CHAPTER 14

Bootstrap Methods and Permutation Tests*

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

Introduction

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

The most common methods for inference about means based on a single sample, matched pairs, or two independent samples are the t procedures described in Chapter 7. For relationships between quantitative variables, we use other t tests and intervals in the correlation and regression setting (Chapter 10). Chapters 11, 12, and 13 present inference procedures for more elaborate settings. All of these methods rest on the use of normal distributions for data. No data are exactly normal. The t procedures are useful in practice because they are *robust*, quite insensitive to deviations from normality in the data. Nonetheless, we cannot use t confidence intervals and tests if the data are strongly skewed, unless our samples are quite large. Inference about spread based on normal distributions is not robust and is therefore of little use in practice. Finally, what should we do if we are interested in, say, a *ratio* of means, such as the ratio of average men's salary to average women's salary? There is no simple traditional inference method for this setting.

The methods of this chapter—bootstrap confidence intervals and permutation tests—apply computing power to relax some of the conditions needed for traditional inference and to do inference in new settings. The big ideas of statistical inference remain the same. The fundamental reasoning is still based on asking, "What would happen if we applied this method many times?" Answers to this question are still given by confidence levels and P-values based on the sampling distributions of statistics. The most important requirement for trustworthy conclusions about a population is still that our data can be regarded as random samples from the population—not even the computer can rescue voluntary response samples or confounded experiments. But the new methods set us free from the need for normal data or large samples. They also set us free from formulas. They work the same way (without formulas) for many different statistics in many different settings. They can, with sufficient computing power, give results that are more accurate than those from traditional methods. What is more, bootstrap intervals and permutation tests are conceptually simpler than confidence intervals and tests based on normal distributions because they appeal directly to the basis of all inference: the sampling distribution that shows what would happen if we took very many samples under the same conditions.

The new methods do have limitations, some of which we will illustrate. But their effectiveness and range of use are so great that they are rapidly becoming the preferred way to do statistical inference. This is already true in high-stakes situations such as legal cases and clinical trials.

Software

Bootstrapping and permutation tests are feasible in practice only with software that automates the heavy computation that these methods require. If you

are sufficiently expert, you can program at least the basic methods yourself. It is easier to use software that offers bootstrap intervals and permutation tests preprogrammed, just as most software offers the various t intervals and tests. You can expect the new methods to become gradually more common in standard statistical software.

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

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

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

14.5 Significance Testing Using Permutation Tests

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

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

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

3. Locate the observed statistic on this distribution. A value in the main body of the distribution could easily occur just by chance. A value in the tail would

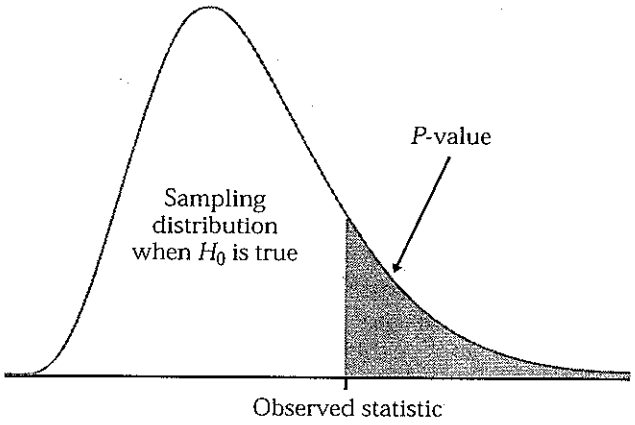

FIGURE 14.19 The *P*-value of a statistical test is found from the sampling distribution the statistic would have if the null hypothesis were true. It is the probability of a result at least as extreme as the value we actually observed.

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

null hypothesis

P-value

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

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

RESAMPLING FOR SIGNIFICANCE TESTS

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

EXAMPLE 14.11 Do new "directed reading activities" improve the reading ability of elementary school students, as measured by their Degree of Reading Power (DRP) scores? A study assigns students at random to either the new method

TABLE 14.3

Degree of Reading Power scores for third-graders

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

(treatment group, 21 students) or traditional teaching methods (control group, 23 students). The DRP scores at the end of the study appear in Table 14.3.[11] In Example 7.14 (page 489) we applied the two-sample t test to these data.

