CHAPTER 14

Bootstrap Methods and Permutation Tests*

*This chapter was written by Tim Hesterberg, David S. Moore, Shaun Monaghan, Ashley Clipson, and Rachel Epstein, with support from the National Science Foundation under grant DMI-0078706. We thank Bob Thurman, Richard Heiberger, Laura Chihara, Tom Moore, and Gudmund Iversen for helpful comments on an earlier version.

Introduction

The continuing revolution in computing is having a dramatic influence on statistics. Exploratory analysis of data becomes easier as graphs and calculations are automated. Statistical study of very large and very complex data sets becomes feasible. Another impact of fast and cheap computing is less obvious: new methods that apply previously unthinkable amounts of computation to small sets of data 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.

The most common methods for inference about means based on a single sample, matched pairs, or two independent samples are the *t* procedures described in Chapter 7. For relationships between quantitative variables, we use other *t* tests and intervals in the correlation and regression setting (Chapter 10). Chapters 11, 12, and 13 present inference procedures for more elaborate settings. All of these methods rest on the use of normal distributions for data. No data are exactly normal. The *t* procedures are useful in practice because they are *robust*, quite insensitive to deviations from normality in the data. Nonetheless, we cannot use *t* confidence intervals and tests if the data are strongly skewed, unless our samples are quite large. Inference about spread based on normal distributions is not robust and is therefore 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 computing power 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 also set us free from formulas. They work the same way (without formulas) 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. What is more, bootstrap intervals and permutation tests are conceptually simpler than confidence intervals and tests based on normal distributions 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 rapidly becoming the preferred way to do statistical inference. This is already true in high-stakes situations such as legal cases and clinical trials.

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 gradually more common in standard statistical software.

This chapter uses S-PLUS,[1] the software choice of most statisticians doing research on resampling methods. A free version of S-PLUS is available to students. You will also need two free libraries that supplement S-PLUS: the `S+Resample` library, which provides menu-driven access to the procedures described in this chapter, and the `IPSdata` library, which contains all the data sets for this text. You can find links for downloading this software on the text Web site, `www.whfreeman.com/ipsresample`.

You will find that using S-PLUS is straightforward, especially if you have experience with menu-based statistical software. After launching S-PLUS, load the `IPSdata` library. This automatically loads the `S+Resample` library as well. The `IPSdata` menu includes a guide with brief instructions for each procedure in this chapter. Look at this guide each time you meet something new. There is also a detailed manual for resampling under the `Help` menu. The resampling methods you need are all in the `Resampling` submenu in the `Statistics` menu in S-PLUS. Just choose the entry in that menu that describes your setting.

S-PLUS is highly capable statistical software that can be used for everything in this text. If you use S-PLUS for all your work, you may want to obtain a more detailed book on S-PLUS.

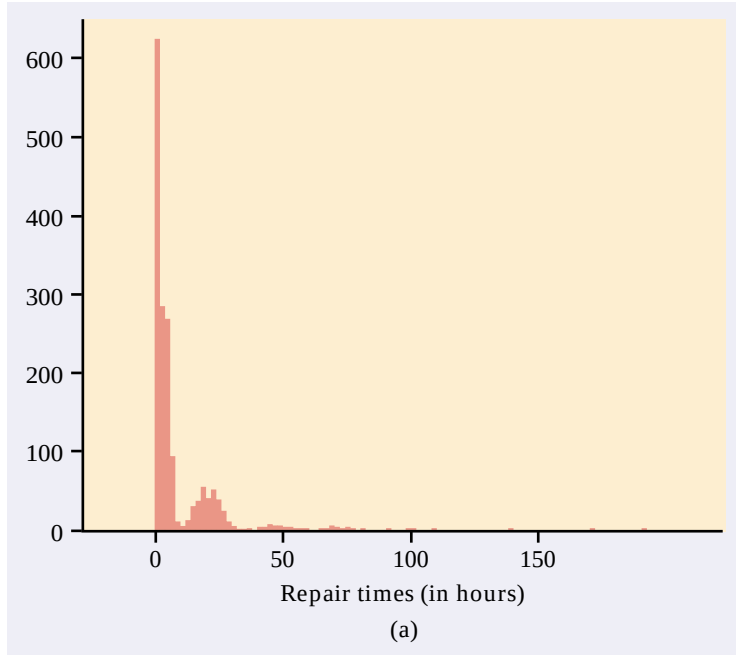
14.1 The Bootstrap Idea

Here is a situation in which the new computer-intensive methods are now being applied. We will use this example to introduce these methods.

EXAMPLE 14.1 In most of the United States, many different companies offer local telephone service. It isn't in the public interest to have all these companies digging up streets to bury cables, so the primary local telephone company in each region must (for a fee) share its lines with its competitors. The legal term for the primary company is Incumbent Local Exchange Carrier, ILEC. The competitors are called Competing Local Exchange Carriers, or CLECs.

Verizon is the ILEC for a large area in the eastern United States. As such, it must provide repair service for the customers of the CLECs in this region. Does Verizon do repairs for CLEC customers as quickly (on the average) as for its own customers? If not, it is subject to fines. The local Public Utilities Commission requires the use of tests of significance to compare repair times for the two groups of customers.

Repair times are far from normal. Figure 14.1 shows the distribution of a random sample of 1664 repair times for Verizon's own customers.[2] The distribution has a very long right tail. The median is 3.59 hours, but the mean is 8.41 hours and the longest repair time is 191.6 hours. We hesitate to use *t* procedures on such data, especially as the sample sizes for CLEC customers are much smaller than for Verizon's own customers.



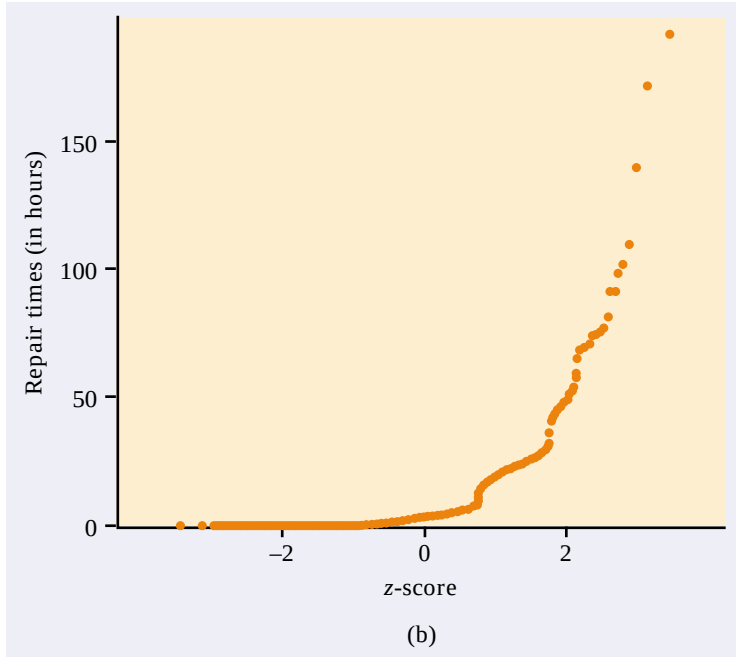


FIGURE 14.1 (a) The distribution of 1664 repair times for Verizon customers. (b) Normal quantile plot of the repair times. The distribution is strongly right-skewed.

The big idea: resampling and the bootstrap distribution

Statistical inference is based on the sampling distributions of sample statistics. The bootstrap is first of all a way of finding the sampling distribution, at least approximately, from just one sample. Here is the procedure:

Step 1: Resampling. A sampling distribution is based on many random samples from the population. In Example 14.1, we have just one random sample. 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.

resamples

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, then putting it back before drawing again. As a result, any number can be drawn more than once, or not at all. If we sampled *without* replacement, we'd get the same set of numbers we started with, though in a different order. Figure 14.2 illustrates three resamples from a sample of six observations. In practice, we draw hundreds or thousands of resamples, not just three.

Step 2: Bootstrap distribution. The sampling distribution of a statistic collects the values of the statistic from many samples. The **bootstrap distribution** of a statistic collects its values from many resamples. The bootstrap distribution gives information about the sampling distribution.

bootstrap distribution

THE BOOTSTRAP IDEA

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

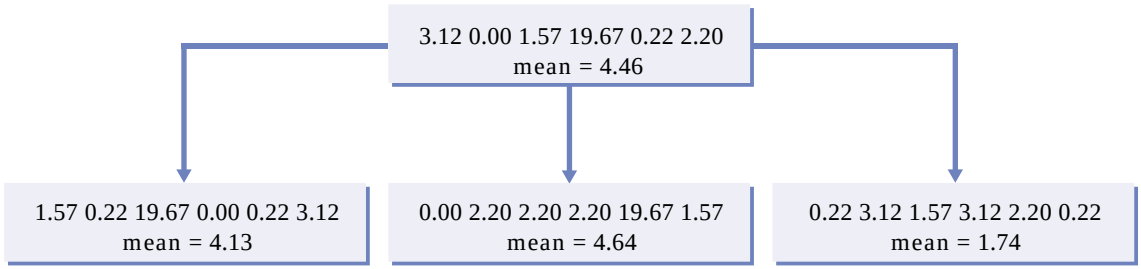


FIGURE 14.2 The resampling idea. The top box is a sample of size $n = 6$ from the Verizon data. The three lower boxes are three resamples from this original sample. Some values from the original are repeated in the resamples because each resample is formed by sampling with replacement. We calculate the statistic of interest—the sample mean in this example—for the original sample and each resample.

EXAMPLE 14.2 In Example 14.1, we want to estimate the population mean repair time μ, so the statistic is the sample mean $\bar{x}$. For our one random sample of 1664 repair times, $\bar{x} = 8.41$ hours. When we resample, we get different values of $\bar{x}$, just as we would if we took new samples from the population of all repair times.

Figure 14.3 displays the bootstrap distribution of the means of 1000 resamples from the Verizon repair time data, using first a histogram and a density curve and then a normal quantile plot. The solid line in the histogram marks the mean 8.41 of the original sample, and the dashed line marks the mean of the bootstrap means. 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.

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. That is, 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).

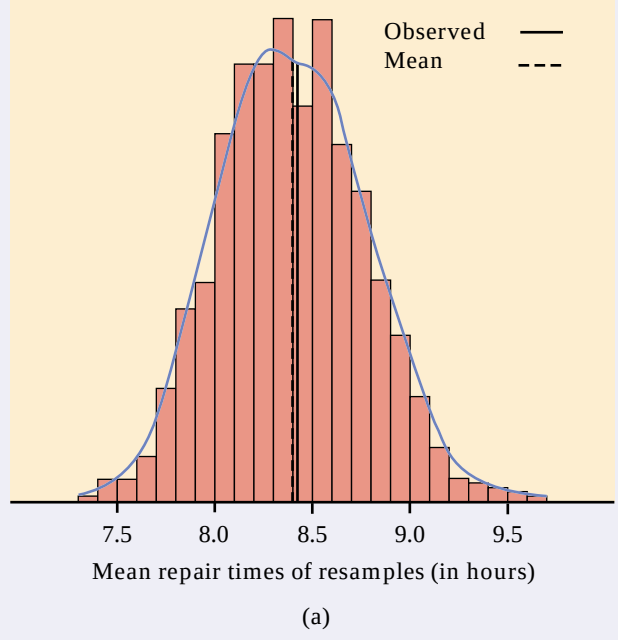
Spread: The histogram and density curve in Figure 14.3 picture the variation among the resample means. We can get a numerical measure by calculating their standard deviation. Because this is the standard deviation of the 1000 values of $\bar{x}$ that make up the bootstrap distribution, we call it the **bootstrap standard error** of $\bar{x}$. The numerical value is 0.367. 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, $s = 14.69$ and

bootstrap standard error

$$\frac{s}{\sqrt{n}} = \frac{14.69}{\sqrt{1664}} = 0.360$$

The bootstrap standard error 0.367 agrees closely with the theory-based estimate 0.360.

In discussing Example 14.2, 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 the 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 fails.* 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.



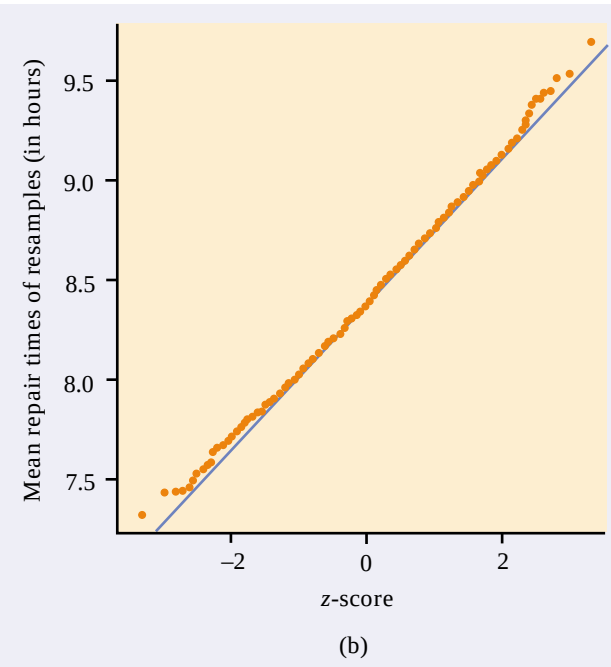


FIGURE 14.3 (a) The bootstrap distribution for 1000 resample means from the sample of Verizon repair times. The solid line marks the original sample mean, and the dashed line marks the average of the bootstrap means. (b) The normal quantile plot confirms that the bootstrap distribution is nearly normal in shape.

Figure 14.4 illustrates the bootstrap idea by comparing three distributions. Figure 14.4(a) shows the idea of the sampling distribution of the sample mean $\bar{x}$: take many random samples from the population, calculate the mean $\bar{x}$ for each sample, and collect these $\bar{x}$-values into a distribution.

Figure 14.4(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, appeal instead to 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.

Figure 14.4(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.

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."[3] 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 the one actual sample of size 1664 would vary because of random sampling.

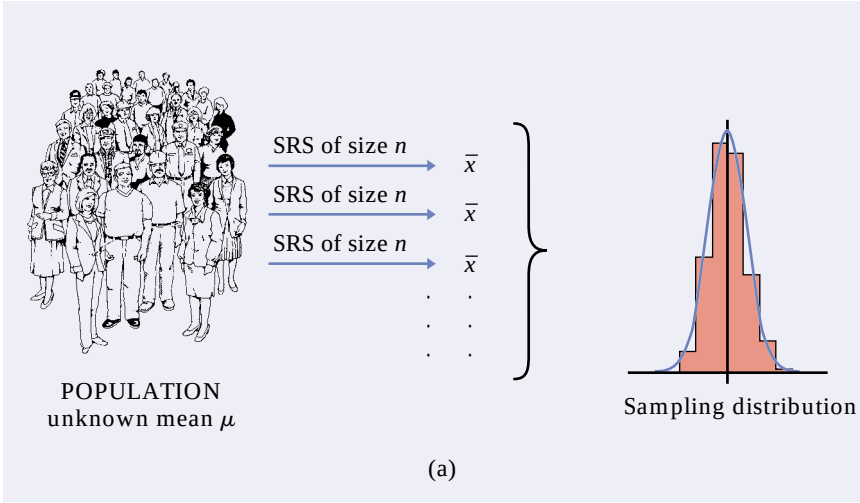
Using the same data for two purposes—to estimate a parameter and also to estimate the variability of the estimate—is perfectly legitimate. We do exactly this when we calculate $\bar{x}$ to estimate μ and then calculate $s/\sqrt{n}$ from the same data to estimate the variability of $\bar{x}$.

