CHAPTER 18

Resampling and Nonparametric Approaches to Data

Objectives
To present resampling and nonparametric (distribution-free) procedures that can be used for testing hypotheses but which rely on less restrictive assumptions about populations than do previously discussed tests.

Contents
18.1 Bootstrapping as a General Approach
18.2 Bootstrapping with One Sample
18.3 Resampling with Two Related Samples
18.4 Resampling with Two Independent Samples
18.5 Bootstrapping Confidence Limits on a Correlation Coefficient
18.6 Wilcoxon’s Rank-Sum Test
18.7 Wilcoxon’s Matched-Pairs Signed-Ranks Test
18.8 The Sign Test
18.9 Kruskall–Wallis One-way Analysis of Variance
18.10 Friedman’s Rank Test for $k$ Correlated Samples
Most of the statistical procedures we have discussed throughout this book have involved estimation of one or more parameters of the distribution of scores in the population(s) from which the data were sampled and assumptions concerning the shape of that distribution. For example, the $t$ test uses the sample variance ($s^2$) as an estimate of the population variance ($\sigma^2$) and also requires the assumption that the population from which the sample was drawn is normal. Tests such as the $t$ test, which involve either assumptions about specific parameters or their estimation, are referred to as parametric tests.

There is a class of tests, however, that does not rely on parameter estimation and/or distribution assumptions. Such tests are usually referred to as nonparametric tests or distribution-free tests. By and large, if a test is nonparametric it is also distribution-free, and in fact it is the distribution-free nature of the test that is most valuable to us. Although the two names are often used interchangeably, the tests will be referred to here as nonparametric tests because that term is somewhat more common.

There is another approach to statistical analysis, which is predominantly nonparametric in nature, that has become considerably more popular in recent years due to the increased computing power we now enjoy. These are called resampling procedures, which can further be broken down into bootstrapping and randomization tests. I will discuss several of these that do not require strict parametric assumptions. These techniques are useful either when we are uncomfortable with the assumptions that a parametric test, such as $t$, would require, or when we just don’t have good parametric procedures to do what we want—such as forming a confidence interval on a median when we doubt that the distribution is normally distributed. I will discuss these procedures first because I believe that in a short time they will overtake what are now the more common nonparametric tests, and may eventually overtake the traditional parametric tests.

The major advantage generally attributed to nonparametric tests is also the most obvious—they do not rely on any very seriously restrictive assumptions concerning the shape of the sampled population(s). This is not to say that nonparametric tests do not make any distribution assumptions, but only that the assumptions they do require are far more general than those required for the parametric tests. The exact null hypothesis being tested may depend, for example, on whether or not two populations are symmetric or have a similar shape. None of these tests, however, makes an a priori assumption about the specific shape of the distribution; that is, the validity of the test is not affected by whether or not the variable is normally distributed in the population. A parametric test, on the other hand, usually includes some type of normality assumption, and, if that assumption is false, the conclusions drawn from that test may be inaccurate. In addition, some violations of parametric test assumptions may cause that test to be less powerful for a specific set of data than the corresponding nonparametric test. Perhaps the most articulate spokesperson for nonparametric/distribution-free tests has been Bradley (1968), who still has one of the clearest descriptions of the underlying assumptions and their role.

Another characteristic of nonparametric tests that often acts as an advantage is the fact that many of them, especially the ones discussed in this chapter, are more sensitive to medians than to means. Thus, if the nature of your data is such that you are interested primarily in medians, the tests presented here may be particularly useful to you.

Those who argue in favor of using parametric tests in almost every case do not deny that nonparametric tests are more liberal in the assumptions they require. They argue, however, that the assumptions normally cited as being required of parametric tests are overly restrictive in practice and that the parametric tests are remarkably unaffected by violations
of distribution assumptions. See Rasmussen (1987) for an example where parametric tests win out even with their assumptions violated.

The major disadvantage generally attributed to nonparametric tests is their (reputed) lower power relative to the corresponding parametric test. In general, when the assumptions of the parametric test are met, the nonparametric test requires somewhat more observations than does the comparable parametric test for the same level of power. Thus, for a given set of data, the parametric test is more likely to lead to rejection of a false null hypothesis than is the corresponding nonparametric test. Moreover, even when the distribution assumptions are violated to a moderate degree, the parametric tests are thought to maintain their advantage. A number of studies, however, have shown that for perfectly reasonable data sets nonparametric tests may have greater power than the corresponding parametric test. The problem is that we generally do not know when the nonparametric test will be more powerful.

Some nonparametric tests have an additional advantage. Since many of them rank the raw scores and operate on those ranks, they offer a test of differences in central tendency that are not affected by one or a few very extreme scores (outliers). An extreme score in a set of data actually can make the parametric test less powerful, because it inflates the variance, and hence the error term, as well as biasing the mean by shifting it toward the outlier (the latter may increase or decrease the mean difference).

Nonparametric tests can be divided into several different approaches. One group of tests, which we will discuss in the second half of the chapter, depends on ranking the data and carrying out the statistical test on the ranks. These are the most commonly known nonparametric procedures, and are particularly useful when the ranking procedure reduces problems with outliers. A second group of tests are broadly known under the title of “resampling statistics,” and these tests rely on drawing repeated samples from some population and evaluating the distribution of the resulting test statistic. Within the resampling statistics, the bootstrapping procedures, to be discussed next, rely on random sampling with replacement, from a population whose characteristics reflect the characteristics of the sample. Bootstrapping procedures are particularly important in those situations where we are interested in statistics, such as the median, whose sampling distribution and standard error cannot be derived analytically (i.e., from a standard formula, such as the formula for the standard error of the mean) unless we are willing to assume a normally distributed population.\(^2\) The next section will be an introduction to bootstrapping.

After looking at the bootstrap, we will move on to other resampling procedures that do not rely on drawing repeated samples, with replacement, from some population. Instead, we will consider all possible permutations, or rearrangements, of the data. These are often called permutation or randomization tests, and they are covered in Sections 18.2–18.4. Whereas bootstrapping involves sampling with replacement, permutation tests involve sampling without replacement.

### 18.1 Bootstrapping as a General Approach

Think for the moment about the standard \(t\) test on the difference between two population means. (Everything that I am about to say would apply, with only the obvious changes, if I had chosen any other parametric test, but the \(t\) test is a good example.) To carry out our \(t\) test we first assumed that we drew our samples from two normal populations and that the populations had the same variance \((\sigma^2)\). We then assume that the null hypothesis was true,

---

\(^2\) If the population is normally distributed, the standard error of the median is approximately 1.25 times the standard error of the mean. If the distribution is skewed, however, the standard error of the median cannot easily be calculated.
and ask what kinds of differences between means (or what values of $t$) we would expect if we drew an infinite number of pairs of samples from these normal populations, calculated the means, and then took their differences. Notice in all of this that we ask about sampling from normal populations with equal variances. To go one step further, if we actually computed all of these samples from the specified population, the resulting sampling distribution of $t$ would be the same as the tabled sampling distribution that we normally use to compute the probability of $t$ under the null hypothesis.

But suppose that we are not willing to assume that our data came from normal populations, or that we are not willing to assume that these populations had equal variances. Perhaps if we knew enough statistics, which neither you nor I do, and we were willing to assume that the populations have some other specified distribution (e.g., an exponential distribution), we could derive something comparable to our $t$ test, and use that for our purposes. Of course that test, if we could derive one, would still only apply when data come from that particular kind of distribution. But suppose that we think that our populations are not distributed according to any of the common distributions. Then what do we do? Bootstrapping gives us a way to solve this problem. Before I talk about how we would perform a bootstrapped hypothesis test, however, let’s look at another problem that we can deal with using the bootstrap.