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

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

The null hypothesis H_0 for the resampling test is that the teaching method has no effect on the distribution of DRP scores. If H_0 is true, the DRP scores in Table 14.3 do not depend on the teaching method. Each student has a DRP score that describes that child and is the same no matter which group the child is assigned to. The observed difference in group means just reflects the accident of random assignment to the two groups.

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

permutation tests

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

- Choose 21 of the 44 students at random to be the treatment group; the other 23 are the control group. This is an ordinary SRS, chosen *without replacement.* It is called a **permutation resample.** Calculate the mean DRP score in each group, using the individual DRP scores in Table 14.3. The difference between these means is our statistic.

permutation resample

- Repeat this resampling from the 44 students hundreds of times. The distribution of the statistic from these resamples estimates the sampling distribution under the condition that H_0 is true. It is called a **permutation distribution.**

permutation distribution

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

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

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

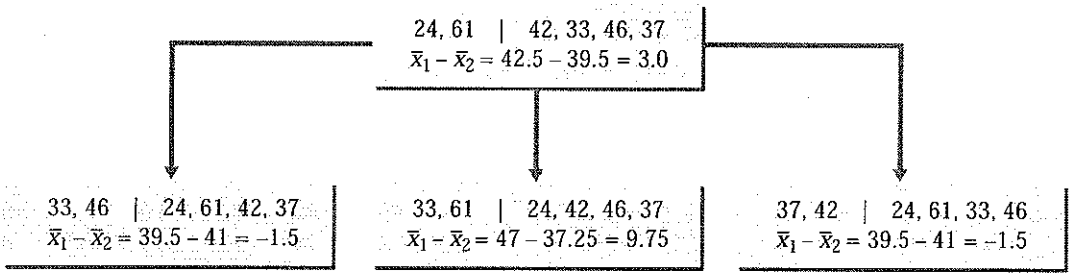

FIGURE 14.20 The idea of permutation resampling. The top box shows the outcomes of a study with four subjects in one group and two in the other. The boxes below show three permutation resamples. The values of the statistic for many such resamples form the permutation distribution.

Figure 14.20 illustrates permutation resampling on a small scale. The top box shows the results of a study with four subjects in the treatment group and two subjects in the control group. A permutation resample chooses an SRS of four of the six subjects to form the treatment group. The remaining two are the control group. The results of three permutation resamples appear below the original results, along with the statistic (difference of group means) for each.

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

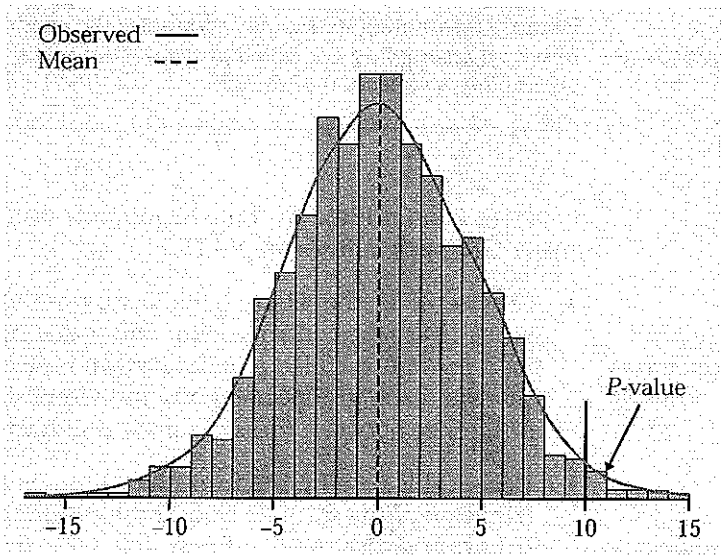

FIGURE 14.21 The permutation distribution of the statistic $\bar{x}_{\text{treatment}} - \bar{x}_{\text{control}}$ based on the DRP scores of 44 students. The dashed line marks the mean of the permutation distribution: it is very close to zero, the value specified by the null hypothesis. The solid vertical line marks the observed difference in means, 9.954. Its location in the right tail shows that a value this large is unlikely to occur when the null hypothesis is true.

when the null hypothesis H_0 is true. As H_0 suggests, the distribution is centered at 0 (no effect). The solid vertical line in the figure marks the location of the statistic for the original sample, 9.954. Use the permutation distribution exactly as if it were the sampling distribution: the P-value is the probability that the statistic takes a value at least as extreme as 9.954 in the direction given by the alternative hypothesis.