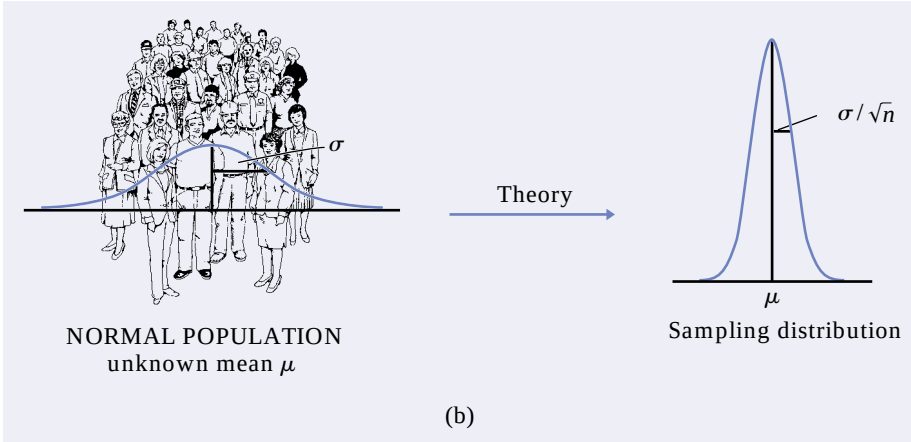
What is new? First of all, we don't rely on the formula $s/\sqrt{n}$ to estimate the standard deviation of $\bar{x}$. Instead, we use the ordinary standard deviation of the many $\bar{x}$-values from our many resamples.[4] Suppose that we take B resamples. Call the means of these resamples $\bar{x}^*$ to distinguish them from the mean $\bar{x}$ of the original sample. 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:

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

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

These formulas go all the way back to Chapter 1. Once we have the values $\bar{x}^*$, we just ask our software for their mean and standard deviation. We will often apply the bootstrap to statistics other than the sample mean. Here is the general definition.





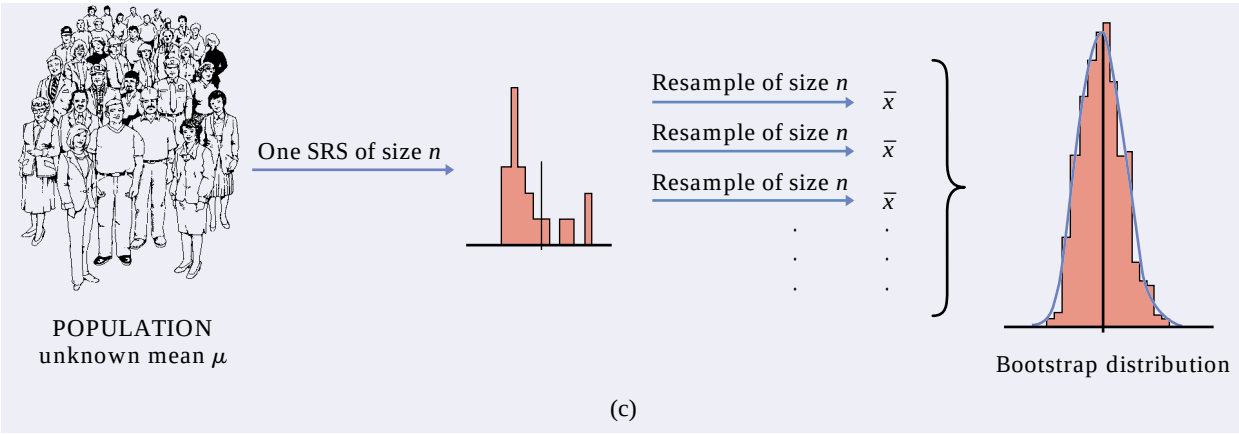


FIGURE 14.4 (a) The idea of the sampling distribution of the sample mean $\bar{x}$: take very many samples, collect the $\bar{x}$-values from each, and look at the distribution of these values. (b) The theory shortcut: if we know that the population values follow a normal distribution, theory tells us 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.

BOOTSTRAP STANDARD ERROR

The **bootstrap standard error** SE_{boot} of a statistic is the standard deviation of the bootstrap distribution of that statistic.

Another thing that is new is that 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 statistic value rather than the parameter value. The bootstrap allows us to calculate standard errors for statistics for which we don't have formulas and to check normality for 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 $\overline{x}$ 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, 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 1000 times {
  Draw a resample with replacement from the data.
  Calculate the resample mean.
  Save the resample mean into a variable.
}
Make a histogram and normal quantile plot of the 1000 means.
Calculate the standard deviation of the 1000 means.
```

```
Number of Replications: 1000

Summary Statistics:
        Observed    Mean      Bias      SE
mean       8.412   8.395  -0.01698  0.3672

Percentiles:
      2.5%  5.0%  95.0%  97.5%
mean 7.717 7.814  9.028  9.114
```

FIGURE 14.5 S-PLUS output for the Verizon data bootstrap, for Example 14.3.

EXAMPLE 14.3 S-PLUS has bootstrap commands built in. If the 1664 Verizon repair times are saved as a variable, we can use menus 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 graphs in Figure 14.3 are part of the S-PLUS output. Figure 14.5 shows some of the text output. The `Observed` entry gives the mean $\bar{x} = 8.412$ of the original sample. `Mean` is the mean of the resample means, $\text{mean}_{\text{boot}}$. `Bias` is the difference between the `Mean` and `Observed` values. The bootstrap standard error is displayed under `SE`. The `Percentiles` are percentiles of the bootstrap distribution, that is, of the 1000 resample means pictured in Figure 14.3. All of these values except `Observed` will differ a bit if you repeat 1000 resamples, because resamples are drawn at random.

SECTION 14.1 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 resampled 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) when 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 get an idea of 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 variation in a statistic based on the original data.

SECTION 14.1 Exercises

Unless an exercise instructs you otherwise, use 1000 resamples for all bootstrap exercises. S-PLUS uses 1000 resamples unless you ask for a different number. Always save your bootstrap results so that you can use them again later.

14.1 To illustrate the bootstrap procedure, let's bootstrap a small random subset of the Verizon data:

3.12 0.00 1.57 19.67 0.22 2.20

(a) Sample *with replacement* from this initial 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.

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 14.2 to 14.5. Use a histogram and 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.

14.2 The distribution of the 60 IQ test scores in Table 1.3 (page 14) is roughly normal (see Figure 1.5) and the sample size is large enough that we expect a normal sampling distribution.

14.3 The distribution of the 64 amounts of oil in Exercise 1.33 (page 37) is strongly skewed, but the sample size is large enough that the central limit theorem may (or may not) result in a roughly normal sampling distribution.

14.4 The amounts of vitamin C in a random sample of 8 lots of corn soy blend (Example 7.1, page 453) are

26 31 23 22 11 22 14 31

The distribution has no outliers, but we cannot assess normality from so small a sample.

14.5 The measurements of C-reactive protein in 40 children (Exercise 7.2, page 472) are very strongly skewed. We were hesitant to use t procedures for inference from these data.

14.6 The "survival times" of machines before a breakdown and of cancer patients after treatment are typically strongly right-skewed. Table 1.8 (page 38) gives the survival times (in days) of 72 guinea pigs in a medical trial.[5]

(a) Make a histogram of the survival times. 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 $n = 72$? To find out, bootstrap these data 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?

14.7 Here is an SRS of 20 of the guinea pig survival times from Exercise 14.6:

92	123	88	598	100	114	89	522	58	191
137	100	403	144	184	102	83	126	53	79

We expect the sampling distribution of $\bar{x}$ to be less close to normal for samples of size 20 than for samples of size 72 from a skewed distribution. These data include some extreme high outliers.

(a) Create and inspect the bootstrap distribution of the sample mean for these data. Is it less close to normal than your distribution from the previous exercise?

(b) Compare the bootstrap standard errors for your two runs. What accounts for the larger standard error for the smaller sample?

14.8 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. Find the sample standard deviation s for the 20 survival times in Exercise 14.7 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 14.7?

14.2 First Steps in Using the Bootstrap

To introduce the big 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 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.

bias

bootstrap estimate of bias

More precisely, the **bias** of a statistic is the difference between the mean of its sampling distribution and the true value of the parameter. 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 14.4 We are interested in the selling prices of residential real estate in Seattle, Washington. Table 14.1 displays the selling prices of a random sample of 50 pieces of real estate sold in Seattle during 2002, as recorded by the county assessor.[6] Unfortunately, the data do not distinguish residential property from commercial property. Most sales are residential, but a few large commercial sales in a sample can greatly increase the sample mean selling price.

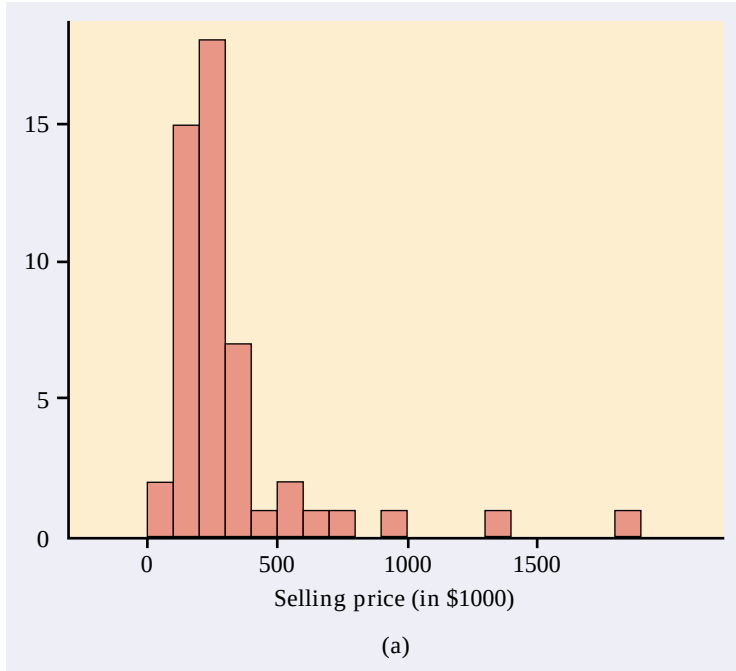
Figure 14.6 shows the distribution of the sample prices. The distribution is far from normal, with a few high outliers that may be commercial sales. The sample is small, and the distribution is highly skewed and "contaminated" by an unknown number of commercial sales. How can we estimate the center of the distribution despite these difficulties?

The first step is to abandon the mean as a measure of center in favor of a statistic that is more resistant to outliers. We might choose the median, but in this case we will use a new statistic, the *25% trimmed mean*.

TABLE 14.1

Selling prices for Seattle real estate, 2002 ($1000s)

142	175	197.5	149.4	705	232	50	146.5	155	1850
132.5	215	116.7	244.9	290	200	260	449.9	66.407	164.95
362	307	266	166	375	244.95	210.95	265	296	335
335	1370	256	148.5	987.5	324.5	215.5	684.5	270	330
222	179.8	257	252.95	149.95	225	217	570	507	190



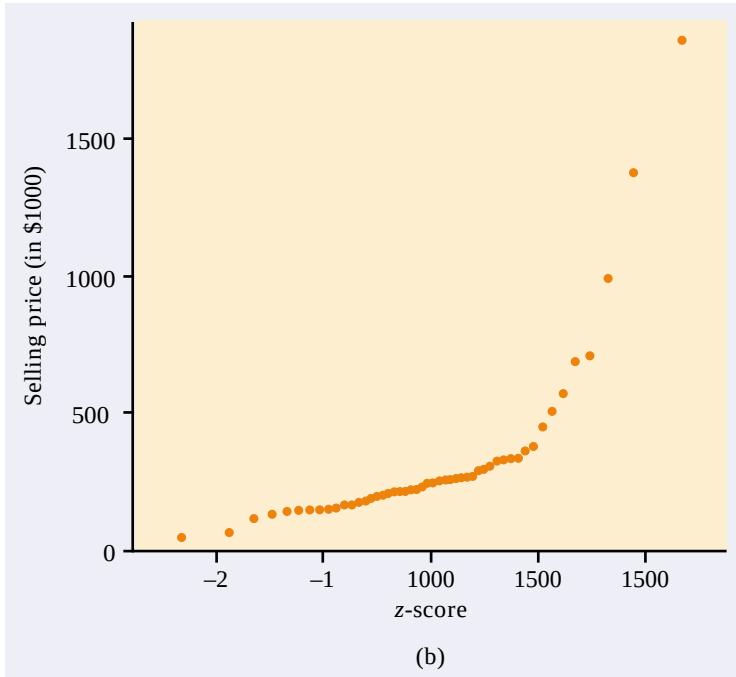


FIGURE 14.6 Graphical displays of the 50 selling prices in Table 14.1. The distribution is strongly skewed, with high outliers.

TRIMMED MEAN

A **trimmed mean** is the mean of only the center observations in a data set. In particular, the **25% trimmed mean** $\overline{x}_{25\%}$ ignores the smallest 25% and the largest 25% of the observations. It is the mean of the middle 50% of the observations.

Recall that the median is the mean of the 1 or 2 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 all real estate sales prices in Seattle in 2002. By the plug-in principle, the *statistic* that estimates this parameter is the 25% trimmed mean of the sample prices in Table 14.1. Because 25% of 50 is 12.5, we drop the 12 lowest and 12 highest prices in Table 14.1 and find the mean of the remaining 26 prices. The statistic is (in thousands of dollars)

$$\overline{x}_{25\%} = 244.0019$$

We can say little about the sampling distribution of the trimmed mean when we have only 50 observations from a strongly skewed distribution. 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 1000 resamples of size 50 from the 50 selling prices in Table 14.1, calculate the 25% trimmed mean for each resample, and form the bootstrap distribution from these 1000 values.

Figure 14.7 shows the bootstrap distribution of the 25% trimmed mean. Here is the summary output from S-PLUS:

```
Number of Replications: 1000

Summary Statistics:
         Observed   Mean    Bias     SE
TrimMean      244  244.7  0.7171  16.83
```

What do we see? **Shape:** The bootstrap distribution is roughly normal. This suggests that the sampling distribution of the trimmed mean is also roughly normal. **Center:** The bootstrap estimate of bias is 0.7171, small relative to the value 244 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). **Spread:** The bootstrap standard error of the statistic is

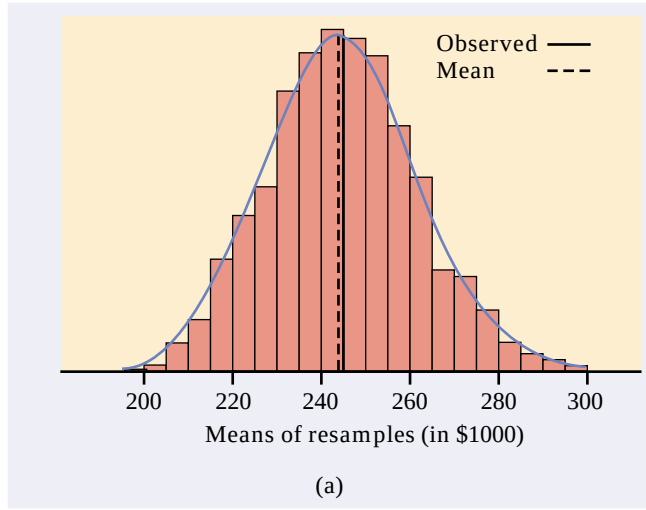
$$\mathrm{SE}_{\mathrm{boot}} = 16.83$$

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 452) for the mean of a normal population:

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

This interval is based on the normal sampling distribution of the sample mean $\overline{x}$ and the formula $\mathrm{SE} = s/\sqrt{n}$ for the standard error of $\overline{x}$. When a bootstrap



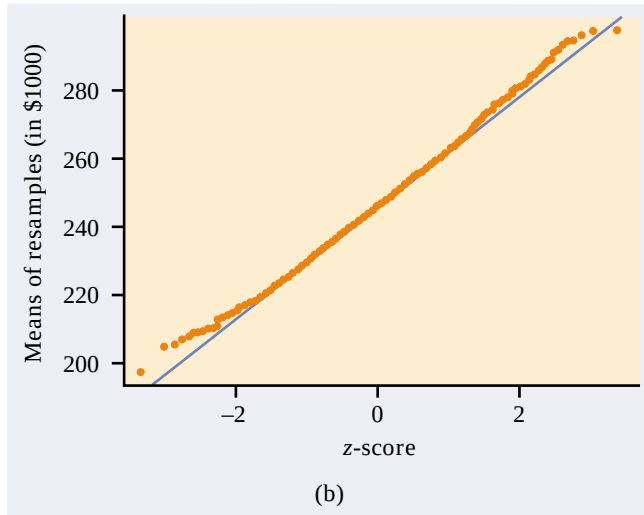


FIGURE 14.7 The bootstrap distribution of the 25% trimmed means of 1000 resamples from the data in Table 14.1. The bootstrap distribution is roughly normal.

distribution is approximately normal and has small bias, we can use essentially the same recipe 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

$$\text{statistic} \pm t^* \text{SE}_{\text{boot}}$$

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 14.5 We want to estimate the 25% trimmed mean of the population of all 2002 Seattle real estate selling prices. Table 14.1 gives an SRS of size $n = 50$. The software output above shows that the trimmed mean of this sample is $\overline{x}_{25\%} = 244$ and that the bootstrap standard error of this statistic is $SE_{boot} = 16.83$. A 95% confidence interval for the population trimmed mean is therefore

$$\begin{aligned}\overline{x}_{25\%} \pm t^* SE_{boot} &= 244 \pm (2.009)(16.83) \\ &= 244 \pm 33.81 \\ &= (210.19, 277.81)\end{aligned}$$

Because Table D does not have entries for $n - 1 = 49$ degrees of freedom, we used $t^* = 2.009$, the entry for 50 degrees of freedom.

We are 95% confident that the 25% trimmed mean (the mean of the middle 50%) for the population of real estate sales in Seattle in 2002 is between \$210,190 and \$277,810.

Bootstrapping to compare two groups

Two-sample problems (Section 7.2) 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 comparison of means. The most important new idea is that bootstrap resampling must mimic the "separate samples" design that produced the original data.

BOOTSTRAP FOR COMPARING TWO POPULATIONS

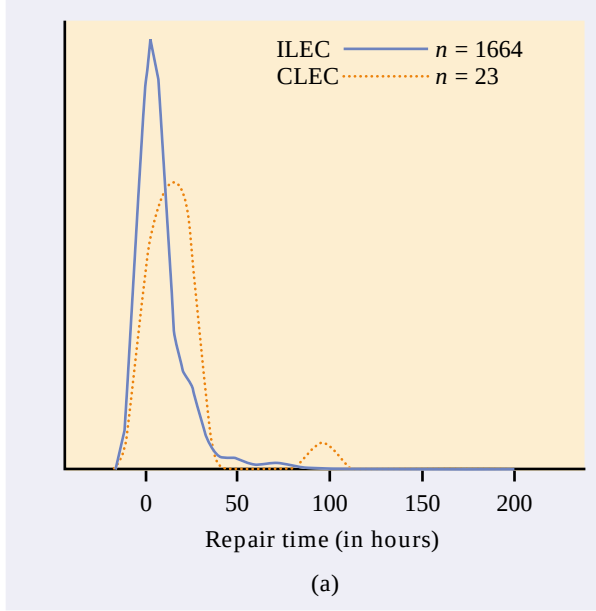
Given independent SRSs of sizes n and m from two populations:

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 hundreds of times.

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

EXAMPLE 14.6 We saw in Example 14.1 that Verizon is required to perform repairs for customers of competing providers of telephone service (CLECs) within its region. How do repair times for CLEC customers compare with times for Verizon's own customers? Figure 14.8 shows density curves and normal quantile plots for the service times (in hours) of 1664 repair requests from customers of Verizon and



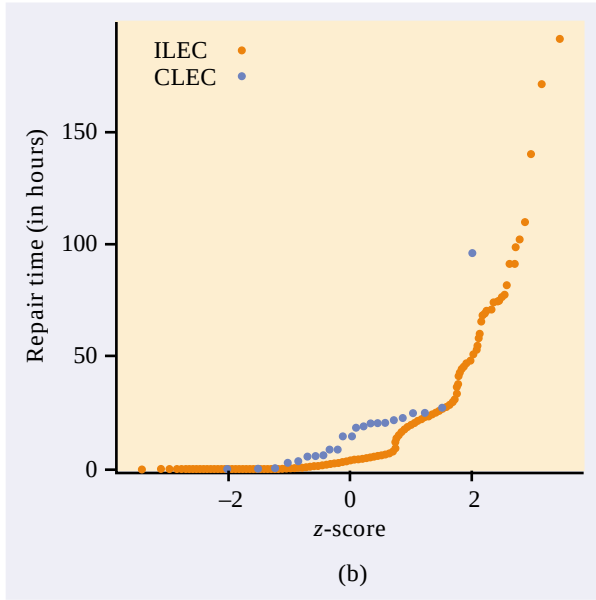


FIGURE 14.8 Density curves and normal quantile plots of the distributions of repair times for Verizon customers and customers of a CLEC. (The density curves extend below zero because they smooth the data. There are no negative repair times.)

23 requests from customers of a CLEC during the same time period. The distributions are both far from normal. Here are some summary statistics:

Service provider	n	$\bar{x}$	s
Verizon	1664	8.4	14.7
CLEC	23	16.5	19.5
Difference		−8.1	

The data suggest that repair times may be longer for CLEC customers. The mean repair time, for example, is almost twice as long for CLEC customers as for Verizon customers.

In the setting of Example 14.6 we want to estimate the difference of population means, $\mu_1 - \mu_2$. We are reluctant to use the two-sample t confidence interval because one of the samples is both small and 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 1000 resamples consists of two group resamples, one of size 1664 drawn with replacement from the Verizon data and one of size 23 drawn with replacement from the CLEC data. For each combined resample, compute the statistic $\bar{x}_1 - \bar{x}_2$. The 1000 differences form the bootstrap distribution. The bootstrap standard error is the standard deviation of the bootstrap distribution.

S-PLUS automates the proper bootstrap procedure. Here is some of the S-PLUS output:

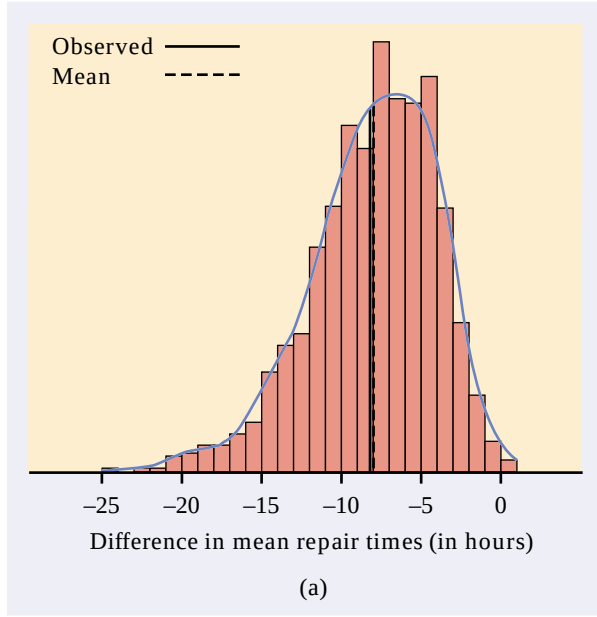
```
Number of Replications: 1000