If I asked you to calculate a confidence interval on a mean, and I told you that the population from which the data came was normal, you could solve the problem. In particular, you know that the standard error of the mean is equal to the population standard deviation (perhaps estimated by the sample standard deviation) divided by the square root of $n$. You could then measure off the appropriate number of standard errors from the mean using the normal (or $t$) distribution, and you would have your answer. But, suppose that I asked you for the confidence limit on the median instead of the mean. Now you are stuck, because you don’t have a nice simple formula to calculate the standard error of the median. So what do you do? Again, you use the bootstrap.

Macauley (1999, personal communication) collected mental status information on older adults. One of her dependent variables was a memory score on the Neurobehavioral Cognitive Status Examination for the 20 participants who were 80–84 years old. As you might expect, these data were negatively skewed, because some, but certainly not all, of her participants had lost some cognitive functioning. Her actual data are shown in Figure 18.1.

Figure 18.1  Sample distribution of memory scores for participants 80–84 years of age
Macauley wanted to establish confidence limits on the population median for this age group. Here she was faced with both problems outlined above. It does not seem reasonable to base that confidence interval on the assumption that the population is normally distributed (it most clearly is not), and we want confidence limits on a median, but don’t have a convenient formula for the standard error of the median. What’s a body to do?

What we will do is to assume that the population is distributed exactly as our sample. In other words, we will assume that the shape of the parent population is as shown in Figure 18.1.

It might seem like a substantial undertaking to create an infinitely large population of numbers such as that seen in Figure 18.1, but, in fact, it is trivially easy. All that we have to do is to take the sample on which it is based, as represented in Figure 18.1, and draw as many observations as we need, with replacement, from that sample. This is the way that all bootstrapping programs work, as you will see. In other words, 20 individual observations from an infinite population shaped as in Figure 18.1 is exactly the same as 20 individual observations drawn with replacement from the sample distribution. In the future when I speak of a population created to exactly mirror the shape of the sample data, I will refer to this as a pseudo-population.

18.2 Bootstrapping with One Sample

Macauley was interested in defining a 95% confidence interval on the median of memory scores of older participants. As I said above, she had reason to doubt that the population of scores was normally distributed, and there is no general formula defining the standard error of the median. But neither of those considerations interferes with computing the confidence interval she sought. All that she had to do was to assume that the shape of the population was accurately reflected in the distribution of her sample, then draw a large number of new samples (each of $n=20$) from that population. For each of these samples she computed the median, and when she was through she examined the distribution of these medians. She could then empirically determine those values that encompassed 95% of the sample medians.

It is quite easy to solve Macauley’s problem using a program named Resampling Stats by Simon and Bruce (1999). The syntax and the results are shown in Figure 18.2, and a histogram of the results is presented in Figure 18.3. There is no particular reason for you to learn the sequence of commands that are required for Resampling Stats, but a cursory look at the program is enlightening. The first two lines of the program describe the problem and set aside sufficient space to store 10,000 sample medians. Then the data are read in to create a pseudo-population from which we can sample with replacement. The next two lines calculate and print the median of the original sample. At this point the program goes into a loop that repeats 10,000 times, each time drawing a sample of 20 observations from our pseudo-population, computing its median, and labeling that median as “bmedian.” After 10,000 medians have been drawn and stored in an array called “medians,” the program prints a frequency distribution and histogram of the results, calculates the standard deviation of these medians, which is the standard error of the median, and prints that. The amazing thing is that it probably took me 5 minutes to compose, type, and revise this paragraph, while it took the program 7.8 seconds to draw those 10,000 samples and print the results.

The results in Figures 18.2 and 18.3 are interesting for several reasons. In the first place, they show you what happens when you try to calculate medians of a large number of relatively small samples. The distribution in Figure 18.3 is quite discrete, because the median is going to be the middle value in a limited set of numbers. You couldn’t get a
median of 9.63, for example, no matter how many samples you drew. For this particular population the medians must be an integer (or the average of two integers in the ordered array) between 5 and 11. There are no other possibilities.

Ideally, to calculate a 95% confidence interval we would like to find those outcomes that cut off 2.5% of the observations at each end of the distribution. With the very discrete distribution we have with medians, there is no point that cuts off the lowest 2.5% of the distribution. At the extreme, $4/10,000 = .04\%$ lie at or below a median of 5, and $(496 + 4)/10,000 = 5.00\%$ lie at or below a median of 6. At the other end of the distribution, $4006/10,000 = 40.6\%$ lie at or below 9, and $9997/10,000 = 99.97\%$ lie at or below 10. To be conservative we would choose the extremes of each of these sets, and put the confidence interval at 5–10, which includes virtually all of the distribution. We really have a 99.97% confidence interval, which is probably close enough for any purpose to which we would be likely to put these data. If we were willing to let the lower bound represent the 5% point, we would have an interval at 6–10. What is important

---

3 This is the simplest approach to obtaining confidence limits, and relies on the 2.5 and 97.5 percentiles of the sampling distribution of the median. There are a number of more sophisticated estimators, but the one given here best illustrates the approach.

---

Figure 18.2 Resampling Stats program and results bootstrapping the sample median 10,000 times
here, and the reason why Macauley wanted these limits in the first place, is that for this
memory test, the lower bound of what is classed as “normal functioning” is a score of 10.
The confidence interval does include 10 as its upper limit, and so we cannot reject the
null hypothesis that people in this age group, on average, fall in the normal range. An
examination of the sampling distribution reinforces this view, and perhaps gives us a
more complete understanding of the performance of this age cohort. The fact that there
are a number of individuals whose scores are well below 10 might lead us to seek a dif-
ferent confidence interval, that being limits on the proportion of people in that age
group who fall below 10. While that would be a perfectly legitimate use of bootstrapp-
ing for these data, we will not pursue that question here.

This may not seem like the most inspiring example of bootstrapping, because it makes
bootstrapping look rather imprecise. It is a good example nonetheless, because it reflects
the sometimes awkward nature of real data. As we will see, however, not all data lead to
such discrete distributions. In addition, the discreteness of the result is inherent in the data,
not merely in the process itself. If we drew 10,000 samples from this population and calcu-
lated \( t \) values, the resulting \( t \) distribution would be almost as discrete. The problem comes
from drawing samples from a distribution with a limited number of different values, in-
stead of modeling the results of drawing from continuous (e.g., normal) distributions. If it
is not reasonable to assume normality, it is not reasonable to draw from normal distribu-
tions just to get a prettier graph.

### 18.3 Resampling with Two Paired Samples

We will now move from the bootstrap, where we drew large numbers of samples from a
pseudo-population using sampling with replacement, to randomization, or permutation,
procedures that involve taking the full set of observations and randomly shuffling them and
assigning them to conditions randomly.
Hoaqlin, Mosteller, and Tukey (1983) looked at the role of beta-endorphins in response to stress. They were interested in testing whether beta-endorphin levels rose in stressful situations. They recorded beta-endorphin levels in 19 patients 12 hours before surgery and again, for the same patients, 10 minutes before surgery. The data follow in fmol/ml.