We seek evidence that the treatment increases DRP scores, so the alternative hypothesis is that the distribution of the statistic $\bar{x}_{\text{treatment}} - \bar{x}_{\text{control}}$ is centered not at 0 but at some positive value. Large values of the statistic are evidence against the null hypothesis in favor of this one-sided alternative. The permutation test P-value is the proportion of the 999 resamples that give a result at least as great as 9.954. A look at the resampling results finds that 14 of the 999 resamples gave a value 9.954 or larger, so the estimated P-value is 14/999, or 0.014.

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

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

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

Figure 14.21 shows that the permutation distribution has a roughly normal shape. Because the permutation distribution approximates the sampling distribution, we now know that the sampling distribution is close to normal. When the sampling distribution is close to normal, we can safely apply the usual two-sample t test. The t test in Example 7.14 gives $P = 0.013$, very close to the P-value from the permutation test.

Using software

In principle, you can program almost any statistical software to do a permutation test. It is more convenient to use software that automates the process of resampling, calculating the statistic, forming the permutation distribution, and finding the P-value. The menus in S-PLUS allow you to request permutation tests along with standard tests whenever they make sense. The permutation distribution in Figure 14.21 is one output. Another is this summary of the test results:

```
Number of Replications: 999

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

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

Permutation tests in practice

Permutation tests versus t tests. We have analyzed the data in Table 14.3 both by the two-sample t test (in Chapter 7) and by a permutation test.

Comparing the two approaches brings out some general points about permutation tests versus traditional formula-based tests.

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

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

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

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

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

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

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

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

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

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

To perform a permutation test, we randomly regroup the total set of repair times into two groups that are the same sizes as the two original samples. This is consistent with the null hypothesis that CLEC versus ILEC has no effect on repair time. Each repair time appears once in the data in each resample, but some repair times from the ILEC group move to CLEC, and vice versa. We calculate the test statistic for each resample and create its permutation distribution. The P-value is the proportion of the resamples with statistics that exceed the observed statistic.

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

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

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

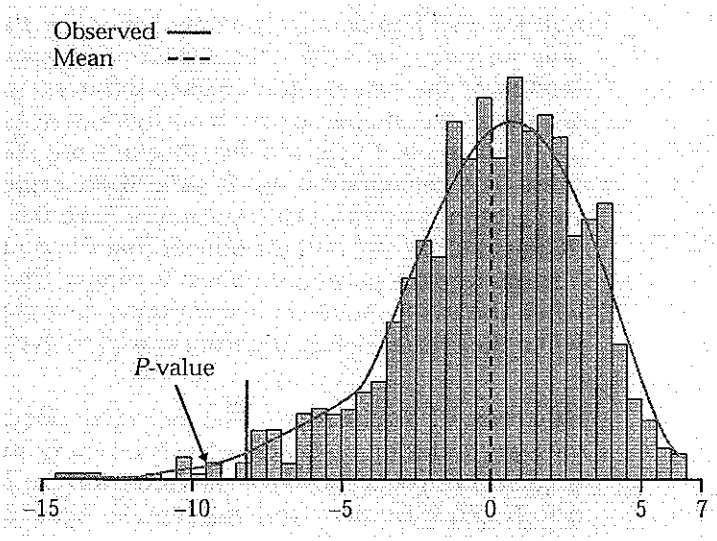

FIGURE 14.22 The permutation distribution of the difference of means $\bar{x}_1 - \bar{x}_2$ for the Verizon repair time data.

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

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

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

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

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

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

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

Sources of variation. Just as in the case of bootstrap confidence intervals, permutation tests are subject to two sources of random variability: the original sample is chosen at random from the population, and the resamples are chosen at random from the sample. Again as in the case of the bootstrap, the added variation due to resampling is usually small and can be made as small as we like by increasing the number of resamples. For example, Verizon uses 500,000 resamples.

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

Permutation tests in other settings

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

GENERAL PROCEDURE FOR PERMUTATION TESTS

To carry out a permutation test based on a statistic that measures the size of an effect of interest:

1. Compute the statistic for the original data.

2. Choose permutation resamples from the data without replacement in a way that is consistent with the null hypothesis of the test and with the study design. Construct the permutation distribution of the statistic from its values in a large number of resamples.