Summary Statistics:
         Observed   Mean    Bias    SE
meanDiff   -8.098 -8.251 -0.1534 4.052
```

Figure 14.9 shows that the bootstrap distribution is not close to normal. It has a short right tail and a long left tail, so that it is skewed to the left. *Because the bootstrap distribution is nonnormal, we can't trust the bootstrap t confidence interval.* When the sampling distribution is nonnormal, no method based on normality is safe. 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 14.4, are the next step in practical use of the bootstrap.

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 110).



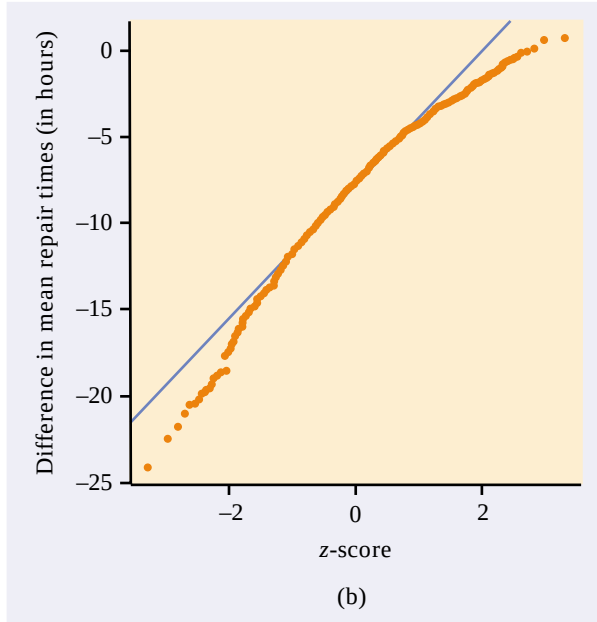


FIGURE 14.9 The bootstrap distribution of the difference in means for the Verizon and CLEC repair time data.

EXAMPLE 14.7 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.[7] Buying a ticket entitles a player to pick a number between 000 and 999. Half of 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.

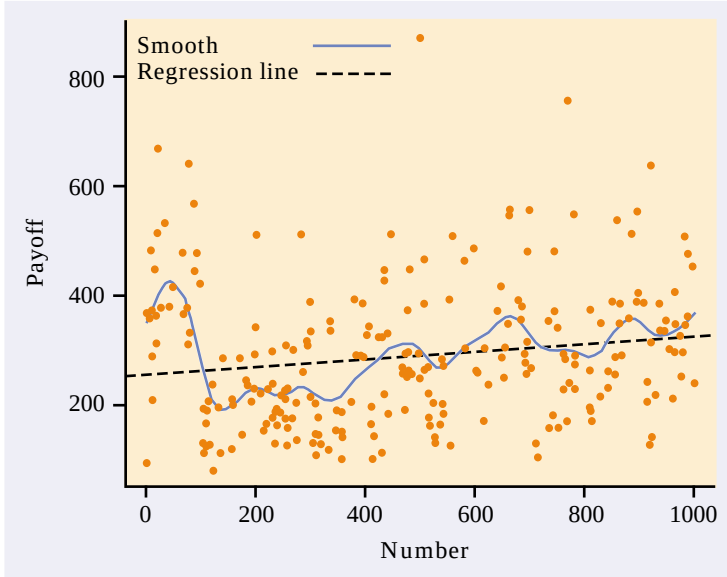


FIGURE 14.10 The first 254 winning numbers in the New Jersey Pick-It Lottery and the payoffs for each. To see patterns we use least-squares regression (line) and a scatterplot smoother (curve).

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 14.10 is a scatterplot of the first 254 winning numbers and their payoffs. What patterns can we see?

The straight line in Figure 14.10 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. 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.

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 14.11 shows the curves that result from applying the smoother to 20 resamples from the 254 data points in Figure 14.10. 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.

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.

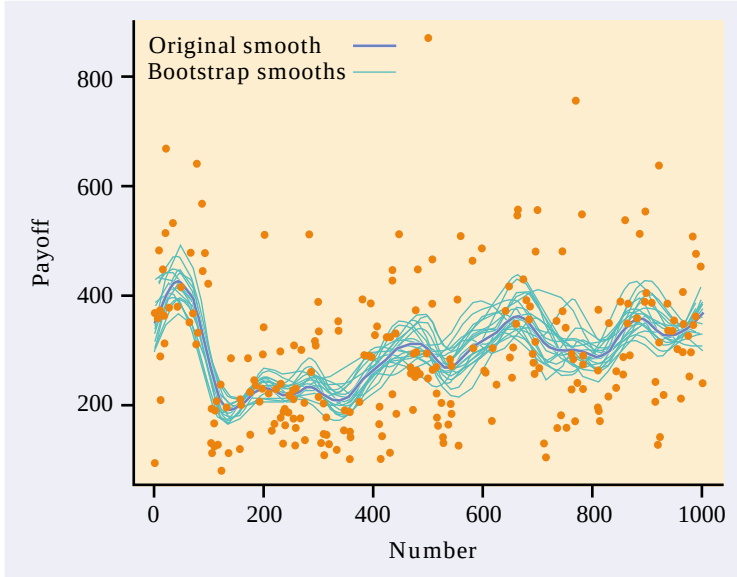


FIGURE 14.11 The curves produced by the scatterplot smoother for 20 resamples from the data displayed in Figure 14.10. The curve for the original sample is the heavy line.

SECTION 14.2 | 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 $\pm$ t^*SE_{boot},** for the parameter. Do not use this *t* interval if the bootstrap distribution is not normal or shows substantial bias.

SECTION 14.2 | Exercises

14.9 Return to or re-create the bootstrap distribution of the sample mean for the 72 guinea pig lifetimes in Exercise 14.6.

(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 $\overline{x}$. 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 (b).

14.10 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 14.7 shows one bootstrap distribution for the trimmed mean selling price for Seattle real estate. Repeat the resampling of the data in Table 14.1 to get another bootstrap distribution for the trimmed mean.

(a) Plot the bootstrap distribution and compare it with Figure 14.7. Are the two bootstrap distributions similar?

(b) What are the values of the mean statistic, bias, and bootstrap standard error for your new bootstrap distribution? How do they compare with the previous values given on page 14-16?

(c) Find the 95% bootstrap t confidence interval based on your bootstrap distribution. Compare it with the previous result in Example 14.5.

14.11 For Example 14.5 we bootstrapped the 25% trimmed mean of the 50 selling prices in Table 14.1. 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.

14.12 We will see in Section 14.3 that bootstrap methods often work poorly for the median. To illustrate this, bootstrap the sample median of the 50 selling prices in Table 14.1. Why is the bootstrap t confidence interval not justified?

14.13 We have a formula (page 488) for the standard error of $\overline{x}_1 - \overline{x}_2$. This formula does not depend on normality. How does this formula-based standard error for the data of Example 14.6 compare with the bootstrap standard error?

14.14 Table 7.4 (page 491) 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 on page 492.

14.15 Table 7.6 (page 512) contains the ratio of current assets to current liabilities for random samples of healthy firms and failed firms. Find the difference in means (healthy minus failed).

(a) Bootstrap the difference in means $\bar{x}_1 - \bar{x}_2$ and look at the bootstrap distribution. Does it meet the conditions for a bootstrap t confidence interval?

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

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

14.16 Explain the difference between the standard deviation of a sample and the standard error of a statistic such as the sample mean.

14.17 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.

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.

14.18 Because the shape and bias of the bootstrap distribution approximate the shape and bias of the sampling distribution, bootstrapping helps check whether the sampling distribution allows use of the usual t procedures. In Exercise 14.4 you bootstrapped the mean for the amount of vitamin C in a random sample of 8 lots of corn soy blend. Return to or re-create your work.

(a) The sample is very small. Nonetheless, the bootstrap distribution suggests that t inference is justified. Why?

(b) Give SE_{boot} and the bootstrap t 95% confidence interval. How do these compare with the formula-based standard error and usual t interval given in Example 7.1 (page 453)?

14.19 Exercise 7.5 (page 473) gives data on 60 children who said how big a part they thought luck played in solving a puzzle. The data have a discrete 1 to 10 scale. Is inference based on t distributions nonetheless justified? Explain your answer. If t inference is justified, compare the usual t and bootstrap t 95% confidence intervals.

14.20 Your company sells exercise clothing and equipment on the Internet. To design clothing, you collect data on the physical characteristics of your customers. Here are the weights in kilograms for a sample of 25 male runners. Assume these runners are a random sample of your potential male customers.

67.8	61.9	63.0	53.1	62.3	59.7	55.4	58.9	60.9
69.2	63.7	68.3	92.3	64.7	65.6	56.0	57.8	66.0
62.9	53.6	65.0	55.8	60.4	69.3	61.7		

Because your products are intended for the "average male runner," you are interested in seeing how much the subjects in your sample vary from the average weight.

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

(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?

14.21 Each year, the business magazine *Forbes* publishes a list of the world's billionaires. In 2002, the magazine found 497 billionaires. Here is the wealth, as estimated by *Forbes* and rounded to the nearest $100 million, of an SRS of 20 of these billionaires:[8]

8.6	1.3	5.2	1.0	2.5	1.8	2.7	2.4	1.4	3.0
5.0	1.7	1.1	5.0	2.0	1.4	2.1	1.2	1.5	1.0

You are interested in (vaguely) "the wealth of typical billionaires." Bootstrap an appropriate statistic, inspect the bootstrap distribution, and draw conclusions based on this sample.

14.22 Why is the bootstrap distribution of the difference in mean Verizon and CLEC repair times in Figure 14.9 so skewed? Let's investigate by bootstrapping the mean of the CLEC data and comparing it with the bootstrap distribution for the mean for Verizon customers. The 23 CLEC repair times (in hours) are

26.62	8.60	0	21.15	8.33	20.28	96.32	17.97
3.42	0.07	24.38	19.88	14.33	5.45	5.40	2.68
0	24.20	22.13	18.57	20.00	14.13	5.80	

(a) Bootstrap the mean for the CLEC data. Compare the bootstrap distribution with the bootstrap distribution of the Verizon repair times in Figure 14.3.

(b) Based on what you see in (a), what is the source of the skew in the bootstrap distribution of the difference in means $\overline{x}_1 - \overline{x}_2$?

14.3 How Accurate Is a Bootstrap Distribution?*

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

We understand that a statistic will vary from sample to sample, so that 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. Now we have used the bootstrap distribution as a substitute for the sampling distribution. This introduces a second source of random variation: resamples are chosen at random from the original sample.

SOURCES OF VARIATION AMONG BOOTSTRAP DISTRIBUTIONS

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 14.12 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 $\overline{x}$ of that sample. These vary from sample to sample. The distribution of the $\overline{x}$-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.

Now draw 1000 resamples from an original sample, calculate $\overline{x}$ for each resample, and present the 1000 $\overline{x}$'s in a histogram. This is a bootstrap distribution for $\overline{x}$. The middle column in Figure 14.12 displays five bootstrap distributions based on 1000 resamples from each of the five samples. The right

*This section is optional.

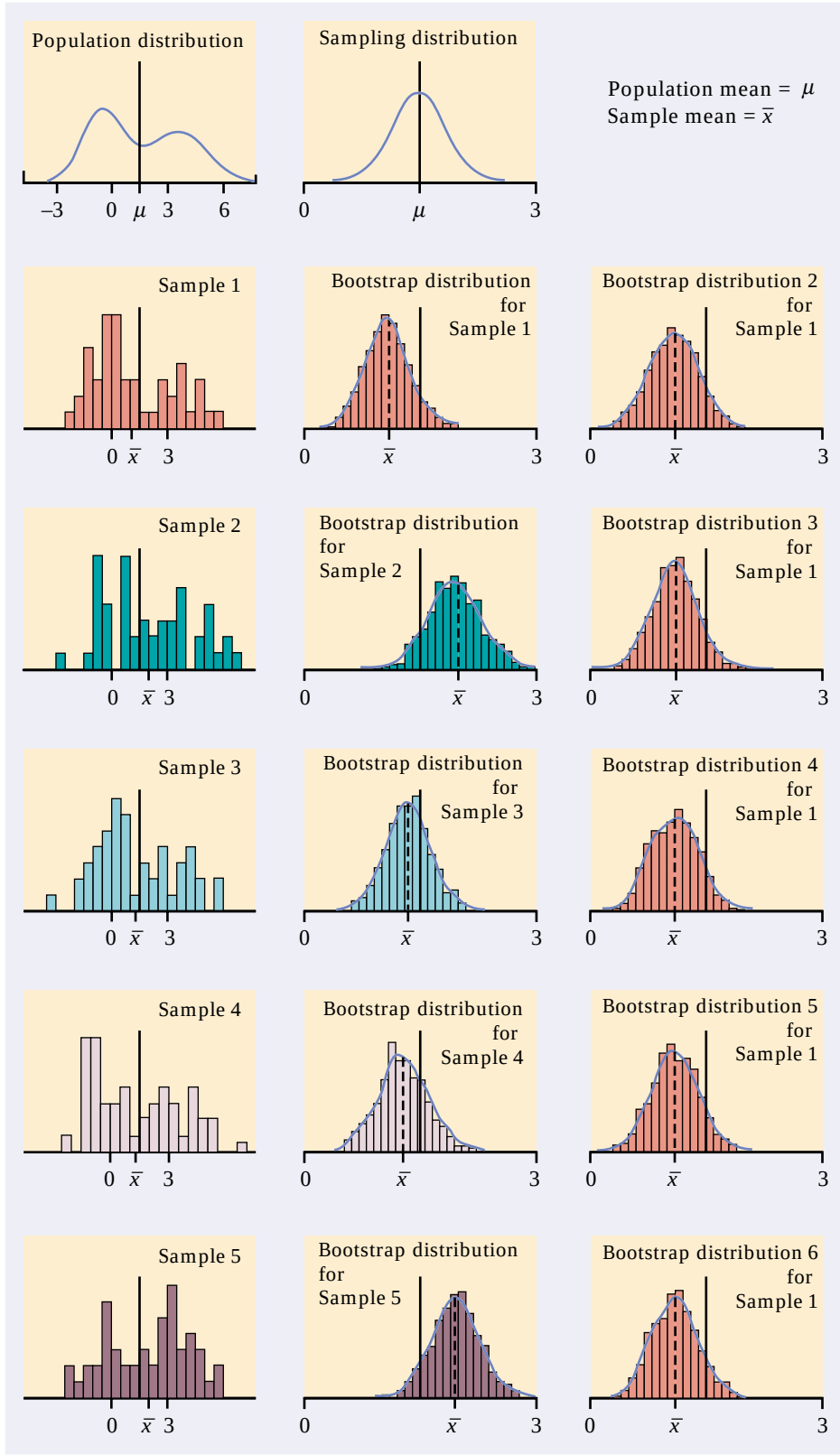


FIGURE 14.12 Five random samples ($n = 50$) from the same population, with a bootstrap distribution for the sample mean formed by resampling from each of the five samples. At the right are five more bootstrap distributions from the first sample.

column shows the results of repeating the resampling from the first sample five more times. Compare the five bootstrap distributions in the middle column to see the effect of the random choice of the original samples. Compare the six bootstrap 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 $\overline{x}$ for its original sample. That is, the bootstrap estimate of bias is small in all five cases. Of course, the five $\overline{x}$-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 do 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 14.12 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 of the variation among 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 statisticians prefer large samples because small samples give more variable results. This general fact is also true for bootstrap procedures.

Figure 14.13 repeats Figure 14.12, with two important differences. The five original samples are only of size $n = 9$, rather than the $n = 50$ of Figure 14.12. The population distribution (top left) is normal, so that the sampling distribution of $\overline{x}$ is normal despite the small sample size. The bootstrap distributions in the middle column show much more variation in shape and spread than those for larger samples in Figure 14.12. 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

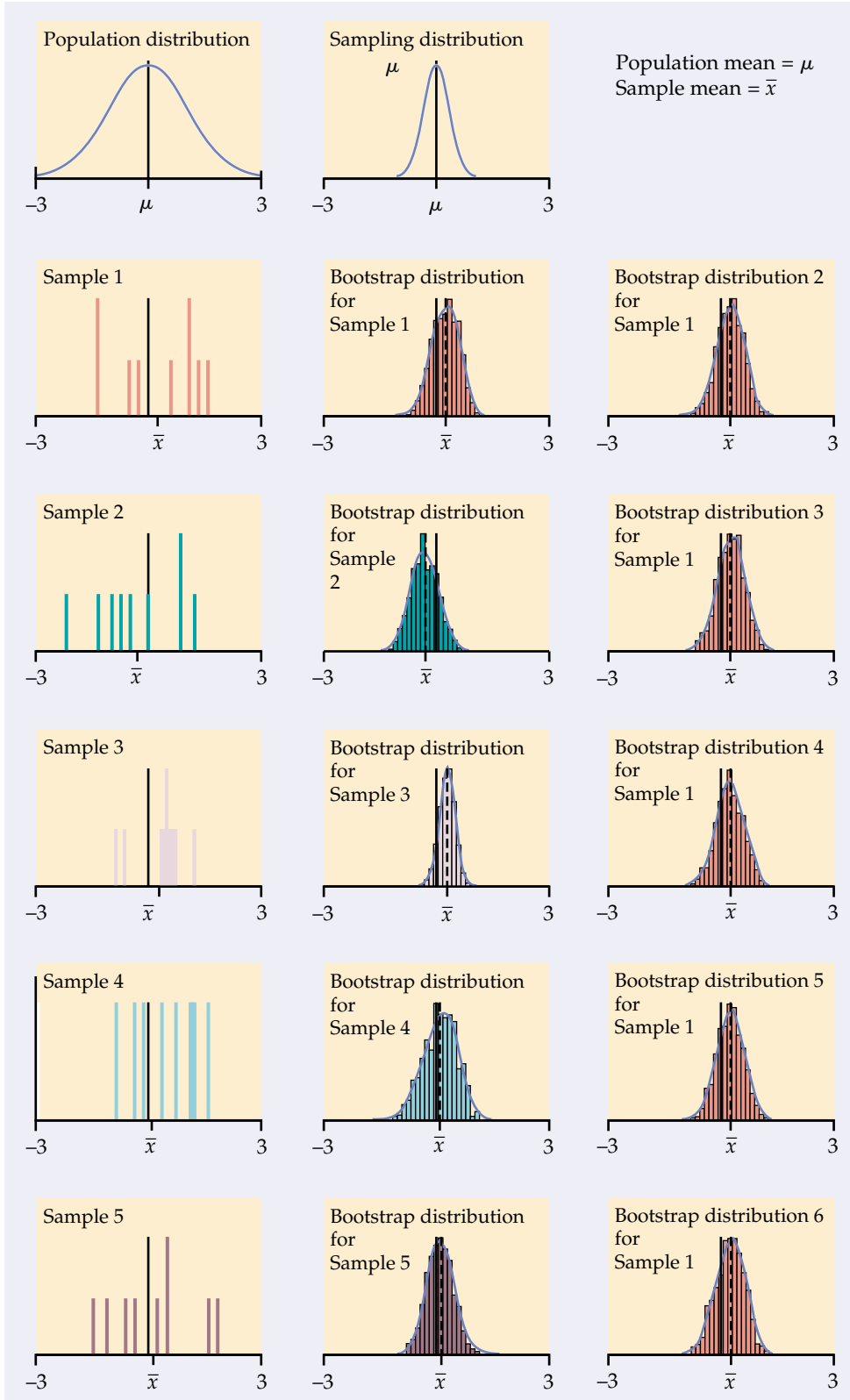


FIGURE 14.13 Five random samples ($n = 9$) from the same population, with a bootstrap distribution for the sample mean formed by resampling from each of the five samples. At the right are five more bootstrap distributions from the first sample.

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.

VARIATION IN BOOTSTRAP DISTRIBUTIONS

For most statistics, almost all the variation among 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 real estate sales prices in Example 14.4, 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 14.14 follows the format of Figures 14.12 and 14.13. 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 for 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 8th-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 therefore 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.

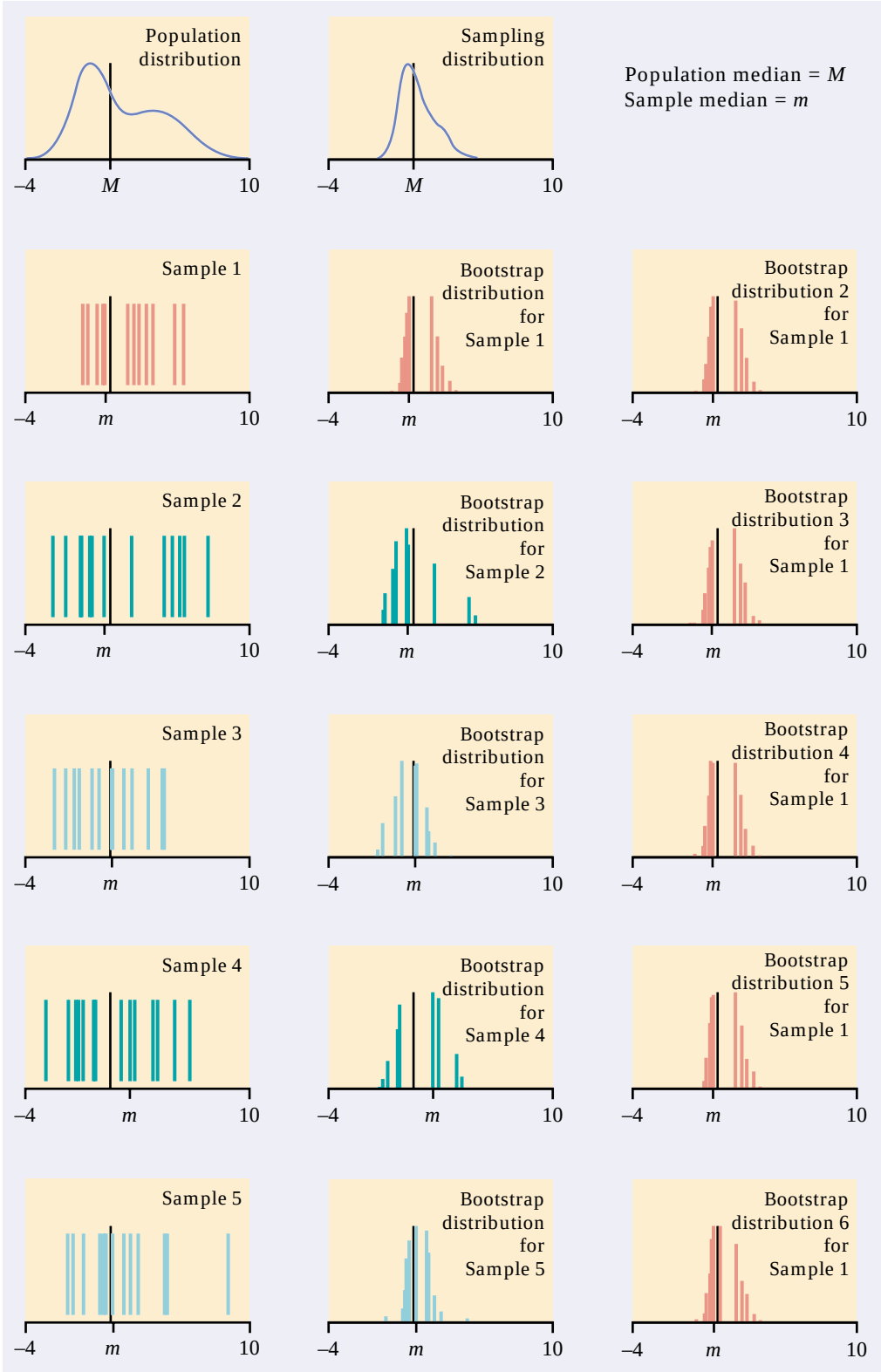


FIGURE 14.14 Five random samples ($n = 15$) from the same population, with a bootstrap distribution for the sample median formed by resampling from each of the five samples. At the right are five more bootstrap distributions from the first sample.

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.*

SECTION 14.3 | Summary

Almost all of the variation among bootstrap distributions 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 14.3 | Exercises