<table>
<thead>
<tr>
<th></th>
<th>12 hours</th>
<th>10 min.</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10.0</td>
<td>20.0</td>
<td>10.0</td>
</tr>
<tr>
<td></td>
<td>6.5</td>
<td>14.0</td>
<td>7.5</td>
</tr>
<tr>
<td></td>
<td>8.0</td>
<td>13.5</td>
<td>5.5</td>
</tr>
<tr>
<td></td>
<td>12.0</td>
<td>18.0</td>
<td>6.0</td>
</tr>
<tr>
<td></td>
<td>5.0</td>
<td>9.0</td>
<td>2.5</td>
</tr>
<tr>
<td></td>
<td>11.5</td>
<td>18.0</td>
<td>2.0</td>
</tr>
<tr>
<td></td>
<td>5.0</td>
<td>6.5</td>
<td>-4.0</td>
</tr>
<tr>
<td></td>
<td>3.5</td>
<td>7.4</td>
<td>2.9</td>
</tr>
<tr>
<td></td>
<td>7.5</td>
<td>6.0</td>
<td>1.5</td>
</tr>
<tr>
<td></td>
<td>5.8</td>
<td>25.0</td>
<td>19.2</td>
</tr>
<tr>
<td></td>
<td>4.7</td>
<td>20.3</td>
<td>15.6</td>
</tr>
</tbody>
</table>

Because these are paired scores, we are primarily interested in the difference scores. We want to test the null hypothesis that the average difference score was 0.0, which would indicate that there was no change in endorphin levels on average. The difference scores are shown in the bottom line of the table, where it is clear that most differences are positive, and those that are negative are relatively small. If you were to plot the differences in this example, you would find that they are very positively skewed, which might discourage us from using a standard parametric \( t \) test. Moreover, if we were particularly interested in the median of the differences, a \( t \) test would not be appropriate. We will solve our problem by drawing on resampling statistics.

Our resampling procedure is based on the idea that if the null hypothesis is true, a patient’s 10-minute score was just as likely to be larger than his 12-hour score as it was to be smaller. If a patient has scores of 8.0 and 13.5, and if the null hypothesis is true, the 13.5 could just as likely come from the 12-hour measurement as from the 10-minute measurement. Under \( H_0 \), each difference had an equal chance of being positive or negative. This tells us how to model what the data would look like under \( H_0 \). We will simply draw a very large number of samples of 19 difference scores each, in such a way that the difference score has a 50:50 chance of being positive or negative. For each sample we will calculate the median of the differences, and then plot the sampling distribution of these differences. Remember, this is the sampling distribution of the differences when \( H_0 \) is true. We can compare our obtained median difference against this distribution to test \( H_0 \).

The way that we will conduct this test using Simon and Bruce’s *Resampling Stats* is to take all 19 difference scores and randomly attach the sign of the difference. (*Assigning the sign at random is exactly equivalent to randomly assigning one score to the 12-hour condition and the other to the 10-minute condition.*) We will then calculate the median difference and store that. This procedure will be repeated many times (in this case, 10,000 times). The program and results are shown in Figure 18.4, with the resulting histogram in Figure 18.5.

From Figure 18.4 we can see that the obtained median difference score was 6. From either the frequency distribution in Figure 18.4 or the histogram in Figure 18.5 we see the results of drawing 10,000 samples from a model in which the null hypothesis is true.

---

\(^4\) I have made two very trivial changes to avoid difference scores of 0.0, just to make the explanation easier. With differences of zero, we normally simply remove those cases from the data.
Figure 18.4  Resampling program and results for beta-endorphin data
Figure 18.5 is reassuring because it shows us that when the null is true, the resampled medians are distributed symmetrically about 0, which is what we would expect. From that figure we can also see that our obtained median of 6 is certainly extreme under \( H_0 \). Going back to Figure 18.4 we see that there were 10 resampled medians as large as 6, and 13 resampled values as low as −6. If we want a two-tailed test, the probability of a median as extreme as the one we obtained is \((10 + 13)/10,000 = .0023\), which is certainly a small probability. These results, thus, tell us that if we were sampling from a model where \( H_0 \) is true, the probability is very small that we would obtain a sample median as extreme as the one we obtained. Therefore we will reject the null hypothesis and conclude that beta-endorphin levels do increase as the time for surgery approaches. This is really a very good thing, because endorphins act as the body’s pain pills.

18.4 Resampling with Two Independent Samples

Now we will move on to the resampling equivalent of the \( t \) test for two independent samples. The example we will use involves data collected by Epping-Jordan, Compas, and Howell (1994) on the effect of avoidance on women’s recovery from breast cancer. Epping-Jordan was interested in examining the question of whether people who try to actively avoid thinking about their cancer have a poorer prognosis over a one-year period than those who do not report high levels of avoidance behavior. She collected data on the incidence of avoidance shortly after patients had been diagnosed with breast cancer. At the end of one year she sorted patients into those who were in remission (49 cases) and those who were no better or who had died (28 cases). These groups were labeled Success and Fail, respectively. The data are shown in Table 18.1. Epping-Jordan then compared the earlier reported level of avoidance for the two groups.

For this example we will compare the medians of the two groups, although we could just as easily compare their means.

If the null hypothesis is true in Epping-Jordan’s case, the two samples (Success and Fail) can be thought of as having been drawn from one population. Any particular Avoidance
A score would be as likely to belong to the Success group as to belong to the Fail group. We could model this null situation by assigning a random sample of 49 of the scores to the Success group and the remaining 18 scores to the Fail group. (Notice here that we are sampling without replacement.) The difference between those two groups’ medians would be an example of a median difference that we might reasonably obtain under $H_0$. We could repeat this procedure (randomly assigning 49 scores to the Success group and 18 scores to the Fail group) many times, and look at the median differences we obtain. Finally, we could compare the difference we actually found with those we obtained when we modeled the null hypothesis.

The above procedure is quite easy to do, because we simply shuffle the complete data set, split the result into the first 49 cases and the last 18 cases, compute and record the medians and the median differences, shuffle the data again, and repeat this process 10,000 times. The result of such a procedure is shown in Figures 18.6 and 18.7. I have omitted the program syntax because it would not add to the presentation.

### Table 18.1 Data on avoidance from Epping-Jordon et al. (1994)

<table>
<thead>
<tr>
<th>Success</th>
<th>Fail</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>14</td>
</tr>
<tr>
<td>23</td>
<td>17</td>
</tr>
<tr>
<td>20</td>
<td>18</td>
</tr>
<tr>
<td>8</td>
<td>23</td>
</tr>
<tr>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>13</td>
<td>15</td>
</tr>
<tr>
<td>13</td>
<td>8</td>
</tr>
<tr>
<td>16</td>
<td>12</td>
</tr>
<tr>
<td>10</td>
<td>17</td>
</tr>
<tr>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td>20</td>
<td>14</td>
</tr>
<tr>
<td>9</td>
<td>23</td>
</tr>
<tr>
<td>13</td>
<td>11</td>
</tr>
</tbody>
</table>

Median

<table>
<thead>
<tr>
<th>Success</th>
<th>Fail</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>17</td>
</tr>
</tbody>
</table>

$n$

<table>
<thead>
<tr>
<th>Success</th>
<th>Fail</th>
</tr>
</thead>
<tbody>
<tr>
<td>49</td>
<td>18</td>
</tr>
</tbody>
</table>

### MEDSUCC = 13
Median for Success group

### MEDFAIL = 17
Median for Fail group

### MEDDIFF = 4
Difference in Medians

### GREATER = 222
Number of difference $\geq 4$

### LESS = 264
Number of differences $< -4$

### MOREEXT = 486
486/10,000 = .0486 = Probability of this difference under the null hypothesis

### Figure 18.6 Summary results of resampling from Epping-Jordan et al. data
From Figure 18.6 we can see that the median Avoidance score for the Success group was 13, and the median for the Fail group was 17. The group who failed to improve exhibited more avoidance behavior early in treatment. The difference in median avoidance is 24. From the output you can also see that when we model the null hypothesis, 222 of the resamples were greater than a difference of 4, 264 results were less than a difference of 4, and 486 results were greater than ±4. Out of 10,000 samples, this represents 486/10,000 = 4.86% of the cases. Thus only 4.9% of the resampling statistics were more extreme than our result, and we can reject the null hypothesis at α = .05. We can conclude that those in the Fail group experienced significantly more avoidance behavior early in treatment than those who later were classed as successes.5

18.5 Bootstrapping Confidence Limits on a Correlation Coefficient

The standard approach to correlation problems is to calculate a correlation coefficient and then to apply a hypothesis test with the hope of showing that the correlation is significantly different from 0.00. However, there are a lot of significant correlations that are so low that they are not particularly important, even if they are significantly different from 0.00. Along with the recent emphasis on effect size measures should be an increase in the use of confidence limits.

