 is in this interval, we cannot conclude that the two population means are different. We discuss hypothesis testing in Section 16.5.

In this example, the bootstrap distribution of the difference is close to Normal. *When the bootstrap distribution is non-Normal, we can't trust the bootstrap* t *confidence interval.* Fortunately, there are more general ways of using the bootstrap to get confidence intervals that can be safely applied when the bootstrap distribution is not Normal. These methods, which we discuss in Section 16.4, are the next step in practical use of the bootstrap.
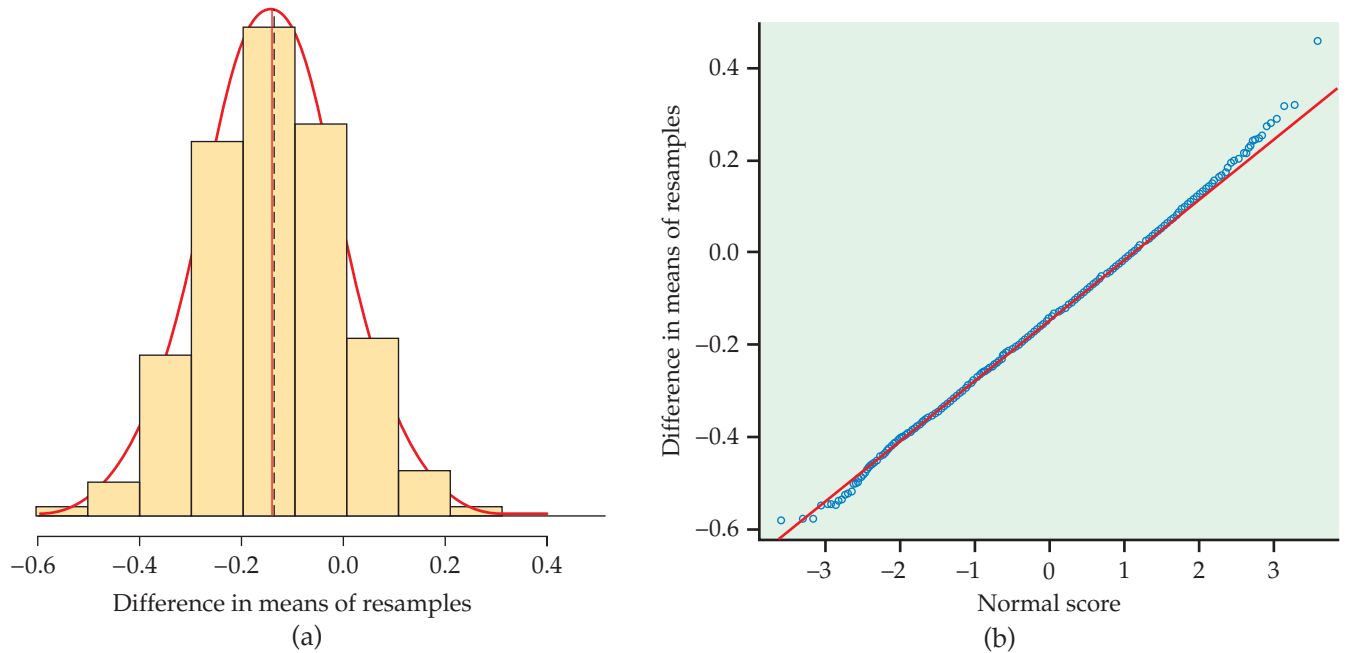
FIGURE 16.11 The bootstrap distribution and Normal quantile plot for the differences in means for the GPA data.

USE YOUR KNOWLEDGE

DRP

16.18 **Bootstrap comparison of average reading abilities.** Table 7.3 (page 437) gives the scores on a test of reading ability for two groups of third-grade students. The treatment group used "directed reading activities" and the control group followed the same curriculum without the activities.

(a) Bootstrap the difference in means $\bar{x}_1 - \bar{x}_2$ and report the bootstrap standard error.

(b) Inspect the bootstrap distribution. Is a bootstrap t confidence interval appropriate? If so, give a 95% confidence interval.

(c) Compare the bootstrap results with the two-sample t confidence interval reported in Example 7.14 on page 441.

GPA

16.19 **Formula-based versus bootstrap standard error.** We have a formula (page 436) for the standard error of $\bar{x}_1 - \bar{x}_2$. This formula does not depend on Normality. How does this formula-based standard error for the data of Example 16.6 compare with the bootstrap standard error?

BEYOND THE BASICS

The Bootstrap for a Scatterplot Smoother

The bootstrap idea can be applied to quite complicated statistical methods, such as the scatterplot smoother illustrated in Chapter 2 (page 94).

EXAMPLE 16.7

Do all daily numbers have an equal payoff? The New Jersey Pick-It Lottery is a daily numbers game run by the state of New Jersey. We'll analyze the first 254 drawings after the lottery was started in 1975.[4] Buying a ticket entitles a player to pick a number between 000 and 999. Half the money bet each day goes into the prize pool. (The state takes the other half.) The state picks a winning number at random, and the prize pool is shared equally among all winning tickets.

Although all numbers are equally likely to win, numbers chosen by fewer people have bigger payoffs if they win because the prize is shared among fewer tickets. Figure 16.12 is a scatterplot of the first 254 winning numbers and their payoffs. What patterns can we see?
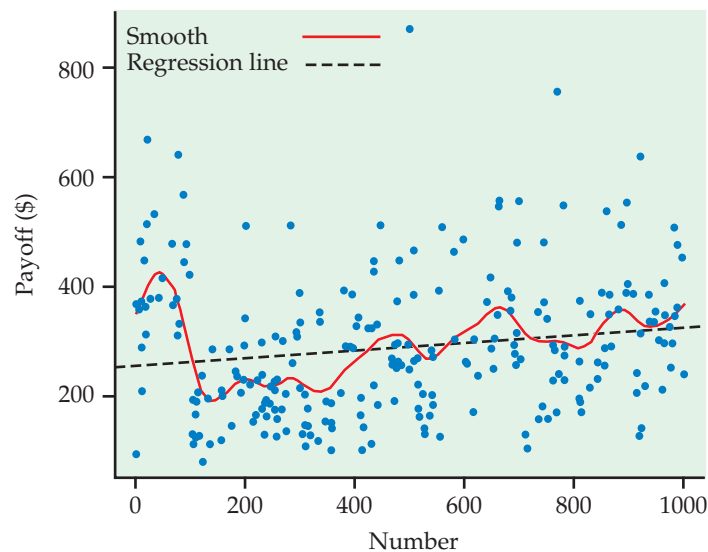

FIGURE 16.12 The first 254 winning numbers in the New Jersey Pick-It Lottery and the payoffs for each, Example 16.7. To see patterns, we use least-squares regression (dashed line) and a scatterplot smoother (curve).

The straight line in Figure 16.12 is the least-squares regression line. The line shows a general trend of higher payoffs for larger winning numbers. The curve in the figure was fitted to the plot by a scatterplot smoother that follows local patterns in the data rather than being constrained to a straight line. The curve suggests that there were larger payoffs for numbers in the intervals 000 to 100, 400 to 500, 600 to 700, and 800 to 999.

Are the patterns displayed by the scatterplot smoother just chance? We can use the bootstrap distribution of the smoother's curve to get an idea of how much random variability there is in the curve. Each resample "statistic" is now a curve rather than a single number. Figure 16.13 shows the curves that result from applying the smoother to 20 resamples from the 254 data points in Figure 16.12. The original curve is the thick line. The spread of the resample curves about the original curve shows the sampling variability of the output of the scatterplot smoother.
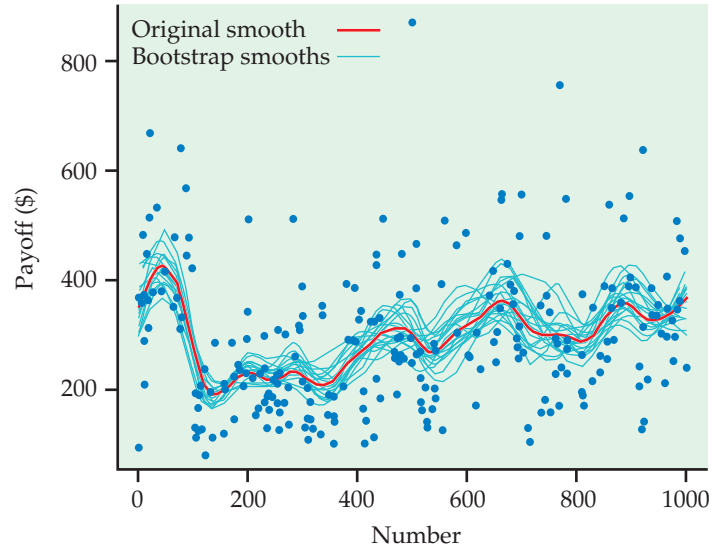
FIGURE 16.13 The curves produced by the scatterplot smoother for 20 resamples from the data displayed in Figure 16.12. The curve for the original sample is the heavy line.

Nearly all the bootstrap curves mimic the general pattern of the original smoother curve, showing, for example, the same low average payoffs for numbers in the 200s and 300s. This suggests that these patterns are real, not just chance. In fact, when people pick "random" numbers, they tend to choose numbers starting with 2, 3, 5, or 7, so these numbers have lower payoffs. This pattern disappeared after 1976; it appears that players noticed the pattern and changed their number choices.

SECTION 16.2 SUMMARY

• Bootstrap distributions mimic the shape, spread, and bias of sampling distributions.

• The **bootstrap standard error, SE$_{boot}$**, of a statistic is the standard deviation of its bootstrap distribution. It measures how much the statistic varies under random sampling.

• The bootstrap estimate of the **bias** of a statistic is the mean of the bootstrap distribution minus the statistic for the original data. Small bias means that the bootstrap distribution is centered at the statistic of the original sample and suggests that the sampling distribution of the statistic is centered at the population parameter.

• The bootstrap can estimate the sampling distribution, bias, and standard error of a wide variety of statistics, such as the **trimmed mean**, whether or not statistical theory tells us about their sampling distributions.

• If the bootstrap distribution is approximately Normal and the bias is small, we can give a **bootstrap t confidence interval, statistic ± t^*SE$_{boot}$,**

for the parameter. Do not use this t interval if the bootstrap distribution is not Normal or shows substantial bias.

• To use the bootstrap **to compare two populations**, draw separate resamples from each sample and compute a statistic that compares the two groups. Repeat many times and use the bootstrap distribution for inference.

SECTION 16.2 EXERCISES

For Exercises 16.16 and 16.17, see page 16-16; and for Exerscises 16.18 and 16.19, see page 16-19.

16.20 Should you use the bootstrap standard error and the t distribution for the confidence interval? For each of the following situations, explain whether or not you would use the bootstrap standard error and the t distribution for the confidence interval. Give reasons for your answers.

(a) The bootstrap distribution of the mean is approximately Normal, and the difference between the mean of the data and the mean of the bootstrap distribution is large relative to the mean of the data.

(b) The bootstrap distribution of the mean is approximately Normal, and the difference between the mean of the data and the mean of the bootstrap distribution is small relative to the mean of the data.

(c) The bootstrap distribution of the mean is clearly skewed, and the difference between the mean of the data and the mean of the bootstrap distribution is large relative to the mean of the data.

(d) The bootstrap distribution of the mean is clearly skewed, and the difference between the mean of the data and the mean of the bootstrap distribution is small relative to the mean of the data.

16.21 Use the bootstrap standard error and the t distribution for the confidence interval. The observed mean is 108.7, the mean of the bootstrap distribution is 109.8, the standard error is 8.2, and $n = 41$. Use the t distribution to find the 95% confidence interval.

16.22 Bootstrap t confidence interval for the StubHub! prices. In Exercise 16.9 (page 16-11), we examined the bootstrap for the prices of tickets to the NCAA Women's Final Four Basketball Championship. STUB1

(a) Find the bootstrap t 95% confidence interval for these data.

(b) Compare the interval you found in part (a) with the usual t interval.

(c) Which interval do you prefer? Give reasons for your answer.

16.23 Bootstrap t confidence interval for the ages of the *Titanic* passengers. In Exercise 16.11 (page 16-12), we examined the bootstrap for the ages of the *Titanic* passengers. TITANIC

(a) Find the bootstrap t 95% confidence interval for these data.

(b) Compare the interval you found in part (a) with the usual t interval.

(c) Which interval do you prefer? Give reasons for your answer.

16.24 Bootstrap t confidence interval for time spent watching traditional television. Return to or re-create the bootstrap distribution of the sample mean for the eight times spent watching television in Exercise 16.10 (page 16-11). TVTIME

(a) Although the sample is small, verify using graphs and numerical summaries of the bootstrap distribution that the distribution is reasonably Normal and that the bias is small relative to the observed $\bar{x}$.

(b) The bootstrap t confidence interval for the population mean μ is, therefore, justified. Give the 95% bootstrap t confidence interval for μ.

(c) Give the usual t 95% interval and compare it with your interval from part (b).

16.25 Bootstrap t confidence interval for service center call lengths. Return to or re-create the bootstrap distribution of the sample mean for the 80 service center call lengths in Exercise 16.14 (page 16-12). CALLS80

(a) What is the bootstrap estimate of the bias? Verify from the graphs of the bootstrap distribution that the distribution is reasonably Normal (some right-skew remains) and that the bias is small relative to the observed $\bar{x}$. The bootstrap t confidence interval for the population mean μ is, therefore, justified.

(b) Give the 95% bootstrap t confidence interval for μ.

(c) The only difference between the bootstrap t and usual one-sample t confidence intervals is that the

bootstrap interval uses SE_{boot} in place of the formula-based standard error $s/\sqrt{n}$. What are the values of the two standard errors? Give the usual t 95% interval and compare it with your interval from part (b).

16.26 Another bootstrap distribution of the trimmed mean. Bootstrap distributions and quantities based on them differ randomly when we repeat the resampling process. A key fact is that they do not differ very much if we use a large number of resamples. Figure 16.9 (page 16-15) shows one bootstrap distribution of the trimmed mean of the GPA data. Repeat the resampling of these data to get another bootstrap distribution of the trimmed mean. **GPA**

(a) Plot the bootstrap distribution and compare it with Figure 16.9. Are the two bootstrap distributions similar?

(b) What are the values of the bias and bootstrap standard error for your new bootstrap distribution? How do they compare with the previous values given on page 16-15?

(c) Find the 95% bootstrap t confidence interval based on your bootstrap distribution. Compare it with the previous result in Example 16.5 (page 16-16).