14.23 Most statistical software includes a function to generate samples from normal distributions. Set the mean to 8.4 and the standard deviation to 14.7. You can think of all the numbers that would be produced by this function if it ran forever as a population that has the $N(8.4, 14.7)$ distribution. Samples produced by the function are samples from this population.

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

(b) Draw an SRS of size $n = 10$ from this population. Bootstrap the sample mean $\overline{x}$ using 1000 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?

14.24 The data for Example 14.1 are 1664 repair times for customers of Verizon, the local telephone company in their area. In that example, these observations formed a sample. Now we will treat these 1664 observations as a population. The population distribution is pictured in Figures 14.1 and 14.8. It is very nonnormal. The population mean is $\mu = 8.4$, and the population standard deviation is $\sigma = 14.7$.

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

(b) Draw an SRS of size $n = 10$ from this population. Bootstrap the sample mean $\overline{x}$ using 1000 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. What are the effects of increasing the sample size?

14.25 The populations in the two previous exercises have the same mean and standard deviation, but one is very close to normal and the other is strongly nonnormal. Based on your work in these exercises, how does nonnormality of the population affect the bootstrap distribution of $\overline{x}$? How does it affect the bootstrap standard error? Does 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 $\overline{x}$ and the way in which bootstrap distributions mimic the sampling distribution.

14.4 Bootstrap Confidence Intervals

To this point, 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 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%. The 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 much more accurate bootstrap intervals, our recommendation is that *when bootstrap t and bootstrap percentile intervals do not agree closely, neither type of interval should be used.*

EXAMPLE 14.8 In Example 14.5 (page 14-18) we found that a 95% bootstrap t confidence interval for the 25% trimmed mean of Seattle real estate sales prices is 210.2 to 277.8. The bootstrap distribution in Figure 14.7 shows a small bias and, though roughly normal, is a bit skewed. Is the bootstrap t confidence interval accurate for these data?

The S-PLUS bootstrap output includes the 2.5% and 97.5% percentiles of the bootstrap distribution. They are 213.1 and 279.4. 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 real estate sales in Example 14.8 is

$$\overline{x}_{25\%} \pm t^* \mathrm{SE}_{\mathrm{boot}} = 244 \pm 33.81$$

We can learn something by also writing the percentile interval starting at the statistic $\overline{x}_{25\%} = 244$. In this form, it is

$$244.0 - 30.9, \quad 244.0 + 35.4$$

Unlike the t interval, the percentile interval is not symmetric—its endpoints are different distances from the statistic. The slightly greater distance to the 97.5% percentile reflects the slight right skewness of the bootstrap distribution.

Confidence intervals for the correlation

The bootstrap allows us to find standard errors and confidence intervals for a wide variety of statistics. We have done this for the mean and the trimmed mean. We also learned how to find the bootstrap distribution for a difference of means, but that distribution for the Verizon data (Example 14.6,

page 14-19) is so far from normal that we are reluctant to use the bootstrap t or percentile confidence intervals. 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 of means, we must pay attention to how we should resample.

EXAMPLE 14.9 Major League Baseball (MLB) owners claim they need limitations on player salaries to maintain competitiveness among richer and poorer teams. This argument assumes that higher salaries attract better players. Is there a relationship between an MLB player's salary and his performance?

Table 14.2 contains the names, 2002 salaries, and career batting averages of 50 randomly selected MLB players (excluding pitchers).[9] The scatterplot in Figure 14.15 suggests that the relationship between salary and batting average is weak. The sample correlation is $r = 0.107$. Is this small correlation significantly different from 0? To find out, we can calculate a 95% confidence interval for the population correlation and see whether or not it covers 0. If the confidence interval does not cover 0, the observed correlation is significant at the 5% level.

TABLE 14.2

Major League Baseball salaries and batting averages

Name	Salary	Average	Name	Salary	Average
Matt Williams	$9,500,000	0.269	Greg Colbrunn	$1,800,000	0.307
Jim Thome	$8,000,000	0.282	Dave Martinez	$1,500,000	0.276
Jim Edmonds	$7,333,333	0.327	Einar Diaz	$1,087,500	0.216
Fred McGriff	$7,250,000	0.259	Brian L. Hunter	$1,000,000	0.289
Jermaine Dye	$7,166,667	0.240	David Ortiz	$950,000	0.237
Edgar Martinez	$7,086,668	0.270	Luis Alicea	$800,000	0.202
Jeff Cirillo	$6,375,000	0.253	Ron Coomer	$750,000	0.344
Rey Ordonez	$6,250,000	0.238	Enrique Wilson	$720,000	0.185
Edgardo Alfonzo	$6,200,000	0.300	Dave Hansen	$675,000	0.234
Moises Alou	$6,000,000	0.247	Alfonso Soriano	$630,000	0.324
Travis Fryman	$5,825,000	0.213	Keith Lockhart	$600,000	0.200
Kevin Young	$5,625,000	0.238	Mike Mordecai	$500,000	0.214
M. Grudzielanek	$5,000,000	0.245	Julio Lugo	$325,000	0.262
Tony Batista	$4,900,000	0.276	Mark L. Johnson	$320,000	0.207
Fernando Tatis	$4,500,000	0.268	Jason LaRue	$305,000	0.233
Doug Glanville	$4,000,000	0.221	Doug Mientkiewicz	$285,000	0.259
Miguel Tejada	$3,625,000	0.301	Jay Gibbons	$232,500	0.250
Bill Mueller	$3,450,000	0.242	Corey Patterson	$227,500	0.278
Mark McLemore	$3,150,000	0.273	Felipe Lopez	$221,000	0.237
Vinny Castilla	$3,000,000	0.250	Nick Johnson	$220,650	0.235
Brook Fordyce	$2,500,000	0.208	Thomas Wilson	$220,000	0.243
Torii Hunter	$2,400,000	0.306	Dave Roberts	$217,500	0.297
Michael Tucker	$2,250,000	0.235	Pablo Ozuna	$202,000	0.333
Eric Chavez	$2,125,000	0.277	Alexis Sanchez	$202,000	0.301
Aaron Boone	$2,100,000	0.227	Abraham Nunez	$200,000	0.224

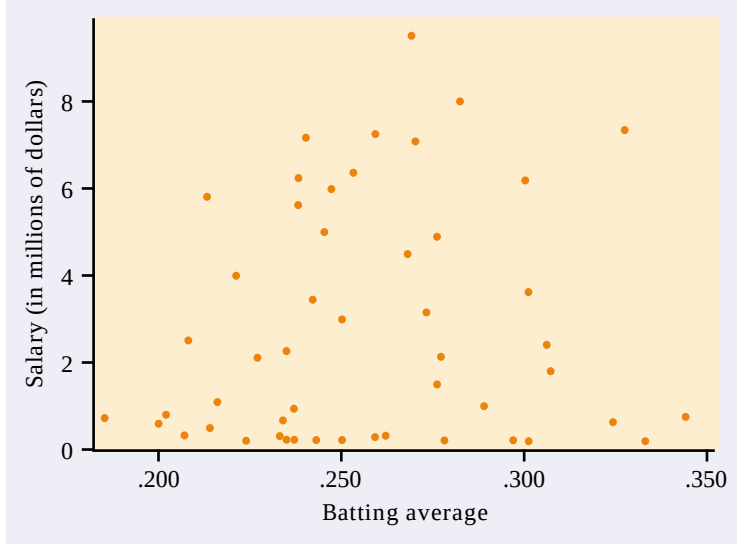


FIGURE 14.15 Career batting average and 2002 salary for a random sample of 50 Major League Baseball players.

How shall we resample from Table 14.2? Because each observation consists of the batting average and salary for one player, we resample players (that is, observations). Resampling batting averages and salaries separately would lose the tie between a player's batting average and his salary. Software such as S-PLUS automates proper resampling. Once we have produced a bootstrap distribution by resampling, we can examine the distribution and form a confidence interval in the usual way. We need no special formulas or procedures to handle the correlation.

Figure 14.16 shows the bootstrap distribution and normal quantile plot for the sample correlation for 1000 resamples from the 50 players in our sample. The bootstrap distribution is close to normal and has small bias, so a 95% bootstrap t confidence interval appears reasonable.

The bootstrap standard error is $\mathrm{SE}_{\mathrm{boot}} = 0.125$. The t interval using the bootstrap standard error is

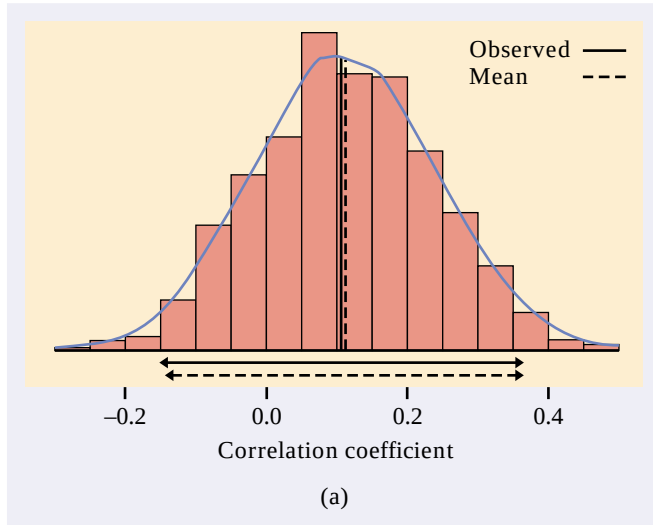
$$\begin{aligned} r \pm t^* \mathrm{SE}_{\mathrm{boot}} &= 0.107 \pm (2.009)(0.125) \\ &= 0.107 \pm 0.251 \\ &= (-0.144, 0.358) \end{aligned}$$

The 95% bootstrap percentile interval is

$$\begin{aligned} (2.5\% \text{ percentile}, 97.5\% \text{ percentile}) &= (-0.128, 0.356) \\ &= (0.107 - 0.235, 0.107 + 0.249) \end{aligned}$$

The two confidence intervals are in reasonable agreement.

The confidence intervals give a wide range for the population correlation, and both include 0. These data do not provide significant evidence that there is a relationship between salary and batting average. A larger sample might result in a significant relationship, but the evidence from this sample suggests



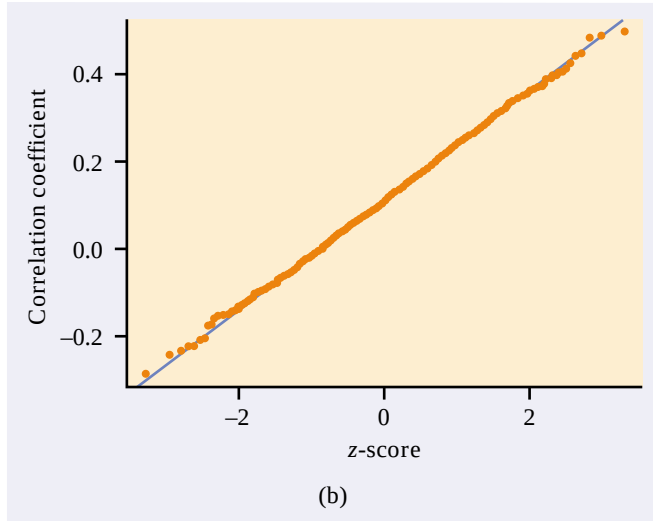


FIGURE 14.16 The bootstrap distribution and normal quantile plot for the correlation r for 1000 resamples from the baseball player data in Table 14.2. The solid double-ended arrow below the distribution is the t interval, and the dashed arrow is the percentile interval.

that any relationship is quite weak. Of course, batting average is only one facet of a player's performance. It is possible that there may be a significant salary-performance relationship if we include several measures of performance.

More accurate bootstrap confidence intervals: BCa and tilting

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 in-

terval 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

accurate in a particular setting if 95% of the time it produces intervals that capture the parameter and if the 5% misses are equally shared between high and low misses. Perfect accuracy isn't available in practice, but some methods are more accurate than others.

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 intervals in Examples 14.8 and 14.9 reveal some right skewness, but not enough to invalidate inference. The bootstrap distribution in Figure 14.9 (page 14-21) for comparing two means, on the other hand, is so skewed that we hesitate to use the *t* or percentile intervals. 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; or
- we require high accuracy because the stakes are high (large sums of money or public welfare).

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 AND TILTING 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.

The **bootstrap tilting interval** adjusts the process of randomly forming resamples (though a clever implementation allows use of the same resamples as other bootstrap methods).

These two methods are accurate in a wide variety of settings, have reasonable computation requirements (by modern standards), and do not produce excessively wide intervals. The BCa intervals are more widely used. Both are based on the big ideas of resampling and the bootstrap distribution. Now that you understand the big ideas, you should always use one of these more accurate methods if your software offers them. We did not meet them earlier

because the details of producing the confidence intervals are quite technical.[10] The BCa method requires more than 1000 resamples for high accuracy. Use 5000 or more resamples if the accuracy of inference is very important. Tilting is more efficient, so that 1000 resamples are generally enough. Don't forget that *even BCa and tilting 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 14.10 The 2002 Seattle real estate sales data are strongly skewed (Figure 14.6). Figure 14.17 shows the bootstrap distribution of the sample mean $\bar{x}$. We see that the skewness persists in the bootstrap distribution and therefore in the sampling distribution. Inference based on a normal sampling distribution is not appropriate.

We generally prefer resistant measures of center such as the median or trimmed mean for skewed data. Accordingly, in Example 14.5 (page 14-18) we bootstrapped the 25% trimmed mean. However, the mean is easily understood by the public and is needed for some purposes, such as projecting taxes based on total sales value.

The bootstrap t and percentile intervals aren't reliable when the sampling distribution of the statistic is skewed. Figure 14.18 shows software output that includes all four of the confidence intervals we have mentioned, along with the traditional one-sample t interval. The BCa interval is

$$(329.3 - 62.2, 329.3 + 127.0) = (267.1, 456.3)$$

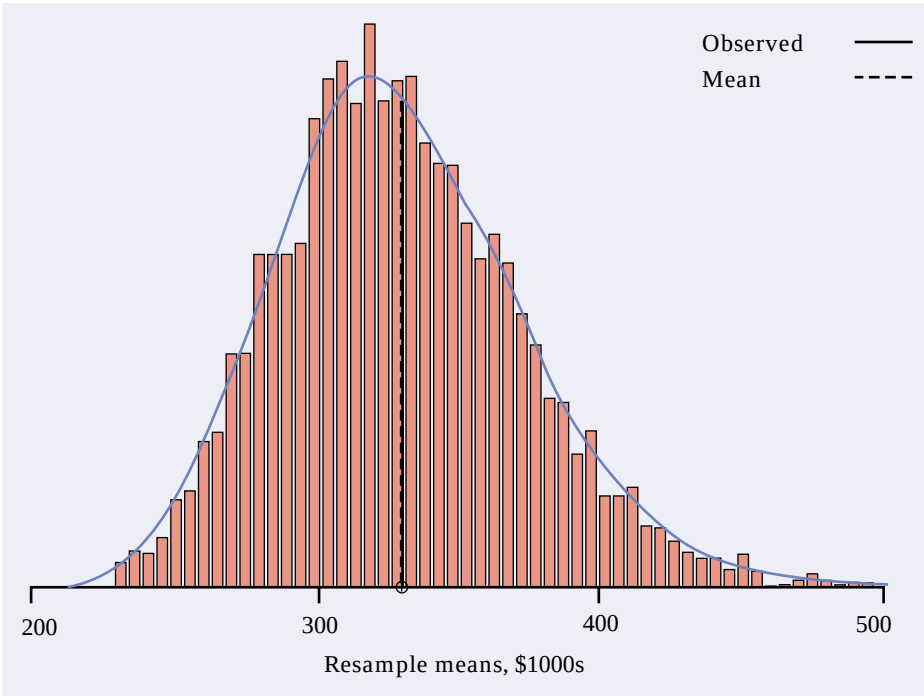


FIGURE 14.17 The bootstrap distribution of the sample means of 5000 resamples from the data in Table 14.1, for Example 14.10. The bootstrap distribution is right-skewed, so we conclude that the sampling distribution of $\bar{x}$ is right-skewed as well.