As we saw in Chapter 9, Fisher’s arcsine transformation

\[ r' = (0.5) \log_e \left( \frac{1 + r}{1 - r} \right) \]

provides one way to adjust for the skewed sampling distribution of \( r \) when \( \rho \neq 0 \). An attractive alternative is to draw bootstrapped samples on the assumption that the bivariate

5 If we had run a standard \( t \) test on the means of these data, that probability would have been .0397.
data reflect the relationship in the population, and then to obtain confidence limits simply by taking the cutoffs for the $\alpha/2$ percent of each end of the distribution.

As an example, we can look at the data from Macauley on the mental status scores of older adults. Macauley’s data included 123 adults between the ages of 60 and 97, and we can look at the relationship between memory performance and age. We would probably expect to see a negative correlation between the two variables, but the significance of the correlation is not as useful as confidence limits on this correlation, which give us a better sense of how strong the relationship really is.

The bootstrap approach to obtaining these confidence limits would involve sampling 123 cases, with replacement, from the $XY$ pairs in the sample, computing the correlation between the variables, and repeating this a large number of times. We then find the 2.5 and 97.5 percentile of the sampling distribution, and that gives us our 95% confidence limits.

I have written a Windows program, which is available at www.uvm.edu/~dhowell/StatPages/ that will carry out this procedure. (It will also calculate a number of other resampling procedures.) The results of drawing 2000 resamples with replacement from the pseudo-population of pairs of scores are shown in Figure 18.8.

![Bootstrapping Correlations](image)

**Figure 18.8** Sampling distribution and confidence limits on correlation between age and memory performance in older adults
In the center of this figure you can see the sampling distribution of $r$. To the left is the obtained correlation ($-0.268$) and upper and lower confidence limits. These are $-0.43$ and $-0.11$. Because they are both on the same side of 0.00, we also know that our correlation is significant. The confidence interval may strike you as surprisingly wide, but confidence intervals on correlation coefficients often are.

The example from Macauley involved a fairly low correlation coefficient that, because it was only $-0.268$, was nearly symmetrically distributed around 0.00. If we run the same analysis on the beta-endorphin data that we used earlier, we can easily see the skewed nature of the sampling distribution for large correlations. This result is shown in Figure 18.9.

Figure 18.9 presents two interesting results. In the first place, notice that, because the correlation is fairly large ($r = 0.699$), the sampling distribution is very negatively skewed. In addition, notice how asymmetrical the confidence limits are. The upper limit is 0.91, which is a bit more than 20 points higher than $r$. However, the lower limit is 0.11, which is approximately 59 points lower. Whenever we have large correlations the sampling distribution will be skewed and our confidence limits will be asymmetrical.
An excellent discussion of bootstrapped estimates of confidence limits can be found in Mooney and Duval (1983). They discuss corrections for bias that are relatively easy to apply. Excellent sources on both bootstrapping and randomization tests can be found in Edgington (1995), Manly (1997), and Efron and Tibshirani (1993). Efron has probably been the most influential developer of the bootstrap approach, and his book with Tibshirani is an important source. Good (2000) has a presentation of permutation tests, and Lunnenborg (2000) addresses resampling methods at a sophisticated, but very readable, level.

Additional information on resampling and bootstrapping is available from the website that I maintain at http://www.uvm.edu/~dhowell/StatPages/StatHomePage.html. These particular pages cover the whole philosophy behind resampling procedures and the ways in which they differ from parametric procedures. This is a rapidly expanding field, and a wealth of new results are being published on a regular basis.

Although I happen to like my own programs best, for obvious personal reasons, the R programming environment, which is free and can be downloaded at www.r-project.org, and its commercial application S-Plus, do an excellent job of handling resampling procedures because of their flexibility and the way they implement repetitive sampling. However the language is not easy to learn.

### 18.6 Wilcoxon’s Rank-Sum Test

We will now move away from bootstrapping and randomization to the more traditional non-parametric tests. One of the most common and best-known of these tests is the Wilcoxon rank-sum test for two independent samples. This test is often thought of as the nonparametric analogue of the $t$ test for two independent samples, although it tests a slightly different, and broader, null hypothesis. Its null hypothesis is the hypothesis that the two samples were drawn at random from identical populations (not just populations with the same mean), but it is especially sensitive to population differences in central tendency. Thus, rejection of $H_0$ is generally interpreted to mean that the two distributions had different central tendencies, but it is possible that rejection actually resulted from some other difference between the populations. Notice that when we gain one thing (freedom from assumptions) we pay for it with something else (loss of specificity).

The logical basis of Wilcoxon’s rank-sum test is particularly easy to understand. Assume that we have two independent treatment groups, with $n_1$ observations in group 1 and $n_2$ observations in group 2. Further assume that the null hypothesis is false to a very substantial degree and that the population from which group 1 scores have been sampled contains values generally lower than the population from which group 2 scores were drawn. Then, if we were to rank all $n_1 + n_2 = N$ scores from lowest to highest without regard to group membership, we would expect that the lower ranks would fall primarily to group 1 scores and the higher ranks to group 2 scores. Going one step further, if we were to sum the ranks assigned to each group, the sum of the ranks in group 1 would be expected to be appreciably smaller than the sum of the ranks in group 2.

Now consider the opposite case, in which the null hypothesis is true and the scores for the two groups were sampled from identical populations. In this situation if we were to rank all $N$ scores without regard to group membership, we would expect some low ranks and some high ranks in each group, and the sum of the ranks assigned to group 1 would be roughly equal to the sum of the ranks assigned to group 2. These situations are illustrated in Table 18.2.

Wilcoxon bused his test on the logic just described, using the sum of the ranks in one of the groups as his test statistic. If that sum is too small relative to the other sum, we will reject the null hypothesis. More specifically, we will take as our test statistic the sum of the
Given this value, we can use tables of the Wilcoxon statistic \( W_S \) to test the null hypothesis. To take a specific example, consider the following hypothetical data on the number of recent stressful life events reported by a group of Cardiac Patients in a local hospital and a control group of Orthopedic Patients in the same hospital. It is well known that stressful life events (marriage, new job, death of spouse, and so on) are associated with illness, and it is reasonable to expect that, on average, many cardiac patients would have experienced more recent stressful events than would orthopedic patients (who just happened to break an ankle while tearing down a building or a leg while playing touch football). It would appear from the data that this expectation is borne out. Since we have some reason to suspect that life stress scores probably are not symmetrically distributed in the population (especially for cardiac patients, if our research hypothesis is true), we will choose to use a nonparametric test. In this case, we will use the Wilcoxon rank-sum test because we have two independent groups.

