



Jenkedco/Shutterstock

CHAPTER

32

Resampling: Permutation Tests and the Bootstrap

Sampling distributions provide an important tool for statistical inference, and answer the question, “What would happen in many samples?” When our data were quantitative measurements from a single population, we answered this question by first assuming that our data were a random sample from a Normal population. Because our inferences were about means, they were based on the sampling distribution of \bar{x} , and to construct this sampling distribution, we thought of our data as one of many possible samples that we could have obtained from this Normal population. Under this set of assumptions, there are simple formulas for inference based on the t distribution. These inferences are useful in practice because they are *robust*, although we cannot use these methods for data that are strongly skewed unless the samples are large. Strong outliers can also affect our conclusions.

The techniques of this chapter allow us to weaken some of these assumptions. Both **permutation tests** and the **bootstrap** are examples of **resampling** methods. The common element in both of these methods is the use of the individual observations in the sample to construct the relevant sampling distribution for inference. Without further assumptions about the populations from which the data were drawn, the construction of these sampling distributions is specific to the *observed* sample data and requires software to automate the required computations in all but the smallest examples.

In this chapter, we cover...

- 32.1 Randomization in experiments as a basis for inference
- 32.2 Permutation tests for comparing two treatments with software
- 32.3 Generating bootstrap samples
- 32.4 Bootstrap standard errors and confidence intervals

permutation tests
bootstrap
resampling

Permutation tests are useful for designed experiments in which the treatments are assigned at random to the subjects. They use the randomization in the experimental design to construct the sampling distribution, and these sampling distributions are valid *without* the assumption of Normal distributions for any sample sizes. However, there are no longer simple formulas for the test statistics, and software is required except for situations with very small sample sizes.

Bootstrap methods also allow us to weaken some of the traditional assumptions, and we can make inferences on the population from which the data are a random sample, regardless of the shape of the population distribution. In addition, the bootstrap is a very general method that allows us to make inferences about parameters other than means without the need to modify the basic bootstrap technique.

32.1 Randomization in Experiments as a Basis for Inference

An experiment that uses both comparison of two or more treatments and random assignment of subjects to treatments is called a *randomized comparative experiment* (see Section 9.4). We make no assumptions about how subjects are selected to take part in the experiment and only make use of the randomness in the assignment of the subjects to the treatments. Here is an example of a small experiment in which we have six subjects that are randomly assigned to two treatments.

EXAMPLE 32.1 A Completely Randomized Design

Suppose you have three men—Ari, Luis, and Troy—and three women—Ana, Deb, and Hui—for an experiment. Three of the six subjects are to be assigned completely at random to a new experimental weight loss treatment and three to a placebo. Here are all 20 possible ways of selecting three of these subjects for the treatment group (the remaining three are in the placebo group).

Treatment Group	Treatment Group
Ari, Luis, Troy	Luis, Troy, Ana
Ari, Luis, Ana	Luis, Troy, Deb
Ari, Luis, Deb	Luis, Troy, Hui
Ari, Luis, Hui	Luis, Ana, Deb
Ari, Troy, Ana	Luis, Ana, Hui
Ari, Troy, Deb	Luis, Deb, Hui
Ari, Troy, Hui	Troy, Ana, Deb
Ari, Ana, Deb	Troy, Ana, Hui
Ari, Ana, Hui	Troy, Deb, Hui
Ari, Deb, Hui	Ana, Deb, Hui

With a completely randomized design, each of these 20 possible treatment groups is equally likely; thus each has probability 1/20 of being the actual group assigned to the treatment. Notice that the chance that all the men are assigned to the treatment group is 1/20, and the chance that the treatment group consists of either all men or all women is 2/20.

Although we have made no assumptions other than the random assignment of the subjects to the treatments, our inferences do require one additional assumption. If there is *no difference* in the effects of the new treatment and the placebo on weight loss, then the weight lost by any subject should be the same regardless of whether

the subject received the new treatment or a placebo. Here are the observed weight losses (in pounds) for each subject:

Subject	Ari	Luis	Troy	Ana	Deb	Hui
Weight loss	2	15	8	1	12	9

We emphasize that if the effects of the new treatment and the placebo on weight loss did not differ, we would have observed these same weight losses regardless of who was assigned the treatment and who the placebo. Here is how we can use this fact plus the randomization to test the hypothesis that there is no difference in the effects of the new treatment and the placebo against the alternative that the new treatment produces a larger weight loss than the placebo. Notice that this is a one-sided alternative.

EXAMPLE 32.2 A Completely Randomized Design, continued

We will use the mean weight lost by those receiving the new treatment minus the mean weight lost by those receiving the placebo to test the null hypothesis that there is no difference in the effects of the new treatment and the placebo on weight loss. If there is no difference between the new treatment and the placebo, the observed difference in means is the result of the “luck of the draw”—namely which three subjects happened, by chance, to be assigned to the treatment group. In Example 32.1, we listed all 20 possible ways of selecting three of these subjects for the treatment group (the remaining three are in the placebo group). With a completely randomized design, each of these 20 possible treatment groups is equally likely; thus each has probability $1/20$ of being the actual group assigned to the treatment. We can use this information to determine the sampling distribution of the possible observed differences in mean weight loss. For example, if Ari, Luis, and Troy are assigned to the treatment group, the mean weight loss for the group is $\frac{2+15+8}{3} = 8.33$. The mean weight loss for the placebo group (Ana, Deb, and Hui) is then $\frac{1+12+9}{3} = 7.33$. The difference in mean weight losses (treatment group weight loss minus control group weight loss) is $8.33 - 7.33 = 1.00$.

We can repeat this calculation for each possible assignment of subjects to experimental groups. For the 20 possible ways we can assign subjects to the treatment and placebo groups, the differences in the mean weight losses for the two groups are:

Treatment Group	Control Group	Difference in Mean Weight Loss
Ari, Luis, Troy	Ana, Deb, Hui	1.00
Ari, Luis, Ana	Troy, Deb, Hui	−3.67
Ari, Luis, Deb	Troy, Ana, Hui	3.67
Ari, Luis, Hui	Troy, Ana, Deb	1.67
Ari, Troy, Ana	Luis, Deb, Hui	−8.33
Ari, Troy, Deb	Luis, Ana, Hui	−1.00
Ari, Troy, Hui	Luis, Ana, Deb	−3.00
Ari, Ana, Deb	Luis, Troy, Hui	−5.67
Ari, Ana, Hui	Luis, Troy, Deb	−7.67
Ari, Deb, Hui	Luis, Troy, Ana	−0.33
Luis, Troy, Ana	Ari, Deb, Hui	0.33

(Continued)



Treatment Group	Control Group	Difference in Mean Weight Loss
Luis, Troy, Deb	Ari, Ana, Hui	7.67
Luis, Troy, Hui	Ari, Ana, Deb	5.67
Luis, Ana, Deb	Ari, Troy, Hui	3.00
Luis, Ana, Hui	Ari, Troy, Deb	1.00
Luis, Deb, Hui	Ari, Troy, Ana	8.33
Troy, Ana, Deb	Ari, Luis, Hui	−1.67
Troy, Ana, Hui	Ari, Luis, Deb	−3.67
Troy, Deb, Hui	Ari, Luis, Ana	3.67
Ana, Deb, Hui	Ari, Luis, Troy	−1.00

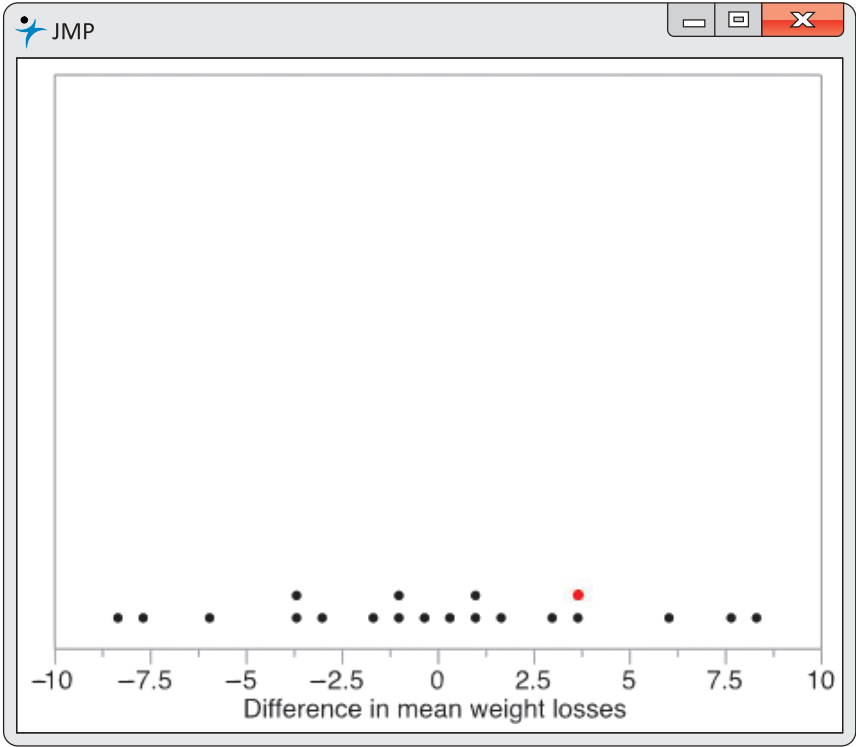
Ordering the differences in mean weight losses from low to high, recalling that each assignment of subjects to the treatment groups has probability $1/20 = 0.05$, and combining duplicates, we obtain:

Weight loss	−8.33	−7.67	−5.67	−3.67	−3.00	−1.67	−1.00	−0.33
Probability	0.05	0.05	0.05	0.10	0.05	0.05	0.10	0.05
Weight loss	0.33	1.00	1.67	3.00	3.67	5.67	7.67	8.33
Probability	0.05	0.10	0.05	0.05	0.10	0.05	0.05	0.05

This is the sampling distribution of the differences in mean weight losses for the two groups under the assumption that weight lost by a subject does not depend on the group to which the subject was assigned. It is derived from considering all possible random assignments of subjects to experimental groups and is referred to as the **permutation distribution**. From this permutation distribution, we can determine whether the actual observed difference is statistically significant. Figure 32.1 is a dotplot of this sampling distribution.

permutation distribution

FIGURE 32.1
Output from JMP, for Example 32.2. The output gives a dotplot of the difference in mean weight losses for the 20 possible ways of selecting three of the subjects for the treatment group. The red dot corresponds to the case where Troy, Deb, and Hui are assigned to the treatment group. Statistical analysis relies heavily on statistical software, and JMP is one of the most popular software choices both in industry and in colleges and schools of business. Computer output from other statistical packages like Minitab, SPSS, and R is similar, so you can feel comfortable using any one of these packages.



What would we conclude if we conducted this experiment and Troy, Deb, and Hui were assigned to the group receiving the new treatment? In this case, the difference in mean weight losses for the two groups would be 3.67. This point is in red in Figure 32.1. The mean weight lost for the treatment group is greater than for the placebo group, and this might be regarded as evidence that the new treatment is effective. However, if there is no difference in the effects of the new treatment and the placebo, Example 32.2 suggests that the chance of observing a difference as large as or larger than 3.67 is 0.25. Would you regard an effect that has probability 0.25 of occurring by chance as being rare? If not, then you would not consider the results to be statistically significant. The sample sizes are too small for all but the most extreme data to achieve statistical significance.

Assumptions for a Simple Permutation Test

- Treatments are assigned to experimental units by a randomized design. **The randomization method is important when assessing the significance of the result.**
- We test the null hypothesis of no difference in the effect of the treatments on a response.

If there is no difference in the effects of the treatments, then the response measured for a unit will be the same regardless of the treatment received.

A Simple Permutation Test Procedure

We run a randomized experiment to compare the effects of treatments on a response. To assess these effects, we use a statistic whose magnitude increases as the difference in the effects of the treatments increases. To carry out a permutation test of the null hypothesis of no difference in the effects of the treatments:

- Determine the probability of every possible assignment of treatments to experimental units.
- For each assignment, calculate the value of the statistic under the null hypothesis.
- The probability of each possible assignment and the value of the statistic for that assignment is the permutation distribution of the statistic under the null hypothesis. This permutation distribution determines the *P*-value of the data resulting from the assignment actually obtained.

In Example 32.2, we assumed the treatments were assigned to units using a completely randomized design and computed the permutation distribution of the difference in the sample means of the treatments under the null hypothesis of no difference in the effect of the treatments. We now apply these principles to conduct a permutation test for data from a matched pairs experiment.

EXAMPLE 32.3 A Permutation Test for Paired Data

West Nile virus, chikungunya, Rocky Mountain spotted fever, and Lyme disease are becoming increasingly common insect-borne diseases in North America. Insect repellents can provide protection from bites of insects that carry these diseases, but which are the most effective? To investigate, we compare two insect repellents. The active ingredient in one is 15% DEET. The active ingredient in the other is oil of lemon eucalyptus. Repellents are tested on four volunteers. For each volunteer, the





left arm is sprayed with one of the repellents and the right arm with the other. Which arm receives which repellent is determined randomly. Beginning 30 minutes after applying the repellents, once every hour volunteers put both arms in separate 8-cubic-foot cages containing 200 disease-free female mosquitoes in need of a blood meal to lay their eggs. Volunteers leave their arms in the cage for five minutes. The repellent is considered to have failed if a volunteer was bitten two or more times in the five-minute period during a one-hour session or at least once in the five-minute periods in two consecutive one-hour sessions. The response is the number of one-hour sessions until a repellent fails.¹ The four volunteers are Erin, Raj, Todd, and Wanda. Here are the number of hours until failure for the two repellents.