16.27 Bootstrap distribution of the standard deviation s. For Example 16.5 (page 16-16), we bootstrapped the 25% trimmed mean of 150 GPAs. Another statistic whose sampling distribution is unfamiliar to us is the standard deviation s. Bootstrap s for these data. Discuss the shape and bias of the bootstrap distribution. Is the bootstrap t confidence interval for the population standard deviation σ justified? If it is, give a 95% confidence interval. **GPA**

16.28 Bootstrap comparison of tree diameters. In Exercise 7.79 (page 459), you were asked to compare the mean diameter at breast height (DBH) for trees from the northern and southern halves of a land tract using a random sample of 30 trees from each region. **NSPINES**

(a) Use a back-to-back stemplot or side-by-side boxplots to examine the data graphically. Does it appear reasonable to use standard t procedures?

(b) Bootstrap the difference in means $\bar{x}_{North} - \bar{x}_{South}$ and look at the bootstrap distribution. Does it meet the conditions for a bootstrap t confidence interval?

(c) Report the bootstrap standard error and the 95% bootstrap t confidence interval.

(d) Compare the bootstrap results with the usual two-sample t confidence interval.

16.29 Bootstrapping a Normal data set. The following data are "really Normal." They are an SRS from the standard Normal distribution $N(0, 1)$, produced by a software Normal random number generator. **NORMALD**

0.01	−0.04	−1.02	−0.13	−0.36	−0.03	−1.88	0.34	−0.00	1.21
−0.02	−1.01	0.58	0.92	−1.38	−0.47	−0.80	0.90	−1.16	0.11
0.23	2.40	0.08	−0.03	0.75	2.29	−1.11	−2.23	1.23	1.56
−0.52	0.42	−0.31	0.56	2.69	1.09	0.10	−0.92	−0.07	−1.76
0.30	−0.53	1.47	0.45	0.41	0.54	0.08	0.32	−1.35	−2.42
0.34	0.51	2.47	2.99	−1.56	1.27	1.55	0.80	−0.59	0.89
−2.36	1.27	−1.11	0.56	−1.12	0.25	0.29	0.99	0.10	0.30
0.05	1.44	−2.46	0.91	0.51	0.48	0.02	−0.54		

(a) Make a histogram and Normal quantile plot. Do the data appear to be "really Normal"? From the histogram, does the $N(0, 1)$ distribution appear to describe the data well? Why?

(b) Bootstrap the mean. Why do your bootstrap results suggest that t confidence intervals are appropriate?

(c) Give both the bootstrap and the formula-based standard errors for $\bar{x}$. Give both the bootstrap and usual t 95% confidence intervals for the population mean μ.

16.30 Bootstrap distribution of the median. We will see in Section 16.3 that bootstrap methods often work poorly for the median. To illustrate this, bootstrap the sample median of the 21 viewing times we studied in Example 16.1 (page 16-3). Why is the bootstrap t confidence interval not justified? **FACE4**

16.31 Bootstrap distribution of the mpg standard deviation. Computers in some vehicles calculate various quantities related to performance. One of these is the fuel efficiency, or gas mileage, usually expressed as miles per gallon (mpg). For one vehicle equipped in this way, the mpg were recorded each time the gas tank was filled, and the computer was then reset. We studied these data in Exercise 7.32 (page 429) using methods based on Normal distributions.[5] **MPG20**

41.5	50.7	36.6	37.3	34.2	45.0	48.0	43.2	47.7	42.2
43.2	44.6	48.4	46.4	46.8	39.2	37.3	43.5	44.3	43.3

In addition to the average mpg, the driver is also interested in how much variability there is in the mpg.

(a) Calculate the sample standard deviation s for these mpg values.

(b) We have no formula for the standard error of s. Find the bootstrap standard error for s.

(c) What does the standard error indicate about how accurate the sample standard deviation is as an estimate of the population standard deviation?

(d) Would it be appropriate to give a bootstrap t interval for the population standard deviation? Why or why not?

16.3 How Accurate Is a Bootstrap Distribution?

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

- Describe the effect of the size of the original sample on the variation in bootstrap distributions.
- Describe the effect of the number of resamples on the variation in bootstrap distributions.

We said earlier that "When can I safely bootstrap?" is a somewhat subtle issue. Now we will give some insight into this issue.

Because a statistic varies sample to sample, inference about the population must take this random variation into account. The sampling distribution of a statistic displays the variation in the statistic due to selecting samples at random from the population. For example, the margin of error in a confidence interval expresses the uncertainty due to sampling variation. In this chapter, we have used the bootstrap distribution as a substitute for the sampling distribution. This introduces a second source of random variation: choosing resamples at random from the original sample.

SOURCES OF VARIATION IN A BOOTSTRAP DISTRIBUTION

Bootstrap distributions and conclusions based on them include two sources of random variation:

1. Choosing an original sample at random from the population.

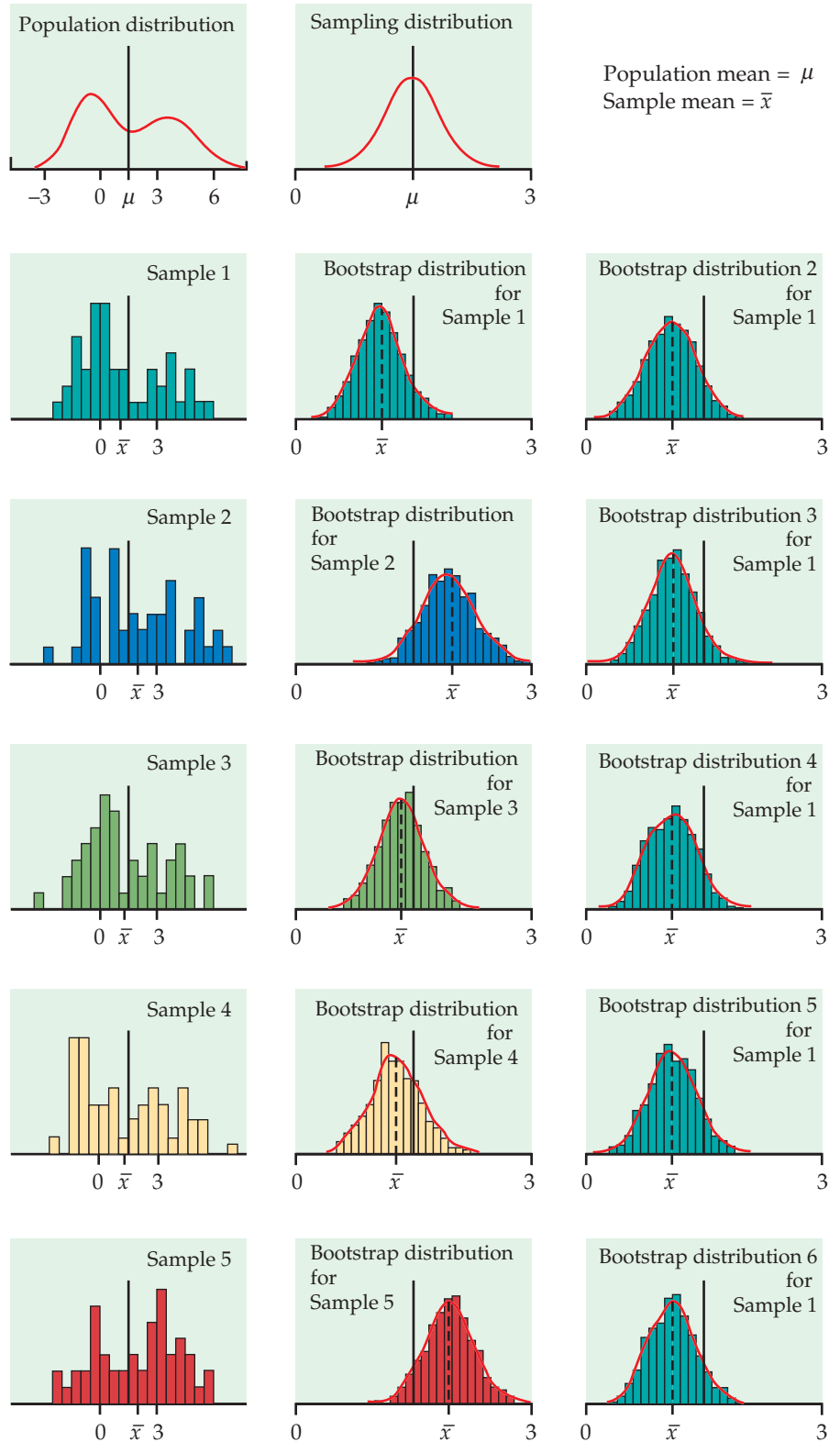2. Choosing bootstrap resamples at random from the original sample.

A statistic in a given setting has only one sampling distribution. It has many bootstrap distributions, formed by the two-step process just described. Bootstrap inference generates one bootstrap distribution and uses it to tell us about the sampling distribution. Can we trust such inference?

Figure 16.14 displays an example of the entire process. The population distribution (top left) has two peaks and is far from Normal. The histograms in the left column of the figure show five random samples from this population, each of size 50. The line in each histogram marks the mean $\bar{x}$ of that sample. These vary from sample to sample. The distribution of the $\bar{x}$-values from all possible samples is the sampling distribution. This sampling distribution appears to the right of the population distribution. It is close to Normal, as we expect because of the central limit theorem.

The middle column in Figure 16.14 displays the bootstrap distribution of $\bar{x}$ for each of the five samples. Each distribution was created by drawing 1000 resamples from the original sample, calculating $\bar{x}$ for each resample, and presenting the 1000 $\bar{x}$'s in a histogram. The right column shows the bootstrap distribution of the first sample, repeating the resampling five more times.

Compare the five bootstrap distributions in the middle column to see the effect of the random choice of the original sample. Compare the six bootstrap

FIGURE 16.14 Five random samples of $n = 50$ from the same population, with a bootstrap distribution of the sample mean formed by resampling from each of the five samples. At the right are five more bootstrap distributions from the first sample.

distributions drawn from the first sample to see the effect of the random resampling. Here's what we see:

• Each bootstrap distribution is centered close to the value of $\bar{x}$ for its original sample. That is, the bootstrap estimate of bias is small in all five cases. Of course, the five $\bar{x}$-values vary, and not all are close to the population mean μ.

• The shape and spread of the bootstrap distributions in the middle column vary a bit, but all five resemble the sampling distribution in shape and spread. That is, the shape and spread of a bootstrap distribution depend on the original sample, but the variation from sample to sample is not great.

• The six bootstrap distributions from the same sample are very similar in shape, center, and spread. That is, *random resampling adds very little variation to the variation due to the random choice of the original sample from the population.*

Figure 16.14 reinforces facts that we have already relied on. If a bootstrap distribution is based on a moderately large sample from the population, its shape and spread don't depend heavily on the original sample and do mimic the shape and spread of the sampling distribution. Bootstrap distributions do not have the same center as the sampling distribution; they mimic bias, not the actual center.

The figure also illustrates a fact that is important for practical use of the bootstrap: the bootstrap resampling process (using 1000 or more resamples) introduces very little additional variation. We can rely on a bootstrap distribution to inform us about the shape, bias, and spread of the sampling distribution.

Bootstrapping small samples

We now know that almost all the variation in bootstrap distributions for a statistic such as the mean comes from the random selection of the original sample from the population. We also know that, in general, larger samples are preferred because small samples give more variable results. This general fact is also true for bootstrap procedures.
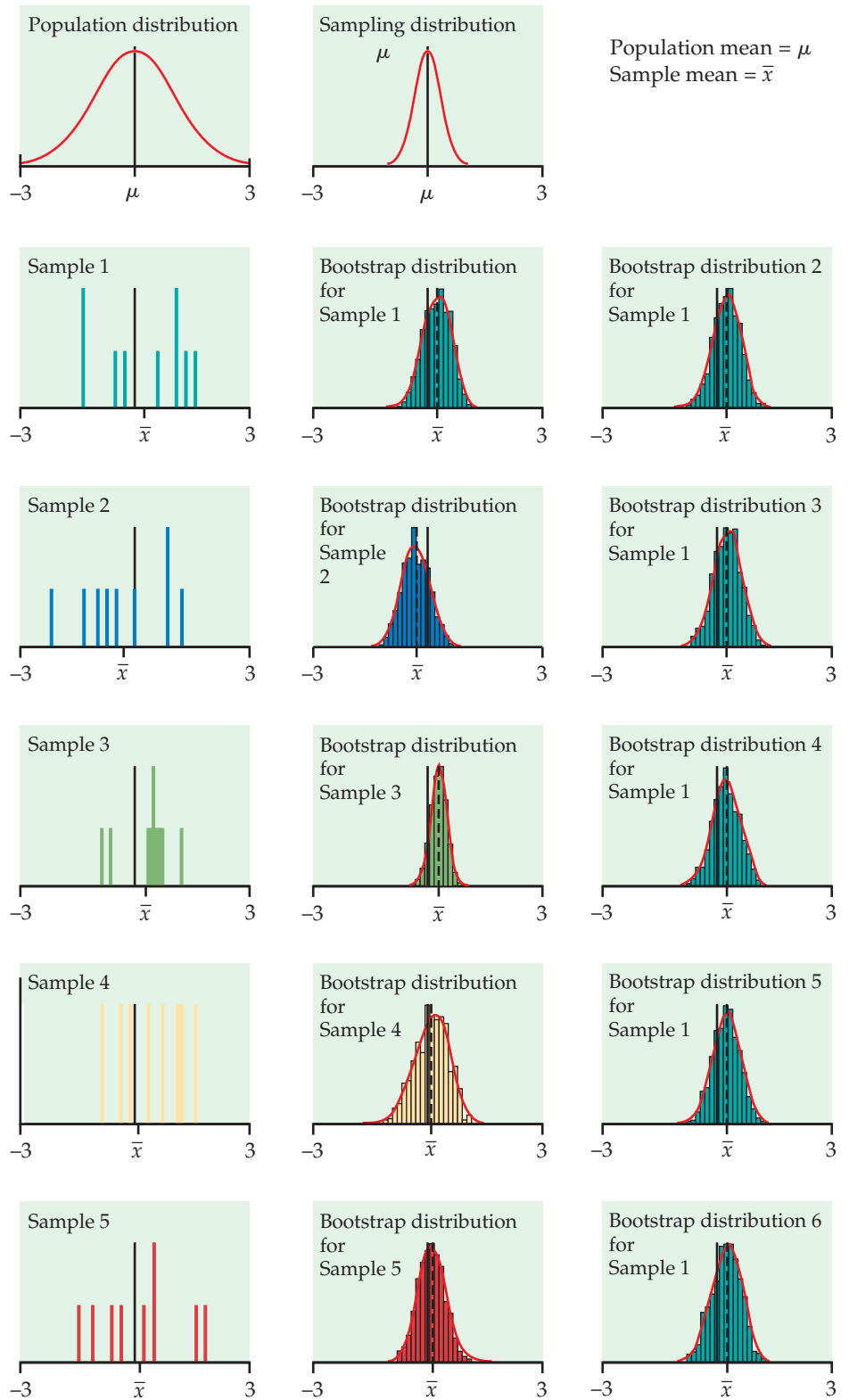
Figure 16.15 repeats Figure 16.14, with two important differences. The five original samples are only of size $n = 9$, rather than the $n = 50$ of Figure 16.14. Also, the population distribution (top left) is Normal, so that the sampling distribution of $\bar{x}$ is Normal despite the small sample size.