<table>
<thead>
<tr>
<th>Cardiac Patients</th>
<th>Orthopedic Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw Data</td>
<td>32 8 7 29 5 0</td>
</tr>
<tr>
<td>Ranks</td>
<td>11 9 8 10 6 1</td>
</tr>
</tbody>
</table>

To apply Wilcoxon’s test we first rank all 11 scores from lowest to highest, assigning tied ranks to tied scores (see the discussion on ranking in Chapter 10). The orthopedic group is the smaller of the two and, if those patients generally have had fewer recent stressful life events, then the sum of the ranks assigned to that group should be relatively low. Letting \( W_S \) stand for the sum of the ranks in the smaller group (the orthopedic group), we find

\[
W_S = 2 + 3.5 + 3.5 + 5 + 7 = 21
\]

We can evaluate the obtained value of \( W_S \) by using Wilcoxon’s table (Appendix \( W_S \)), which gives the smallest value of \( W_S \) that we would expect to obtain by chance if the null hypothesis were true. From Appendix \( W_S \) we find that for \( n_1 = 5 \) subjects in the smaller group and \( n_2 = 6 \) subjects in the larger group (\( n_1 \) is always the number of subjects in the smaller group if group sizes are unequal), the entry for \( \alpha = .025 \) (one-tailed) is 18. This means that for a difference between groups to be significant at the one-tailed .025 level, or the two-tailed .05 level, \( W_S \) must be less than or equal to 18. Since we found \( W_S \) to be equal to 21, we cannot reject \( H_0 \). (By way of comparison, if we ran a \( t \) test on these data, ignoring

---

6 Because the sum of the ranks in the smaller group and the sum of the ranks in the larger group sum to a constant, we only need to use one of those sums.
the fact that one sample variance is almost 50 times the other and that the data suggest that our prediction of the shape of the distribution of cardiac scores may be correct, \( t \) would be 1.92 on 9 \( df \), a nonsignificant result with \( p = .110 \). Using a resampling program on the means of the raw data, the probability of an outcome this extreme would be .059. A similar test on medians would yield \( p = .059 \).

The entries in Appendix \( W_S \) are for a one-tailed test and will lead to rejection of the null hypothesis only if the sum of the ranks for the smaller group is sufficiently small. It is possible, however, that the larger ranks could be congregated in the smaller group, in which case if \( H_0 \) is false, the sum of the ranks would be larger than chance expectation rather than smaller. One rather awkward way around this problem would be to rank the data all over again, this time ranking from high to low. If we did this, then the smaller ranks would now appear in the smaller group and we could proceed as before. We do not have to go through the process of reranking the data, however. We can accomplish the same thing by using the symmetric properties of the distribution of the rank sum by calculating a statistic called \( W'_S \). The statistic \( W'_S \) is the sum of the ranks for the smaller group that we would have found if we had reversed our ranking and ranked from highest to lowest:

\[
W'_S = 2\overline{W} - W_S
\]

where \( 2\overline{W} = n_1(n_1 + n_2 + 1) \) and is shown in the table in Appendix \( W_S \). We can then evaluate \( W'_S \) against the tabled value and have a one-tailed test on the upper tail of the distribution. For a two-tailed test of \( H_0 \) (which is what we normally want), we calculate \( W_S \) and \( W'_S \), enter the table with whichever is smaller, and double the listed value of \( \alpha \).

To illustrate \( W_S \) and \( W'_S \), consider the two sets of data in Table 18.3. Notice that the two data sets exhibit the same degree of extremeness, in the sense that for the first set four of the five lowest ranks are in group 1, and in the second set four of the five highest ranks are in group 1. Moreover, \( W_S \) for set 1 is equal to \( W'_S \) for set 2, and vice versa. Thus, if we establish the rule that we will calculate both \( W_S \) and \( W'_S \) for the smaller group and refer the smaller of \( W_S \) and \( W'_S \) to the tables, we will come to the same conclusion with respect to the two data sets.

### The Normal Approximation

Appendix \( W_S \) is suitable for all cases in which \( n_1 \) and \( n_2 \) are less than or equal to 25. For larger values of \( n_1 \) and/or \( n_2 \), we can make use of the fact that the distribution of \( W_S \) approaches a normal distribution as sample sizes increase. This distribution has

\[
\text{Mean} = \frac{n_1(n_1 + n_2 + 1)}{2}
\]

### Table 18.3 Sample data for Wilcoxon’s rank-sum test

<table>
<thead>
<tr>
<th>Set 1</th>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X )</td>
<td>2 15 16 19</td>
<td>18 23 25 37 82</td>
</tr>
<tr>
<td>Ranks</td>
<td>1 2 3 5</td>
<td>4 6 7 8 9</td>
</tr>
<tr>
<td>( W_S = 11 )</td>
<td>( W'_S = 29 )</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Set 2</th>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X )</td>
<td>60 40 24 21</td>
<td>23 18 15 14 4</td>
</tr>
<tr>
<td>Ranks</td>
<td>9 8 7 5</td>
<td>6 4 3 2 1</td>
</tr>
<tr>
<td>( W_S = 29 )</td>
<td>( W'_S = 11 )</td>
<td></td>
</tr>
</tbody>
</table>
and

\[ \text{Standard error} = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} \]

Since the distribution is normal and we know its mean and standard deviation (the standard error), we can calculate

\[ z = \frac{\text{Statistic} - \text{Mean}}{\text{Standard deviation}} = \frac{W_S - \frac{n_1 (n_1 + n_2 + 1)}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} \]

and obtain from the tables of the normal distribution an approximation of the true probability of a value of \( W_S \) at least as low as the one obtained. (It is immaterial whether we use \( W_S \) or \( W'_S \) in this situation, since they will produce equal values of \( z \), differing only in sign.)

To illustrate the computations for the case in which the larger ranks fall in the smaller groups and to illustrate the use of the normal approximation (although we do not really need to use an approximation for such small sample sizes), consider the data in Table 18.4. These

<table>
<thead>
<tr>
<th>Table 18.4</th>
<th>Hypothetical data on birthweight of infants born to mothers with different levels of prenatal care</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Beginning of Care</strong></td>
<td><strong>Third Trimester</strong></td>
</tr>
<tr>
<td>Birthweight Rank</td>
<td>Birthweight Rank</td>
</tr>
<tr>
<td>1680 2</td>
<td>2940 10</td>
</tr>
<tr>
<td>3830 17</td>
<td>3380 16</td>
</tr>
<tr>
<td>3110 14</td>
<td>4900 18</td>
</tr>
<tr>
<td>2760 5</td>
<td>2810 9</td>
</tr>
<tr>
<td>1700 3</td>
<td>2800 8</td>
</tr>
<tr>
<td>2790 7</td>
<td>3210 15</td>
</tr>
<tr>
<td>3050 12</td>
<td>3080 13</td>
</tr>
<tr>
<td>2660 4</td>
<td>2950 11</td>
</tr>
<tr>
<td>1400 1</td>
<td></td>
</tr>
<tr>
<td>2775 6</td>
<td></td>
</tr>
</tbody>
</table>

\[ W_S = \sum (\text{Ranks in Group 2}) = 100 \]
\[ W'_S = 2\bar{W} - W_S = 152 - 100 = 52 \]
\[ W_S = \frac{n_1 (n_1 + n_2 + 1)}{2} \]
\[ z = \frac{W_S - \frac{n_1 (n_1 + n_2 + 1)}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} = \frac{8(10)(8 + 10 + 1)}{\sqrt{126.6667}} = 2.13 \]
data are hypothetical (but not particularly unreasonable) data on the birthweight (in grams) of children born to mothers who did not seek prenatal care until the third trimester and those born to mothers who received prenatal care starting in the first trimester.