Subject	Erin	Raj	Todd	Wanda
DEET	5	7	4	4
Oil of lemon eucalyptus	8	7	6	7
Difference (DEET minus oil of lemon eucalyptus)	−3	0	−2	−3

This is a matched pairs design with the two observations on each subject forming the matched pair. We use the difference between the paired responses (DEET minus oil of lemon eucalyptus) to determine if there is a difference in the effect of the treatments.

Under the assumption of no difference in the effects of the two treatments, the number of hours until failure for any arm of any subject would be the same regardless of the repellent applied to the arm. For example, suppose in the experiment Erin's left arm received DEET and her right arm received oil of lemon eucalyptus. If there were no difference in the effects of the two treatments, we would have obtained the same responses if her left arm had received oil of lemon eucalyptus and her right arm DEET. In other words, the number of hours until failure for the left arm would still be five (even though it received oil of lemon eucalyptus rather than DEET) and would still be eight for the right arm (even though it received DEET rather than oil of lemon eucalyptus). However, the observed difference would be 3 rather than −3. The observed difference of −3 is simply the result of the random assignment of repellents to arms. It was just as likely to have been 3.

Assignment of repellents to arms was determined by chance, and under the assumption of no difference in the effect of the two treatments, the responses actually observed were one of 16 equally likely possible outcomes. These 16 possible outcomes correspond to the 16 possible ways in which the repellents could have been assigned to the left and right arms of the four volunteers (two possibilities for Erin, two for Raj, two for Todd, and two for Wanda).

For each subject, there are two possible ways repellents could have been assigned to the left and right arm. Both are equally likely and the absolute value of the difference between the response to DEET and the response to oil of lemon eucalyptus will be the same. All that changes is whether the difference is negative or positive. Thus, a simpler way of thinking about the 16 equally likely assignments of repellents to the two arms is to think of the magnitudes of the differences as fixed numbers 0, 2, 3, and 3, with the signs of the differences being determined by the randomization performed on each subject. The largest possible mean of the differences would be $\frac{0+2+3+3}{4} = 2.0$ if the arm with the larger response had received DEET for each subject, while the smallest mean of the difference would be $\frac{(0) + (-2) + (-3) + (-3)}{4} = -2.0$ if the arm with the smaller response had received DEET for each subject. Here are all 16 possible randomizations with the mean of the difference for each possible randomization. The first row gives the magnitude of the difference for each subject in parentheses following the name.

Erin (3)	Raj (0)	Todd (2)	Wanda (3)	Mean of the Differences
+	+	+	+	2.0
+	+	+	−	0.5
+	+	−	+	1.0
+	−	+	+	2.0
−	+	+	+	0.5
+	+	−	−	−0.5
+	−	+	−	0.5
−	+	+	−	−1.0
+	−	−	+	1.0
−	+	−	+	−0.5
−	−	+	+	0.5
+	−	−	−	−0.5
−	+	−	−	−2.0
−	−	+	−	−1.0
−	−	−	+	−0.5
−	−	−	−	−2.0

Ordering the differences in mean hours until failure from low to high, recalling that each assignment of treatments to the arms of subjects has probability $\frac{1}{16} = 0.0625$, and combining duplicates, we obtain:

Mean difference	−2.0	−1.0	−0.5	0.5	1.0	2.0
Probability	0.125	0.125	0.250	0.250	0.125	0.125

This is the permutation distribution of the differences in mean hours until failure for the two repellents under the assumption that hours until failure for the arm of a subject do not depend on the treatment to which the arm of the subject was assigned. From this permutation distribution, we can determine whether the observed difference in the treatment means is statistically significant.

For a two-sided test of no difference in the effect of the two repellents, the *P*-value is the probability of obtaining a mean difference as or more extreme (as far or farther from 0) than we actually observed. In the experiment, the observed mean difference was -2.0 . From Example 32.3, the possible outcomes as far as or farther from 0 are -2.0 and 2.0 . The probability that we would observe these values by chance if there were no difference in the effects of the treatments is the sum of the probabilities of these two values—namely, $0.125 + 0.125 = 0.250$. An outcome that has probability 0.25 of occurring by chance would not be considered statistically significant, and we would not regard these data as evidence of a statistically significant difference in the effect of DEET and oil of lemon eucalyptus.

Notice that the most extreme outcomes possible are differences of either -2.0 or 2.0 . From Example 32.3, we see that the probability of observing one of these outcomes is $0.125 + 0.125 = 0.250$. A *P*-value of 0.250 would not typically be considered statistically significant. With only four subjects in a matched pairs experiment, we are not able to demonstrate statistical significance even at the 0.1 level

using a permutation test. More units are needed, but with more units the number of possible permutations grows rapidly and it becomes difficult to enumerate all possibilities. Software is needed to handle larger sample sizes, and we will discuss this in the next section.

Macmillan Learning Online Resources

- The StatBoards video, *Permutation Tests*, discusses additional examples.
- The Snapshots video, *Resampling Procedures*, includes some discussion of permutation tests.

APPLY YOUR KNOWLEDGE

32.1 A Very Simple Setting. Does taking notes by hand in a statistics course improve performance? Some recent research suggests that this may be the case.² To explore this, six volunteers (Doug, Elizabeth, Oksana, Sebastian, Vishal, and Xinyi) agree to take part in an experiment. Four are assigned completely at random to take handwritten notes in class, and the other two are assigned to take notes on their laptops. Total points earned on the two in-class exams and final exam are used to determine course performance. The results are (out of a possible total of 500 points):

Handwritten Notes (Person)	Notes on Laptop (Person)
380 (Doug)	370 (Elizabeth)
400 (Oksana)	310 (Xinyi)
420 (Sebastian)	
360 (Vishal)	

- There are 15 possible ways the six subjects can be assigned to the two groups, with the handwritten notes group having size 4 and the laptop notes group size 2. List these.
- For each, determine the difference in mean points (mean number of points for the handwritten notes group minus mean number of points for the laptop notes group). Combine any duplicates and make a table of the possible mean differences and the corresponding probability of each under the null hypothesis of no difference in the effect of the treatments on total points earned. (Each of the 15 possible assignments of subjects to treatments has probability $1/15$ under the null hypothesis.) This is the permutation distribution.
- Compute the P -value of the data. Assume the two-sided alternative hypothesis is that the mean number of points is different for the two groups.
- In this example, is it possible to demonstrate significance at the 5% level using the permutation test? Explain.
- Assume that total number of points is Normally distributed for both groups. Use the two-sample t procedure to test the hypotheses. Use Option 1 if you have access to software.

32.2 Growing Trees Faster. The concentration of carbon dioxide (CO_2) in the atmosphere is increasing rapidly due to our use of fossil fuels. Because plants use CO_2 to fuel photosynthesis, more CO_2 may cause trees and other plants to grow faster. An elaborate apparatus allows researchers to pipe extra CO_2 to a 30-meter circle of forest. They selected two nearby circles in each of three parts of a pine forest and randomly chose one of each pair to receive extra CO_2 .

The response variable is the mean increase in base area for 30 to 40 trees in a circle during a growing season. We measure this in percent increase per year. Here are one year's data:³

Pair	Control Plot	Treated Plot	Treated–Control
1	9.752	10.587	0.835
2	7.263	9.244	1.981
3	5.742	8.675	2.933

- Explain why this is a matched pairs design.
- State the null and alternative hypotheses. Explain clearly why the investigators used a one-sided alternative.
- Under the assumption of no difference in the effects of extra CO₂ and no extra CO₂ on the response, list all possible outcomes that could occur as a result of the possible random assignments of treatments to the different matched pairs. From this list, construct the sampling (permutation) distribution.
- What is the *P*-value for the one-sided test of no difference in the effects of extra CO₂ and no extra CO₂ on the response?

32.2 Permutation Tests for Two Treatments with Software

The calculations necessary to obtain the permutation distributions in the previous section are dependent on the individual observations in the samples, and even for small sample sizes require considerable computation. In this section, we look at some larger examples and illustrate the use of software to obtain the required results. In addition, although previous comparisons between the treatments were based on the means, there is nothing in the methodology that prevents us from comparing the groups using another statistic such as the median. For two independent samples, we would just take the difference in the medians for the groups in each possible experiment, rather than the difference in means, to construct the permutation distribution. This would result in a more robust comparison of the treatments in the presence of outliers. Similarly, for paired data, we could compute the median of the differences rather than the mean when constructing the permutation distribution.

EXAMPLE 32.4 A Completely Randomized Design with Software

Recapping the example of the last section, suppose you have three men—Ari, Luis, and Troy—and three women—Ana, Deb, and Hui—for an experiment. Three of the six subjects are to be assigned completely at random to a new experimental weight loss treatment and three to a placebo. Troy, Deb, and Hui are assigned to the group receiving the new treatment with these resulting weight losses:

Subject	Ari	Luis	Troy	Ana	Deb	Hui
Weight Loss	2	15	8	1	12	9

In this case, the difference (treatment – placebo) in mean weight losses for the two groups is $\frac{8 + 12 + 9}{3} - \frac{2 + 15 + 1}{3} = 3.67$. Using the permutation distribution worked



WTLOSS2

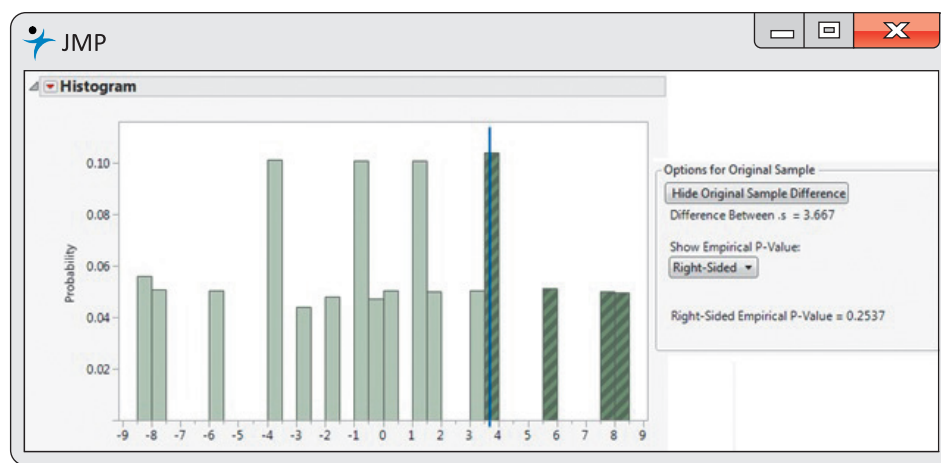
out in Example 32.2, we have shown that the P -value, the chance of observing a difference as large as or larger than 3.67, is 0.25.

Some specialized statistical software for a permutation test uses the full enumeration of all possible permutations for smaller sample sizes to obtain the permutation distribution and P -value. However, with larger sample sizes, the number of permutations becomes so large that the permutation distribution must be simulated using a random sample of permutations. Software, such as JMP, simulates the permutation distribution for all sample sizes when calculating a P -value. If we simulate enough permutations, the answer will be very close to the one that would be obtained by a full enumeration of all permutations. To see this, we use the simulation process on this small data set and observe how the simulated resampling distribution we obtain compares with the exact resampling distribution we computed in Example 32.2.

Figure 32.2 gives the JMP output for this example. The histogram estimating the permutation distribution was obtained by simulating 10,000 random permutations and agrees very closely with the permutation distribution derived in Example 32.2 based on all possible permutations. The shaded bars in the histogram include values for the difference in means that are greater than or equal to 3.67, and the empirical P -value of 0.2537 is very close to the theoretical value of 0.25.

FIGURE 32.2

Output from JMP, for Example 32.4. The output gives the histogram estimating the permutation distribution for the difference in mean weight losses and the associated P -value using 10,000 random permutations.



The use of a random sample of permutations to simulate the permutation distribution and calculate the P -value means that the answers obtained will vary from one simulation to another, but choosing a large number of permutations will decrease the variability from one simulation to another. Both aspects of this will be explored in the next example.

EXAMPLE 32.5 The Effect of the Number of Simulated Permutations

Example 32.4 used 10,000 random permutations to simulate the permutation distribution and calculate the P -value for the weight loss data. Because the sample sizes for the two groups were small, in Example 32.2 we were able to list all possible permutations and compute the exact P -value. Comparison of the simulated value to the exact value showed very good agreement. In general, the simulated permutation distribution and P -value will tend to be closer to the true permutation distribution and P -value when we use a larger number of simulated permutations. Figure 32.3(a) shows a second resampling using 10,000 random permutations for the weight loss data of Example 32.2. We see that the P -value is again close

to the true value of 0.25, and the simulated permutation distribution again agrees well with the exact permutation distribution. The use of 10,000 simulations—or even 1000 simulations—will generally be large enough to give sufficiently accurate results for most purposes. If greater accuracy is required, then the number of simulations can be increased.