3. Find the P-value by locating the original statistic on the permutation distribution.

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

EXAMPLE 14.14 Can the full moon influence behavior? A study observed 15 nursing home patients with dementia. The number of incidents of aggressive behavior was recorded each day for 12 weeks. Call a day a "moon day" if it is the day of a full moon or the day before or after a full moon. Table 14.4 gives the average number of aggressive incidents for moon days and other days for each subject.[12] These are matched pairs data. In Example 7.7, the matched pairs t test found evidence that the mean number of aggressive incidents is higher on moon days ($t = 6.45$, df $= 14$, $P < 0.001$). The data show some signs of nonnormality. We want to apply a permutation test.

The null hypothesis says that the full moon has no effect on behavior. If this is true, the two entries for each patient in Table 14.4 are two measurements of aggressive behavior made under the same conditions. There is no distinction between "moon days" and "other days." Resampling in a way consistent with this null hypothesis randomly assigns one of each patient's two scores to "moon" and the other to "other." We don't mix results for different subjects, because the original data are paired.

TABLE 14.4

Aggressive behaviors of dementia patients

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

The permutation test (like the matched pairs t test) uses the difference of means $\bar{x}_{\text{moon}} - \bar{x}_{\text{other}}$. Figure 14.23 shows the permutation distribution of this statistic from 9999 resamples. None of these resamples produces a difference as large as the observed difference, $\bar{x}_{\text{moon}} - \bar{x}_{\text{other}} = 2.433$. The estimated one-sided P-value is therefore

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

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

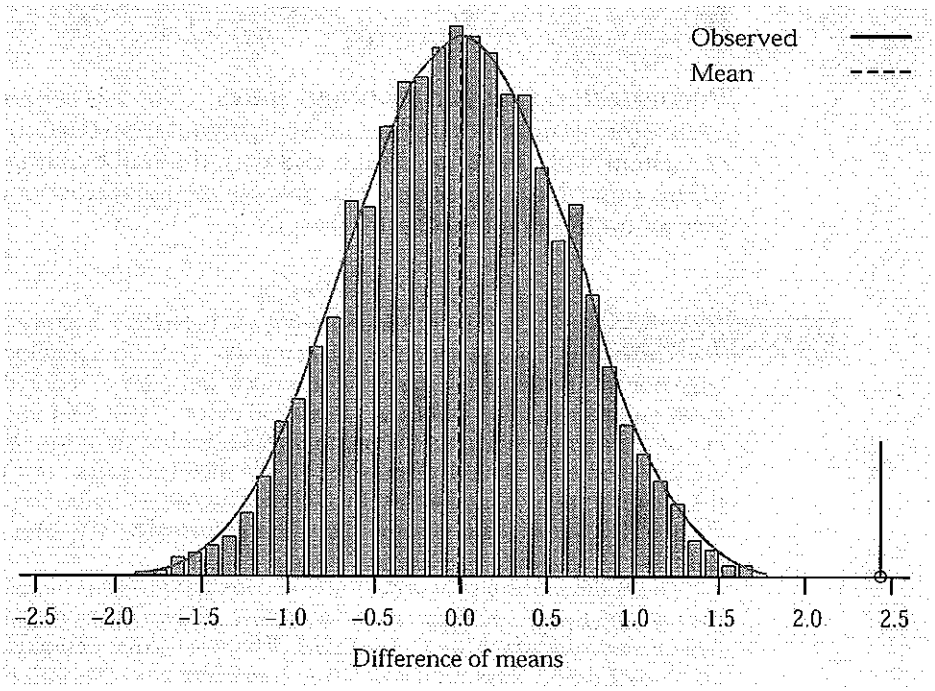

FIGURE 14.23 The permutation distribution for the mean difference (moon days versus other days) from 9999 paired resamples from the data in Table 14.5, for Example 14.14.

The permutation distribution in Figure 14.23 is close to normal, as a normal quantile plot confirms. The paired sample t test is therefore reliable and agrees with the permutation test that the P-value is very small.

Permutation test for the significance of a relationship. Permutation testing can be used to test the significance of a relationship between two variables. For example, in Example 14.9 we looked at the relationship between baseball players' batting averages and salaries.

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

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

When can we use permutation tests? We can use a permutation test only when we can see how to resample in a way that is consistent with the study design and with the null hypothesis. We now know how to do this for the following types of problems:

- **Two-sample problems** when the null hypothesis says that the two populations are identical. We may wish to compare population means, proportions, standard deviations, or other statistics. You may recall from Section 7.3 that traditional tests for comparing population standard deviations work very poorly. Permutation tests are a much better choice.
- **Matched pairs designs** when the null hypothesis says that there are only random differences within pairs. A variety of comparisons is again possible.
- **Relationships between two quantitative variables** when the null hypothesis says that the variables are not related. The correlation is the most common measure of association, but not the only one.

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

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

SECTION 14.5 | Summary

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

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

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

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

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

SECTION 14.5 | Exercises

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

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

Treatment group	24	61		
Control group	42	33	46	37

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

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

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

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

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

TABLE 14.5

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

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

(a) State the null and alternative hypotheses.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