For the data in Table 18.4 the sum of the ranks in the smaller group equals 100. From Appendix we find , and thus . Since 52 is smaller than 100, we enter Appendix with , , and . ( is defined as the smaller sample size.) Since we want a two-tailed test, we will double the tabled value of . The critical value of (or ) for a two-tailed test at is 53, meaning that only 5% of the time would we expect a value of or less than or equal to 53 if is true. Our obtained value of is 52, which thus falls in the rejection region, and we will reject . We will conclude that mothers who do not receive prenatal care until the third trimester tend to give birth to smaller babies. This probably does not mean that not having care until the third trimester causes smaller babies, but only that variables associated with delayed care (e.g., young mothers, poor nutrition, or poverty) are also associated with lower birthweight.

The use of the normal approximation for evaluating is illustrated in the bottom part of Table 18.3. Here we find that . From Appendix we find that the probability of as large as 100 or as small as 52 (a as extreme as ) is 2(.0166) = .033. Since this value is smaller than our traditional cutoff of , we will reject and again conclude that there is sufficient evidence to say that failing to seek early prenatal care is related to lower birthweight. Note that both the exact solution and the normal approximation lead to the same conclusion with respect to . However, a resampling test on the means using randomization would yield (two-tailed). (It would be instructive for you to calculate for the same set of data.)

The Treatment of Ties

When the data contain tied scores, any test that relies on ranks is likely to be somewhat distorted. Ties can be dealt with in several different ways. You can assign tied ranks to tied scores (as we have been doing), you can flip a coin and assign consecutive ranks to tied scores, or you can assign untied ranks in whatever way will make it hardest to reject . In actual practice, most people simply assign tied ranks. Although that may not be the best way to proceed statistically, it is clearly the most common and is the method that we will use here.

The Null Hypothesis

Wilcoxon’s rank-sum test evaluates the null hypothesis that the two sets of scores were sampled from identical populations. This is broader than the null hypothesis tested by the corresponding test, which dealt specifically with means (primarily as a result of the underlying assumptions that ruled out other sources of difference). If the two populations are assumed to have the same shape and dispersion, then the null hypothesis tested by the rank-sum test will actually deal with the central tendency (in this case the medians) of the two populations, and if the populations are also symmetric, the test will be a test of means. In any event, the rank-sum test is particularly sensitive to differences in central tendency.

Wilcoxon’s Test and Resampling Procedures

An interesting feature of Wilcoxon’s test is that it is actually not anything you haven’t seen before. Wilcoxon derived his test as a permutation test on ranked data, and such tests are often referred to as rank-randomization tests. In other words, if you took the data we had earlier, converted them to ranks, and ran a standard permutation tests (which is really a
randomization test where we draw every possible permutation once and only once), you would obtain the same result that Wilcoxon’s test produces. The reason that Wilcoxon was able to derive his test many years before computers could reasonably do the calculations, and why he could create tables for it, is that he uses ranks. We know a good many things about ranks, such as their sum and mean, without having to do the calculations. If we have five numbers, we know that their ranks will be the numbers 1 – 5, and the sum of the ranks will be 15, regardless of what their individual values are. This allowed Wilcoxon to derive the resulting sampling distributions once, and only once, and thus create his tables.

The Mann–Whitney U statistic

Mann–Whitney U test

A common competitor to the Wilcoxon rank-sum test is the Mann–Whitney U test. We do not need to discuss the Mann–Whitney test at any length, however, because the two are equivalent tests, and there is a perfect linear relationship between \( W_S \) and \( U \). The only reason for its inclusion here is that you may run across a reference to \( U \), and therefore you should know what it is. Very simply,

\[
U = \frac{n_1(n_1 + 2n_2 + 1)}{2} - W_S
\]

where \( n_1 \) is the smaller of the two sample sizes. From this formula we can see that for any given set of sample sizes, \( U \) and \( W_S \) differ by only a constant (as do their critical values). Since we have this relationship between the two statistics, we can always convert \( U \) to \( W_S \) and evaluate \( W_S \) using Appendix \( W_S \).

18.7 Wilcoxon’s Matched-Pairs Signed-Ranks Test

Wilcoxon is credited with developing not only the most popular nonparametric test for independent groups, but also the most popular test for matched groups (or paired scores). This test is the nonparametric analogue of the \( t \) test for related samples, and it tests the null hypothesis that two related (matched) samples were drawn either from identical populations or from symmetric populations with the same mean. More specifically, it tests the null hypothesis that the distribution of difference scores (in the population) is symmetric about zero. This is the same hypothesis tested by the corresponding \( t \) test when that test’s normality assumption is met.

The development of the logic behind the Wilcoxon matched-pairs signed-ranks test is as straightforward as it was for his rank-sum test and can be illustrated with a simple example. Assume that we want to test the often-stated hypothesis that a long-range program of running will reduce blood pressure. To test this hypothesis, we measure the blood pressure of a number of participants, ask them to engage in a systematic program of running for 6 months, and again test their blood pressure at the end of that period. Our dependent variable will be the change in blood pressure over the 6-month interval. If running does reduce blood pressure, we would expect most of the participants to show a lower reading the second time, and thus a positive pre–post difference. We also would expect that those whose blood pressure actually went up (and thus have a negative pre–post difference) would be only slightly higher. On the other hand, if running is worthless as a method of controlling blood pressure, then about one-half of the difference scores will be positive and one-half will be negative, and the positive differences will be about as large as the negative ones. In other words, if \( H_0 \) is really true, we would no longer expect most changes to be in the predicted direction with only small changes in the unpredicted direction. Notice that we have two expectations here: (1) Most of the changes will be in the same
direction; (2) Those that are in the opposite direction will be small ones. We will relax that second expectation when we shortly come to the Sign test, but with a concomitant loss in power.

As is illustrated in the following numerical example, in carrying out the Wilcoxon matched-pairs signed ranks test we first calculate the difference score for each pair of measurements. We then rank all difference scores without regard to the sign of the difference, then assign the algebraic sign of the differences to the ranks themselves, and finally sum the positive and negative ranks separately. The test statistic \( T \) is taken as the smaller of the absolute values (i.e., ignoring the sign) of the two sums, and is evaluated against the tabled entries in Appendix \( T \). (It is important to note that in calculating \( T \) we attach algebraic signs to the ranks only for convenience. We could just as easily, for example, circle those ranks that went with improvement and underline those that went with deterioration. We are merely trying to differentiate between the two cases.)

Assume that the study previously described produced the following data on systolic blood pressure before and after the six-month training session:

<table>
<thead>
<tr>
<th>Before:</th>
<th>130</th>
<th>170</th>
<th>125</th>
<th>170</th>
<th>130</th>
<th>130</th>
<th>145</th>
<th>160</th>
</tr>
</thead>
<tbody>
<tr>
<td>After:</td>
<td>120</td>
<td>163</td>
<td>120</td>
<td>135</td>
<td>143</td>
<td>136</td>
<td>144</td>
<td>120</td>
</tr>
<tr>
<td>Difference ( B - A ):</td>
<td>10</td>
<td>7</td>
<td>5</td>
<td>35</td>
<td>-13</td>
<td>-6</td>
<td>1</td>
<td>40</td>
</tr>
<tr>
<td>Rank of Difference:</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>7</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>Signed Rank:</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>7</td>
<td>-6</td>
<td>-3</td>
<td>1</td>
<td>8</td>
</tr>
</tbody>
</table>

\[
T_+ = \sum \text{(positive ranks)} = 27
\]
\[
T_- = \sum \text{(negative ranks)} = -9
\]

The first two rows contain the participants’ blood pressures as measured before and after a six-month program of running. The third row contains the difference scores, obtained by subtracting the “after” score from the “before.” Notice that only two participants showed a negative change—increased blood pressure. Since these difference scores do not appear to reflect a population distribution that is anywhere near normal, we have chosen to use a nonparametric test. In the fourth row, all the difference scores have been ranked without regard to the direction of the change; in the fifth row, the appropriate sign has been appended to the ranks to discriminate those participants whose blood pressure decreased from those whose blood pressure increased. At the bottom of the table we see the sum of the positive and negative ranks \( T_+ \) and \( T_- \). Since \( T \) is defined as the smaller absolute value of \( T_+ \) and \( T_- \), \( T = 9 \).