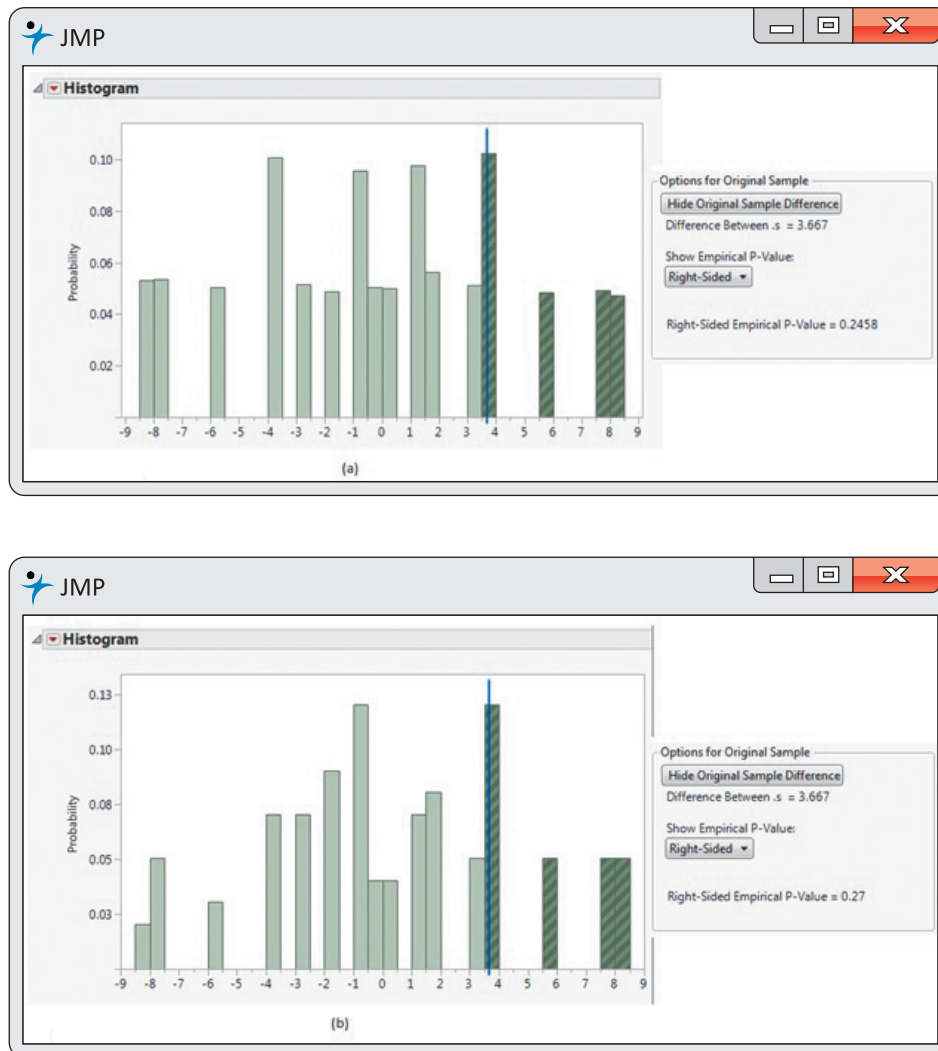


FIGURE 32.3

Output from JMP, for Example 32.5.
 (a) Histogram estimating the permutation distribution for the difference in mean weight losses and the associated P -value using 10,000 random permutations.
 (b) Histogram estimating the permutation distribution for the difference in mean weight losses and the associated P -value using 100 random permutations.

Figure 32.3(b) uses 100 random permutations to simulate the permutation distribution and P -value. In this case, we see that the estimated P -value is farther from the true value of 0.25, and the simulated permutation distribution shows less overall agreement with the true permutation distribution. The use of 100 resamples tends to be too small to produce accurate results as the answers will vary considerably among different simulations.

The next example has sample sizes of 27 in each group, which results in 1.95×10^{15} possible permutations, going beyond the limitations of software to enumerate all permutations. The example is again worked out using JMP software with 10,000 random permutations generated to approximate the permutation distribution.

**EXAMPLE 32.6 Shared Pain and Bonding**

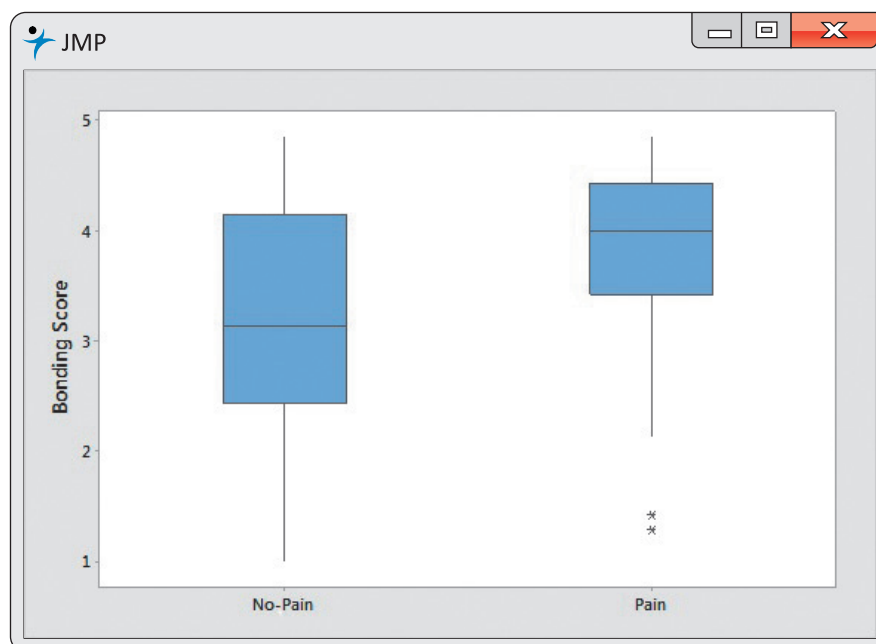
Although painful experiences are involved in social rituals in many parts of the world, little is known about the social effects of pain. Will sharing a painful experience in a small group lead to greater bonding of group members than sharing a similar nonpainful experience? Fifty-four university students in South Wales were divided at random into a pain group containing 27 students, and a no-pain group containing the remaining 27 students. Pain was induced by two tasks. In the first task, students submerged their hands in freezing water for as long as possible, moving metal balls at the bottom of the vessel into a submerged container, and in the second task, students performed a standing wall squat with back straight and knees at 90° for as long as possible. The no-pain group completed the first task using room temperature water for 90 seconds and the second task by balancing on one foot for 60 seconds, changing feet if necessary. In both the pain and nonpain settings, the students completed the tasks in small groups, which typically consisted of four students and contained similar levels of group interaction. Afterward, each student completed a questionnaire to create a bonding score based on responses to seven statements such as, “I feel the participants in this study have a lot in common” or “I feel I can trust the other participants.” Each response was scored on a five-point scale (1 = strongly agree, 5 = strongly disagree) and the scores on the seven statements were averaged to create a bonding score for each subject. Here are the bonding scores for the subjects in the two groups:⁴

No-pain group:	3.43	4.86	1.71	1.71	3.86	3.14	4.14	3.14	4.43	3.71
	3.00	3.14	4.14	4.29	2.43	2.71	4.43	3.43	1.29	1.29
	3.00	3.00	2.86	2.14	4.71	1.00	3.71			
Pain group:	4.71	4.86	4.14	1.29	2.29	4.43	3.57	4.43	3.57	3.43
	4.14	3.86	4.57	4.57	4.29	1.43	4.29	3.57	3.57	3.43
	2.29	4.00	4.43	4.71	4.71	2.14	3.57			

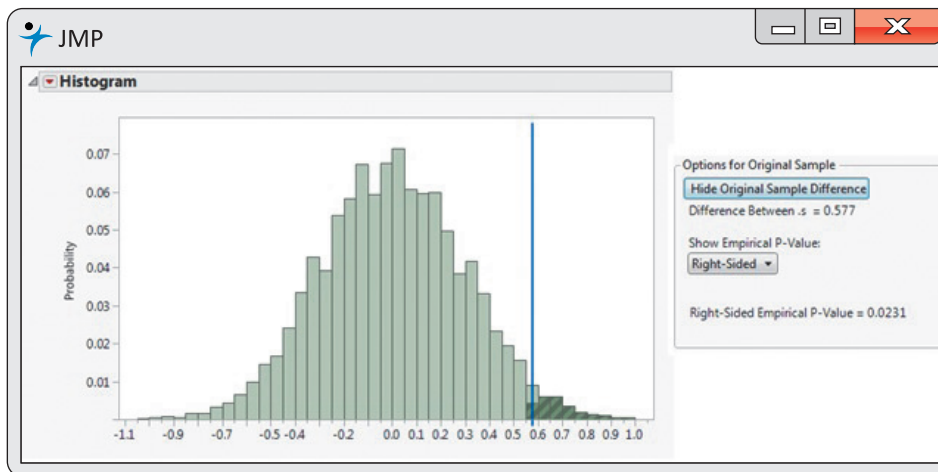
Do the data show that sharing a painful experience in a small group leads to higher bonding scores for group members than sharing a similar nonpainful experience? Figure 32.4 is a comparative boxplot of the two samples.

FIGURE 32.4

Output from JMP, for the data of Example 32.6. The output provides a comparative boxplot for the bonding scores for the no-pain and pain groups.

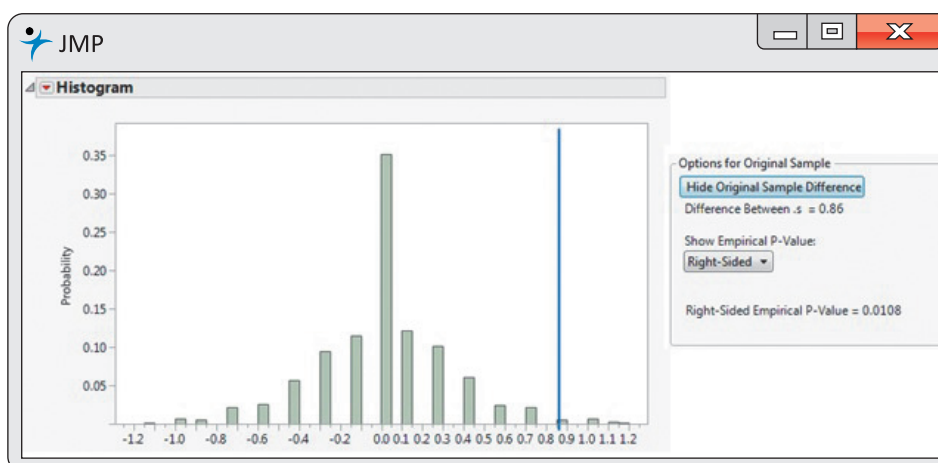


As hypothesized, the “Pain” group tends to have higher bonding scores than the “No-Pain” group, with two low outliers. Figure 32.5 gives the JMP output for the permutation test, which compares the sample means in the two treatments. The estimated histogram for the permutation distribution is based on 10,000 simulated permutations, and the output gives an observed difference in means of 0.577 and a one-sided P -value of 0.0231. The two-sample t -test applied to these data has a one-sided P -value of 0.0240, in very close agreement with the permutation test. The agreement between the two-sample t and the permutation test that uses the difference between the two sample means to compare the treatments is not a coincidence. Theory tells us that, for larger sample sizes, both procedures tend to yield similar results.

**FIGURE 32.5**

Output from JMP, for Example 32.6. The output gives the histogram estimating the permutation distribution for the difference in sample means and the associated P -value using 10,000 random permutations.

We know that the sample mean is not robust in the presence of outliers, which suggests that in this example, we might prefer to use a more robust statistic such as the median to compare the two treatments. For the permutation distribution, no new theory is required. For each permutation, rather than computing $\bar{x}_{\text{Pain}} - \bar{x}_{\text{NoPain}}$, we compute instead $M_{\text{Pain}} - M_{\text{NoPain}}$, where M indicates the sample median. We then compare the observed difference in the two medians to the estimated permutation distribution for the difference in the two medians to find the P -value. Figure 32.6 gives the JMP output for this permutation distribution and reports a P -value of 0.0108. Additionally, as expected, the estimate of the difference $M_{\text{Pain}} - M_{\text{NoPain}} = 0.86$ is larger than $\bar{x}_{\text{Pain}} - \bar{x}_{\text{NoPain}} = 0.577$ because the two low outliers in the pain group have little effect on the median of the pain group but a large effect in reducing the mean of the pain group.

**FIGURE 32.6**

Output from JMP, for Example 32.6. The output gives the histogram estimating the permutation distribution for the difference in sample medians and the associated P -value using 10,000 random permutations.

The presence of outliers in the pain group is not due to an error. Rather, it is likely that there are a small portion of subjects who would experience little bonding regardless of which treatment they were assigned to. When these subjects are assigned to the “Pain” group, they appear as outliers because most bonding scores are higher, while when assigned to the “No-Pain” group, their scores do not appear unusually low. For these reasons, the use of means to compare these treatments is not the best option, and an analysis using medians is preferable.

The subjects in the experiment of Example 32.6 consist of both male and female students. Is there a difference in the level of bonding experienced by male and female students when having a painful experience in a small group? While this is a reasonable question, we need to see if a permutation test is still appropriate for answering it. We no longer have a comparative experiment because it was the treatments that were assigned at random to the subjects, not their sex. Here is a way of thinking about this question that leads directly to the use of a permutation test to compare the sexes, but has a slightly different justification than in the comparison of treatments in a randomized comparative experiment.

Think of the 27 bonding scores in the pain group as fixed numbers that are to be randomly assigned to the 27 students in the “Pain” group. If these scores are assigned at random to the 27 students, then the sex variable should be unrelated to the resulting bonding scores, whereas if the sex variable were related to the bonding score, then we would expect to find either the male or female group to have larger bonding scores. Here is how we can use permutation tests to see if there is a relationship between sex and bonding score.

For each random assignment of the scores to the students, compute the difference in median scores for males and females. The resulting sampling distribution for this difference in medians is exactly the same as the permutation distribution we would have obtained if males and females were considered treatments. Comparing the observed difference in median bonding scores for the two sexes to this permutation distribution produces a valid test of the null hypothesis that the scores are randomly assigned to the students against the alternative that there is a difference in average bonding scores for the males and females in the “Pain” group.