Even with a Normal population distribution, the bootstrap distributions in the middle column show much more variation in shape and spread than those for larger samples in Figure 16.14. Notice, for example, how the skewness of the fourth sample produces a skewed bootstrap distribution. The bootstrap distributions are no longer all similar to the sampling distribution at the top of the column.

We can't trust a bootstrap distribution from a very small sample to closely mimic the shape and spread of the sampling distribution. Bootstrap confidence intervals will sometimes be too long or too short or too long in one direction and too short in the other. The six bootstrap distributions based on the first sample are again very similar. Because we used 1000 resamples, resampling adds very little variation. There are subtle effects that can't be seen from a few pictures, but the main conclusions are clear.

FIGURE 16.15 Five random samples of $n = 9$ from the same population, with a bootstrap distribution of the sample mean formed by resampling from each of the five samples. At the right are five more bootstrap distributions from the first sample.

VARIATION IN BOOTSTRAP DISTRIBUTIONS

For most statistics, almost all the variation in bootstrap distributions comes from the selection of the original sample from the population. You can reduce this variation by using a larger original sample.

Bootstrapping does not overcome the weakness of small samples as a basis for inference. We will describe some bootstrap procedures that are usually more accurate than standard methods, but even they may not be accurate for very small samples. Use caution in any inference—including bootstrap inference—from a small sample.

The bootstrap resampling process using 1000 or more resamples introduces very little additional variation.

Bootstrapping a sample median

In dealing with the grade point averages in Example 16.5, we chose to bootstrap the 25% trimmed mean rather than the median. We did this in part because the usual bootstrapping procedure doesn't work well for the median unless the original sample is quite large. Now we will bootstrap the median in order to understand the difficulties.

Figure 16.16 follows the format of Figures 16.14 and 16.15. The population distribution appears at top left, with the population median M marked. Below in the left column are five samples of size $n = 15$ from this population, with their sample medians m marked. Bootstrap distributions of the median based on resampling from each of the five samples appear in the middle column. The right column again displays five more bootstrap distributions from resampling the first sample. The six bootstrap distributions from the same sample are, once again, very similar to each other—resampling adds little variation—so we concentrate on the middle column in the figure.
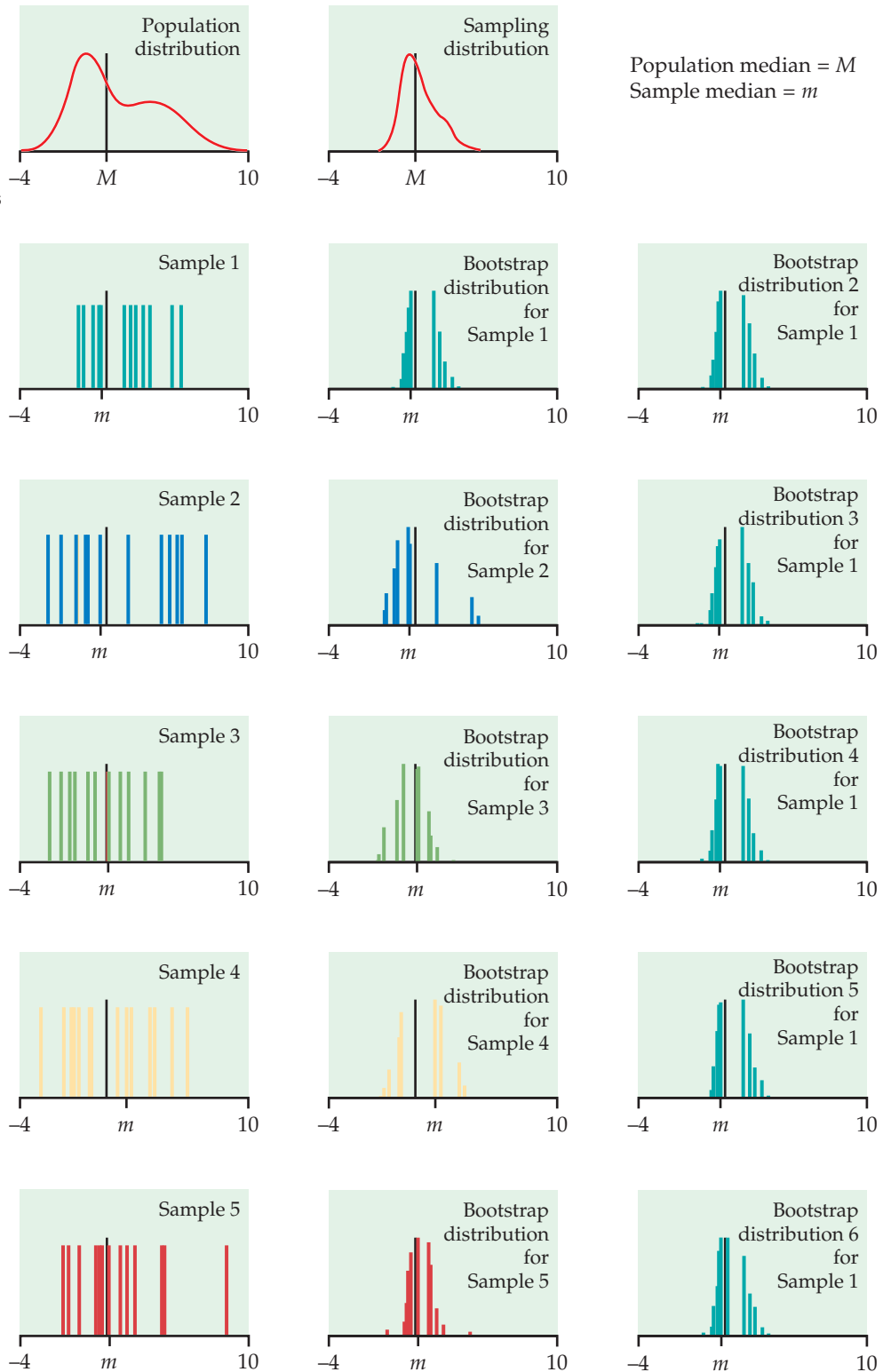
Bootstrap distributions from the five samples differ markedly from each other and from the sampling distribution at the top of the column. Here's why. The median of a resample of size 15 is the eighth-largest observation in the resample. This is always one of the 15 observations in the original sample and is usually one of the middle observations. Each bootstrap distribution repeats the same few values, and these values depend on the original sample. The sampling distribution, on the other hand, contains the medians of all possible samples and is not confined to a few values.

The difficulty is somewhat less when n is even because the median is then the average of two observations. It is much less for moderately large samples—say, $n = 100$ or more. Bootstrap standard errors and confidence intervals from such samples are reasonably accurate, though the shapes of the bootstrap distributions may still appear odd. You can see that the same difficulty will occur for small samples with other statistics, such as the quartiles, that are calculated from just one or two observations from a sample.

There are more advanced variations of the bootstrap idea that improve performance for small samples and for statistics such as the median and quartiles. *Unless you have expert advice or undertake further study, avoid bootstrapping the median and quartiles unless your sample is rather large.*

FIGURE 16.16 Five random samples of $n = 15$ from the same population, with a bootstrap distribution of the sample median formed by resampling from each of the five samples. At the right are five more bootstrap distributions from the first sample.

SECTION 16.3 SUMMARY

- Almost all the variation in a bootstrap distribution for a statistic is due to the selection of the original random sample from the population. Resampling introduces little additional variation.

- Bootstrap distributions based on small samples can be quite variable. Their shape and spread reflect the characteristics of the sample and may not accurately estimate the shape and spread of the sampling distribution. Bootstrap inference from a small sample may therefore be unreliable.

- Bootstrap inference based on samples of moderate size is unreliable for statistics like the median and quartiles that are calculated from just a few of the sample observations.

SECTION 16.3 EXERCISES

16.32 Variation in the bootstrap distributions.
Consider the variation in the bootstrap for each of the following situations with two scenarios, S1 and S2. In comparing the variation, do you expect, in general, that S1 will have less variation than S2, that S2 will have less variation than S1, or that the variation for S1 and S2 will be approximately the same? Give reasons for your answers. Here, we use n for the size of the original sample and B for the number of resamples.

(a) S1: $n = 50$, $B = 2000$; S2: $n = 50$, $B = 4000$.

(b) S1: $n = 10$, $B = 2000$; S2: $n = 50$, $B = 2000$.

(c) S1: $n = 50$, $B = 200$; S2: $n = 50$, $B = 2000$.

(d) S1: $n = 10$, $B = 2000$; S2: $n = 50$, $B = 4000$.

16.33 Bootstrap versus sampling distribution. Most statistical software includes a function to generate samples from Normal distributions. Set the mean to 26 and the standard deviation to 27. You can think of all the numbers that would be produced by this function if it ran forever as a population that has the $N(26,27)$ distribution. Samples produced by the function are samples from this population.

(a) What is the exact sampling distribution of the sample mean $\bar{x}$ for a sample of size n from this population?

(b) Draw an SRS of size $n = 10$ from this population. Bootstrap the sample mean $\bar{x}$ using 2000 resamples from your sample. Give a histogram of the bootstrap distribution and the bootstrap standard error.

(c) Repeat the same process for samples of sizes $n = 40$ and $n = 160$.

(d) Write a careful description comparing the three bootstrap distributions and also comparing them with the exact sampling distribution. What are the effects of increasing the sample size?

16.34 The effect of increasing the sample size. In Example 1.36, we discussed the distribution of the time to start a business for 189 countries. The distribution of times is very non-Normal. A histogram with a smooth density curve is given in Figure 1.20(a) (page 52). However, for this histogram we excluded one country, Suriname, where it takes 208 days to start a business. Exclude Suriname from the data set and use the remaining data for the remaining 188 countries. **📊 TTS**

(a) Let's think of the 188 countries as the population for this exercise. Find the mean μ and the standard deviation σ for this population.

(b) Although we don't know the shape of the sampling distribution of the sample mean $\bar{x}$ for a sample of size n from this population, we do know the mean and standard deviation of this distribution. What are they?

(c) Draw an SRS of size $n = 10$ from this population. Bootstrap the sample mean $\bar{x}$ using 2000 resamples from your sample. Give a histogram of the bootstrap distribution and the bootstrap standard error.

(d) Repeat the same process for samples of sizes $n = 40$ and $n = 160$.

(e) Write a careful description comparing the three bootstrap distributions. What are the effects of increasing the sample size?

16.35 The effect of non-Normality. The populations in the two previous exercises have the same mean and standard deviation, but one is Normal and the other is strongly non-Normal. Based on your work in these exercises, how does non-Normality of the population affect the bootstrap distribution of $\bar{x}$? How does it affect the bootstrap standard error? Do either of these effects diminish when we start with a larger sample? Explain what you have observed based on what you know about the sampling distribution of $\bar{x}$ and the way in which bootstrap distributions mimic the sampling distribution.

16.4 Bootstrap Confidence Intervals

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

- Use the bootstrap distribution to find a bootstrap percentile confidence interval.
- Read software output to find the BCa confidence interval.

Until now, we have met just one type of inference procedure based on resampling, the bootstrap t confidence intervals. We can calculate a bootstrap t confidence interval for any parameter by bootstrapping the corresponding statistic. We don't need conditions on the population or special knowledge about the sampling distribution of the statistic.

The flexible and almost automatic nature of bootstrap t intervals is appealing— but there is a catch. These intervals work well only when the bootstrap distribution tells us that the sampling distribution is approximately Normal and has small bias. How well must these conditions be met? What can we do if we don't trust the bootstrap t interval? In this section, we will see how to quickly check t confidence intervals for accuracy, and we will learn alternative bootstrap confidence intervals that can be used more generally than the bootstrap t.

Bootstrap percentile confidence intervals

Confidence intervals are based on the sampling distribution of a statistic. If a statistic has no bias as an estimator of a parameter, its sampling distribution is centered at the true value of the parameter. We can then get a 95% confidence interval by marking off the central 95% of the sampling distribution. The t critical values in a t confidence interval are a shortcut to marking off the central 95%.

This shortcut doesn't work under all conditions—it depends both on lack of bias and on Normality. One way to check whether t intervals (using either bootstrap or formula-based standard errors) are reasonable is to compare them with the central 95% of the bootstrap distribution. The 2.5 and 97.5 percentiles mark off the central 95%. The interval between the 2.5 and 97.5 percentiles of the bootstrap distribution is often used as a confidence interval in its own right. It is known as a *bootstrap percentile confidence interval*.

BOOTSTRAP PERCENTILE CONFIDENCE INTERVALS

The interval between the 2.5 and 97.5 percentiles of the bootstrap distribution of a statistic is a 95% **bootstrap percentile confidence interval** for the corresponding parameter. Use this method when the bootstrap estimate of bias is small.

The conditions for safe use of bootstrap t and bootstrap percentile intervals are a bit vague. We recommend that you check whether these intervals are reasonable by comparing them with each other. If the bias of the bootstrap distribution is small and the distribution is close to Normal, the bootstrap t and percentile confidence intervals will agree closely.

Percentile intervals, unlike t intervals, do not ignore skewness. Percentile intervals are therefore usually more accurate, as long as the bias is small. Because we will soon meet a much more accurate bootstrap interval, our recommendation is that *when bootstrap t and bootstrap percentile intervals do not agree closely, neither type of interval should be used.*

EXAMPLE 16.8

Bootstrap percentile confidence interval for the trimmed mean. In Example 16.5 (page 16-16), we found that a 95% bootstrap t confidence interval for the 25% trimmed mean of GPA for the population of college students after three semesters at this large university is between 2.794 and 3.106. The bootstrap distribution in Figure 16.9 shows a small bias and, though closely Normal, is a bit skewed. Is the bootstrap t confidence interval accurate for these data?

We can use the quantile function in R to compute the needed percentiles of our 3000 resamples. For this bootstrap distribution, the 2.5 and 97.5 percentiles are 2.793 and 3.095, respectively. These are the endpoints of the 95% bootstrap percentile confidence interval. This interval is quite close to the bootstrap t interval. We conclude that both intervals are reasonably accurate.

The bootstrap t interval for the trimmed mean of GPA in Example 16.8 is

$$\bar{x}_{25\%} \pm t^*\mathrm{SE}_{\mathrm{boot}} = 2.950 \pm 0.156$$

We can learn something by also writing the percentile interval starting at the statistic $\bar{x}_{25\%} = 2.950$. In this form, it is

$$2.950 - 0.157, \quad 2.950 + 0.145$$

Unlike the t interval, the percentile interval is not symmetric—its endpoints are different distances from the statistic. The slightly greater distance to the 2.5 percentile reflects the slight left-skewness of the bootstrap distribution.