To evaluate \( T \) we refer to Appendix \( T \), a portion of which is shown in Table 18.5. This table has a format somewhat different from that of the other tables we have seen. The easiest way to understand what the entries in the table represent is by way of an analogy. Suppose that to test the fairness of a coin you were going to flip it eight times and reject the null hypothesis, at \( \alpha = .05 \) (one-tailed), if there were too few heads. Out of eight flips of a coin there is no set of outcomes that has a probability of exactly .05 under \( H_0 \). The probability of one or fewer heads is .0352, and the probability of two or fewer heads is .1445. Thus, if we want to work at \( \alpha = .05 \), we can either reject for one or fewer heads, in which case the probability of a Type I error is actually .0352 (less than .05), or we can reject for two or fewer heads, in which case the probability of a Type I error is actually .1445 (very much greater than .05). The same kind of problem arises with \( T \) because it, like the binomial distribution that gave us the probabilities of heads and tails, is a discrete distribution.\(^7\)

\(^7\) A similar situation arises for the Wilcoxon rank-sum test, but the standard tables for that test give only the conservative cutoff.
In Appendix T we find that for a one-tailed test at $\alpha = 0.025$ (or a two-tailed test at $\alpha = 0.05$) with $n = 8$, the entries are 3 (0.0195) and 4 (0.0273). This tells us that if we want to work at a (one-tailed) $\alpha = 0.025$, which is the equivalent of a two-tailed test at $\alpha = 0.05$, we can either reject for $T = 3$ (in which case $\alpha$ actually equals 0.0195) or we can reject for $T = 4$ (in which case the true value of $\alpha$ is 0.0273). Since we want a two-tailed test, the probabilities should be doubled to 3 (0.0390) and 4 (0.0546). Since we obtained a $T$ value of 9, we would not reject, whichever cutoff we chose. We will conclude therefore that we have no reason to doubt that blood pressure is unaffected by a short (6-month) period of daily running. It is going to take a lot more than six months to make up for a lifetime of dissipated habits.

### Ties

Ties can occur in the data in two different ways. One way would be for a participant to have the same before and after scores, leading to a difference score of 0, which has no sign. In this case, we normally eliminate that participant from consideration and reduce the sample size accordingly, although this leads to some bias in the data.

In addition, we could have tied difference scores that lead to tied rankings. If both the tied scores are of the same sign, we can break the ties in any way we wish (or assign tied...
ranks) without affecting the final outcome. If the scores are of opposite signs, we normally assign tied ranks and proceed as usual.

The Normal Approximation

When the sample size is larger than 50, which is the limit for Appendix $T$, a normal approximation is available to evaluate $T$. For larger sample sizes, we know that the sampling distribution of $T$ is approximately normally distributed with

$$\text{Mean} = \frac{n(n + 1)}{4} \quad \text{and} \quad \text{Standard error} = \sqrt{\frac{n(n + 1)(2n + 1)}{24}}$$

Thus, we can calculate

$$z = \frac{T - \frac{n(n + 1)}{4}}{\sqrt{\frac{n(n + 1)(2n + 1)}{24}}}$$

and evaluate $z$ using Appendix $z$. The procedure is directly analogous to that used with the rank-sum test and will not be repeated here.

Another interesting example of the use of Wilcoxon’s signed-ranks matched-pairs test is found in a study by Manning, Hall, and Gold (1990). These investigators were interested in studying the role of glucose in memory, in particular its effects on performance of memory tasks for elderly people. There has been considerable suggestion in the literature that participants with poor glucose regulation show poor memory and decreased performance on other kinds of neuropsychological tests.

Manning et al. asked 17 elderly volunteers to perform a battery of tests early in the morning after having drunk an 8-ounce lemon-flavored drink sweetened with either glucose or saccharin. Saccharin would taste as sweet but would not elevate blood glucose levels. Participants performed these tasks under both conditions, so we have matched sets of data. On one of these tasks, for which they had data on only 16 people, participants were read a narrative passage and were asked for recall of that passage 5 minutes later. The dependent variable was not explicitly defined, but we will assume that it was the number of specific propositions recalled from the passage.

The data given in Table 18.6 were generated to produce roughly the same means, standard deviations, and test results as the data found by Manning et al. From Appendix $T$ with $N = 16$ and a two-tailed test at $\alpha = .05$, we find that the critical value of $T$ is 35 or 36, depending on whether you prefer to err on the liberal or conservative side. Our value of $T_{\text{obs}} = 14.5$ is less than either and is therefore significant. This is the same conclusion that Manning et al. came to when they reported improved recall in the Glucose condition.

As an example of using the normal approximation, we can solve for the normal variate ($z$ score) associated with a $T$ of 14.5 for $N = 16$. In this case,

$$z = \frac{T - \frac{n(n + 1)}{4}}{\sqrt{\frac{n(n + 1)(2n + 1)}{24}}} = \frac{14.5 - \frac{(16)(17)}{4}}{\sqrt{\frac{(16)(17)(33)}{24}}} = -2.77$$

which has a two-tailed probability under $H_0$ of .0056. A resampling procedure on the means would produce $p = .002$ (two-tailed).
The Sign Test

The Wilcoxon matched-pairs signed-ranks test is an excellent distribution-free test for differences with matched samples. Unlike Student’s $t$ test, it makes less than maximum use of the data, in that it substitutes ranks for raw score differences, thus losing some of the subtle differences among the data points. When the assumptions of Student’s $t$ hold, it also has somewhat less power. When those assumptions do not hold, however, it may have greater power. A test that goes even further in the direction of gaining freedom from assumptions at the cost of power is the sign test. This test loses even more information by ignoring the values altogether and looking only at the sign of the differences. As a result, it loses even more power. We discussed the test briefly in Chapter 6 but will give a second example here for completeness.

We can use the example from Manning et al. (1990) in the preceding section. It might be argued that this is a good candidate for such a test because the Wilcoxon test was forced to rely on a large number of tied ranks. This argument is not all that persuasive because the results would have been the same no matter how you had broken the tied ranks, but it would be comforting to know that Manning et al.’s results are sufficiently solid that a sign test would also reveal their statistical significance.

The data from Manning et al. are repeated in Table 18.7. From these data you can see that 13 out of 16 participants showed higher recall under the Glucose condition, whereas only 3 of the 16 showed higher recall under the Saccharin condition. The sign test consists simply of asking the question of whether a 3-to-13 split would be likely to occur if recall under the two conditions were equally good.

This test could be set up in several ways. We could solve for the binomial probability of 13 or more successes out of 16 trials given $p = .50$. From standard tables, or the binomial formula, we would find

\[ p(13) = .0085 \]
\[ p(14) = .0018 \]
\[ p(15) = .0002 \]
\[ p(16) = .0000 \]
\[ \text{Sum} = .0105 \]
Since the binomial distribution is symmetric for \( p = 0.5 \), we would then double this probability to obtain the two-tailed probability, which in this case is 0.021. Since this probability is less than 0.05, we would reject the null hypothesis and conclude that recall is greater in the Glucose condition.

We could also solve for this probability by using the normal approximation given in Chapter 5. We would again come to essentially the same result, differing only by the accuracy of the approximation.