EXAMPLE 32.7 Shared Pain and Bonding, continued

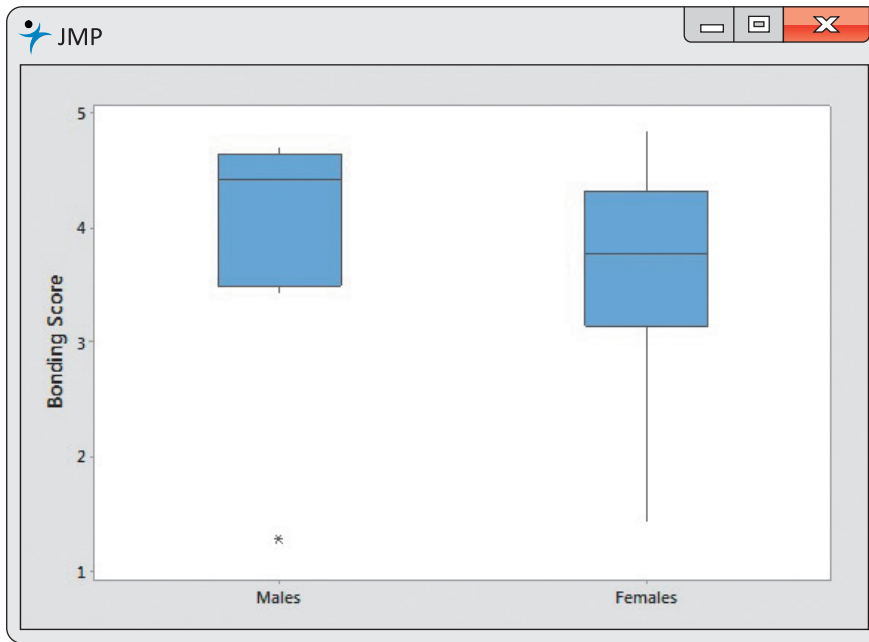


The “Pain” group contains nine males and 18 females. Here are the bonding scores for the two groups:

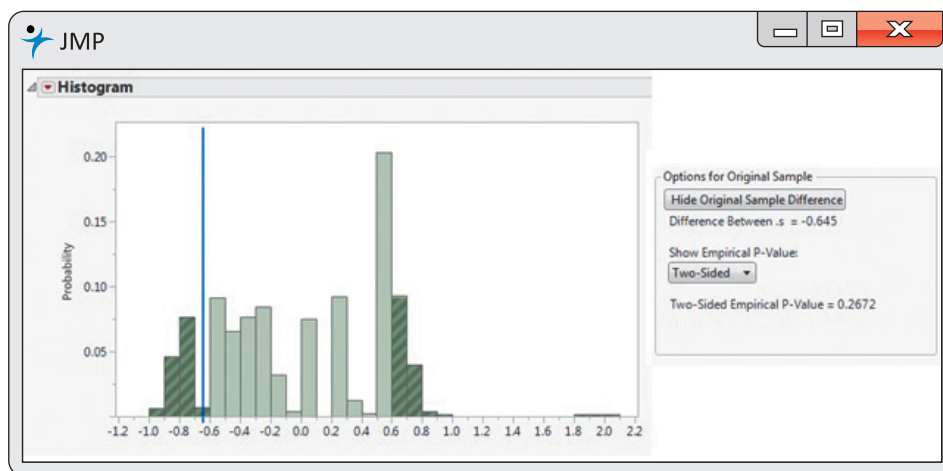
Males:	1.29	4.43	4.43	3.57	3.86	4.57	3.43	4.71	4.71
Females:	4.71	4.86	4.14	2.29	3.57	3.43	4.14	4.57	4.29
	1.43	4.29	3.57	3.57	2.29	4.00	4.43	2.14	3.57

Do the data show that sharing a painful experience in a small group leads to higher bonding scores for either males or females?

The comparative boxplot in Figure 32.7 suggests that males may have higher bonding scores than females, although the males have a low outlier, which suggests we compare the groups using medians rather than means. The JMP output in Figure 32.8 gives the estimated histogram for the permutation distribution and has a two-sided *P*-value of 0.2672. However, the sample sizes are small, and a larger study designed to test the hypothesis that males tend to have higher bonding scores than females when experiencing pain in small groups might be of further interest.

**FIGURE 32.7**

Output from JMP, for the data of Example 32.7. The output provides a comparative boxplot for the bonding scores for males and females in the pain group.

**FIGURE 32.8**

Output from JMP, for Example 32.7. The output gives the histogram estimating the permutation distribution for the difference in sample medians and the associated P -value using 10,000 random permutations.

The next example uses R software to carry out the calculations for a permutation test for paired data. As in the two-sample problem, the software simulates the permutation distribution to approximate the P -value.

EXAMPLE 32.8 Golf Scores

Here are the golf scores of 12 members of a college women's golf team in two rounds of tournament play. (A golf score is the number of strokes required to complete the course, so low scores are better.)

	Player											
	1	2	3	4	5	6	7	8	9	10	11	12
Round 2	94	85	89	89	81	76	107	89	87	91	88	80
Round 1	89	90	87	95	86	81	102	105	83	88	91	79
Difference	5	-5	2	-6	-5	-5	5	-16	4	3	-3	1



GOLF

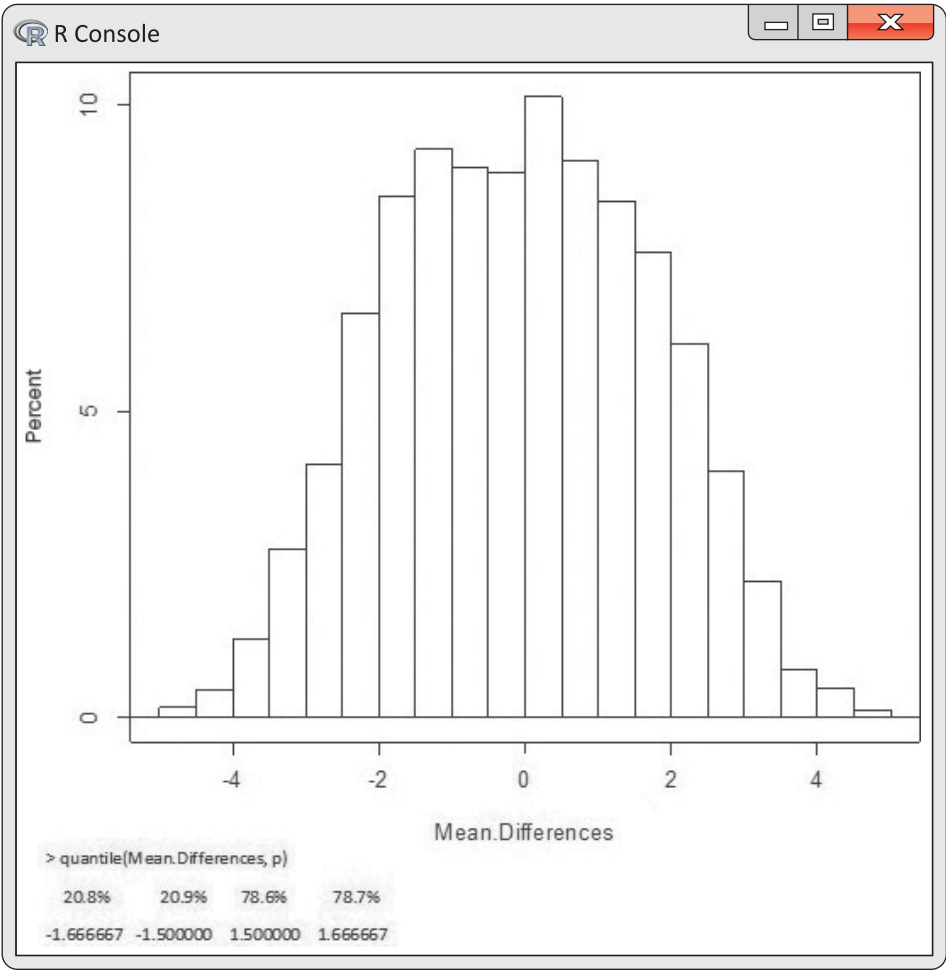


FIGURE 32.9
Output from R, for Example 32.8. The output gives the histogram estimating the permutation distribution for the difference in sample means and selected percentiles of this distribution. Statistical analysis relies heavily on statistical software, and R is an extremely powerful statistical environment that is free to anyone; it relies heavily on members of the academic and general statistical communities for support. Computer output from other statistical packages like JMP, Minitab, and SPSS is similar, so you can feel comfortable using any one of these packages.

Negative differences indicate better (lower) scores on the second round. Based on this sample, can we conclude that this team's golfers perform differently in the two rounds of a tournament?

This is a matched pairs design with the two observations on each player forming the matched pair. We use the average of the differences between the paired responses (Round 2 minus Round 1) to determine if there is evidence that the golfers perform differently in the two rounds of the tournament. As in Example 32.3, the permutation distribution is obtained by assigning plus and minus signs to the absolute values of the differences in the two scores in all possible ways and for each assignment computing the average of the resulting differences. Because the number of possible ways of assigning plus and minus signs gets large very quickly, we again simulate a random sample of possible assignments of plus and minus signs to approximate the permutation distribution.

Figure 32.9 gives a histogram of the averages of the differences obtained from 10,000 simulated assignments of plus and minus signs to the absolute values of the differences using R software. This should provide a good approximation to the permutation distribution of the average difference. The observed average difference for this data is -1.67 , and because this is a two-sided test, the P -value is the area under the permutation distribution less than or equal to -1.67 plus the area greater than or equal to 1.67 . Several percentiles, obtained by R, are displayed in Figure 32.9 (the percentile is the value in the second line, and the first line is the area less than or equal to that value written as a percentage). The area less than or equal to -1.67 is 0.208, and the area greater than or equal to 1.67 is 1 minus



the area less than or equal to 1.5, or $1 - 0.786 = 0.214$. This gives a P -value of 0.422, suggesting little difference in the scores between the two rounds. The matched pairs t test has $P = 0.3716$, showing very similar conclusions to the permutation test.

In this and the previous section, we have demonstrated how to compute P -values for permutation tests using both enumeration of all possible randomizations to obtain the exact permutation distribution and simulation to approximate this distribution. The bootstrap is another example of a resampling method. It can be used to estimate standard errors of statistics and construct confidence intervals. We discuss the bootstrap in the next two sections.

Macmillan Learning Online Resources

- The technology manuals for JMP and R explain how to use software to do a permutation test.

APPLY YOUR KNOWLEDGE

32.3 A Very Simple Setting with Medians. In Exercise 32.1, data were given comparing course performance by students taking handwritten notes with those taking notes on their laptops. The comparison was based on the means of the two treatments and the permutation distribution for the difference in means was constructed.

- Give the permutation distribution for the difference in the medians for the two treatments using the 15 possible assignments of subjects to treatments.
- Compute the P -value of the data. Assume the two-sided alternative hypothesis is that the median number of points is different for the two groups.

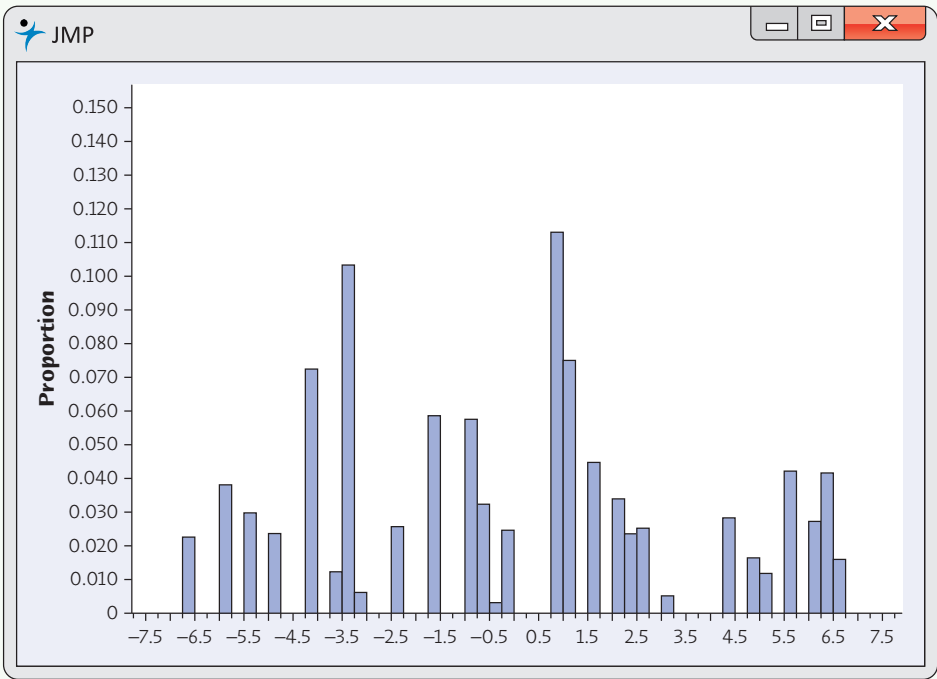
32.4 Do Birds Learn to Time Their Breeding? The exercises in Chapter 21 discuss a study of whether supplementing the diet of blue titmice with extra caterpillars will prevent them from adjusting their breeding date the following year to obtain a better food supply. Thirteen pairs of birds were randomly assigned to either a supplemental diet or their natural diet (the control group), with seven pairs assigned to the supplemental diet and the remainder to the control diet. Here are the data (days after the caterpillar peak):⁵

Control	4.6	2.3	7.7	6.0	4.6	−1.2	
Supplemented	15.5	11.3	5.4	16.5	11.3	11.4	7.7

The null hypothesis is no difference in timing; the alternative hypothesis is that the supplemented birds miss the peak by more days because they don't adjust their breeding date. Figure 32.10 gives the estimated histogram for the permutation distribution of 10,000 simulated permutations for the permutation test that compares the sample medians (median for the supplemental group minus the median for the control group) for the two treatments. Use Figure 32.10 to estimate the P -value of the test.