CLEC										
1	1	5	5	5	1	5	5	5	5	

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

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

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

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

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

14.51 You want to test the equality of the means of two populations. Sketch density curves for two populations for which

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

(b) both permutation and t tests are valid.

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

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

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

14.53 Exercise 7.41 (page 482) gives data on a study of the effect of a summer language institute on the ability of high school language teachers to understand spoken French. This is a matched pairs study, with scores for 20 teachers at the beginning (pretest) and end (posttest) of the institute. We conjecture that the posttest scores are higher on the average.

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

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

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

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

(a) State the null and alternative hypotheses.

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

14.55 In Exercise 14.39, we assessed the significance of the *correlation* between returns on Treasury bills and common stocks by creating bootstrap confidence intervals. If a 95% confidence interval does not cover 0, the observed correlation is significantly different from 0 at the $\alpha = 0.05$ level. We would prefer to do a test that gives us a P-value. Carry out a permutation test and give the P-value. What do you conclude? Is your conclusion consistent with your work in Exercise 14.39?

14.56 The formal medical term for vitamin A in the blood is serum retinol. Serum retinol has various beneficial effects, such as protecting against fractures. Medical researchers working with children in Papua New Guinea asked whether recent infections reduce the level of serum retinol. They classified children as recently infected or not on the basis of other blood tests, then measured serum retinol. Of the 90 children in the sample, 55 had been recently infected. Table 14.6 gives the serum retinol levels for both groups, in micromoles per liter.[13]

TABLE 14.6

Serum retinol levels in two groups of children

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

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

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

14.57 In Exercise 14.56, we did a permutation test for the hypothesis "no difference between infected and uninfected children" using the ratio of mean serum retinol levels to measure "difference." We might also want a bootstrap confidence interval for the ratio of population means for infected and uninfected children. Describe carefully how resampling is done for the permutation test and for the bootstrap, paying attention to the difference between the two resampling methods.

14.58 Here is one conclusion from the data in Table 14.6, described in Exercise 14.56: "The mean serum retinol level in uninfected children was 1.255 times the mean level in the infected children. A 95% confidence interval for the ratio of means in the population of all children in Papua New Guinea is. . . ."

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

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

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

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

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

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

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

We want to use these sample data to test *equality of the population proportions* of successes. Carry out a permutation test. Describe the permutation distribution. The permutation test does not depend on a "nice" distribution shape. Give the P-value and report your conclusion.

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

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

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

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

TABLE 14.7

Effect of calcium and placebo on blood pressure

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

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

(a) State the null and alternative hypotheses.

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

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

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

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

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

CHAPTER 14 | Exercises

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