```
        One-sample t-Test

data:  Price in Seattle2002
t = 7.3484, df = 49, p-value = 0
alternative hypothesis:  mean is not equal to 0
95 percent confidence interval:
 239.2150 419.2992
sample estimates:
 mean of x
  329.2571


Number of Replications: 5000

Summary Statistics:
     Observed   Mean      Bias     SE
mean    329.3  328.4   -0.8679   43.68

Percentiles:
         2.5%       5%      95%   97.5%
mean 253.5264 263.1985 406.4151 425.513

BCa Confidence Intervals:
         2.5%       5%      95%    97.5%
mean 267.0683 275.5542 433.4044 456.2938

Tilting Confidence Intervals:
         2.5%       5%      95%    97.5%
mean 263.1428 272.4917 430.7042 455.2483

T Confidence Intervals using Bootstrap Standard Errors:
         2.5%       5%     95%    97.5%
mean 241.4652 256.0183 402.496 417.0491
```

FIGURE 14.18 S-PLUS output for bootstrapping the mean of the Seattle real estate selling price data, for Example 14.10. The output includes four types of confidence intervals for the population mean.

and the tilting interval is

$$(329.3 - 66.2, 329.3 + 125.9) = (263.1, 455.2)$$

These intervals agree closely. Both are strongly asymmetrical: the upper endpoint is about twice as far from the sample mean as the lower endpoint. This reflects the strong right skewness of the bootstrap distribution.

The output in Figure 14.18 also shows that both endpoints of the less-accurate intervals (one-sample t, bootstrap t, and percentile) are too low. These intervals miss the population mean on the low side too often (more than 2.5%) and miss on the high side too seldom. They give a biased picture of where the true mean is likely to be.

While the BCa and tilting calculations are radically different, the results tend to be about the same, except for random variation in the BCa if the number of resamples is less than about 5000. Both procedures are accurate, so we expect them to produce similar results unless a small sample size makes any inference dubious.

SECTION 14.4 | 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 either a **BCa** or **bootstrap tilting** interval. The t and percentile intervals are inaccurate under these circumstances unless the sample sizes are very large. The tilting and BCa confidence intervals adjust for bias and skewness and are generally accurate except for small samples.

SECTION 14.4 | Exercises

Many of these exercises require software that will calculate accurate bootstrap confidence intervals. If your software finds BCa but not tilting intervals, ignore requests for tilting intervals. S-PLUS supplies both types.

14.26 What percentiles of the bootstrap distribution are the endpoints of a 90% bootstrap percentile confidence interval?

14.27 In Exercise 14.17 (page 14-25) you bootstrapped the mean of a simulated SRS from the standard normal distribution $N(0, 1)$ and found the standard t and bootstrap t 95% confidence intervals for the mean.

(a) Find the bootstrap percentile 95% 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?

14.28 Bootstrapping is a good way to check if traditional inference methods are accurate for a given sample. Consider the following data:

109	123	118	99	121	134	126	114	129	123	171	124
111	125	128	154	121	123	118	106	108	112	103	125
137	121	102	135	109	115	125	132	134	126	116	105
133	111	112	118	117	105	107					

(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?

14.29 The graphs in Figure 14.16 do not appear to show any important skewness in the bootstrap distribution of the correlation for Example 14.9. Compare the bootstrap percentile and bootstrap t intervals for the correlation, given in the discussion of Example 14.9. Does the comparison suggest any skewness?

14.30 Continue to work with the data given in Exercise 14.28.

(a) Find the bootstrap BCa or tilting 95% confidence interval. We believe that these intervals are quite accurate.

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

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

14.31 Find the BCa and tilting 95% confidence intervals for the correlation between baseball salaries and batting averages, from the data in Table 14.2. Are these more accurate intervals in general agreement with the bootstrap t and percentile intervals? Do you still agree with the judgment in the discussion of Example 14.9 that the simpler intervals are adequate?

14.32 The distribution of the 60 IQ test scores in Table 1.3 (page 14) is roughly normal (see Figure 1.5), 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.

(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.

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

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

14.33 The distribution of the 72 guinea pig lifetimes in Table 1.8 (page 38) is strongly skewed. In Exercise 14.9 (page 14-23) you found a bootstrap t confidence interval for the population mean μ, even though some skewness remains in the bootstrap distribution. Bootstrap the mean lifetime and give all four bootstrap 95% confidence intervals: t, percentile, BCa, and tilting. Make a graphical comparison by drawing a vertical line at the original sample mean $\overline{x}$ and displaying the four intervals horizontally, one above the other. Discuss what you see: Do bootstrap t and percentile agree? Do the more accurate intervals agree with the two simpler methods?

14.34 We would like a 95% confidence interval for the standard deviation σ of Seattle real estate prices. Your work in Exercise 14.11 probably suggests that it is risky to bootstrap the sample standard deviation s from the sample in Table 14.1 and use the bootstrap t interval. Now we have more accurate methods. Bootstrap s and report all four bootstrap 95% confidence intervals: t, percentile, BCa, and tilting. Make a graphical comparison by drawing a vertical line at the original s and displaying the four intervals horizontally, one above the other. Discuss what you see: Do bootstrap t and percentile agree? Do the more accurate intervals agree with the two simpler methods? What interval would you use in a report on real estate prices?

14.35 Exercise 14.7 (page 14-13) gives an SRS of 20 of the 72 guinea pig survival times in Table 1.8. The bootstrap distribution of $\overline{x}$ from this sample is clearly right-skewed. 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 14.33, which use the full sample of 72 lifetimes.

14.36 The CLEC data for Example 14.6 are strongly skewed to the right. The 23 CLEC repair times appear in Exercise 14.22 (page 14-26).

(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 four 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 repair time for all CLEC customers by using a t interval or percentile interval instead of a tilting or BCa interval?

14.37 Example 14.6 (page 14-19) considers the mean difference between repair times for Verizon (ILEC) customers and customers of competing carriers (CLECs). The bootstrap distribution is nonnormal with strong left skewness, so that any t confidence interval is inappropriate. Give the BCa 95% confidence interval for the mean difference in service times for all customers. In practical terms, what kind of error would you make by using a t interval or percentile interval instead of a BCa interval?

14.38 Figure 2.3 (page 108) is a scatterplot of field versus laboratory measurements of the depths of 100 defects in the Trans-Alaska Oil Pipeline. The correlation is $r = 0.944$. Bootstrap the correlation for these data. (The data are in the file *ex14_038.dat*.)

(a) Describe the shape and bias of the bootstrap distribution. Do the simpler bootstrap confidence intervals (*t* and percentile) appear to be justified?

(b) Find all four bootstrap 95% confidence intervals: *t*, percentile, BCa, and tilting. Make a graphical comparison by drawing a vertical line at the original correlation *r* and displaying the four intervals horizontally, 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 comparing field and laboratory measurements?

14.39 Figure 2.7 (page 114) shows a very weak relationship between returns on Treasury bills and returns on common stocks. The correlation is $r = -0.113$. We wonder if this is significantly different from 0. To find out, bootstrap the correlation. (The data are in the file *ex14_039.dat*.)

(a) Describe the shape and bias of the bootstrap distribution. It appears that even simple bootstrap inference (*t* and percentile confidence intervals) is justified. Explain why.

(b) Give the BCa and bootstrap percentile 95% 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?

14.40 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. (Software such as S-PLUS automates resampling methods for regression inference.)

14.41 Continue your study of historical returns on Treasury bills and common stocks, begun in Exercise 14.39, by regressing stock returns on T-bill returns.

(a) Request a plot of the residuals against the explanatory variable and a normal quantile plot of the residuals. The residuals are somewhat nonnormal. In what way? It is hard to predict the accuracy of the usual *t* confidence interval for the slope β_1 of the population regression line.

(b) Examine the shape and bias of the bootstrap distribution of the slope b_1 of the least-squares line. The distribution suggests that even the bootstrap *t* interval will be accurate. Why?

(c) Give the standard *t* confidence interval for β_1 and also the BCa, bootstrap *t*, and bootstrap percentile 95% confidence intervals. What do you conclude about the accuracy of the two *t* intervals? Do the data provide evidence at the 5% level that the population slope β_1 is not 0?

14.42 Continue your study of field measurements versus laboratory measurements of defects in the Trans-Alaska Oil Pipeline, begun in Exercise 14.38, by regressing field measurement result on laboratory measurement result.

(a) Request a plot of the residuals against the explanatory variable and a normal quantile plot of the residuals. These plots suggest that inference based on the usual simple linear regression model (Chapter 10, page 638) may be inaccurate. Why?

(b) Examine the bootstrap distribution of the slope b_1 of the least-squares regression line. The distribution shows some departures from normality. In what way is the bootstrap distribution nonnormal? What is the bootstrap estimate of bias? Based on what you see, would you consider use of bootstrap t or bootstrap percentile intervals?

(c) Give the BCa 95% 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?

14.43 Table 14.2 gives data on a sample of 50 baseball players.

(a) Find the least-squares regression line for predicting salary from batting average.

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

(c) In the discussion of Example 14.9 we found bootstrap confidence intervals for the correlation between salary and batting average. Does your interval for the slope of the population line agree with the conclusion of that example that there may be no relation between salary and batting average? Explain.

14.44 We know that outliers can strongly influence statistics such as the mean and the least-squares line. Example 7.7 (page 459) describes a matched pairs study of disruptive behavior by dementia patients. The differences in Table 7.2 show several low values that may be considered outliers.

(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 or tilting confidence interval from both bootstrap distributions. Discuss the differences.

14.5 Significance Testing Using Permutation Tests

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

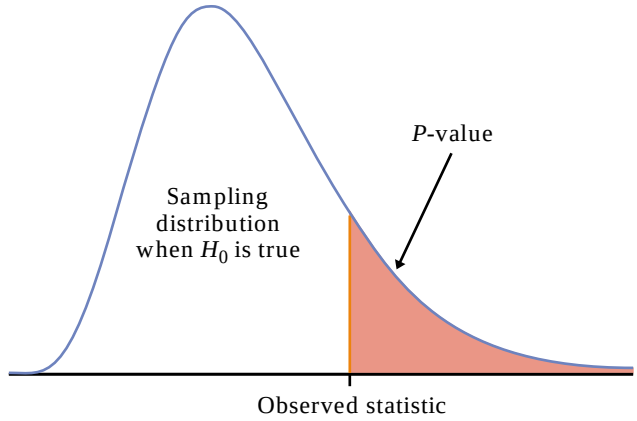


FIGURE 14.19 The *P*-value of a statistical test is found from the sampling distribution the statistic would have if the null hypothesis were true. It is the probability of a result at least as extreme as the value we actually observed.

rarely occur by chance and so is evidence that something other than chance is operating.

null hypothesis

P-value

The statement that the effect we seek is *not* present in the population is the **null hypothesis,** H_0. The probability, calculated taking the null hypothesis to be true, that we would observe a statistic value as extreme or more extreme than the one we did observe is the ***P*-value.** Figure 14.19 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. The reasoning of statistical tests is indirect and a bit subtle but is by now familiar. Tests based on resampling don't change this reasoning. They find *P*-values by resampling calculations rather than from formulas and so can be used in settings where traditional tests don't apply.

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 this 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 14.11 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

TABLE 14.3

Degree of Reading Power scores for third-graders

Treatment group						Control group					
24	61	59	46	43	53	42	33	46	37	62	20
43	44	52	43	57	49	43	41	10	42	53	48
58	67	62	57	56	33	55	19	17	55	37	85
71	49	54				26	54	60	28	42	

(treatment group, 21 students) or traditional teaching methods (control group, 23 students). The DRP scores at the end of the study appear in Table 14.3.[11] In Example 7.14 (page 489) 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 14.3 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.

permutation tests

Here is an outline of the permutation test procedure for comparing the mean DRP scores in Example 14.11:

- 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.** Calculate the mean DRP score in each group, using the individual DRP scores in Table 14.3. The difference between these means is our statistic.

permutation resample

- Repeat this resampling from the 44 students 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

- The value of the statistic actually observed in the study was

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

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

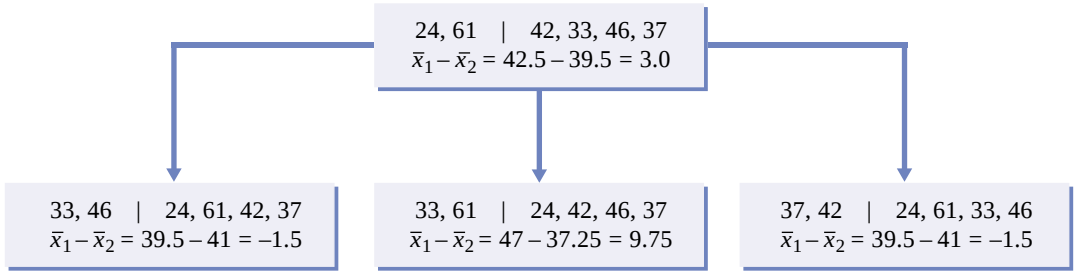


FIGURE 14.20 The idea of permutation resampling. The top box shows the outcomes of a study with four subjects in one group and two in the other. The boxes below show three permutation resamples. The values of the statistic for many such resamples form the permutation distribution.

Figure 14.20 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 of group means) for each.

EXAMPLE 14.12 Figure 14.21 shows the permutation distribution of the difference of means based on 999 permutation resamples from the DRP data in Table 14.3. This is a resampling estimate of the sampling distribution of the statistic

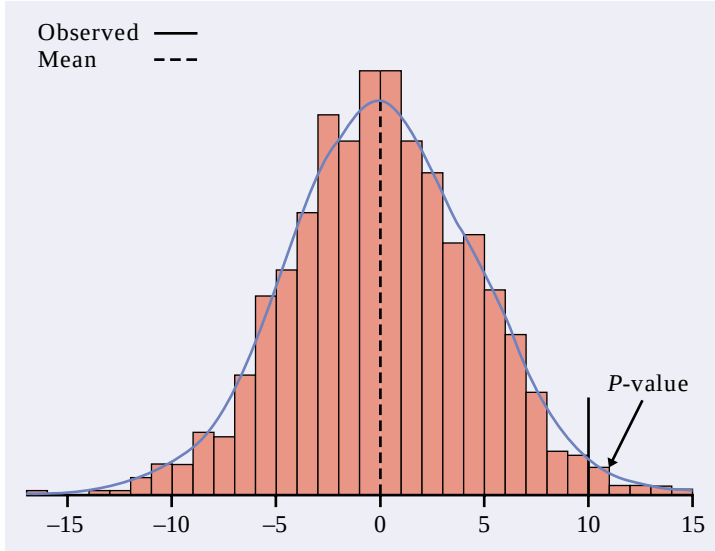


FIGURE 14.21 The permutation distribution of the statistic $\bar{x}_{\text{treatment}} - \bar{x}_{\text{control}}$ based on the DRP scores of 44 students. The dashed line marks the mean of the permutation distribution: it is very close to zero, the value specified by the null hypothesis. The solid vertical 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.

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 $\overline{x}_{\text{treatment}} - \overline{x}_{\text{control}}$ 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 999 resamples that give a result at least as great as 9.954. A look at the resampling results finds that 14 of the 999 resamples gave a value 9.954 or larger, so the estimated *P*-value is 14/999, or 0.014.

Here is a last refinement. Recall from Chapter 8 that we can improve the estimate of a population proportion by adding two successes and two failures to the sample. It turns out that we can similarly improve the estimate of the *P*-value by adding one sample result more extreme than the observed statistic. The final permutation test estimate of the *P*-value is

$$\frac{14+1}{999+1} = \frac{15}{1000} = 0.015$$

The data give good evidence that the new method beats the standard method.

Figure 14.21 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 *t* test in Example 7.14 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 menus in S-PLUS allow you to request permutation tests along with standard tests whenever they make sense. The permutation distribution in Figure 14.21 is one output. Another is this summary of the test results:

```
Number of Replications: 999

Summary Statistics:
      Observed    Mean    SE alternative p.value
score    9.954 0.07153 4.421     greater   0.015
```

By giving "greater" as the alternative hypothesis, the output makes it clear that 0.015 is the one-sided *P*-value.

Permutation tests in practice

Permutation tests versus *t* tests. We have analyzed the data in Table 14.3 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 of sample means estimates the difference of 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 of 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 of 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 of 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 of 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 14.13 In Example 14.6, we looked at the difference in means between repair times for 1664 Verizon (ILEC) customers and 23 customers of competing companies (CLECs). Figure 14.8 (page 14-19) shows both distributions. Penalties are assessed if a significance test concludes at the 1% significance level that CLEC customers are receiving inferior service. The alternative hypothesis is one-sided because the Public Utilities Commission wants to know if CLEC customers are disadvantaged.

Because the distributions are strongly skewed and the sample sizes are very different, two-sample t tests are inaccurate. An inaccurate testing procedure might declare 3% of tests significant at the 1% level when in fact the two groups of customers are treated identically, so that only 1% of tests should in the long run be significant. Errors like this would cost Verizon substantial sums of money.

Verizon performs permutation tests with 500,000 resamples for high accuracy, using custom software based on S-PLUS. Depending on the preferences of each state's regulators, one of three statistics is chosen: the difference in means, $\overline{x}_1 - \overline{x}_2$; the pooled-variance t statistic; or a modified t statistic in which only the standard deviation of the larger group is used to determine the standard error. The last statistic prevents the large variation in the small group from inflating the standard error.

To perform a permutation test, we randomly regroup the total set of repair times into two groups that are the same sizes as the two original samples. This is consistent with the null hypothesis that CLEC versus ILEC has no effect on repair time. Each repair time appears once in the data in each resample, but some repair times from the ILEC group move to CLEC, 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.

Here are the P-values for the three tests on the Verizon data, using 500,000 permutations. The corresponding t test P-values, obtained by comparing the t statistic with t critical values, are shown for comparison.

Test statistic	t test P-value	Permutation test P-value
$\overline{x}_1 - \overline{x}_2$		0.0183
Pooled t statistic	0.0045	0.0183
Modified t statistic	0.0044	0.0195

The t test results are off by a factor of more than 4 because they do not take skewness into account. The t test suggests that the differences are significant at the 1% level, but the more accurate P-values from the permutation test indicate otherwise. Figure 14.22 shows the permutation distribution of the first

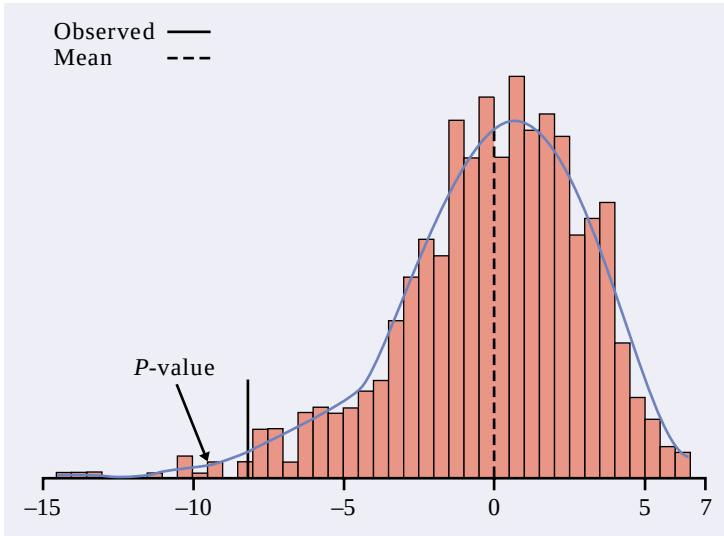


FIGURE 14.22 The permutation distribution of the difference of means $\overline{x}_1 - \overline{x}_2$ for the Verizon repair time data.

statistic, the difference in sample means. The strong skewness implies that *t* tests will be inaccurate.

If you read Chapter 15, on nonparametric tests, you will find there more comparison 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 gender did not matter.

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 nonnormal populations because of the central limit theorem. The central limit theorem helps explain the robustness of the two-sample *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. But *resampling in a way that moves observations between the two groups requires that the two populations are identical when the null hypothesis is true—not only are their means the same, but also their spreads and shapes.* Our preferred version of the two-sample *t* allows different standard deviations in the two groups, so the shapes are both normal but need not have the same spread.

In Example 14.13, the distributions are strongly skewed, ruling out the *t* test. The permutation test is valid if the repair time distributions for Verizon customers and CLEC customers have the same shape, so that they are identical under the null hypothesis that the centers (the means) are the same. Fortunately, the permutation test is robust. That is, it gives accurate *P*-values when the two population distributions have somewhat different shapes, say, 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. For example, Verizon uses 500,000 resamples.

For most purposes, 999 resamples are sufficient. If stakes are high or *P*-values are near a critical value (for example, near 0.01 in the Verizon case), increase the number of resamples. Here is a helpful guideline: If the true (one-sided) *P*-value is p, the standard deviation of the estimated *P*-value is approximately $\sqrt{p(1-p)/B}$, where B is the number of resamples. You can choose B to obtain a desired level of accuracy.

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 14.14 Can the full moon influence behavior? 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 14.4 gives the average number of aggressive incidents for moon days and other days for each subject.[12] These are matched pairs data. In Example 7.7, the 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 nonnormality. 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 14.4 are two measurements of aggressive behavior made under the same conditions. There is no distinction between "moon days" and "other days." Resampling 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.

TABLE 14.4

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			

The permutation test (like the matched pairs t test) uses the difference of means $\bar{x}_{\text{moon}} - \bar{x}_{\text{other}}$. Figure 14.23 shows the permutation distribution of this statistic from 9999 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 therefore

$$P = \frac{0+1}{9999+1} = \frac{1}{10{,}000} = 0.0001$$

There is strong evidence that aggressive behavior is more common on moon days.

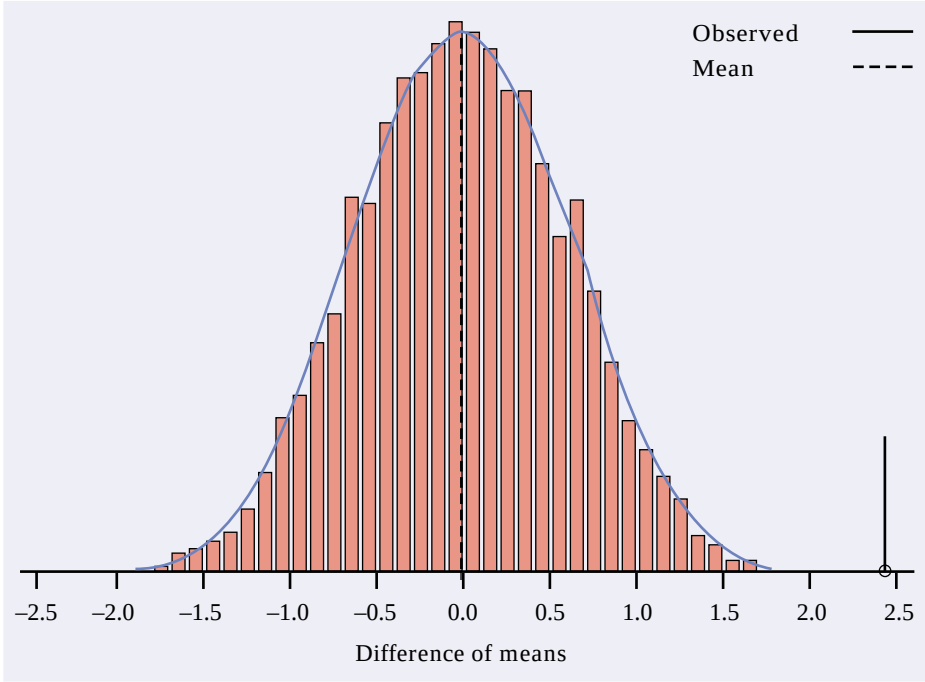


FIGURE 14.23 The permutation distribution for the mean difference (moon days versus other days) from 9999 paired resamples from the data in Table 14.5, for Example 14.14.

The permutation distribution in Figure 14.23 is close to normal, as a normal quantile plot confirms. The paired sample 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 14.9 we looked at the relationship between baseball players' batting averages and salaries.

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

Take the correlation as the test statistic. For every resample, calculate the correlation between the batting averages (in their original order) and salaries (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 14.5 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 the null hypothesis is true, so we use bootstrap resampling rather than permutation resampling.

SECTION 14.5 Exercises

The number of resamples on which a permutation test is based determines the number of decimal places and accuracy in the resulting P-value. Tests based on 999 resamples give P-values to three places (multiples of 0.001), with a margin of error $2\sqrt{P(1-P)/999}$ *equal to 0.014 when the true one-sided P-value is 0.05. If high accuracy is needed or your computer is sufficiently fast, you may choose to use 9999 or more resamples.*

14.45 To illustrate the process, let's perform a permutation test by hand for a small random subset of the DRP data (Example 14.12). Here are the data:

Treatment group	24	61		
Control group	42	33	46	37

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

(b) Resample: Start with the 6 scores and choose an SRS of 2 scores to form the treatment group for the first resample. You can do this by labeling the scores 1 to 6 and using consecutive random digits from Table B or by rolling a die to choose from 1 to 6 at random. Using either method, be sure to skip repeated digits. A resample is an ordinary SRS, without replacement. The remaining 4 scores are the control group. What is the difference of group means for this resample?

(c) Repeat step (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.

14.46 Table 14.1 contains the selling prices for a random sample of 50 Seattle real estate transactions in 2002. Table 14.5 contains a similar random sample of sales in 2001. Test whether the means of the two random samples of the 2001 and 2002 real estate sales data are significantly different.

TABLE 14.5

Selling prices for Seattle real estate, 2001 ($1000s)

419	55.268	65	210	510.728	212.2	152.720	266.6	69.427	125
191	451	469	310	325	50	675	140	105.5	285
320	305	255	95.179	346	199	450	280	205.5	135
190	452.5	335	455	291.905	239.9	369.95	569	481	475
495	195	237.5	143	218.95	239	710	172	228.5	270

(a) State the null and alternative hypotheses.

(b) Perform a two-sample *t* test. What is the *P*-value?

(c) Perform a permutation test on the difference in means. What is the *P*-value? Compare it with the *P*-value you found in part (b). What do you conclude based on the tests?

(d) Find a bootstrap BCa 95% confidence interval for the difference in means. How is the interval related to your conclusion in (c)?

14.47 Here are heights (inches) of professional female basketball players who are centers and forwards. We wonder if the two positions differ in average height.

Forwards

69	72	71	66	76	74	71	66	68	67	70	65	72
70	68	73	66	68	67	64	71	70	74	70	75	75
69	72	71	70	71	68	70	75	72	66	72	70	69

Centers

72	70	72	69	73	71	72	68	68	71	66	68	71
73	73	70	68	70	75	68						

(a) Make a back-to-back stemplot of the data. How do the two distributions compare?

(b) State null and alternative hypotheses. Do a permutation test for the difference in means of the two groups. Give the *P*-value and draw a conclusion.

14.48 A customer complains to the owner of an independent fast-food restaurant that the restaurant is discriminating against the elderly. The customer claims

that people 60 years old and older are given fewer french fries than people under 60. The owner responds by gathering data, collected without the knowledge of the employees so as not to affect their behavior. Here are data on the weight of french fries (grams) for the two groups of customers:

Age < 60:	75	77	80	69	73	76	78	74	75	81
Age ≥ 60:	68	74	77	71	73	75	80	77	78	72

(a) Display the two data sets in a back-to-back stemplot. Do they appear substantially different?

(b) If we perform a permutation test using the mean for "< 60" minus the mean for "≥ 60," should the alternative hypothesis be two-sided, greater, or less? Explain.

(c) Perform a permutation test using the chosen alternative hypothesis and give the *P*-value. What should the owner report to the customer?

14.49 Verizon uses permutation testing for hundreds of comparisons, such as between different time periods, between different locations, and so on. Here is a sample from another Verizon data set, containing repair times in hours for Verizon (ILEC) and CLEC customers.

ILEC															
1	1	1	1	2	2	1	1	1	1	2	2	1	1	1	1
2	2	1	1	1	1	2	3	1	1	1	1	2	3	1	1
1	1	2	3	1	1	1	1	2	3	1	1	1	1	2	3
1	1	1	1	2	3	1	1	1	1	2	4	1	1	1	1
2	5	1	1	1	1	2	5	1	1	1	1	2	6	1	1
1	1	2	8	1	1	1	1	2	15	1	1	1	2	2	
CLEC															
1	1	5	5	5	1	5	5	5	5						

(a) Choose and make data displays. Describe the shapes of the samples and how they differ.

(b) Perform a *t* test to compare the population mean repair times. Give hypotheses, the test statistic, and the *P*-value.

(c) Perform a permutation test for the same hypotheses using the pooled-variance *t* statistic. Why do the two *P*-values differ?

(d) What does the permutation test *P*-value tell you?

14.50 The estimated *P*-value for the DRP study (Example 14.12) based on 999 resamples is $P = 0.015$. For the Verizon study (Example 14.13) the estimated *P*-value for the test based on $\overline{x}_1 - \overline{x}_2$ is $P = 0.0183$ based on 500,000 resamples. What is the approximate standard deviation of each of these estimated *P*-values? (Use each *P* as an estimate of the unknown true *P*-value *p*.)

14.51 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.

Exercises 14.52 to 14.63 concern permutation tests for hypotheses stated in terms of a variety of parameters. In some cases, there are no standard formula-based tests for the hypotheses in question. These exercises illustrate the flexibility of permutation tests.

14.52 Because distributions of real estate prices are typically strongly skewed, we often prefer the median to the mean as a measure of center. We would like to test the null hypothesis that Seattle real estate sales prices in 2001 and 2002 have equal medians. Sample data for these years appear in Tables 14.1 and 14.5. Carry out a permutation test for the *difference in medians*, find the *P*-value, and explain what the *P*-value tells us.

14.53 Exercise 7.41 (page 482) 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.

(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 of means in matched pairs*, using 9999 resamples. The normal quantile plot shows that the permutation distribution is reasonably normal, but the histogram looks a bit odd. What explains the appearance of the histogram? What is the *P*-value for the permutation test? Do your tests in (a) and (c) lead to the same practical conclusion?

14.54 Table 14.2 contains the salaries and batting averages of a random sample of 50 Major League Baseball players. Can we conclude that the *correlation* between these variables is greater than 0 in the population of all players?

(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.

14.55 In Exercise 14.39, we assessed the significance of the *correlation* between returns on Treasury bills and common stocks 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. We would prefer to do a test that gives us 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 14.39?

14.56 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, then measured serum retinol. Of the 90 children in the sample, 55 had been recently infected. Table 14.6 gives the serum retinol levels for both groups, in micromoles per liter.[13]

TABLE 14.6

Serum retinol levels 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					

(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?

14.57 In Exercise 14.56, 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.

14.58 Here is one conclusion from the data in Table 14.6, described in Exercise 14.56: "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. . . ."

(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?

14.59 In Exercise 14.49 we compared the mean repair times for Verizon (ILEC) and CLEC customers. We might also wish to compare the variability of repair times. For the data in Exercise 14.49, the F statistic for comparing sample variances is 0.869 and the corresponding P-value is 0.67. We know that this test is very sensitive to lack of normality.

(a) Perform a two-sided permutation test on the *ratio of standard deviations*. What is the P-value and what does it tell you?

(b) What does a comparison of the two P-values say about the validity of the F test for these data?

14.60 Does added calcium intake reduce the blood pressure of black men? In a randomized comparative double-blind trial, 10 men were given a calcium supplement for twelve weeks and 11 others received a placebo. For each subject, record whether or not blood pressure dropped. Here are the data:[14]