FIGURE 32.10
The estimated histogram for the permutation distribution for the difference in sample medians, for Exercise 32.4.



32.5 Nintendo and Laparoscopic Skills (Software Required). In laparoscopic surgery, a video camera and several thin instruments are inserted into the patient’s abdominal cavity. The surgeon uses the image from the video camera positioned inside the patient’s body to perform the procedure by manipulating the instruments that have been inserted. It has been found that the Nintendo Wii, with its motion-sensing interface, reproduces the movements required in laparoscopic surgery more closely than other video games. If training with a Nintendo Wii can improve laparoscopic skills, it can complement the more expensive training on a laparoscopic simulator. Forty-two medical residents were chosen, and all were tested on a set of basic laparoscopic skills. Twenty-one were selected at random to undergo systematic Nintendo Wii training for one hour per day, five days per week, for four weeks. The remaining 21 residents were given no Nintendo Wii training and were asked to refrain from video games during this period. At the end of four weeks, all 42 residents were tested again on the same set of laparoscopic skills. One of the skills involved a virtual gall bladder removal with several performance measures, including time to complete the task recorded. Here are the improvement (before–after) times in seconds after four weeks for the two groups:⁶



Treatment						Control					
291	134	186	128	84	243	21	66	54	85	229	92
212	121	134	221	59	244	43	27	77	−29	−14	88
79	333	−13	−16	71	−16	145	110	32	90	45	−81
71	77	144				68	61	44			

Does the Nintendo Wii training significantly increase the mean improvement time?

- (a) Use software to estimate the histogram for the permutation distribution for the difference in the mean improvements for the two groups (mean improvement for the treatment group minus the mean improvement for the control group). Use 10,000 simulated permutations.
- (b) From your estimated histogram, compute the *P*-value of the data.

32.6 Comparing Two Insect Repellants (Software Required). In Example 32.3 we compared two insect repellants using a permutation test for a matched pairs experiment. Because of the small sample size, we were able to obtain the exact permutation distribution as:



Mean difference	-2.0	-1.0	-0.5	0.5	1.0	2.0
Probability	0.125	0.125	0.250	0.250	0.125	0.125

In this example, the observed mean difference in treatments (DEET – oil of lemon eucalyptus) is -2 . Using this permutation distribution, we have shown that the two-sided P -value, the chance of observing a difference this extreme, is 0.25 .

- Simulate the permutation distribution using 100 simulations and give the estimated P -value. Repeat this with a second simulation. How close are the answers to the exact permutation distribution and P -value?
- Simulate the permutation distribution using 10,000 simulations and give the estimated P -value. Repeat this with a second simulation. How close are the answers to the exact permutation distribution and P -value?
- What do the results in parts (a) and (b) show about the effect of the number of simulations on the estimated permutation distribution and P -value? Explain briefly.

32.3 Generating Bootstrap Samples

In Chapter 2, we looked at samples of travel times to work in both North Carolina and New York. Here is a stemplot of the travel times in minutes for the 15 workers in North Carolina, chosen at random by the U.S. Census Bureau:⁷

```

0 | 5
1 | 000025
2 | 005
3 | 00
4 | 00
5 |
6 | 0

```

The stemplot of the 20 travel times of the random sample of workers in New York state is

```

0 | 5
1 | 005555
2 | 0005
3 | 00
4 | 005
5 |
6 | 005
7 |
8 | 5

```

Suppose we want confidence intervals for the average travel time in New York and the average travel time in North Carolina, or we want to compare the travel times for New York and North Carolina by finding a confidence interval for the ratio of the average travel time in New York to the average travel time in North Carolina. The travel times in both North Carolina and New York are skewed to the right with high outliers. The sample sizes are moderate, raising some concerns about the use of the t procedures for confidence intervals on the individual means, and we have no current method for finding a confidence interval for the ratio of two means.

The statistics we would use to estimate the average travel times in New York and North Carolina are the means of the samples from each state, while for the statistic to estimate the ratio we could use the ratio of these sample means. But inference requires the sampling distributions of these statistics. The **bootstrap method** provides a way to approximate these sampling distributions using the information in the original samples. This is done by generating many samples by **sampling with replacement** from the original sample. These samples with replacement are called **bootstrap samples**.

bootstrap method

sampling with replacement

bootstrap samples

Sampling with Replacement

To sample with replacement from a population, after an observation is selected, return it to the population before the next observation is drawn. Thus, the same observation can be selected multiple times when taking a random sample with replacement. If you are sampling with replacement using Table B, when you come to a label that has previously been used, *do not* ignore it but include the observation corresponding to this label in the sample again. When using software, you often have the option to select a sample either with or without replacement.

A bootstrap sample from the North Carolina travel times is obtained by treating the original sample of 15 travel times as representing the population. The phrase “representing the population” is important because by sampling with replacement from the sample, we mimic what would happen if we took a random sample from a population consisting of many, many copies of the original sample. To sample from this population, we select a random sample with replacement of the 15 travel times. This process of selecting random samples with replacement from the original sample is referred to as resampling. Here are three bootstrap samples obtained by resampling from the North Carolina travel times, along with the means of these bootstrap samples. Note that the mean of the 15 original North Carolina travel times is 22.47 minutes.

30	25	20	10	30	5	40	25	25	30	10	10	40	30	25	mean = 23.67
40	10	10	12	10	20	60	10	10	10	30	10	30	20	10	mean = 19.47
5	40	60	15	20	12	10	25	40	30	25	25	40	60	60	mean = 31.13

In the first sample, the observation of 25 minutes was selected four times, and in the third sample, the observation of 60 minutes was selected three times. This can occur because we are sampling with replacement. As expected, the means of the three bootstrap samples fluctuate about the original sample mean of 22.47 minutes. In practice, we will select a large number of bootstrap samples for statistical inference.

Because the original random sample of 15 North Carolina times should be representative of the population of all North Carolina travel times, the bootstrap

samples should mimic selecting random samples from this population. And the **bootstrap distribution** of a statistic computed from these bootstrap samples should give an approximation to the sampling distribution of this statistic.

bootstrap distribution

Bootstrap Method

- Create many bootstrap samples, and for each bootstrap sample, compute the statistic of interest.
- The distribution of this statistic from the bootstrap samples is known as the bootstrap distribution and provides an approximation to the sampling distribution of the statistic.

EXAMPLE 32.9 Bootstrap Distribution of the Sample Mean: North Carolina Travel Times

We wish to estimate the population mean of the North Carolina travel times and will use \bar{x} , the sample mean, as our estimate. To approximate the sampling distribution of the sample mean for the North Carolina travel times, take 1000 bootstrap samples from the North Carolina travel times, and for each bootstrap sample, compute the sample mean. Figure 32.11(a) provides a histogram of the 1000 sample means obtained, which is the bootstrap distribution of the sample mean.

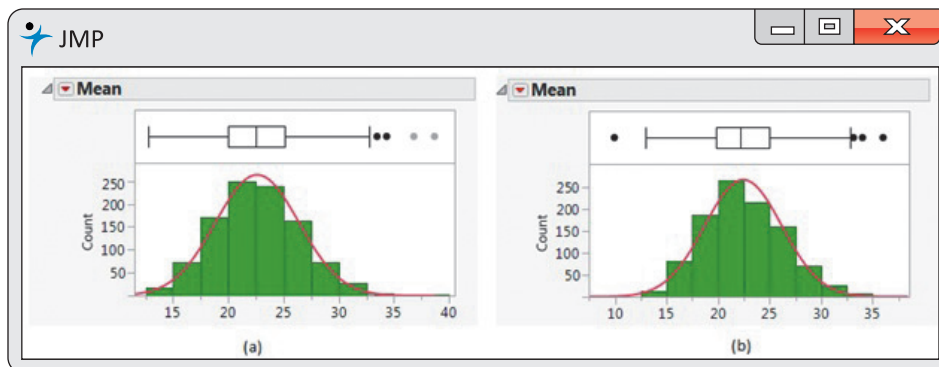


FIGURE 32.11

Output from JMP, for Example 32.9.
(a) Histogram of the bootstrap distribution for the mean travel times using 1000 random bootstrap samples.
(b) Histogram of the bootstrap distribution for the mean travel times using a second set of 1000 random bootstrap samples.

The mean of these 1000 bootstrap sample means is 22.61, and the standard deviation is 3.78. This bootstrap distribution is an approximation to the sampling distribution of \bar{x} . The shape of the sampling distribution is fairly well approximated by the normal curve superimposed on the histogram, although there is a slight skew to the right. This suggests that our t methods from Chapter 20 should work fairly well for making inferences about the population mean of the North Carolina travel times, because the t methods assume that the sampling distribution of \bar{x} is approximately normal. We will return to this in the next section, where we discuss confidence intervals based on the bootstrap method.

As with any simulation method, the accuracy of the answers is improved as the number of simulations is increased. Figure 32.11(b) provides the bootstrap distribution for a second resampling of 1000 bootstrap samples from the North Carolina travel times. The mean of these 1000 bootstrap sample means is 22.43, and the standard deviation is 3.75. The shape of the bootstrap distribution is similar to that given in Figure 32.11(a), with the mean being slightly smaller and the variability being virtually identical. In general, 1000 bootstrap samples are sufficient to approximate the sampling distribution of a statistic unless very high accuracy is required.

The next example looks at the bootstrap distribution of the ratio of two sample means, a situation for which we currently have no methods for inference.

EXAMPLE 32.10 Bootstrap Distribution of the Ratio of Means: North Carolina and New York Travel Times



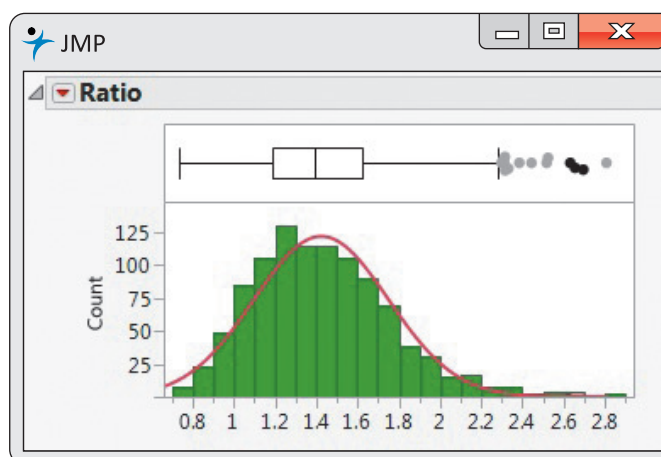
We wish to estimate the ratio of the population means of the travel times in New York and North Carolina—specifically, the mean of the New York travel times divided by the mean of the North Carolina travel times. As our estimate we will use $\bar{x}_{NY}/\bar{x}_{NC}$, the ratio of the sample means. In this case, the bootstrap method requires that we resample from both the New York and the North Carolina travel times to obtain the bootstrap distribution of the ratio. Here is the process we need to follow:

- Select a resample of the New York travel times and a resample of the North Carolina travel times. After obtaining the mean for each resample, compute the ratio of the mean of the New York travel times to the mean of the North Carolina travel times.
- Repeat this process 1000 times to obtain 1000 bootstrap estimates of the ratio.

Figure 32.12 provides a histogram of the 1000 bootstrap estimates of the ratio of the means of New York to North Carolina travel times. This is the bootstrap distribution for the ratio of the two means.

FIGURE 32.12

Output from JMP, for Example 32.10. The output gives the histogram of the bootstrap distribution for the ratio of mean travel times using 1000 random bootstrap samples.




The mean of these 1000 bootstrap sample ratios is 1.42, and the standard deviation is 0.33. This bootstrap distribution is an approximation to the sampling distribution of $\bar{x}_{NY}/\bar{x}_{NC}$. Unlike Example 32.9, this bootstrap distribution is not that well approximated by the normal curve superimposed on the histogram and shows considerable right skewness.

The bootstrap distribution depends on both the number of bootstrap samples selected and the particular bootstrap samples obtained. This means that, for a fixed number of bootstrap samples, two individuals using this method will obtain different bootstrap distributions. As seen in Example 32.9, the choice of 1000 bootstrap samples is generally sufficiently large that the differences obtained in the bootstrap distributions will be fairly small. In the next section, we show how to use the bootstrap distribution to provide confidence intervals. There are several methods to generate confidence intervals from the bootstrap distribution, and we provide the details for the bootstrap percentile confidence interval.

Macmillan Learning Online Resources


- The technology manuals describe how to sample with replacement using several software packages.

APPLY YOUR KNOWLEDGE


32.7 Generating Bootstrap Samples. Here are the SAT mathematics scores of a random sample of seven students selected from the freshman class at Georgia Southern University:  SAT

470 690 540 570 470 680 710

- Compute the mean SAT mathematics score for this sample.
- Generate three bootstrap samples by resampling from the original sample of Georgia Southern University SAT mathematics scores. For each bootstrap sample, compute the mean. If you are using Table B, begin at line 116.
- How do the means of the bootstrap samples compare to the mean of the original sample? Is this what you would expect? Explain.