USE YOUR KNOWLEDGE

FACE4

16.36 Determining the percentile endpoints. What percentiles of the bootstrap distribution are the endpoints of a 99% bootstrap percentile confidence interval? How do they change for a 90% bootstrap percentile confidence interval?

16.37 Bootstrap percentile confidence interval for profile viewing time. Consider the sample of Facebook viewing times in Exercise 16.1 (page 16-7). Bootstrap the sample mean using 2000 resamples.

(a) Make a histogram and a Normal quantile plot. Does the bootstrap distribution appear close to Normal? Is the bias small relative to the observed sample mean?

(b) Find the 95% bootstrap t confidence interval.

(c) Give the 95% confidence percentile interval and compare it with the interval in part (b).

A more accurate bootstrap confidence interval: BCa

Any method for obtaining confidence intervals requires some conditions in order to produce exactly the intended confidence level. These conditions (for example, Normality) are never exactly met in practice. So a 95% confidence

interval in practice will not capture the true parameter value exactly 95% of the time.

In addition to "hitting" the parameter 95% of the time, a good confidence interval should divide its 5% of "misses" equally between high misses and low misses. We will say that a method for obtaining 95% confidence intervals is **accurate** in a particular setting if 95% of the time it produces intervals that capture the parameter and if the 5% of misses are equally shared between high and low misses. Perfect accuracy isn't available in practice, but some methods are more accurate than others.

accurate

One advantage of the bootstrap is that we can to some extent check the accuracy of the bootstrap t and percentile confidence intervals by examining the bootstrap distribution for bias and skewness and by comparing the two intervals with each other. The interval in Example 16.8 reveals a slight left-skewness, but not enough to invalidate inference.

In general, the t *and percentile intervals may not be sufficiently accurate when*

• The statistic is strongly biased, as indicated by the bootstrap estimate of bias.

• The sampling distribution of the statistic is clearly skewed, as indicated by the bootstrap distribution and by comparing the t and percentile intervals.

Most confidence interval procedures are more accurate for larger sample sizes. The t and percentile procedures improve only slowly: they require 100 times more data to improve accuracy by a factor of 10. (Recall the $\sqrt{n}$ in the formula for the usual one-sample t interval.) These intervals may not be very accurate except for quite large sample sizes. There are more elaborate bootstrap procedures that improve faster, requiring only 10 times more data to improve accuracy by a factor of 10. These procedures are quite accurate unless the sample size is very small.

> ### BCa CONFIDENCE INTERVALS
>
> The **bootstrap bias-corrected accelerated (BCa) interval** is a modification of the percentile method that adjusts the percentiles to correct for bias and skewness. This interval should always be used over the percentile and t intervals if software provides it.

This method is accurate in a wide variety of settings, has reasonable computation requirements (by modern standards), and does not produce excessively wide intervals. The BCa intervals are among the most widely used intervals. Because this interval is related to the percentile method, it is still based on the key ideas of resampling and the bootstrap distribution.

You should always use this more accurate method (or an alternative like tilting intervals) if your software offers it. The details of producing confidence intervals are quite technical.[6] The BCa method requires more than 1000 resamples for high accuracy. We recommend that you use 5000 or more resamples. *Don't forget that even BCa confidence intervals should be used cautiously when sample sizes are small because there are not enough data to accurately determine the necessary corrections for bias and skewness.*

EXAMPLE 16.9

GPA

The BCa confidence interval for the ratio of variances. In Example 16.6 (page 16-17), we compared the GPA means of men and women using a 95% bootstrap t confidence interval. Because 0 was contained in the interval, we concluded that there was not enough evidence to state that the two means were different. Suppose we also want to compare the variances. Figure 16.10 (page 16-17) suggests that the spread among the male GPAs is larger than that of the females. The ratio of the male sample variance to the female sample variance is 1.321. Can we conclude there is a difference?

In Section 12.1 (page 665), we discussed the modified Levene's test for equality of spread. Let's instead use the bootstrap. Specifically, we'll form a 95% confidence interval for σ_1^2/σ_2^2.

Figure 16.17 shows the bootstrap distribution of the ratio of sample variances s_1^2/s_2^2. We see strong skewness in the bootstrap distribution and, therefore, in the sampling distribution. This is not unexpected. Recall that if the data are Normal and the variances are equal, we'd expect this ratio to follow an F distribution.

The bootstrap t and percentile intervals aren't reliable when the sampling distribution of the statistic is skewed. Figure 16.18 shows software output that includes the percentile and BCa intervals. The bootstrap t interval is closely related to the Normal interval that is also supplied. The basic confidence interval is another method based on the percentiles of the bootstrap distribution that we will not discuss here.

The BCa interval is

$$(1.321 - 0.456, 1.321 + 0.914) = (0.865, 2.235)$$

and the percentile interval is

$$(1.321 - 0.468, 1.321 + 0.880) = (0.853, 2.201)$$

In this case, the percentile and BCa intervals are similar, but the BCa is shifted slightly as it has adjusted for the bias, which was estimated at 0.054. Both intervals are strongly asymmetrical: the upper endpoint is about twice as far from the sample ratio as the lower endpoint. This reflects the strong right-skewness of the bootstrap distribution.
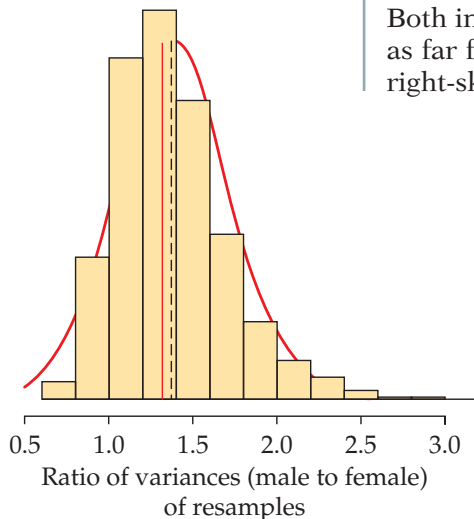

FIGURE 16.17 The bootstrap distribution of the ratio of sample variances of 5000 resamples from the data in Example 16.6. The bootstrap distribution is right-skewed, so we conclude that the sampling distribution of the ratio of sample variances is right-skewed, Example 16.9.
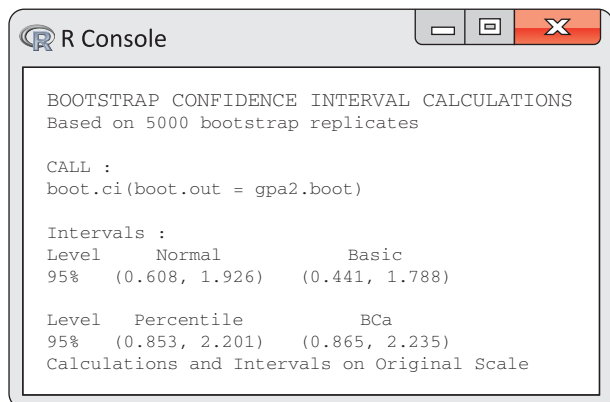


```
R Console

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 5000 bootstrap replicates

CALL :
boot.ci(boot.out = gpa2.boot)

Intervals :
Level      Normal              Basic
95%   (0.608, 1.926)     (0.441, 1.788)

Level    Percentile           BCa
95%    (0.853, 2.201)     (0.865, 2.235)
Calculations and Intervals on Original Scale
```

FIGURE 16.18 R output for bootstrapping the ratio of variances for the GPA data, Example 16.9.

The output in Figure 16.18 also shows that both endpoints of the less-accurate intervals (bootstrap t via the Normal interval and the percentile interval) are too low. These intervals miss the population ratio on the low side too often (more than 2.5% of the time) and miss on the high side too seldom. They give a biased picture of where the true ratio is likely to be.

Confidence intervals for the correlation

The bootstrap allows us to find confidence intervals for a wide variety of statistics. So far, we have looked at the sample mean, trimmed mean, the difference between two means, and the ratio of sample variances using a variety of different bootstrap confidence intervals. The choice of interval depended on the shape of the bootstrap distribution and the desired accuracy.

Now we will bootstrap the correlation coefficient. This is our first use of the bootstrap for a statistic that depends on two related variables. As with the difference between two means, we must pay attention to how we should resample.

EXAMPLE 16.10

LAUND24

Correlation between price and rating. Consumers Union provides ratings on a large variety of consumer products. They use sophisticated testing methods as well as surveys of their members to create these ratings. The ratings are published in their magazine, *Consumer Reports*.

An article in *Consumer Reports* rated laundry detergents on a scale from 1 to 100. Here are the ratings along with the price per load, in cents, for 24 laundry detergents:

Rating	Price (cents)	Rating	Price (cents)	Rating	Price (cents)	Rating	Price (cents)
61	17	59	22	56	22	55	16
55	30	52	23	51	11	50	15
50	9	48	16	48	15	48	18
46	13	46	13	45	17	36	8
35	8	34	12	33	7	32	6
32	5	29	14	26	11	26	13

In Example 2.8 (page 85), we examined the relationship between rating and price per load for 53 laundry detergents. Based on that study, we expect that the higher-priced detergents will tend to have higher ratings. The scatterplot in Figure 16.19 shows that the higher-priced products do tend to have better ratings, but the relationship is not particularly strong. The correlation is 0.671. Let's use the bootstrap to find a 95% confidence interval for the population correlation.

Our confidence interval will also provide a test of the null hypothesis that the population correlation is zero. If the 95% confidence interval does not include zero, we can reject the null hypothesis in favor of the two-sided alternative. Although we would expect the correlation to be positive, we could be surprised and
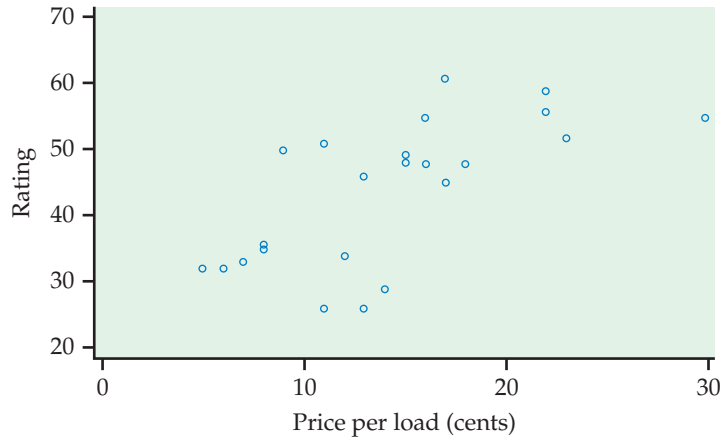
FIGURE 16.19 Scatterplot of price per load (in cents) versus rating for 24 laundry detergents, Example 16.10.

find that it is negative. It is important to keep in mind that *we cannot use what we learned by looking at the scatterplot to formulate our alternative hypothesis.*

How shall we resample from the laundry detergent data? Because each observation consists of the price and the rating for one product, we resample products. Resampling prices and ratings separately would lose the connection between a product's price and its rating. Software such as R automates proper resampling. Once we have produced a bootstrap distribution by resampling, we can examine the distribution and construct a confidence interval in the usual way. We need no special formulas or procedures to handle the correlation.

Figure 16.20 shows the bootstrap distribution and Normal quantile plot for the sample correlation for 5000 resamples from the 24 laundry detergents in our sample. The bootstrap distribution is skewed to the left with relatively small bias. We'll need to check whether a 95% bootstrap t confidence interval is reasonable here.

FIGURE 16.20. The bootstrap distribution and Normal quantile plot for the correlation r for 5000 resamples from the laundry detergent data set.

The bootstrap standard error is $SE_{boot} = 0.086$. The t interval using the bootstrap standard error is

$$r \pm t^*SE_{boot} = 0.671 \pm (2.074)(0.086)$$
$$= 0.671 \pm 0.178$$
$$= (0.493, 0.849)$$

The 95% bootstrap percentile interval is

$$(2.5 \text{ percentile}, 97.5 \text{ percentile}) = (0.485, 0.827)$$
$$= (0.671 - 0.186, 0.671 + 0.156)$$

The two confidence intervals are not too different. If you feel this discrepancy is acceptable, you might want to use the percentile interval to account for the skewness in the bootstrap distribution.

While the confidence intervals give a wide range for the population correlation, both of them include only positive values. Thus, these data provide significant evidence that there is a positive relationship between a laundry detergent's rating and its price per load.

SECTION 16.4 SUMMARY

- Both bootstrap t and (when they exist) traditional z and t confidence intervals require statistics with small bias and sampling distributions close to Normal. We can check these conditions by examining the bootstrap distribution for bias and lack of Normality.

- The **bootstrap percentile confidence interval** for 95% confidence is the interval from the 2.5 percentile to the 97.5 percentile of the bootstrap distribution. Agreement between the bootstrap t and percentile intervals is an added check on the conditions needed by the t interval. Do not use t or percentile intervals if these conditions are not met.

- When bias or skewness is present in the bootstrap distribution, use a **BCa interval**. The t and percentile intervals are inaccurate under these circumstances unless the sample sizes are very large. The BCa confidence intervals adjust for bias and skewness and are generally accurate except for small samples.

SECTION 16.4 EXERCISES

For Exercises 16.36 and 16.37, see page 16-32.

16.38 Find the 95% bootstrap percentile confidence interval. The mean of a sample is $\bar{x} = 226.0$ and the standard deviation is $s = 75.6$. The mean of the bootstrap distribution is $\bar{x} = 226.2$ and the standard deviation is $s = 11.8$. A bootstrap distribution has the following percentiles:

				Percentile				
0.01	0.025	0.05	0.10	0.50	0.90	0.95	0.975	0.99
199.07	203.55	207.38	210.97	225.90	241.63	245.93	248.68	252.28

Find the 95% bootstrap percentile confidence interval.

16.39 Summarize the output. Figures 16.21 and 16.22 show software output from R with information about a bootstrap analysis. Summarize the information in the output. Be sure to include the BCa confidence interval.