Treatment	Subjects	Successes	Proportion
Calcium	10	6	0.60
Placebo	11	4	0.36
Total	21	10	0.48

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. The permutation test does not depend on a "nice" distribution shape. Give the P-value and report your conclusion.

14.61 We want a 95% confidence interval for the difference in the proportions of reduced blood pressure between a population of black men given calcium and a similar population given a placebo. Summary data appear in Exercise 14.60.

(a) Give the plus four confidence interval. Because the sample sizes are both small, we may wish to use the bootstrap to check this interval.

(b) Bootstrap the sample data. We recommend tilting confidence intervals for proportions based on small samples. Other bootstrap intervals may be inaccurate. Give all four bootstrap confidence intervals (t, percentile, BCa, tilting). How do the other three compare with tilting? How does the tilting interval compare with the plus four interval?

14.62 We prefer measured data to the success/failure data given in Exercise 14.60. Table 14.7 gives the actual values of seated systolic blood pressure for this experiment. Example 7.20 (page 501) applies the pooled t test (a procedure that we do not recommend) to these data. Carry out a permutation test to discover whether the calcium group had a significantly greater decrease in blood pressure.

TABLE 14.7

Effect of calcium and placebo on blood pressure

Calcium group			Placebo group		
Begin	End	Decrease	Begin	End	Decrease
107	100	7	123	124	−1
110	114	−4	109	97	12
123	105	18	112	113	−1
129	112	17	102	105	−3
112	115	−3	98	95	3
111	116	−5	114	119	−5
107	106	1	119	114	5
112	102	10	114	112	2
136	125	11	110	121	−11
102	104	−2	117	118	−1
			130	133	−3

14.63 Are the variances of decreases in blood pressure equal in populations of black men given calcium and given a placebo? Example 7.22 (page 518) applied the *F* test for equality of variances to the data in Table 14.7. This test is unreliable because it is sensitive to nonnormality in the data. The permutation test does not suffer from this drawback.

(a) State the null and alternative hypotheses.

(b) Perform a permutation test using the *F* statistic (ratio of sample variances) as your statistic. What do you conclude?

(c) Compare the permutation test *P*-value with that in Example 7.22. What do you conclude about the *F* test for equality of variances for these data?

14.64 Exercise 7.27 (page 478) gives these data on a delicate measurement of total body bone mineral content made by two operators on the same 8 subjects:[15]

	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 14 | Exercises

14.65 The bootstrap distribution of the 25% trimmed mean for the Seattle real estate sales (Figure 14.7) is not strongly skewed. We were therefore willing in

Examples 14.5 and 14.8 to use the bootstrap t and bootstrap percentile confidence intervals for the trimmed mean of the population. Now we can check these against more accurate intervals. Bootstrap the trimmed mean and give all of the bootstrap 95% confidence intervals: t, percentile, BCa, and tilting. Make a graphical comparison by drawing a vertical line at the original sample mean $\overline{x}$ and displaying the four intervals horizontally, one above the other. Describe how the intervals compare. Do you still regard the bootstrap t and percentile intervals as adequately accurate?

14.66 Exercise 7.29 (page 479) reports the changes in reasoning scores of 34 preschool children after six months of piano lessons. Here are the changes:

2	5	7	−2	2	7	4	1	0	7	3	4	3	4	9	4	5
2	9	6	0	3	6	−1	3	4	6	7	−2	7	−3	3	4	4

(a) Make a histogram and normal quantile plot of the data. Is the distribution approximately normal?

(b) Find the sample mean and its standard error using formulas.

(c) Bootstrap the mean and find the bootstrap standard error. Does the bootstrap give comparable results to theoretical methods?

14.67 Your software can generate random numbers that have the uniform distribution on 0 to 1. Figure 4.9 (page 283) 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 bootstrap t 95% confidence interval.

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

14.68 A fitness center employs 20 personal trainers. Here are the ages in years of the female and male personal trainers working at this center:

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?

14.69 Exercise 2.9 (page 116) describes a study that suggests that the "pain" caused by social rejection really is pain, in the sense that it causes activity in brain

areas known to be activated by physical pain. Here are data for 13 subjects on degree of social distress and extent of brain activity:[16]

Subject	Social distress	Brain activity	Subject	Social distress	Brain activity
1	1.26	−0.055	8	2.18	0.025
2	1.85	−0.040	9	2.58	0.027
3	1.10	−0.026	10	2.75	0.033
4	2.50	−0.017	11	2.75	0.064
5	2.17	−0.017	12	3.33	0.077
6	2.67	0.017	13	3.65	0.124
7	2.01	0.021			

Make a scatterplot of brain activity against social distress. There is a positive linear association, with correlation $r = 0.878$. Is this correlation significantly greater than 0? Use a permutation test.

14.70 Use the bootstrap to obtain a 95% confidence interval for the correlation in the population of all similar subjects from the data in the previous exercise.

(a) The permutation distribution in the previous exercise was reasonably normal, with somewhat short tails. The bootstrap distribution is very nonnormal: most resample correlations are not far from 1, the largest value that a correlation can have. Explain carefully why the two distributions differ in shape. Also explain why we might expect a bootstrap distribution to have this shape when the observed correlation is $r = 0.878$.

(b) Choose an appropriate bootstrap confidence interval for the population correlation and state your conclusion.

14.71 We have compared the selling prices of Seattle real estate in 2002 (Table 14.1) and 2001 (Table 14.5). Let's compare 2001 and 2000. Here are the prices (thousands of dollars) for 20 random sales in Seattle in the year 2000:

333	126.5	207.5	199.5	1836	360	175	133	1100	203
194.5	140	280	475	185	390	242	276	359	163.95

(a) Plot both the 2000 and the 2001 data. Explain what conditions needed for a two-sample t test are violated.

(b) Perform a permutation test to find the P-value for the difference in means. What do you conclude about selling prices in 2000 versus 2001?

14.72 Exercise 7.37 (page 481) gives the following readings for 12 home radon detectors when exposed to 105 picocuries per liter of radon:

91.9	97.8	111.4	122.3	105.4	95.0
103.8	99.6	96.6	119.3	104.8	101.7

Part (a) of Exercise 7.37 judges that a *t* confidence interval can be used despite the skewness of the data.

(a) Give a formula-based 95% *t* interval for the population mean.

(b) Find the bootstrap 95% BCa or tilting interval for the mean.

(c) Look at the bootstrap distribution. Is it approximately normal in appearance?

(d) Do you agree that the *t* interval is robust enough in this case? Why or why not? Keep in mind that whether the confidence interval covers 105 is important for the study's purposes.

14.73 The study described in the previous exercise used a one-sample *t* test to see if the mean reading of all detectors of this type differs from the true value 105. Can you replace this test by a permutation test? If so, carry out the test and compare results. If not, explain why not.

14.74 In financial theory, the standard deviation of returns on an investment is used to describe the risk of the investment. The idea is that an investment whose returns are stable over time is less risky than one whose returns vary a lot. The data file *ex14_074.dat* contains the returns (in percent) on 1129 consecutive days for a portfolio that weights all U.S. common stocks according to their market value.[17]

(a) What is the standard deviation of these returns?

(b) Bootstrap the standard deviation. What is its bootstrap standard error?

(c) Find the 95% bootstrap *t* confidence interval for the population standard deviation.

(d) Find the 95% tilting or BCa confidence interval for the standard deviation. Compare the confidence intervals and give your conclusions about the appropriateness of the bootstrap *t* interval.

14.75 Nurses in an inner-city hospital were unknowingly observed on their use of latex gloves during procedures for which glove use is recommended.[18] The nurses then attended a presentation on the importance of glove use. One month after the presentation, the same nurses were observed again. Here are the proportions of procedures for which each nurse wore gloves:

Nurse	Before	After	Nurse	Before	After
1	0.500	0.857	8	0.000	1.000
2	0.500	0.833	9	0.000	0.667
3	1.000	1.000	10	0.167	1.000
4	0.000	1.000	11	0.000	0.750
5	0.000	1.000	12	0.000	1.000
6	0.000	1.000	13	0.000	1.000
7	1.000	1.000	14	1.000	1.000

(a) Why is a one-sided alternative proper here? Why must matched pairs methods be used?

(b) Do a permutation test for the difference in means. Does the test indicate that the presentation was helpful?

14.76 In the previous exercise, you did a one-sided permutation test to compare means before and after an intervention. If you are mainly interested in whether or not the effect of the intervention is significant at the 5% level, an alternative approach is to give a bootstrap confidence interval for the mean difference within pairs. If zero (no difference) falls outside the interval, the result is significant. Do this and report your conclusion.

14.77 Examples 8.9 (page 557) and 8.11 (page 562) examine survey data on binge drinking among college students. Here are data on the prevalence of frequent binge drinking among female and male students:[19]

Gender	Sample size	Binge drinkers
Men	7,180	1,630
Women	9,916	1,684
Total	17,096	3,314

The sample is large, so that traditional inference should be accurate. Nonetheless, use resampling methods to obtain

(a) a 95% confidence interval for the proportion of all students who are frequent binge drinkers.

(b) a test of the research hypothesis that men are more likely than women to be frequent binge drinkers.

(c) a 95% confidence interval for the difference in the proportions of men and of women who are frequent binge drinkers.

14.78 Is there a difference in the readability of advertisements in magazines aimed at people with varying educational levels? Here are word counts in randomly selected ads from magazines aimed at people with high and middle education levels.[20]

Education level	Word count								
High	205	203	229	208	146	230	215	153	205
	80	208	89	49	93	46	34	39	88
Medium	191	219	205	57	105	109	82	88	39
	94	206	197	68	44	203	139	72	67

(a) Make histograms and normal quantile plots for both data sets. Do the distributions appear approximately normal? Find the means.

(b) Bootstrap the means of both data sets and find their bootstrap standard errors.

(c) Make histograms and normal quantile plots of the bootstrap distributions. What do the plots show?

(d) Find the 95% percentile and tilting intervals for both data sets. Do the intervals for high and medium education level overlap? What does this indicate?

(e) Bootstrap the difference in means and find a 95% percentile confidence interval. Does it contain 0? What conclusions can you draw from your confidence intervals?

14.79 The researchers in the study described in the previous exercise expected higher word counts in magazines aimed at people with high education level. Do a permutation test to see if the data support this expectation. State hypotheses, give a *P*-value, and state your conclusions. How do your conclusions here relate to those from the previous exercise?

14.80 The following table gives the number of burglaries per month in the Hyde Park neighborhood of Chicago for a period before and after the commencement of a citizen-police program:[21]

Before										
60	44	37	54	59	69	108	89	82	61	47
72	87	60	64	50	79	78	62	72	57	57
61	55	56	62	40	44	38	37	52	59	58
69	73	92	77	75	71	68	102			
After										
88	44	60	56	70	91	54	60	48	35	49
44	61	68	82	71	50					

(a) Plot both sets of data. Are the distributions skewed or roughly normal?

(b) Perform a one-sided (which side?) *t* test on the data. Is there statistically significant evidence of a decrease in burglaries after the program began?

(c) Perform a permutation test for the difference in means, using the same alternative hypothesis as in part (b). What is the *P*-value? Is there a substantial difference between this *P*-value and the one in part (b)? Use the shapes of the distributions to explain why or why not. What do you conclude from your tests?

(d) Now do a permutation test using the opposite one-sided alternative hypothesis. Explain what this is testing, why it is not of interest to us, and why the *P*-value is so large.

14.81 The previous exercise applied significance tests to the Hyde Park burglary data. We might also apply confidence intervals.

(a) Bootstrap the difference in mean monthly burglary counts. Make a histogram and a normal quantile plot of the bootstrap distribution and describe the distribution.

(b) Find the bootstrap standard error, and use it to create a 95% bootstrap t confidence interval.

(c) Find the 95% percentile confidence interval. Compare this with the t interval. Does the comparison suggest that these intervals are accurate? How do the intervals relate to the results of the tests in the previous exercise?

CHAPTER 14 | Notes

1. S-PLUS is a registered trademark of Insightful Corporation.

2. Verizon repair time data used with the permission of Verizon.

3. 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."

4. 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.

5. T. Bjerkedal, "Acquisition of resistance in guinea pigs infected with different doses of virulent tubercle bacilli," *American Journal of Hygiene*, 72 (1960), pp. 130–148.

6. Seattle real estate sales data provided by Stan Roe of the King County Assessor's Office.

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

8. From the *Forbes* Web site, `www.forbes.com`.

9. From `www.espn.com`, July 2, 2002.

10. 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.

11. 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.

12. 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.

13. Data provided by Francisco Rosales of the Department of Nutritional Sciences, Penn State University. See 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.

14. Roseann M. Lyle, et al., "Blood pressure and metabolic effects of calcium supplementation in normotensive white and black men," *Journal of the American Medical Association,* 257 (1987), pp. 1772–1776.

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

16. Data from a plot in Naomi I. Eisenberger, Matthew D. Lieberman, and Kipling D. Williams, "Does rejection hurt? An fMRI study of social exclusion," *Science,* 302 (2003), pp. 290–292.

17. These are daily returns from January 1990 through part of May 2004 for the CREF Equity Index Fund, which is a good proxy for all U.S. common stocks. The returns were calculated from net asset values appearing on the TIAA-CREF Web site, `www.tiaa-cref.org`.

18. L. Friedland et al., "Effect of educational program on compliance with glove use in a pediatric emergency department," *American Journal of Diseases of Childhood,* 146 (1992), 1355–1358.

19. Results of this survey are reported in Henry Wechsler et al., "Health and behavioral consequences of binge drinking in college," *Journal of the American Medical Association*, 272 (1994), pp. 1672–1677.

20. F. K. Shuptrine and D. D. McVicker, "Readability levels of magazine ads," *Journal of Advertising Research,* 21, No. 5 (1981), p. 47.

21. G. V. Glass, V. L. Wilson, and J. M. Gottman, *Design and Analysis of Time Series Experiments,* Colorado Associated University Press, 1975.