32.8 New York Travel Times (Software Required). Generate 1000 bootstrap samples by resampling from the New York travel times. For each bootstrap sample, compute the bootstrap sample mean.  TRAVNY

- What are the mean and standard deviation of the 1000 bootstrap sample means?
- Draw a histogram of the bootstrap distribution of the sample mean using the 1000 bootstrap sample means. If your software allows this, superimpose a normal curve on the histogram. Describe the shape of this bootstrap distribution.

32.9 Pulling Wood Apart (Software Required). How heavy a load (pounds) is needed to pull apart pieces of Douglas fir 4 inches long and 1.5 inches square? Here are data from students doing a laboratory exercise:  WOOD

33,190	31,860	32,590	26,520	33,280
32,320	33,020	32,030	30,460	32,700
23,040	30,930	32,720	33,650	32,340
24,050	30,170	31,300	28,730	31,920

Generate 1000 bootstrap samples by resampling from these data. For each bootstrap sample, compute the bootstrap sample mean.

- What are the mean and standard deviation of the 1000 bootstrap sample means?
- Draw a histogram of the bootstrap distribution of the sample mean using the 1000 bootstrap sample means. If your software allows this, superimpose a normal curve on the histogram. Describe the shape of this bootstrap distribution.

32.4 Bootstrap Standard Errors and Confidence Intervals

The variability in the bootstrap distribution reflects how the statistic of interest varies from sample to sample. A numerical measure of this variability is the standard deviation of the bootstrap distribution known as the **bootstrap standard error**. For the sample mean, we know from theory that the standard error of the mean is given

bootstrap standard error

by the formula s/\sqrt{n} . Referring to the North Carolina travel times in Example 32.9, the standard deviation of the original sample of 15 travel times is $s = 15.231$, and the standard error of the mean is $s/\sqrt{15} = 3.93$. The bootstrap standard errors for the two sets of 1000 resamples of the North Carolina travel times reported in Example 32.9 are 3.78 and 3.75, respectively, fairly close to the value based on theory. This illustrates the advantage of the bootstrap method: It can produce an estimate of the standard error of a statistic in situations where we do not have a simple formula.

**bootstrap percentile
confidence interval**

Our method of constructing a confidence interval for a parameter has been to use the endpoints of the central area of the sampling distribution of the estimate. Confidence intervals for proportions or means have found this central area by using the appropriate percentile of either the z or the t distribution multiplied by the standard error of the estimate. In a similar manner, the **bootstrap percentile confidence interval** treats the bootstrap distribution as the sampling distribution of the estimate and, for a specific confidence coefficient, uses the appropriate percentiles that mark off the central area of this bootstrap distribution to form the confidence interval.

Bootstrap Percentile Confidence Interval

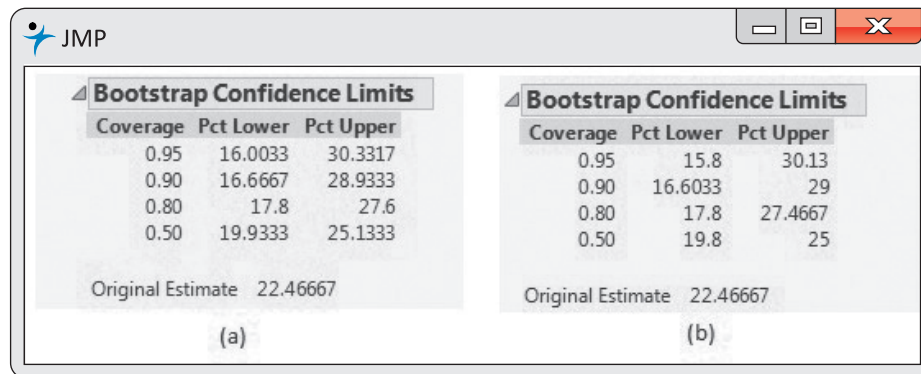
For a 95% bootstrap confidence interval, use the lower 2.5 percentile and the upper 97.5 percentile of the bootstrap distribution as the endpoints of the confidence interval. For a 90% confidence interval, use the lower 5 percentile and the upper 95 percentile.

Unlike our previous confidence intervals, the bootstrap percentile confidence interval does not assume that the sampling distribution of the estimate is normal or even symmetric. Many software packages will produce a bootstrap confidence interval, but there are several variations of the bootstrap confidence interval that lead to different answers. We will have more to say about these other intervals and the situations for which the bootstrap percentile confidence interval is appropriate, but first here are some examples of the bootstrap percentile confidence interval using software.

EXAMPLE 32.11 Bootstrap Percentile Confidence Interval: North Carolina Travel Times



We wish to estimate the population mean of the North Carolina travel times using a 95% confidence interval. In Example 32.9, we saw that the bootstrap distribution of the sample mean is fairly well approximated by a normal curve, suggesting that despite the skewness and single outlier present, the sampling distribution of \bar{x} is approximately normal. Figure 32.13(a) and Figure 32.13(b) provide the JMP output giving bootstrap percentile confidence intervals associated with the bootstrap distributions in Figure 32.11(a) and Figure 32.11(b), respectively. In Figure 32.13(a), the 95% bootstrap confidence interval is 16.0033 minutes to 30.3317 minutes. The original sample mean is 22.47 minutes, and because of the slight asymmetry in the bootstrap distribution, the confidence interval is not centred about the sample mean as the usual t interval is. The 95% confidence interval using the t is 14.03 minutes to 30.90 minutes. The agreement between these intervals is reasonable, with the bootstrap interval making a small adjustment for the skewness of the sampling distribution.

**FIGURE 32.13**

Output from JMP, for Example 32.11. (a) Bootstrap confidence intervals for the bootstrap distribution for the mean travel times using 1000 random bootstrap samples. (b) Bootstrap confidence intervals for the bootstrap distribution for the mean travel times using a second set of 1000 random bootstrap samples.

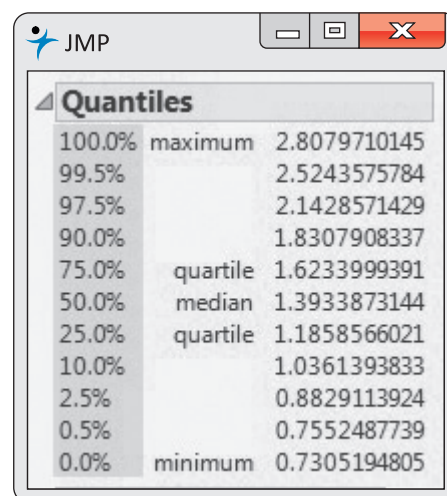
As noted in Example 32.9, the bootstrap distribution in Figure 32.11 (b) is similar to that in Figure 32.11 (a), although the distribution has a slightly smaller mean. This suggests that there should be good agreement between the bootstrap percentile confidence intervals obtained from the two bootstrap distributions. Comparing the bootstrap percentile confidence intervals in Figures 32.13(a) and 32.13(b), there are small differences in the 95% intervals, while the intervals are almost identical for the remaining confidence coefficients reported.

Although we have done multiple resampling in this example, it is important to realize that this was done for illustration and is not appropriate statistical practice. Taking multiple resamples and using the “best” answer invalidates the properties of the procedures, whether it be a bootstrap confidence interval or a permutation test.

EXAMPLE 32.12 Bootstrap Percentile Confidence Interval: Ratio of Means

The bootstrap distribution for the ratio of the population means of the travel times in New York and North Carolina is given in Example 32.10. The bootstrap distribution is skewed to the right and less well approximated by a normal curve than in the previous example. Figure 32.14 gives the JMP output containing several percentiles of this bootstrap distribution. The estimate of the ratio based on the original samples is $\bar{x}_{NY}/\bar{x}_{NC} = 31.250/22.467 = 1.39$.

The 95% confidence interval for the ratio is 0.89 to 2.14, showing that the mean travel time to New York is more than 90% of and less than double the mean travel time to North Carolina. Again we see that the interval is not centered around the estimate.

**FIGURE 32.14**

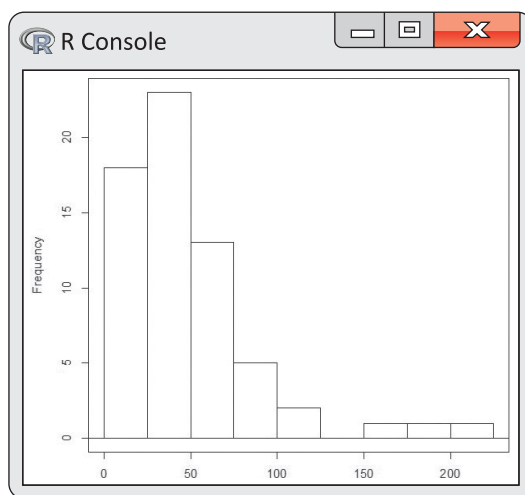
Output from JMP, for Example 32.12. The output gives several percentiles of the bootstrap distribution for the ratio of mean travel times using 1000 random bootstrap samples.

EXAMPLE 32.13 Bootstrap Percentile Confidence Intervals: Mean and Median

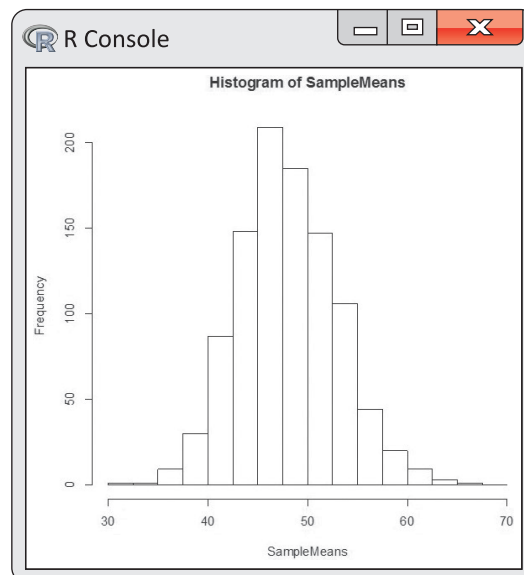
OIL

Exercise 20.43 gives the estimated total amounts of oil (in thousands of barrels) for a random sample of 64 wells in the Devonian Richmond Dolomite area of the Michigan basin. Figure 32.15(a) is a histogram of the original data from R, and Figure 32.15(b) is a histogram from R of the bootstrap distribution for the mean of 1000 resamples. The original distribution is very skewed with several high outliers. The bootstrap distribution appears to be fairly well approximated by a normal curve. The endpoints of the 95% bootstrap percentile confidence interval are determined by the 2.5% and 97.5% quantiles given in Figure 32.15(c), and the 95% bootstrap percentile confidence interval is 39.24 thousand barrels to 58.26 thousand barrels. The 95% t confidence interval is 38.20 thousand barrels to 58.30 thousand barrels. The good agreement between these two confidence intervals is not surprising given the approximate normality of the bootstrap distribution, which results from the fairly large sample size.

Although we have provided a confidence interval for the mean, because of the extreme skewness of the original distribution, a better measure of the center of the distribution is the median because it is less affected by the long right tail and the outliers. For the original data, we have $\bar{x} = 48.25$ thousand barrels, and the sample median is $M = 37.8$ thousand barrels. The R output for the bootstrap distribution and confidence interval for the median are given in Figure 32.16(a) and Figure 32.16(b), respectively. The bootstrap distribution is fairly irregular and less well approximated by a normal curve. The 95% confidence interval is 32.7 thousand barrels to 47.6 thousand barrels. Because the bootstrap confidence intervals for the mean and median are for two *different* parameters, they cannot be compared directly with each



(a)

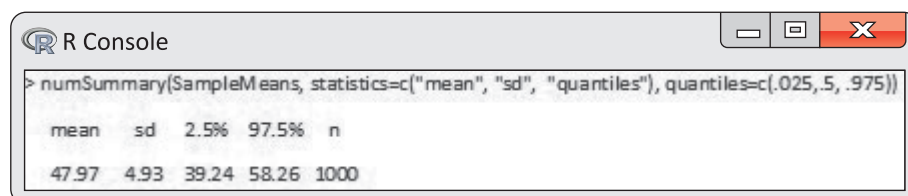


(b)

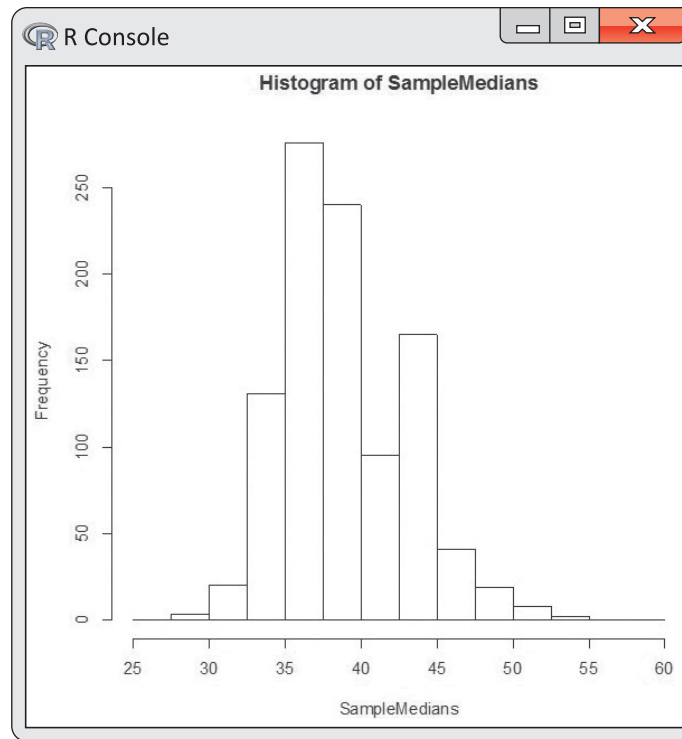
FIGURE 32.15

Output from R, for Example 32.13.