16.40 Confidence interval for the average IQ score. The distribution of the 60 IQ test scores in Table 1.1 (page 14) is roughly Normal, and the sample size is large enough that we expect a Normal sampling distribution. We will compare confidence intervals for the population mean IQ μ based on this sample. IQ

(a) Use the formula $s/\sqrt{n}$ to find the standard error of the mean. Give the 95% t confidence interval based on this standard error.
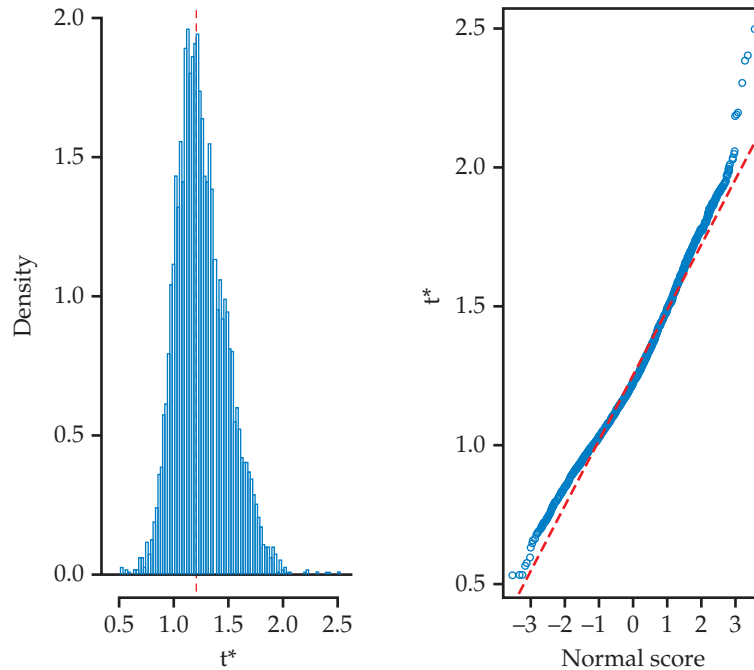
FIGURE 16.21 R graphical output, Exercise 16.39.

FIGURE 16.22 Output from R with bootstrap confidence intervals, Exercise 16.39.

(b) Bootstrap the mean of the IQ scores. Make a histogram and a Normal quantile plot of the bootstrap distribution. Does the bootstrap distribution appear Normal? What is the bootstrap standard error? Give the 95% bootstrap t confidence interval.

(c) Give the 95% confidence percentile and BCa intervals. Make a graphical comparison by drawing a vertical line at the original sample mean $\bar{x}$ and displaying the three intervals vertically, one above the other. How well do your four confidence intervals agree? Was bootstrapping needed to find a reasonable confidence interval, or was the formula-based confidence interval good enough?

16.41 Confidence interval for a Normal data set. In Exercise 16.29 (page 16-23), you bootstrapped the mean of a simulated SRS from the standard Normal distribution $N(0, 1)$ and found the 95% standard t and bootstrap t confidence intervals for the mean. **NORMALD**

(a) Find the 95% bootstrap percentile confidence interval. Does this interval confirm that the t intervals are acceptable?

(b) We know that the population mean is 0. Do the confidence intervals capture this mean?

16.42 Using bootstrapping to check traditional methods. Bootstrapping is a good way to check if traditional inference methods are accurate for a given sample. Consider the following data: **DATA30**

98	107	113	104	94	100	107	98	112	97
99	95	97	90	109	102	89	101	93	95
95	87	91	101	119	116	91	95	95	104

(a) Examine the data graphically. Do they appear to violate any of the conditions needed to use the one-sample t confidence interval for the population mean?

(b) Calculate the 95% one-sample t confidence interval for this sample.

(c) Bootstrap the data, and inspect the bootstrap distribution of the mean. Does it suggest that a t interval should be reasonably accurate? Calculate the bootstrap t 95% interval.

(d) Find the 95% bootstrap percentile interval. Does it agree with the two t intervals? What do you conclude about the accuracy of the one-sample t interval here?

16.43 Comparing bootstrap confidence intervals. The graphs in Figure 16.9 (page 16-15) do not appear to show any important skewness in the bootstrap distribution of the trimmed mean for Example 16.4. Compare the bootstrap percentile and bootstrap t intervals for the trimmed mean, given in the discussion of Example 16.8 (page 16-32). Does the comparison suggest any skewness? **GPA**

16.44 More on using bootstrapping to check traditional methods. Continue to work with the data given in Exercise 16.42. **DATA30**

(a) Find the 95% BCa confidence interval.

(b) Does your opinion of the robustness of the one-sample t confidence interval change when comparing it with the BCa interval?

(c) To check the accuracy of the one-sample t confidence interval, would you generally use the bootstrap percentile or the BCa interval? Explain.

16.45 BCa interval for the correlation coefficient. Find the 95% BCa confidence interval for the correlation between price and rating, from the data in Example 16.10 (page 16-35). Is this more accurate interval in general agreement with the 95% bootstrap t and percentile intervals? Do you still agree with the judgment in the discussion of Example 16.10 that the simpler intervals are adequate? **LAUND24**

16.46 Bootstrap confidence intervals for the average audio file length. In Exercise 16.17 (page 16-16), you found a bootstrap t confidence interval for the population mean μ. Careful examination of the bootstrap distribution reveals a slight skewness in the right tail. Is this something to be concerned about? Bootstrap the mean and give all three 95% bootstrap confidence intervals: t, percentile, and BCa. Make a graphical comparison by displaying the three intervals vertically, one above the other. Discuss what you see. **SONGS**

16.47 Bootstrap confidence intervals for service center call lengths. The distribution of the call center lengths that you used in Exercise 16.25 (page 16-22) is strongly skewed.

In that exercise, you found a bootstrap t confidence interval for the population mean μ, even though some skewness remains in the bootstrap distribution. Bootstrap the mean length and give all three bootstrap 95% confidence intervals: t, percentile, and BCa. Make a graphical comparison by drawing a vertical line at the original sample mean $\bar{x}$ and displaying the three intervals horizontally, one above the other. Discuss what you see: Do bootstrap t and percentile agree? Does the more accurate interval agree with the two simpler methods? **CALLS80**

16.48 Bootstrap confidence intervals for the standard deviation. We would like a 95% confidence interval for the standard deviation σ of 150 GPAs. In Exercise 16.27 (page 16-23), we considered the bootstrap t interval. Now we have a more accurate method. Bootstrap s and report all three 95% bootstrap confidence intervals: t, percentile, and BCa. Make a graphical comparison by drawing a vertical line at the original s and displaying the three intervals vertically, one above the other. Discuss what you see: Do bootstrap t and percentile agree? Does the more accurate interval agree with the two simpler methods? What interval would you use in a report on GPAs at this college? **GPA**

16.49 The effect of decreasing the sample size. Exercise 16.15 (page 16-12) gives an SRS of 10 of the service center call lengths from Table 1.2. Describe the bootstrap distribution of $\bar{x}$ from this sample. Give a 95% confidence interval for the population mean μ based on these data and a method of your choice. Describe carefully how your result differs from the intervals in Exercise 16.47, which use the larger sample of 80 call lengths. **CALLS10**

16.50 Bootstrap confidence interval for the GPA data. The GPA data for females from Example 16.6 (page 16-17) are strongly skewed to the left and have a cluster of observations at 4. **GPA**

(a) Bootstrap the mean of the data. Based on the bootstrap distribution, which bootstrap confidence intervals would you consider for use? Explain your answer.

(b) Find all three bootstrap confidence intervals. How do the intervals compare? Briefly explain the reasons for any differences. In particular, what kind of errors would you make in estimating the mean GPA by using a t interval or a percentile interval instead of a BCa interval?

16.51 Bootstrap confidence intervals for the difference in GPAs. Example 16.6 (page 16-17) considers the difference in mean GPAs of men and women. The bootstrap distribution appeared reasonably Normal. Give the 95% BCa confidence interval for the difference in mean GPAs. Is this interval comparable to the bootstrap t interval calculated in the example? **GPA**

16.52 The correlation between GPA and high school math grades. The study described in Example 16.4 (page 16-13) used high school grades to predict GPA. For

this exercise, we will look at the correlation between GPA and high school math grades. 📊 GPA

(a) Describe the distribution of GPAs. Do the same for high school math grades.

(b) Describe the relationship between GPA and high school math grades.

(c) Generate 2000 resamples and use these to obtain the bootstrap distribution for the correlation.

(d) Describe the shape and bias of the bootstrap distribution. Does use of the simpler bootstrap confidence intervals (t and percentile) appear to be justified?

(e) Find all three 95% bootstrap confidence intervals: t, percentile, and BCa. Make a graphical comparison by drawing a vertical line at the original correlation r and displaying the three intervals vertically, one above the other. Discuss what you see. Does it still appear that the simpler intervals are justified? What confidence interval would you include in a report describing the relationship between GPA and high school math grades?

16.53 The correlation between education spending and population. Figure 2.5 (page 90) shows a strong positive relationship between education spending and population in 2015 for 50 states. Use the bootstrap to perform statistical inference for these data. 📊 EDSPEND

(a) Describe the shape and bias of the bootstrap distribution. Do you think that a simple bootstrap inference (t and percentile confidence intervals) is justified? Explain your answer.

(b) Give the 95% BCa and bootstrap percentile confidence intervals for the population correlation. Do they (as expected) agree closely? Do these intervals provide significant evidence at the 5% level that the population correlation is not 0?

16.54 Bootstrap distribution for the slope β_1. Describe carefully how to resample from data on an explanatory variable x and a response variable y to create a bootstrap distribution for the slope b_1 of the least-squares regression line.

16.55 Predicting ratings of laundry detergents. Refer to Example 16.10 (page 16-35). 📊 LAUND24

(a) Find the least-squares regression line for predicting rating from price.

(b) Bootstrap the regression line and give a 95% confidence interval for the slope of the population regression line.

(c) Compare the bootstrap results with the usual method for finding a confidence interval for a regression slope.

16.56 Predicting GPA. Continue your study of GPA and high school math grades, begun in Exercise 16.52,

by performing a regression to predict GPA using high school math grades as the explanatory variable. 📊 GPA

(a) Plot the residuals against the math grades and make a Normal quantile plot of the residuals. Do these plots suggest that inference based on the usual simple linear regression model may be inaccurate? Give reasons for your answer.

(b) Examine the bootstrap distribution of the slope b_1 of the least-squares regression line. Based on what you see, what do you recommend regarding the use of bootstrap t or bootstrap percentile intervals? Give reasons for your recommendation.

(c) Give the 95% BCa confidence interval for the slope β_1 of the population regression line. Compare this with the standard 95% confidence interval based on Normality, the bootstrap t interval, and the bootstrap percentile interval. Using the BCa interval as a standard, which of the other intervals are adequately accurate for practical use?

16.57 Predicting education spending. Continue your study of the relationship between education spending and population, begun in Exercise 16.53. Run the regression to predict education spending using population as the explanatory variable. 📊 EDSPEND

(a) Plot the residuals against the explanatory variable and make a Normal quantile plot of the residuals. Do the residuals appear to be Normal? Explain your answer.

(b) Examine the shape and bias of the bootstrap distribution of the slope b_1 of the least-squares line. Does this distribution suggest that even the bootstrap t interval will be accurate? Give a reason for your answer.

(c) Find the standard 95% t confidence interval for β_1 and also the BCa, bootstrap t, and bootstrap percentile confidence intervals. What do you conclude about the accuracy of the two t intervals?

16.58 The effect of outliers. We know that outliers can strongly influence statistics such as the mean and the least-squares line. A study of dementia patients in nursing homes recorded various types of disruptive behaviors every day for 12 weeks. Days were classified as moon days if they were in a three-day period centered at the day of a full moon. A matched pairs analysis was performed to see if the average number of disruptive behaviors was different on moon days. There were three patients with very small differences that may be considered outliers. 📊 MOON

(a) Bootstrap the mean of the differences with and without the three low values. How do these values influence the shape and bias of the bootstrap distribution?

(b) Give the BCa confidence interval from both bootstrap distributions. Discuss the differences.

16.5 Significance Testing Using Permutation Tests

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

- Describe the steps of a permutation test for comparing two populations.
- Outline the steps needed for a permutation test for a matched pairs study.
- Using software, perform a permutation test for comparing two populations.
- Describe the randomization needed for a permutation test for the relationship between two quantitative variables.

LOOK BACK

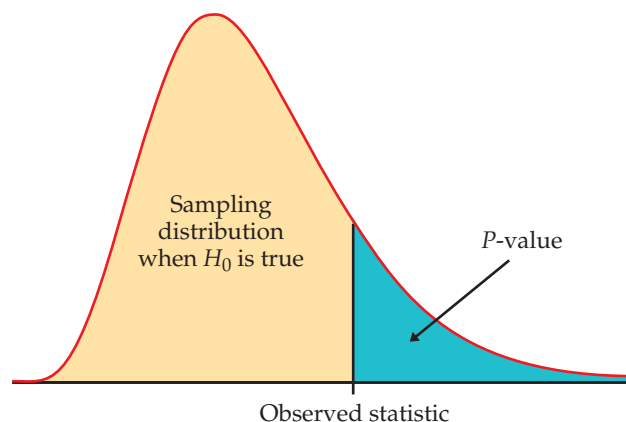tests of significance, p. 361

Significance tests tell us whether an observed effect, such as a difference between two means or a correlation between two variables, could reasonably occur "just by chance" in selecting a random sample. If not, we have evidence that the effect observed in the sample reflects an effect that is present in the population. The reasoning of tests goes like this:

1. Choose a statistic that measures the effect you are looking for.

2. Construct the sampling distribution that this statistic would have if the effect were *not* present in the population.

3. Locate the observed statistic on this distribution. A value in the main body of the distribution could easily occur just by chance. A value in the tail would rarely occur by chance and so is evidence that something other than chance is operating.

LOOK BACK

null hypothesis, p. 363

P-value, p. 366

The statement that the effect we seek is *not* present in the population is the null hypothesis, H_0. Assuming the null hypothesis is true, the probability that we would observe a statistic value as extreme or more extreme than the one we did observe is the *P*-value. Figure 16.23 illustrates the idea of a *P*-value. Small *P*-values are evidence against the null hypothesis and in favor of a real effect in the population. Tests based on resampling don't change this approach. They find *P*-values by resampling calculations rather than from formulas and so can be used in settings where traditional tests don't apply.

FIGURE 16.23 The *P*-value of a statistical test is found from the sampling distribution the statistic would have if the null hypothesis were true. It is the probability of a result at least as extreme as the value we actually observed.

Sampling distribution when H_0 is true

P-value

Observed statistic

Because *P*-values are calculated *acting as if the null hypothesis were true* we cannot resample from the observed sample as we did earlier. In the absence of bias, resampling from the original sample creates a bootstrap distribution centered at the observed value of the statistic. If the null hypothesis is, in fact, not true, this value may be far from the parameter value stated by the null hypothesis. We must estimate what the sampling distribution of the statistic would be if the null hypothesis were true. That is, we must obey the following rule:

RESAMPLING FOR SIGNIFICANCE TESTS

To estimate the *P*-value for a test of significance, estimate the sampling distribution of the test statistic when the null hypothesis is true by resampling in a manner that is consistent with the null hypothesis.

EXAMPLE 16.11

DRP

permutation test

"Directed reading activities." Do new "directed reading activities" improve the reading ability of elementary school students, as measured by their Degree of Reading Power (DRP) scores? A study assigns students at random to either the new method (treatment group, 21 students) or traditional teaching methods (control group, 23 students). The DRP scores at the end of the study appear in Table 16.1.[7] In Example 7.13 (page 440), we applied the two-sample *t* test to these data.

To apply resampling, we will start with the difference between the sample means as a measure of the effect of the new activities:

$$\text{statistic} = \bar{x}_{\text{treatment}} - \bar{x}_{\text{control}}$$

The null hypothesis H_0 for the resampling test is that the teaching method has no effect on the distribution of DRP scores. If H_0 is true, the DRP scores in Table 16.1 do not depend on the teaching method. Each student has a DRP score that describes that child and is the same no matter which group the child is assigned to. The observed difference in group means just reflects the accident of random assignment to the two groups.

Now we can see how to resample in a way that is consistent with the null hypothesis: imitate many repetitions of the random assignment of students to treatment and control groups, with each student always keeping his or her DRP score unchanged. Because resampling in this way scrambles the assignment of students to groups, tests based on resampling are called **permutation tests**, from the mathematical name for scrambling a collection of things.

TABLE 16.1	Degree of Reading Power Scores for Third-Graders						
Treatment group				Control group			
24	61	59	46	42	33	46	37
43	44	52	43	43	41	10	42
58	67	62	57	55	19	17	55
71	49	54		26	54	60	28
43	53	57		62	20	53	48
49	56	33		37	85	42	

24, 61 | 42, 33, 46, 37
$\bar{x}_1 - \bar{x}_2 = 42.5 - 39.5 = 3.0$

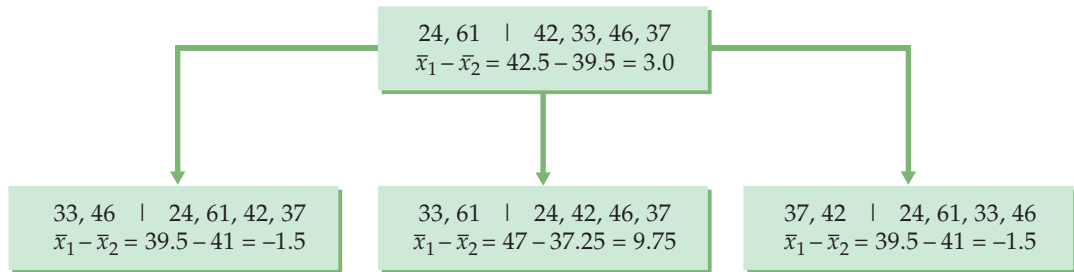| 33, 46 | 24, 61, 42, 37	33, 61 | 24, 42, 46, 37	37, 42 | 24, 61, 33, 46
$\bar{x}_1 - \bar{x}_2 = 39.5 - 41 = -1.5$	$\bar{x}_1 - \bar{x}_2 = 47 - 37.25 = 9.75$	$\bar{x}_1 - \bar{x}_2 = 39.5 - 41 = -1.5$

FIGURE 16.24 The idea of permutation resampling. The top box shows the outcome of a study with four subjects in one group and two in the other. The boxes below show three permutation resamples. The values of the statistic for many such resamples form the permutation distribution.

Here is an outline of the permutation test procedure for comparing the mean DRP scores in Example 16.11:

• Choose 21 of the 44 students at random to be the treatment group; the other 23 are the control group. This is an ordinary SRS, chosen *without replacement*. It is called a **permutation resample**.

permutation resample

• Calculate the mean DRP score in each group, using the students' DRP scores in Table 16.1. The difference between these means is our statistic.

• Repeat this resampling and calculation of the statistic hundreds of times. The distribution of the statistic from these resamples estimates the sampling distribution under the condition that H_0 is true. It is called a **permutation distribution**.

permutation distribution

• Consider the value of the statistic actually observed in the study,

$$\bar{x}_{\text{treatment}} - \bar{x}_{\text{control}} = 51.476 - 41.522 = 9.954$$

Locate this value on the permutation distribution to get the *P*-value.

Figure 16.24 illustrates permutation resampling on a small scale. The top box shows the results of a study with four subjects in the treatment group and two subjects in the control group. A permutation resample chooses an SRS of four of the six subjects to form the treatment group. The remaining two are the control group. The results of three permutation resamples appear below the original results, along with the statistic (difference in group means) for each.
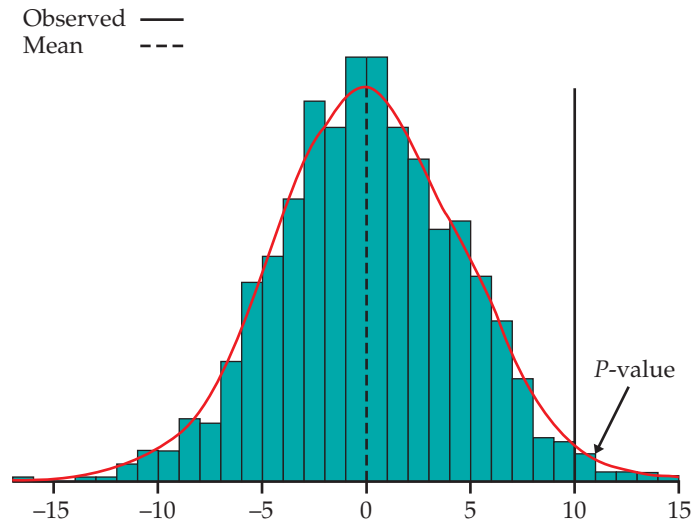
EXAMPLE 16.12

lu

DRP

Permutation test for the DRP study. Figure 16.25 shows the permutation distribution of the difference in means based on 1000 permutation resamples from the DRP data in Table 16.1. This is a resampling estimate of the sampling distribution of the statistic when the null hypothesis H_0 is true. As H_0 suggests, the distribution is centered at 0 (no effect). The solid vertical line in the figure marks the location of the statistic for the original sample, 9.954. Use the permutation distribution exactly as if it were the sampling distribution: the *P*-value is the probability that the statistic takes a value at least as extreme as 9.954 in the direction given by the alternative hypothesis.

We seek evidence that the treatment increases DRP scores, so the alternative hypothesis is that the distribution of the statistic $\bar{x}_{\text{treatment}} - \bar{x}_{\text{control}}$

FIGURE 16.25 The permutation distribution of the difference between the treatment mean and the control mean based on the DRP scores of 44 students, Example 16.12. The dashed line marks the mean of the permutation distribution: it is very close to zero, the value specified by the null hypothesis. The solid vertical line marks the observed difference in means, 9.954. Its location in the right tail shows that a value this large is unlikely to occur when the null hypothesis is true.

is centered not at 0 but at some positive value. Large values of the statistic are evidence against the null hypothesis in favor of this one-sided alternative. The permutation test P-value is the proportion of the 1000 resamples that give a result at least as great as 9.954. A look at the resampling results finds that 14 of the 1000 resamples gave a value of 9.954 or larger, so the estimated P-value is 14/1000, or 0.014.

Figure 16.25 shows that the permutation distribution has a roughly Normal shape. Because the permutation distribution approximates the sampling distribution, we now know that the sampling distribution is close to Normal. When the sampling distribution is close to Normal, we can safely apply the usual two-sample t test. The JMP output in Figure 7.13 (page 441) gives $P = 0.013$, very close to the P-value from the permutation test.

Using software

In principle, you can program almost any statistical software to do a permutation test. It is more convenient to use software that automates the process of resampling, calculating the statistic, forming the permutation distribution, and finding the P-value. The package perm in R contains functions that allow you to request permutation tests. The permutation distribution in Figure 16.25 is one output. Another is this summary of the test results:

Exact Permutation Test Estimated by Monte Carlo

data: trtgrp and ctrlgrp
p-value = 0.0154
alternative hypothesis: true mean trtgrp – mean ctrlgrp is greater than 0 sample estimates:
mean trtgrp – mean ctrlgrp
              9.954451

p-value estimated from 5000 Monte Carlo replications
99 percent confidence interval on p-value:
  0.01110640 0.02024333

By giving "greater" as the alternative hypothesis, the output makes it clear that 0.015 is the one-sided *P*-value. This estimate of the *P*-value is more precise than the 0.014 estimate because it is based on 5000 rather than 1000 resamples.

Permutation tests in practice

Permutation tests versus *t* tests We have analyzed the data in Table 16.1 both by the two-sample *t* test (in Chapter 7) and by a permutation test. Comparing the two approaches brings out some general points about permutation tests versus traditional formula-based tests.

- The hypotheses for the *t* test are stated in terms of the two population means,

$$H_0: \mu_{\text{treatment}} - \mu_{\text{control}} = 0$$
$$H_a: \mu_{\text{treatment}} - \mu_{\text{control}} > 0$$

The permutation test hypotheses are more general. The null hypothesis is "same distribution of scores in both groups," and the one-sided alternative is "scores in the treatment group are systematically higher." These more general hypotheses imply the *t* hypotheses if we are interested in mean scores and the two distributions have the same shape.

- The plug-in principle says that the difference in sample means estimates the difference in population means. The *t* statistic starts with this difference. We used the same statistic in the permutation test, but that was a choice: we could use the difference in 25% trimmed means or any other statistic that measures the effect of treatment versus control.

- The *t* test statistic is based on standardizing the difference in means in a clever way to get a statistic that has a *t* distribution when H_0 is true. The permutation test works directly with the difference in means (or some other statistic) and estimates the sampling distribution by resampling. No formulas are needed.

- The *t* test gives accurate *P*-values if the sampling distribution of the difference in means is at least roughly Normal. The permutation test gives accurate *P*-values even when the sampling distribution is not close to Normal.

The permutation test is useful even if we plan to use the two-sample *t* test. Rather than relying on Normal quantile plots of the two samples and the central limit theorem, we can directly check the Normality of the sampling distribution by looking at the permutation distribution. Permutation tests provide a "gold standard" for assessing two-sample *t* tests. If the two *P*-values differ considerably, it usually indicates that the conditions for the two-sample *t* don't hold for these data. Because permutation tests give accurate *P*-values even when the sampling distribution is skewed, they are often used when accuracy is very important. Here is an example.

EXAMPLE 16.13

GPA

Permutation test for GPAs. In Example 16.6 (page 16-17), we looked at the difference in mean GPAs of male and female students. Figure 16.10 (page 16-17) shows both distributions. Because the distributions are skewed and the sample sizes are somewhat different, a two-sample *t* test might be inaccurate.

Based on the summary statistics,

Sex	n	$\bar{x}$	s
Male	91	2.784	0.859
Female	59	2.933	0.748
Difference		-0.149	

the t statistic is -1.12 with either 58 or 135.73 degrees of freedom. The P-value is roughly 0.26 in either case.

We perform permutation tests with 5000 resamples using R. We use the difference in means, $\bar{x}_1 - \bar{x}_2$, as our test statistic. This is done by randomly regrouping the total set of GPAs into two groups that are the same sizes as the two original samples. This is consistent with the null hypothesis that sex has no effect on GPA. Each GPA appears once in the data of each resample, but some GPAs move from the male to the female group, and vice versa. We calculate the test statistic for each resample and create its permutation distribution. The P-value is the proportion of the resamples with statistics that exceed the observed statistic.

A 99% confidence interval for the P-value based on the 5000 resamples is (0.256, 0.309). This interval contains the P-value for the t test. The skewness and differing sample sizes do not have an impact here primarily because the sample sizes are relatively large.

If you read Chapter 15 on nonparametric tests, you will find more comparisons of permutation tests with rank tests as well as tests based on Normal distributions.

Data from an entire population A subtle difference between confidence intervals and significance tests is that confidence intervals require the distinction between sample and population, but tests do not. If we have data on an entire population—say, all employees of a large corporation—we don't need a confidence interval to estimate the difference between the mean salaries of male and female employees. We can calculate the means for all men and for all women and get an exact answer. But it still makes sense to ask, "Is the difference in means so large that it would rarely occur just by chance?" A test and its P-value answer that question.

Permutation tests are a convenient way to answer such questions. In carrying out the test, we pay no attention to whether the data are a sample or an entire population. The resampling assigns the full set of observed salaries at random to men and women and builds a permutation distribution from repeated random assignments. We can then see if the observed difference in mean salaries is so large that it would rarely occur if sex did not matter.

LOOK BACK

two-sample
t test,
p. 440

When are permutation tests valid? The two-sample t test starts from the condition that the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is Normal. This is the case if both populations have Normal distributions, and it is approximately true for large samples from non-Normal populations because of the central limit theorem. The central limit theorem helps explain the robustness of the two-sample

LOOK BACK

robustness of
two-sample
procedures,
p. 442

CAUTION

t test. The test works well when both populations are symmetric, especially when the two sample sizes are similar.

The permutation test completely removes the Normality condition. However, *resampling in a way that moves observations between the two groups requires that the two populations are identical when the null hypothesis is true—that not only their means are the same but also their spreads and shapes.* Our preferred version of the two-sample *t* test allows different standard deviations in the two groups, so the shapes are both Normal but need not have the same spread.

In Example 16.13, the distributions are skewed but we do not rule out the *t* test because of the central limit theorem. The permutation test is valid if the GPA distributions for males and females have the same shape, so that they are identical under the null hypothesis that the centers (the means) are the same. Based on Figure 16.10 (page 16-17), it appears that the distribution for the males has a little more spread than the distribution for the females. Fortunately, the permutation test is robust. That is, it gives accurate *P*-values when the two population distributions have somewhat different shapes, such as when they have slightly different standard deviations.

Sources of variation Just as in the case of bootstrap confidence intervals, permutation tests are subject to two sources of random variability: the original sample is chosen at random from the population, and the resamples are chosen at random from the sample. Again, as in the case of the bootstrap, the added variation due to resampling is usually small and can be made as small as we like by increasing the number of resamples.

The number of resamples on which a permutation test is based determines the number of decimal places and precision in the resulting *P*-value. Tests based on 1000 resamples give *P*-values to three places (multiples of 0.001), with a margin of error of $2\sqrt{P(1-P)/1000}$ equal to 0.014 when the true one-sided *P*-value is 0.05. If higher precision is needed or your computer is sufficiently fast, you may choose to use 10,000 or more resamples.

USE YOUR KNOWLEDGE

16.59 Is a permutation test valid? Suppose a professor wants to compare the effectiveness of two different instruction methods. By design, one method is more team oriented, so he expects the variability in individual tests scores for this method to be smaller. Is it valid to use a permutation test to compare the mean scores of the two methods? Explain.

16.60 Declaring significance. Suppose that a one-sided permutation test based on 250 permutation resamples resulted in a *P*-value of 0.044. What is the approximate standard deviation of the distribution? Would you feel comfortable declaring the results significant at the 5% level? Explain.

Permutation tests in other settings

The bootstrap procedure can replace many different formula-based confidence intervals, provided that we resample in a way that matches the setting. Permutation testing is also a general method that we can adapt to various settings.