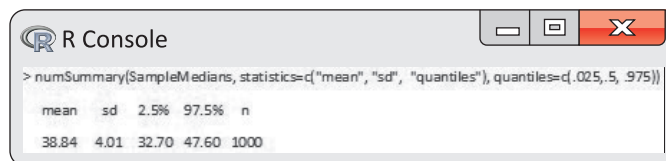
(a) Histogram of the data from Exercise 20.43. (b) Histogram of the bootstrap distribution for the mean of the estimated total amounts of oil using 1000 random bootstrap samples. (c) Summary statistics for the bootstrap distribution for the mean of the estimated total amounts of oil using 1000 random bootstrap samples.



(c)



(a)



(b)

FIGURE 32.16

Output from R, for Example 32.13.

(a) Histogram of the bootstrap distribution for the median of the estimated total amounts of oil using 1000 random bootstrap samples. (b) Summary statistics for the bootstrap distribution for the median of the estimated total amounts of oil using 1000 random bootstrap samples.

other. The bootstrap standard error for the mean is 4.93 thousand barrels, and for the median the bootstrap standard error is 4.01 thousand barrels. Based on a comparison of the standard errors, we are not surprised that the bootstrap confidence interval for the median is shorter than the bootstrap confidence interval for the mean.

In our discussion of methods for inference about means based on the Normal distribution, and for our inferences for proportions, we were able to be fairly specific about the conditions under which the methods were valid. Unfortunately, for bootstrap methods, our recommendations will be a little more vague. An important issue is how to correctly carry out the resampling. If the sample is from a single population, then we just resample from the original sample. In Example 32.9, our resampling required that we resample independently from both samples to estimate a ratio, while in a simple linear regression setting, we would need to resample the X and Y values in pairs when bootstrapping the estimate of the slope or intercept.

The bootstrap method does not work equally well for all statistics. The bootstrap procedure we have described requires a fairly large sample size when bootstrapping the median. For odd sample sizes, the median is the middle observation, and when resampling from a small sample such as the North Carolina travel times, the bootstrap medians take only a few distinct values, and the resulting bootstrap distribution is not an accurate reflection of the sampling distribution of the median. Although the situation is somewhat better for even sample sizes due to the averaging


of the two middle observations, more advanced bootstrapping techniques have been developed that work better for the median when the sample sizes are small.


Finally, certain statistics may show a bias that can be estimated using the bootstrap distribution. When substantial bias is found to be present or the bootstrap distribution is highly skewed, more sophisticated confidence intervals than the percentile method presented in this section are generally recommended. Our treatment of the bootstrap should be viewed as an introduction to a very powerful technique, and further reading is advised for those planning on using the bootstrap in their data analysis.⁸

Macmillan Learning Online Resources

- The technology manual for JMP explains how to use software to compute bootstrap standard errors and confidence intervals.

APPLY YOUR KNOWLEDGE

32.10 New York Travel Times: Mean (Software Required). In Exercise 32.8, you found the bootstrap distribution for the mean of the New York travel times. Use this distribution to find the 90% bootstrap percentile confidence interval for the population mean of the New York travel times. Compare this interval to the 90% confidence interval based on the t distribution. Is this what you would expect? Explain briefly.  TRAVNY

32.11 New York Travel Times: Standard Deviation (Software Required). Unlike inference about the mean of a Normal population, inference about the standard deviation of a Normal population is not robust. That is, it does not remain valid even for small departures from the Normal distribution. The bootstrap method can be used to find a confidence interval for the standard deviation that does not assume the population of New York travel times is Normal.  TRAVNY

- For the New York travel times, generate 1000 bootstrap samples, and for each bootstrap sample compute the standard deviation. What are the mean and standard deviation of the 1000 bootstrap sample standard deviations?
- Draw a histogram of the 1000 bootstrap sample standard deviations. Describe the shape of this bootstrap distribution.
- Find a 95% bootstrap percentile confidence interval for the population standard deviation of the New York travel times.

CHAPTER 32 SUMMARY

Chapter Specifics

- The **permutation distribution** for a test statistic is determined by computing, from the observed data, all possible values of the statistic and the probability of these values under the assumption that treatments are assigned to experimental units using a randomized design. To compute the possible values of the test statistic, assume the null hypothesis is of no treatment effect so that the responses associated with each experimental unit would have been observed regardless of which treatment the unit received. The probability of the possible values is computed from the experimental design used to assign units to treatments. For a completely randomized design, all possible assignments of treatments

to experimental units are equally likely. The permutation distribution for a test statistic is also known as the **randomization distribution**.

- For a **simple permutation test**, we assume treatments are assigned to units by a randomized design and that we wish to test the null hypothesis of no difference in the effect of treatments on a response. We determine the permutation distribution and find the P -value by computing the probability of observing a response as or more extreme than the response actually observed according to the permutation distribution.
- To **sample with replacement** from a population, after each observation is selected, return it to the population before the next unit is sampled. Thus, the same unit can be selected multiple times when taking a random sample with replacement.
- The **bootstrap method** provides a way to approximate the sampling distribution of a statistic using the information in the original sample. This is done by generating many samples by sampling with replacement from the original sample. These samples with replacement are called **bootstrap samples**. For each bootstrap sample, compute the statistic of interest. The distribution of this statistic, computed from the bootstrap samples, is known as the **bootstrap distribution** and provides an approximation to the sampling distribution of the statistic.
- The **bootstrap standard error** for a statistic is the standard deviation of the bootstrap distribution of the statistic. In particular, it is the standard deviation of all values of the statistic generated by the bootstrap method.
- The **bootstrap percentile confidence interval** for a parameter treats the bootstrap distribution of the statistic used to estimate the parameter as the sampling distribution of the statistic. For a specific confidence coefficient, use the appropriate upper and lower percentiles, and mark off the central area of this distribution to form the confidence interval.

Statistics in Summary

Here are the most important skills you should have acquired from reading this chapter.

A. Permutation Tests

1. Compute the permutation distribution when a very small number of subjects are assigned, using a completely randomized design, to two treatment groups, for both paired and unpaired data.
2. From the permutation distribution, compute the P -value for a one-sided and two-sided test of significance for comparing two treatments.
3. Use software to simulate the permutation distribution when subjects are assigned, using a completely randomized design, to two treatment groups for paired or unpaired data.

B. Bootstrap Methods

1. Use software to generate many bootstrap samples from the original sample.
2. From bootstrap samples, generate the bootstrap distribution to approximate the sampling distribution of a statistic.
3. From the bootstrap distribution for a statistic, compute the bootstrap standard error.
4. From the bootstrap distribution for a statistic, compute the bootstrap percentile confidence interval for the corresponding parameter.

Link It

In Chapter 9, we discussed randomized designs. In Chapter 12, we discussed how to compute the probability of any possible assignment of treatments to experimental units for a completely randomized design. In Chapter 15, we introduced sampling distributions and showed how to compute P -values when a completely randomized design is used and we wish to test the null hypothesis of no treatment effects. In Chapter 17, we demonstrated how to compute P -values by simulating the sampling distribution of a statistic.

In this chapter, we formalize these ideas and discuss two resampling methods. First, we introduce permutation tests by expanding the discussions in Chapters 15 and 17. We also explain how to use software to conduct permutation tests for randomized designs with more than a small number of experimental units. Second, we introduce the bootstrap method as a way to approximate the sampling distribution of a statistic even when we do not know the distribution of the population from which the sample was selected. From this bootstrap distribution, we can compute P -values, standard errors of statistics, and confidence intervals for the parameter estimated by a statistic. Software is required to generate bootstrap distributions in practice.

Macmillan Learning Online Resources

If you are having difficulty with any of the sections of this chapter, these online resources should help prepare you to solve the exercises at the end of this chapter:

- **LearningCurve** provides you with a series of questions about the chapter catered to your level of understanding.

CHECK YOUR SKILLS

32.12 What is the effect of concussions on the brain? Researchers measured the brain sizes (hippocampal volume in microliters) of 25 collegiate football players with a history of clinician-diagnosed concussion and 25 collegiate football players without a history of concussion.⁹ Researchers planned to conduct a hypothesis test to see if there was evidence of a difference in the mean brain size between football players with a history of concussion and those without concussion. Which of the following statements is true?

- We must use a permutation test because we cannot be sure if the data come from a Normal population.
- We should not use a permutation test because permutation tests assume subjects are a random sample from some population and the football players in the study were not selected by random sampling.
- Neither (a) nor (b) is true.

32.13 The changing climate will probably bring more rain to California, but we don't know whether the additional rain will come during the winter wet season or extend into the long dry season in spring and summer. Kenwyn Suttle of the University of California at Berkeley and his coworkers carried out a randomized controlled experiment to study the effects of more rain in either season. They randomly assigned 12 plots of open grassland to two treatments: added water equal to 20% of annual rainfall during January to March (winter) or no added water (control). One response variable was total plant biomass, in grams per square meter, produced in

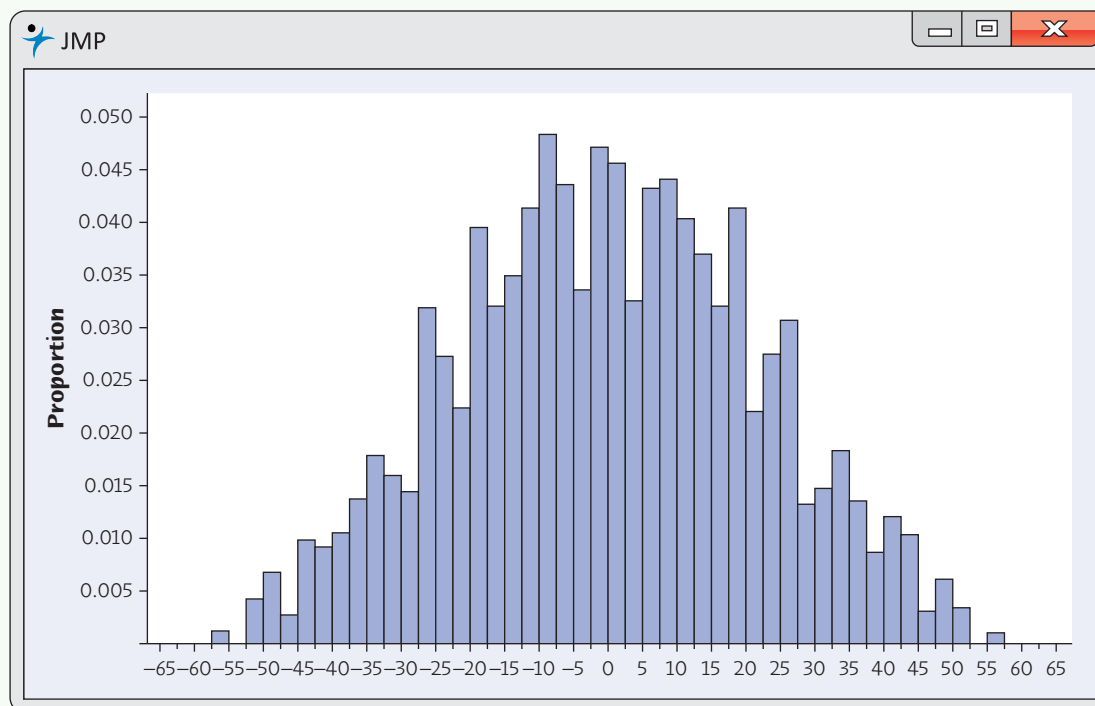
a plot over a year.¹⁰ Here are data for 2004 (mass in grams per square meter):

Winter	Control
254.6453	178.9988
233.8155	205.5165
253.4506	242.6795
228.5882	231.7639
158.6675	134.9847
212.3232	212.4862

We wish to test whether there is a difference in mean biomass between the two treatment groups. Which of the following is true?

- This is a randomized controlled experiment, hence a permutation test is more appropriate than a t test.
- This is a randomized controlled experiment, and we should try both the permutation test and the t test and always report only the one with the smaller P -value.
- We might prefer using a permutation test for these data rather than a t test, because the sample sizes are small and the data contain some possible outliers.

32.14 Figure 32.17 gives the estimated histogram for the permutation distribution, based on 10,000 simulated permutations, of the difference in sample means (mean for winter minus mean for the control) for the data in

**FIGURE 32.17**

The estimated histogram for the permutation distribution for the difference in sample medians, for Exercise 32.14.

Exercise 32.13. For purposes of testing whether there is a difference in mean biomass between the two treatment groups, we would estimate the P -value to be (assume a two-sided alternative)

- (a) larger than 0.05.
- (b) between 0.05 and 0.01.
- (c) smaller than 0.01.

32.15 Our bodies have a natural electrical field that is known to help wounds heal. Does changing the field strength slow healing? A series of experiments with newts investigated this question. In one experiment, the two hind limbs of four newts were assigned at random to either experimental or control groups. This is a matched pairs design. The electrical field in the experimental limbs was reduced to zero by applying a voltage. The control limbs were left alone. Here are the rates at which new cells closed a razor cut in each limb, in micrometers per hour:¹¹

Newt	1	2	3	4
Control limb	36	41	39	42
Experimental limb	28	31	27	33

The number of possible random assignments of treatments to the different matched pairs is

- (a) 4.
- (b) 8.
- (c) 16.

32.16 In Exercise 32.15, suppose we wish to test whether the healing rate is significantly lower in the experimental limbs using a permutation test. Even without listing all possible random assignments of treatments to the different matched pairs, we can conclude that the P -value from the permutation test is

- (a) larger than 0.05.
- (b) smaller than 0.05 but larger than 0.01.
- (c) smaller than 0.01.