> ### GENERAL PROCEDURE FOR PERMUTATION TESTS
>
> To carry out a permutation test based on a statistic that measures the size of an effect of interest:
>
> 1. Compute the statistic for the original data.
>
> 2. Choose permutation resamples from the data without replacement in a way that is consistent with the null hypothesis of the test and with the study design. Construct the permutation distribution of the statistic from its values in a large number of resamples.
>
> 3. Find the *P*-value by locating the original statistic on the permutation distribution.

Permutation test for matched pairs   The key step in the general procedure for permutation tests is to form permutation resamples in a way that is consistent with the study design and with the null hypothesis. Our examples to this point have concerned two-sample settings. How must we modify our procedure for a matched pairs design?

EXAMPLE 16.14

Permutation test for full-moon study. Can the full moon influence be-havior? A study observed 15 nursing-home patients with dementia. The number of incidents of aggressive behavior was recorded each day for 12 weeks. Call a day a "moon day" if it is the day of a full moon or the day before or after a full moon. Table 16.2 gives the average number of aggres-sive incidents for moon days and other days for each subject.[8] These are matched pairs data. A matched pairs *t* test found evidence that the mean number of aggressive incidents is higher on moon days ($t = 6.45$, df $= 14$, $P < 0.001$). The data show some signs of non-Normality. We want to apply a permutation test.

   The null hypothesis says that the full moon has no effect on behavior. If this is true, the two entries for each patient in Table 16.2 are two measurements of aggressive behavior made under the same conditions. There is no distinction between "moon days" and "other days." Resampling

MOON

TABLE 16.2	Aggressive Behaviors of Dementia Patients				
Patient	Moon days	Other days	Patient	Moon days	Other days
1	3.33	0.27	9	6.00	1.59
2	3.67	0.59	10	4.33	0.60
3	2.67	0.32	11	3.33	0.65
4	3.33	0.19	12	0.67	0.69
5	3.33	1.26	13	1.33	1.26
6	3.67	0.11	14	0.33	0.23
7	4.67	0.30	15	2.00	0.38
8	2.67	0.40			

istockphoto

in a way consistent with this null hypothesis randomly assigns one of each patient's two scores to "moon" and the other to "other." We don't mix results for different subjects, because the original data are paired.

The permutation test (like the matched pairs t test) uses the difference in means $\bar{x}_{\text{moon}} - \bar{x}_{\text{other}}$. Figure 16.26 shows the permutation distribution of this statistic from 10,000 resamples. None of these resamples produces a difference as large as the observed difference, $\bar{x}_{\text{moon}} - \bar{x}_{\text{other}} = 2.433$. The estimated one-sided P-value is less than 1 in 1000. We report this result as $P < 0.0001$. There is strong evidence that aggressive behavior is more common on moon days.

FIGURE 16.26 The permutation distribution for the mean difference (moon days minus other days) from 10,000 paired resamples from the data in Table 16.2, Example 16.14.

The permutation distribution in Figure 16.26 is close to Normal, as a Normal quantile plot confirms. The matched pairs t test is, therefore, reliable and agrees with the permutation test that the P-value is very small.

Permutation test for the significance of a relationship Permutation testing can be used to test the significance of a relationship between two variables. For example, in Example 16.10, we looked at the relationship between price and rating of laundry detergents.

The null hypothesis is that there is no relationship. In that case, prices are assigned to detergents for reasons that have nothing to do with rating. We can resample in a way consistent with the null hypothesis by permuting the observed ratings among the detergents at random.

Take the correlation as the test statistic. For every resample, calculate the correlation between the prices (in their original order) and ratings (in the reshuffled order). The P-value is the proportion of the resamples with correlation larger than the original correlation.

When can we use permutation tests? We can use a permutation test only when we can see how to resample in a way that is consistent with the study

design and with the null hypothesis. We now know how to do this for the following types of problems:

• **Two-sample problems** when the null hypothesis says that the two populations are identical. We may wish to compare population means, proportions, standard deviations, or other statistics. You may recall from Section 7.3 that traditional tests for comparing population standard deviations work very poorly. Permutation tests are a much better choice.

• **Matched pairs designs** when the null hypothesis says that there are only random differences within pairs. A variety of comparisons is again possible.

• **Relationships between two quantitative variables** when the null hypothesis says that the variables are not related. The correlation is the most common measure of association, but not the only one.

These settings share the characteristic that the null hypothesis specifies a simple situation such as two identical populations or two unrelated variables. We can see how to resample in a way that matches these situations. *Permutation tests can't be used for testing hypotheses about a single population, comparing populations that differ even under the null hypothesis, or testing general relationships.* In these settings, we don't know how to resample in a way that matches the null hypothesis. Researchers are developing resampling methods for these and other settings, so stay tuned.

When we can't do a permutation test, we can often calculate a bootstrap confidence interval instead. If the confidence interval fails to include the null hypothesis value, then we reject H_0 at the corresponding significance level. This is not as accurate as doing a permutation test, but a confidence interval estimates the size of an effect as well as giving some information about its statistical significance. Even when a test is possible, it is often helpful to report a confidence interval along with the test result. Confidence intervals don't assume that a null hypothesis is true, so we use bootstrap resampling with replacement rather than permutation resampling without replacement.

SECTION 16.5 SUMMARY

• **Permutation tests** are significance tests based on **permutation resamples** drawn at random from the original data. Permutation resamples are drawn **without replacement**, in contrast to bootstrap samples, which are drawn with replacement.

• Permutation resamples must be drawn in a way that is consistent with the null hypothesis and with the study design. In a **two-sample design**, the null hypothesis says that the two populations are identical. Resampling randomly reassigns observations to the two groups. In a **matched pairs** design, randomly permute the two observations within each pair separately. To test the hypothesis of **no relationship** between two variables, randomly reassign values of one of the two variables.

• The **permutation distribution** of a suitable statistic is formed by the values of the statistic in a large number of resamples. Find the *P*-value of the test by locating the original value of the statistic on the permutation distribution.

• When they can be used, permutation tests have great advantages. They do not require specific population shapes such as Normality. They apply to a variety of statistics, not just to statistics that have a simple distribution under the null hypothesis. They can give very accurate *P*-values, regardless of the shape and size of the population (if enough permutations are used).

• It is often useful to give a confidence interval along with a test. To create a confidence interval, we no longer assume that the null hypothesis is true, so we use bootstrap resampling rather than permutation resampling.

SECTION 16.5 EXERCISES

For Exercises 16.59 and 16.60, see page 16-47.

16.61 Marketing cell phones. You have two prototypes of a new cell phone and designed an experiment to help you decide which one to market. Forty students were randomly assigned to use one of the two phones for two weeks. Their overall satisfaction with the phone is recorded on a subjective scale with a range of 1 to 100. Outline the steps needed to compare the means for the two phones using a permutation test.

16.62 Marketing cell phones. Refer to the previous exercise. Suppose that you had each of the 40 students use both phones. Outline the steps needed to compare the means for the two phones using a permutation test.

16.63 Characteristics of cell phones. Refer to Exercise 16.61. Before asking the students to provide an overall satisfaction rating, they were asked to provide ratings for several characteristics of the cell phone. Two of these were satisfaction with the screen and satisfaction with the keyboard. Outline the steps needed to evaluate the relationship between these two variables for the first phone using a permutation test.

16.64 Compare the correlations. Refer to the previous exercise. Suppose that you calculate the correlation between satisfaction with the screen and satisfaction with the keyboard for each phone. Outline the steps needed to compare these two correlations using a permutation test.

16.65 A small-sample permutation test. To illustrate the process, let's perform a permutation test by hand for a small random subset of the DRP data (Example 16.11, page 16-42). Here are the data: **DRP6**

Treatment group	57	61		
Control group	42	62	41	28

(a) Calculate the difference in means $\bar{x}_{\text{treatment}} - \bar{x}_{\text{control}}$ between the two groups. This is the observed value of the statistic.

(b) Resample: Start with the six scores and choose an SRS of two scores to form the treatment group for the first

resample. You can do this by labeling the scores from 1 to 6 and using consecutive random digits from Table B or by rolling a die. Using either method, be sure to skip repeated digits. A resample is an ordinary SRS, without replacement. The remaining four scores are the control group. What is the difference in group means for this resample?

(c) Repeat part (b) 20 times to get 20 resamples and 20 values of the statistic. Make a histogram of the distribution of these 20 values. This is the permutation distribution for your resamples.

(d) What proportion of the 20 statistic values were equal to or greater than the original value in part (a)? You have just estimated the one-sided *P*-value for the original 6 observations.

(e) For this small data set, there are only 15 possible permutations of the data. As a result, we can calculate the exact *P*-value by counting the number of permutations with a statistic value greater than or equal to the original value and then dividing by 15. What is the exact *P*-value here? How close was your estimate?

16.66 Product labels with animals? Participants in a study were asked to indicate their attitude toward a product on a seven-point scale (from 1 = dislike very much to 7 = like very much). A bottle of MagicCoat pet shampoo, with a picture of a collie on the label, was the product. Prior to indicating this preference, subjects were randomly assigned to two groups and were asked to do a word find. Four of the words were common to both groups, and four were either related to the product image or conflicted with the image. The group with the words related to the product image were considered primed. Let's use a permutation test for the comparison. Here are the data: **BRANDPR**

Group	Brand attitude
Primed	2 2 3 3 3 4 4 4 4 4 4 4 4 4 4 5 5 5 5 5 5 5
Nonprimed	1 1 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 4 4 4 5

(a) Examine the scores of each group graphically. Is it appropriate to use the two-sample *t* procedures? Explain your answer.

(b) Perform the two-sample t test to compare the group means. Use a two-sided alternative hypothesis and a significance level of 5%.

(c) Perform a permutation test to compare the group means. Summarize your results and conclusions.

(d) Write a short summary comparing your results in parts (b) and (c). Which method do you recommend for these data? Give reasons for your answer.

16.67 Timing of food intake. Examples 7.15 and 7.16 (pages 443 and 444) examine data on an experiment to compare weight loss in subjects who were classified as early eaters or late eaters, based on the timing of their main meal. In Example 7.16, the following data were analyzed: **In EATER**

Group	Weight loss (kg)				
Early eater	6.3	15.1	9.4	16.8	10.2
Late eater	7.8	0.2	1.5	11.5	4.6

(a) State appropriate null and alternative hypotheses for these data.

(b) Report the result of the pooled two-sample t test.

(c) Perform a permutation test to compare the two means and report the results. Compare the P-value for this test with the P-value for the t test in part (b).

(d) Find a BCa confidence interval for the difference in means. How is this interval related to your results in part (c)?

16.68 Standard deviation of the estimated P-value. The estimated P-value for the DRP study (Example 16.12, page 16-43) based on 1000 resamples is $P = 0.015$. Suppose that we obtained the same P-value based on 4000 resamples. What is the approximate standard deviation of each of these P-values?

16.69 When is a permutation test valid? You want to test the equality of the means of two populations. Sketch density curves for two populations for which

(a) a permutation test is valid but a t test is not.

(b) both permutation and t tests are valid.

(c) a t test is valid but a permutation test is not.

16.70 Testing the correlation between education spending and population. In Exercise 16.53 (page 16-40), we assessed the significance of the *correlation* between education spending and population by creating bootstrap confidence intervals. If a 95% confidence interval does not cover 0, the observed correlation is significantly different from 0 at the $\alpha = 0.05$ level. Let's do a test that provides a P-value. Carry out a permutation test and give the P-value.

What do you conclude? Is your conclusion consistent with your work in Exercise 16.53? **In EDSPEND**

16.71 Assessing a summer language institute. Exercise 7.47 (page 432) gives data on a study of the effect of a summer language institute on the ability of high school language teachers to understand spoken French. This is a matched pairs study, with scores for 20 teachers at the beginning (pretest) and end (posttest) of the institute. We conjecture that the posttest scores are higher on the average. **In FRENCH**

(a) Carry out the matched pairs t test. That is, state hypotheses, calculate the test statistic, and give its P-value.

(b) Make a Normal quantile plot of the gains: posttest score−pretest score. The data have a number of ties and a low outlier. A permutation test can help check the t test result.

(c) Carry out the permutation test for the *difference in means in a matched pairs setting*, using 9999 resamples. The Normal quantile plot shows that the permutation distribution is reasonably Normal. What is the P-value for the permutation test? Do your tests in parts (a) and (c) lead to the same practical conclusion?

16.72 Compare the medians. Refer to the previous exercise. Use a permutation test to compare the medians. Write a short summary of your results and conclusions. Include a comparison of what you found here with what you found in the previous exercise. **In FRENCH**

16.73 Testing the correlation between price and rating. Example 16.10 (page 16-35) uses the bootstrap to find a confidence interval for the correlation between price and rating for 24 laundry detergents. Let's use a permutation test to examine this correlation. **In LAUND24**

(a) State the null and alternative hypotheses.

(b) Perform a permutation test based on the sample correlation. Report the P-value and draw a conclusion.

16.74 Comparing mpg calculations. Exercise 7.41 (page 430) gives data on a comparison of driver and computer mpg calculations. This is a matched pairs study, with mpg values for 20 fill-ups. **In MPG20**

(a) Carry out the matched pairs t test. That is, state hypotheses, calculate the test statistic, and give its P-value.

(b) A permutation test can help check the t test result. Carry out the permutation test for the *difference in means in a matched pairs setting,* using 10,000 resamples. Does this test and the test in part (a) lead to the same practical conclusion?

16.75 Comparing standard deviations. In Example 12.16 (page 666), the modified Levene's test was used to compare standard deviations. Let's instead consider

TABLE 16.3	Serum Retinol Levels (μmol/l) in Two Groups of Children

Not infected						Infected					
0.59	1.08	0.88	0.62	0.46	0.39	0.68	0.56	1.19	0.41	0.84	0.37
1.44	1.04	0.67	0.86	0.90	0.70	0.38	0.34	0.97	1.20	0.35	0.87
0.35	0.99	1.22	1.15	1.13	0.67	0.30	1.15	0.38	0.34	0.33	0.26
0.99	0.35	0.94	1.00	1.02	1.11	0.82	0.81	0.56	1.13	1.90	0.42
0.83	0.35	0.67	0.31	0.58	1.36	0.78	0.68	0.69	1.09	1.06	1.23
1.17	0.35	0.23	0.34	0.49		0.69	0.57	0.82	0.59	0.24	0.41
						0.36	0.36	0.39	0.97	0.40	0.40
						0.24	0.67	0.40	0.55	0.67	0.52
						0.23	0.33	0.38	0.33	0.31	0.35
						0.82					

performing a permutation test using the F statistic (ratio of the largest and smallest sample variances) as your statistic. What do you conclude? Are the two tests comparable? **FRIENDS**

16.76 Comparing serum retinol levels. The formal medical term for vitamin A in the blood is serum retinol. Serum retinol has various beneficial effects, such as protecting against fractures. Medical researchers working with children in Papua New Guinea asked whether recent infections reduce the level of serum retinol. They classified children as recently infected or not on the basis of other blood tests and then measured serum retinol. Of the 90 children in the sample, 55 had been recently infected. Table 16.3 gives the serum retinol levels for both groups, in micromoles per liter.[9] **RETINOL**

(a) The researchers are interested in the proportional reduction in serum retinol. Verify that the mean for infected children is 0.620 and that the mean for uninfected children is 0.778.