32.17 We plan to use the bootstrap method to construct a confidence interval for a population median from a sample of 43 subjects from the population. An important assumption for using the bootstrap method is

- (a) the sample is a random sample from the population.
- (b) the sampling distribution for the sample median must not be well approximated by the Normal distribution.

(c) there are no outliers in the sample.

32.18 We select a random sample of six freshman students from the University of California at Santa Cruz and find that their verbal GREs are 480, 510, 590, 670, 520, and 630. Which of the following is not a possible bootstrap sample?

- (a) 480, 480, 480, 480, 480, 480
- (b) 480, 480, 480, 670, 670, 670
- (c) 480, 630, 630, 740, 590, 510

32.19 A 95% bootstrap percentile confidence interval for the population mean

- (a) is centered about the original sample mean.
(b) is centered about the mean of the bootstrap means.

- (c) uses the 2.5 and the 97.5 percentiles of the bootstrap distribution as the endpoints of the confidence interval.

CHAPTER 32 EXERCISES

32.20 How strong are durable press fabrics? “Durable press” cotton fabrics are treated to improve their recovery from wrinkles after washing. Unfortunately, the treatment also reduces the strength of the fabric. A study compared the breaking strengths of fabrics treated by two commercial durable press processes. Five swatches of the same fabric were assigned at random to each process. Here are the data, in pounds of pull needed to tear the fabric:¹²

 FABRICS

Permafresh	29.9	30.7	30.0	29.5	27.6
Hylite	28.8	23.9	27.0	22.1	24.2

There is a mild outlier in the Permafresh group. Perhaps we should use a permutation test to test the hypothesis of no difference in median pounds of pull needed to tear the fabric. Assume a two-sided alternative and estimate the P -value.

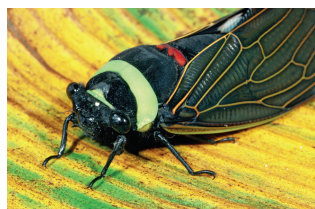
32.21 Do good smells bring good business? (software required) Exercise 21.9 describes an experiment that tested whether background aromas in a restaurant encourage customers to stay longer and spend more. The data on amount spent (in euros) are as follows:

 ODORS4

No Odor									
15.9	18.5	15.9	18.5	18.5	21.9	15.9	15.9	15.9	15.9
15.9	18.5	18.5	18.5	20.5	18.5	18.5	15.9	15.9	15.9
18.5	18.5	15.9	18.5	15.9	18.5	15.9	25.5	12.9	15.9
Lavender Odor									
21.9	18.5	22.3	21.9	18.5	24.9	18.5	22.5	21.5	21.9
21.5	18.5	25.5	18.5	18.5	21.9	18.5	18.5	24.9	21.9
25.9	21.9	18.5	18.5	22.8	18.5	21.9	20.7	21.9	22.5

Is there significant evidence that the lavender odor encourages customers to spend more? Use a permutation test to estimate the appropriate P -value.

32.22 Cicadas as fertilizer? (software required) Exercise 7.48 gives data from an experiment in which some bellflower plants in a forest were “fertilized”



Alastair Staley/Getty Images

with dead cicadas and other plants were not disturbed. The data record the mass of seeds produced by 39 cicada-supplemented plants and 33 undisturbed (control) plants. Here are data (seed mass in milligrams) for 39 cicada plants and 33 undisturbed (control) plants:¹³

 CICADA

Cicada Plants				Control Plants			
0.237	0.277	0.241	0.142	0.212	0.188	0.263	0.253
0.109	0.209	0.238	0.277	0.261	0.265	0.135	0.170
0.261	0.227	0.171	0.235	0.203	0.241	0.257	0.155
0.276	0.234	0.255	0.296	0.215	0.285	0.198	0.266
0.239	0.266	0.296	0.217	0.178	0.244	0.190	0.212
0.238	0.210	0.295	0.193	0.290	0.253	0.249	0.253
0.218	0.263	0.305	0.257	0.268	0.190	0.196	0.220
0.351	0.245	0.226	0.276	0.246	0.145	0.247	0.140
0.317	0.310	0.223	0.229	0.241			
0.192	0.201	0.211					

Do the data show that dead cicadas increase seed mass? Use a permutation test to estimate the appropriate P -value.

32.23 Adolescent obesity (software required). Adolescent obesity is a serious health risk affecting more than 5 million young people in the United States alone. Laparoscopic adjustable gastric banding has the potential to provide a safe and effective treatment. Fifty adolescents between 14 and 18 years old with a body mass index higher than 35 were recruited from Melbourne, Australia, for the study.¹⁴ Twenty-five were randomly selected to undergo gastric banding, and the remaining 25 were assigned to a supervised lifestyle intervention program involving diet, exercise, and behavior modification. All subjects were followed for two years. Here are the weight losses in kilograms for the subjects who completed the study. In the gastric banding group:

 ADOBESE

35.6 81.4 57.6 32.8 31.0 37.6 36.5 -5.4 27.9 49.0 64.8 39.0
43.0 33.9 29.7 20.2 15.2 41.7 53.4 13.4 24.8 19.4 32.3 22.0

In the lifestyle intervention group:

6.0 2.0 -3.0 20.6 11.6 15.5 -17.0 1.4 4.0
-4.6 15.8 34.6 6.0 -3.1 -4.3 -16.7 -1.8 -12.8

Does gastric banding result in significantly greater weight loss than a supervised lifestyle intervention program? Use a permutation test to estimate the appropriate P -value.

32.24 Ancient air (software required). The composition of the earth's atmosphere may have changed over time. To try to discover the nature of the atmosphere long ago, we can examine the gas in bubbles inside ancient amber. Amber is tree resin that has hardened and been trapped in rocks. The gas in bubbles within amber should be a sample of the atmosphere at the time the amber was formed. Measurements on specimens of amber from the late Cretaceous era (75 million to 95 million years ago) give these percentages of nitrogen:¹⁵



ANCTAIR

63.4 65.0 64.4 63.3 54.8 64.5 60.8 49.1 51.0

Assume (this is not yet agreed on by experts) that these observations are an SRS from the late Cretaceous atmosphere.

- Construct a 95% bootstrap confidence interval for the mean percentage of nitrogen in ancient air (the population).
- We wonder if ancient air differs significantly from the present atmosphere, which is 78.1% nitrogen. Based on your confidence interval in part (a), what do you conclude?

32.25 Water quality (software required). To investigate water quality, on August 8, 2010, the Columbus Dispatch took water samples at 20 Ohio State Park swimming areas. Those samples were taken to laboratories and tested for fecal coliform, which are bacteria found in human and animal feces. An unsafe level of fecal coliform means there's a higher chance that disease-causing bacteria are present and more risk that a swimmer will become ill, so it is important to estimate fecal coliform levels in park swimming areas. Here are the fecal coliform levels found by the laboratories:¹⁶



WQUAL

160 40 2800 80 2000 2000 1500 400 150 500
3000 2200 15 80 2000 2000 2600 600 1000 1500

We are willing to regard these particular 20 samples as an SRS from a large population of possible samples. Construct a 95% bootstrap confidence interval for the mean fecal coliform level in Ohio State Park swimming areas.

32.26 Exhaust from school buses (software required). In a study of exhaust emissions from school buses, the pollution intake by passengers was determined for a sample of nine school buses used in the Southern California Air Basin. The pollution intake is the amount of exhaust emissions, in grams per person, that would be breathed in while traveling on the bus during its usual 18-mile trip on congested freeways from South Central LA to a magnet school in West LA. (As a reference, the average intake of motor emissions of carbon monoxide in the LA area is estimated to be about 0.000046 gram per person.) Here are the amounts for the nine buses when driven with the windows open:¹⁷



EMIT

1.15 0.33 0.40 0.33 1.35 0.38 0.25 0.40 0.35



MARK J. TERRILL/AP Images

- Make a stemplot. Are there outliers or strong skewness that would forbid use of the t procedures?
- Construct a 95% bootstrap confidence interval for the mean pollution intake among all school buses used in the Southern California Air Basin that travel the route investigated in the study.

32.27 Pulling wood apart (software required). Exercise 1.46 gave data on how heavy a load is needed to pull apart pieces of Douglas fir. Construct a 95% bootstrap confidence interval for the mean load required to pull apart pieces of Douglas fir.



WOOD

32.28 Does nature heal best? (software required). Exercise 20.33 (page 472) gives these data on the healing rate (micrometers per hour) for cuts in the hind limbs of 12 newts:



NEWTS

Newt	1	2	3	4	5	6	7	8	9	10	11	12
Control limb	36	41	39	42	44	39	39	56	33	20	49	30
Experimental limb	28	31	27	33	33	38	45	25	28	33	47	23

The electrical field in the experimental limbs was reduced to zero by applying a voltage. The control limbs were not treated, so that they had their natural electrical field. Does changing the electrical field slow healing? Do a permutation test to answer this question.

EXPLORING THE WEB

32.29 Bootstrapping in practice. Bootstrapping is often used to construct confidence intervals for parameters more complicated than means. One example can be found online at the JAMA website (jama.jamanetwork.com/journal.aspx) in David J. Nyweide et al., “Association of pioneer accountable care organizations vs traditional medicare fee for service with spending, utilization, and patient experience,” *Journal of the American Medical Association* 313 (2015), pp. 2152–2162. Look at the end of the section titled “Claims Analyses.” The authors used bootstrapping to find a 95% confidence interval for a difference in differences of means. How many bootstrapped samples were used to determine the 95% confidence interval?

Another example at the JAMA website is Robert A. Fowler et al., “Cost-effectiveness of Dalteparin vs unfractionated Heparin for the prevention of venous thromboembolism in critically ill patients,” *Journal of the American Medical Association* 312 (2014), pp. 2135–2145. Look at the end of the section titled “Analytic Plan.” The authors used bootstrapping to construct a 95% confidence interval for incremental cost differences. How many bootstrap samples did they use?

Notes and Data Sources

1. The description of this experiment is based on the methodology described in “How to win the battle of the bugs,” *Consumer Reports*, July 2015, pp. 34–37. Although artificial, the data given in the example are consistent with the findings of *Consumer Reports*.
2. Pam A. Mueller and Daniel M. Oppenheimer, “The pen is mightier than the keyboard: Advantages of longhand over laptop note taking,” *Psychological Science*, 25, no. 6 (2014), pp. 1159–1168.
3. We thank Jason Hamilton, University of Illinois, for providing the data. The study is reported in Evan H. DeLucia et al., “Net primary production of a forest ecosystem with experimental CO₂ enhancement,” *Science*, 284 (1999), pp. 1177–1179. No method for inference can be trusted with $n = 3$. In this study, each observation is very costly, so the small n is inevitable.
4. Brock Bastian et al., “Pain as social glue: Shared pain increases cooperation,” *Psychological Science*, 25 (2014), pp. 2079–2085.
5. Fabrizio Grieco, Arie J. van Noordwijk, and Marcel E. Visser, “Evidence for the effect of learning on timing of reproduction in blue tits,” *Science*, 296 (2002), pp. 136–138. The data are from a graph in this paper.
6. Domenico Giannotti et al., “Play to become a surgeon: Impact of Nintendo Wii training on laparoscopic skills,” *PLOS ONE*, V8, e5272, February 2013, at www.plosone.org.
7. From the 2003 American Community Survey, at the U.S. Census Bureau website, www.census.gov. The data are a subsample of the 13,194 individuals in the ACS North Carolina sample who had travel times greater than zero.
8. For some discussion of more advanced methods for correcting for bias, see Companion Chapter 18 by T. C. Hesterberg et al. to D.S. Moore et al., *The Practice of Business Statistics*, 2nd ed., W.H. Freeman, 2008. A more advanced discussion can be found in B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, Chapman and Hall, 1993.
9. Rashmi Singh et al., “Relationship of collegiate football experience and concussion with hippocampal volume and cognitive outcomes,” *Journal of the American Medical Association*, 311 (2014), pp. 1883–1888.
10. K. B. Suttle, Meredith A. Thomsen, and Mary E. Power, “Species interactions reverse grassland responses to changing climate,” *Science*, 315 (2007), pp. 640–642. Here we present data for only winter and the control, omitting data for a treatment involving added water in the spring.

11. Data provided by Drina Iglesia, Purdue University. The data are part of a larger study reported in D. D. S. Iglesia, E. J. Cragoe, Jr., and J. W. Venable, "Electric field strength and epithelization in the newt (*Notophthalmus viridescens*)," *Journal of Experimental Zoology*, 274 (1996), pp. 56–62.
12. Sherri A. Buzinski, "The effect of position of methylation on the performance properties of durable press treated fabrics," CSR490 honors paper, Purdue University, 1985.
13. Louie H. Yang, "Periodical cicadas as resource pulses in North American forests," *Science*, 306 (2004), pp. 1565–1567. The data are simulated Normal values that match the means and standard deviations reported in this article.
14. Paul E. O'Brien et al., "Laparoscopic adjustable gastric banding in severely obese adolescents," *Journal of the American Medical Association*, 303 (2010), pp. 519–526. We thank the authors for providing the data.
15. R. A. Berner and G. P. Landis, "Gas bubbles in fossil amber as possible indicators for the major gas composition of ancient air," *Science*, 239 (1988), pp. 1406–1409.
16. This study was found online at www.dispatch.com/content/stories/local/2010/08/15/algae-isnt-only-problem-for-lakes.html.
17. J. D. Marshall et al., "Vehicle self-pollution intake fraction: Children's exposure to school bus emissions," *Environmental Science and Technology*, 39 (2005), pp. 2559–2563.