(b) There is no standard test for the null hypothesis that the *ratio of the population means* is 1. We can do a permutation test on the ratio of sample means. Carry out a one-sided test and report the P-value. Briefly describe the center and shape of the permutation distribution. Why do you expect the center to be close to 1?

16.77 Methods of resampling. In Exercise 16.76, we did a permutation test for the hypothesis "no difference between infected and uninfected children" using the ratio of mean serum retinol levels to measure "difference." We might also want a bootstrap confidence interval for the ratio of population means for infected and uninfected children. Describe carefully how resampling is done for the permutation test and for the bootstrap, paying attention to the difference between the two resampling methods.

16.78 Listening to podcasts. A 2012 Edison Research asked U.S. individuals age 12 and older whether or not

they had ever listened to a podcast. The survey was repeated with different users in 2015. For the 2012 survey, 586 of the 2020 people surveyed reported that they had listened to at least one audio podcast. In the 2015 survey, the results were 660 of the 2002 survey participants. We want to use these sample data to test equality of the population proportions of successes. Carry out a permutation test. Describe the permutation distribution. Give the P-value and report your conclusion.

16.79 Sex and GPA. In Exercise 16.51 (page 16-39), we used the bootstrap to compare the mean GPA scores for men and women. **GPA**

(a) Use permutation methods to compare the means for men and women.

(b) Use permutation methods to compare the standard deviations for men and women.

(c) Write a short paragraph summarizing your results and conclusions.

16.80 Sadness and spending. A study of sadness and spending randomized subjects to watch videos designed to produce sad or neutral moods. Each subject was given $10, and after watching the video, he or she was asked to trade $0.50 increments of their $10 for an insulated bottle of water. Here are the data: **SADNESS**

Group	Purchase price								
Neutral	0.00	2.00	0.00	1.00	0.50	0.00	0.50		
	2.00	1.00	0.00	0.00	0.00	0.00	1.00		
Sad	3.00	4.00	0.50	1.00	2.50	2.00	1.50	0.00	1.00
	1.50	1.50	2.50	4.00	3.00	3.50	1.00	3.50	

(a) Use the two-sample t significance test (page 440) to compare the means of the two groups. Summarize your results.

(b) Use the pooled two-sample t significance test (page 449) to compare the means of the two groups. Summarize your results.

(c) Use a permutation test to compare the two groups. Summarize your results.

(d) Discuss the differences among the results you found for parts (a), (b), and (c). Which method do you prefer? Give reasons for your answer.

16.81 Comparing the variances for sadness and spending. Refer to the previous example. Some treatments in randomized experiments such as this can cause variances to be different. Are the variances of the neutral and sad subjects equal? SADNESS

(a) Use the modified Levene's test for equality of variances (page 665) to answer this question. Summarize your results.

(b) Compare the variances using a permutation test. Summarize your results.

(c) Write a short paragraph comparing the modified Levene's test with the permutation test for these data.

16.82 Comparing two operators. Exercise 7.45 (page 431) gives these data on a delicate measurement of total body bone mineral content made by two operators on the same eight subjects:[10] TBBMC

	Subject							
Operator	1	2	3	4	5	6	7	8
1	1.328	1.342	1.075	1.228	0.939	1.004	1.178	1.286
2	1.323	1.322	1.073	1.233	0.934	1.019	1.184	1.304

Do permutation tests give good evidence that measurements made by the two operators differ systematically? If so, in what way do they differ? Do two tests, one that compares centers and one that compares spreads.

CHAPTER 16 EXERCISES

16.83 Trimmed mean of GPA. In Example 16.5 (page 16-16), you used the bootstrap to find a 95% confidence interval for the 25% trimmed mean of GPA. Let's change the statistic of interest to the 5% trimmed mean. Using Example 16.5 as a guide, find the corresponding 95% confidence interval. Compare this interval with the one in Example 16.5. GPA

16.84 Change the trim. Refer to the previous exercise. Change the statistic of interest to the 10% trimmed mean. Answer the questions in the previous exercise and also compare your new interval with the one you found there. GPA

16.85 Compare the correlations. In Exercise 16.51 (page 16-39), we compared the mean GPA for men and women using the bootstrap. In Exercise 16.52, we used the bootstrap to examine the correlation between GPA and high school math grades. Let's find the correlations for men and women separately and ask whether there is evidence that they differ. GPA

(a) Find the correlation between GPA and high school math grades for the men. Use the bootstrap to find a 95% confidence interval for the population correlation.

(b) Repeat part (a) for the women.

(c) Use the bootstrap to test the null hypothesis that the population correlations for men and women are the same, $\rho_{Men} = \rho_{Women}$.

(d) Summarize your findings.

16.86 Use the regression slope. Refer to the previous exercise, where we used correlations to address the question of whether or not the relationship between GPA and high school math grades is the same for men and women. In Exercise 16.56 (page 16-40), we used the bootstrap to examine the slope of the least-squares regression line for predicting GPA using high school math grades. Let's compute the slope separately for men and women and ask whether or not they differ. This is another way to ask the question about whether or not the relationship between GPA and high school math grades is the same for men and women. Answer the questions from the previous exercise using the slope. Compare the results that you find here with those you found in the previous exercise. GPA

16.87 Bootstrap confidence interval for the difference in proportions. Refer to Exercise 16.78 (page 16-53). We want a 95% confidence interval for the change from 2012 to 2015 in the proportions of U.S. residents who report that they have listened to at least one podcast. Bootstrap the sample data. Give all three bootstrap confidence intervals (t, percentile, and BCa). Compare the three intervals and summarize the results. Which intervals would you recommend? Give reasons for your answer.

16.88 Bootstrap confidence interval for the ratio. Here is one conclusion from the data in Table 16.3, described in Exercise 16.76: "The mean serum retinol

level in uninfected children was 1.255 times the mean level in the infected children. A 95% confidence interval for the ratio of means in the population of all children in Papua New Guinea is…." **In** RETINOL

(a) Bootstrap the data and use the BCa method to complete this conclusion.

(b) Briefly describe the shape and bias of the bootstrap distribution. Does the bootstrap percentile interval agree closely with the BCa interval for these data?

16.89 Poetry: an occupational hazard. According to William Butler Yeats, "She is the Gaelic muse, for she gives inspiration to those she persecutes. The Gaelic poets die young, for she is restless, and will not let them remain long on earth." One study designed to investigate this issue examined the age at death for writers from different cultures and sexes.[11]

In Example 1.31 (page 37), we examined the distributions of the age at death for female novelists, poets, and nonfiction writers. Figure 1.15 shows modified side-by-side boxplots for the three categories of writers. The poets do appear to die young! Note that there is an outlier among the nonfiction writers. This writer died at the age of 40, young for a nonfiction writer, but not for a novelist or a poet! Let's use the methods of this chapter to compare the ages at death for poets and nonfiction writers. **In** POETS

(a) Use numerical and graphical summaries to describe the distribution of age at death for the poets. Do the same for the nonfiction writers.

(b) Use the methods of Chapter 7 (page 440) to compare the means of the two distributions. Summarize your findings.

(c) Use the bootstrap methods of this chapter to compare the means of the two distributions. Summarize your findings.

16.90 Medians for the poets. Refer to the previous exercise. Use the bootstrap methods of this chapter to compare the medians of the two distributions. Summarize your findings and compare them with part (c) of the previous exercise. **In** POETS

16.91 Permutation test for the poets. Refer to Exercise 16.89. Answer part (c) of that exercise using the permutation test. Summarize your findings and compare them with what you found in Exercise 16.89. **In** POETS

16.92 Variance for poets. Refer to Exercises 16.89 and 16.91.

(a) Instead of comparing means, compare variances using the ratio of sample variances as the statistic. Summarize your findings.

(b) Explain how questions about the equality of standard deviations are related to questions about the equality of variances.

(c) Use the results of this exercise and the previous three exercises to address the question of whether or not the distributions of the poets and nonfiction writers are the same. **In** POETS

16.93 Bootstrap confidence interval for the median. Your software can generate random numbers that have the uniform distribution on 0 to 1. Figure 4.9 (page 240) shows the density curve. Generate a sample of 50 observations from this distribution.

(a) What is the population median? Bootstrap the sample median and describe the bootstrap distribution.

(b) What is the bootstrap standard error? Compute a 95% bootstrap t confidence interval.

(c) Find the 95% BCa confidence interval. Compare with the interval in (b). Is the bootstrap t interval reliable here?

16.94 Are female personal trainers, on average, younger? A fitness center employs 20 personal trainers. Here are the ages in years of the female and male personal trainers working at this center: **In** TRAIN

Male	25	26	23	32	35	29	30	28	31	32	29
Female	21	23	22	23	20	29	24	19	22		

(a) Make a back-to-back stemplot. Do you think the difference in mean ages will be significant?

(b) A two-sample t test gives $P < 0.001$ for the null hypothesis that the mean age of female personal trainers is equal to the mean age of male personal trainers. Do a two-sided permutation test to check the answer.

(c) What do you conclude about using the t test? What do you conclude about the mean ages of the trainers?

16.95 Adult gamers versus teen gamers. A Pew survey compared adult and teen gamers on where they played games. For the adults, 54% of 1063 survey participants played on game consoles such as Xbox, PlayStation, and Wii. For teens, 89% of 1064 survey participants played on game consoles. Use the bootstrap to find a 95% confidence interval for the difference between the teen proportion who play on consoles and the adult proportion.

16.96 Use a ratio for adult gamers versus teen gamers. Refer to the previous exercise. In many settings, researchers prefer to communicate the comparison of two proportions with a ratio. For gamers who play on consoles, they would report that teens are 1.65 (89/54) times more likely to play on consoles. Use the bootstrap to give a 95% confidence interval for this ratio.

16.97 Another way to communicate the result.
Refer to the previous two exercises. Here is another way to communicate the result: teen gamers are 65% more likely to play on consoles than adult gamers.

(a) Explain how the 65% is computed.

(b) Use the bootstrap to give a 95% confidence interval for this estimate.

(c) Based on this exercise and the previous two, which of the three ways is most effective for communicating the results? Give reasons for your answer.

16.98 Insurance fraud? Jocko's Garage has been accused of insurance fraud. Data on estimates (in dollars) made by Jocko and another garage were obtained for 10 damaged vehicles. Here is what the investigators found: **GARAGE**

Car	1	2	3	4	5
Jocko's	1375	1550	1250	1300	900
Other	1250	1300	1250	1200	950

Car	6	7	8	9	10
Jocko's	1500	1750	3600	2250	2800
Other	1575	1600	3300	2125	2600

(a) Compute the mean estimate for Jocko and the mean estimate for the other garage. Report the difference in the means and the 95% standard t confidence interval.

Be sure to choose the appropriate t procedure for your analysis and explain why you made this choice.

(b) Use the bootstrap to find the confidence interval. Be sure to give details about how you used the bootstrap, which options you chose, and why.

(c) Compare the t interval with the bootstrap interval.

16.99 Other ways to look at Jocko's estimates.
Refer to the previous exercise. Let's consider some other ways to analyze these data. **GARAGE**

(a) For each damaged vehicle, divide Jocko's estimate by the estimate from the other garage. Perform your analysis on these data. Write a short report that includes numerical and graphical summaries, your estimate, the 95% t confidence interval, the 95% bootstrap confidence interval, and an explanation for all choices (such as whether you chose to examine the mean or the median, bootstrap options, etc.).

(b) Compute the mean of Jocko's estimates and the mean of the estimates made by the other garage. Divide Jocko's mean by the mean for the other garage. Report this ratio and find a 95% confidence interval for this quantity. Be sure to justify choices that you made for the bootstrap.

(c) Using what you have learned in this exercise and the previous one, how would you summarize the comparison of Jocko's estimates with those made by the other garage? Assume that your audience knows very little about statistics but a lot about insurance.

CHAPTER 16 NOTES AND DATA SOURCES

1. Information about this free software is available at **r-project.org**.

2. The origin of this quaint phrase is Rudolph Raspe, *The Singular Adventures of Baron Munchausen,* 1786. Here is the passage, from the edition by John Carswell, Heritage Press, 1952: "I was still a couple of miles above the clouds when it broke, and with such violence I fell to the ground that I found myself stunned, and in a hole nine fathoms under the grass, when I recovered, hardly knowing how to get out again. Looking down, I observed that I had on a pair of boots with exceptionally sturdy straps. Grasping them firmly, I pulled with all my might. Soon I had hoist myself to the top and stepped out on terra firma without further ado."

3. In fact, the bootstrap standard error underestimates the true standard error. Bootstrap standard errors are generally too small by a factor of roughly $\sqrt{1 - 1/n}$. This factor is about 0.95 for $n = 10$ and 0.98 for $n = 25$, so we ignore it in this elementary exposition.

4. The 254 winning numbers and their payoffs are republished here by permission of the New Jersey State Lottery Commission.

5. The vehicle is a 2002 Toyota Prius previously owned by the third author.

6. The standard advanced introduction to bootstrap methods is B. Efron and R. Tibshirani, *An Introduction to the Bootstrap,* Chapman and Hall, 1993. For tilting intervals, see B. Efron, "Nonparametric standard errors and confidence intervals" (with discussion), *Canadian Journal of Statistics,* 36 (1981), pp. 369–401; and T. J. DiCiccio and J. P. Romano, "Nonparametric confidence limits by resampling methods and least favourable families," *International Statistical Review,* 58 (1990), pp. 59–76.

7. This example is adapted from Maribeth C. Schmitt, "The effects of an elaborated directed reading activity on the metacomprehension skills of third graders," PhD dissertation, Purdue University, 1987.

8. These data were collected as part of a larger study of dementia patients conducted by Nancy Edwards, School of Nursing, and Alan Beck, School of Veterinary Medicine, Purdue University.

9. Data provided by Francisco Rosales of the Department of Nutritional Sciences, Pennsylvania State University. See Francisco Rosales et al., "Relation of serum retinol to acute phase proteins and malarial morbidity in Papua New Guinea children," *American Journal of Clinical Nutrition*, 71 (2000), pp. 1580–1588.

10. These data were collected in connection with a bone health study at Purdue University and were provided by Linda McCabe.

11. The data were provided by James Kaufman. The study is described in James C. Kaufman, "The cost of the muse: poets die young," *Death Studies*, 27 (2003), pp. 813–821. The quote from Yeats appears in this article.