Yet a third possibility, which is logically equivalent to the others, is to use a goodness of fit \( \chi^2 \) test. In this case we would take 8 as our expected frequency for each cell, since if the two conditions lead to equal recall we would expect half of our 16 participants to do better by chance under each condition. We would then set up the table:

<table>
<thead>
<tr>
<th></th>
<th>Glucose</th>
<th>Saccharin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>13</td>
<td>3</td>
</tr>
<tr>
<td>Expected</td>
<td>8</td>
<td>8</td>
</tr>
</tbody>
</table>

\[
\chi^2 = \sum \frac{(O - E)^2}{E} = \frac{(13 - 8)^2}{8} + \frac{(3 - 8)^2}{8} = 6.25
\]

The critical value of \( \chi^2 \) on 1 df is 3.84, so we can reject \( H_0 \) and again conclude that the difference is significant. (The probability of \( \chi^2 \approx 6.25 \) is 0.0124, which agrees well enough, given the small sample size, with the exact binomial probability.) All three of these tests are more or less equivalent, and you can use whichever is most convenient.

### 18.9 Kruskal–Wallis One-Way Analysis of Variance

The **Kruskal–Wallis one-way analysis of variance** is a direct generalization of the Wilcoxon rank-sum test to the case in which we have three or more independent groups. As such, it is the nonparametric analogue of the one-way analysis of variance discussed in Chapter 11. It tests the hypothesis that all samples were drawn from identical populations and is particularly sensitive to differences in central tendency.

To perform the Kruskal–Wallis test, we simply rank all scores without regard to group membership and then compute the sum of the ranks for each group. The sums are denoted by \( R_i \). If the null hypothesis is true, we would expect the \( R_i \)’s to be more or less equal (aside from difference due to the size of the samples). A measure of the degree to which the \( R_i \) differ from one another is provided by:

\[
H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)
\]
where

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N + 1)$$

$$= \frac{12}{19(20)} \left( \frac{35^2}{7} + \frac{115^2}{8} + \frac{40^2}{4} \right) - 3(19 + 1)$$

$$= \frac{12}{380} (2228.125) - 60$$

$$= 70.36 - 60$$

$$= 10.36$$

$$\chi_0^{2}(2) = 5.99$$

18.10 Friedman’s Rank Test for k Correlated Samples

The last test to be discussed in this chapter is the nonparametric analogue of the one-way repeated-measures analysis of variance, Friedman’s rank test for k correlated samples. It was developed by the well-known economist Milton Friedman—in the days before he
This test is closely related to a standard repeated-measures analysis of variance applied to ranks instead of raw scores. It is a test on the null hypothesis that the scores for each treatment were drawn from identical populations, and it is especially sensitive to population differences in central tendency.

Assume that we want to test the hypothesis that the judged quality of a lecture is related to the number of visual aids used. The experimenter obtains 17 people who frequently give lectures to local business groups on a variety of topics. Each lecturer delivers the same lecture to three different, but equivalent, audiences—once with no visual aids, once with a few transparencies to illustrate major points, and once with transparencies and flip charts to illustrate every point made. At the end of each lecture, the audience is asked to rate the lecture on a 75-point scale, and the mean rating across all members of the audience is taken as the dependent variable. Since the same lecturers serve under all three conditions, we would expect the data to be correlated. Terrible lecturers are terrible no matter how many visual aids they use. Hypothetical data are presented in Table 18.9, in which a higher score represents a more favorable rating. The ranking of the raw scores within each participant are shown in parentheses.

Table 18.9  Hypothetical data on rated quality of lectures

<table>
<thead>
<tr>
<th>Lecturer</th>
<th>None</th>
<th>Few</th>
<th>Many</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50 (1)</td>
<td>58 (3)</td>
<td>54 (2)</td>
</tr>
<tr>
<td>2</td>
<td>32 (2)</td>
<td>37 (3)</td>
<td>25 (1)</td>
</tr>
<tr>
<td>3</td>
<td>60 (1)</td>
<td>70 (3)</td>
<td>63 (2)</td>
</tr>
<tr>
<td>4</td>
<td>58 (2)</td>
<td>60 (3)</td>
<td>55 (1)</td>
</tr>
<tr>
<td>5</td>
<td>41 (1)</td>
<td>66 (3)</td>
<td>59 (2)</td>
</tr>
<tr>
<td>6</td>
<td>36 (2)</td>
<td>40 (3)</td>
<td>28 (1)</td>
</tr>
<tr>
<td>7</td>
<td>26 (3)</td>
<td>25 (2)</td>
<td>20 (1)</td>
</tr>
<tr>
<td>8</td>
<td>49 (1)</td>
<td>60 (3)</td>
<td>50 (2)</td>
</tr>
<tr>
<td>9</td>
<td>72 (1)</td>
<td>73 (2)</td>
<td>75 (3)</td>
</tr>
<tr>
<td>10</td>
<td>49 (2)</td>
<td>54 (3)</td>
<td>42 (1)</td>
</tr>
<tr>
<td>11</td>
<td>52 (2)</td>
<td>57 (3)</td>
<td>47 (1)</td>
</tr>
<tr>
<td>12</td>
<td>36 (2)</td>
<td>42 (3)</td>
<td>29 (1)</td>
</tr>
<tr>
<td>13</td>
<td>37 (3)</td>
<td>34 (2)</td>
<td>31 (1)</td>
</tr>
<tr>
<td>14</td>
<td>58 (3)</td>
<td>50 (1)</td>
<td>56 (2)</td>
</tr>
<tr>
<td>15</td>
<td>39 (1)</td>
<td>48 (3)</td>
<td>44 (2)</td>
</tr>
<tr>
<td>16</td>
<td>25 (2)</td>
<td>29 (3)</td>
<td>18 (1)</td>
</tr>
<tr>
<td>17</td>
<td>51 (1)</td>
<td>63 (2)</td>
<td>68 (3)</td>
</tr>
</tbody>
</table>

\[
\chi^2_F = \frac{12}{Nk(k + 1)} \sum_{i=1}^{k} R_i^2 - 3N(K + 1)
\]

\[
= \frac{12}{17(3)(4)}(30^2 + 45^2 + 27^2) - 3(17)(4)
\]

\[
= \frac{12}{204}(3654) - 204
\]

\[
= 10.94
\]
If the null hypothesis is true, we would expect the rankings to be randomly distributed within each lecturer. Thus, one lecturer might do best with no visual aids, another might do best with many aids, and so on. If this were the case, the sum of the rankings in each condition (column) would be approximately equal. On the other hand, if a few visual aids were to lead to the most popular lecture, then most lecturers would have their highest rating under that condition, and the sum of the rankings for the three conditions would be decidedly unequal.

To apply Friedman’s test, we rank the raw scores for each lecturer separately and then sum the rankings for each condition. We then evaluate the variability of the sums by computing

\[ \chi^2_F = \frac{12}{Nk(k+1)} \sum_{i=1}^{k} R_i^2 - 3N(k+1) \]

where

- \( R_i \) = the sum of the ranks for the \( i \)th condition
- \( N \) = the number of subjects (lecturers)
- \( k \) = the number of conditions

This value of \( \chi^2_F \) can be evaluated with respect to the standard \( \chi^2 \) distribution on \( k-1 \) df.

For the data in Table 18.9, \( \chi^2_F = 10.94 \) on 2 df. Since \( \chi^2_{15}(2) = 5.99 \), we will reject \( H_0 \) and conclude that the judged quality of a lecture differs as a function of the degree to which visual aids are included. The data suggest that some visual aids are helpful, but that too many of them can detract from what the lecturer is saying. (Note: The null hypothesis we have just tested says nothing about differences among participants [lecturers], and in fact participant differences are completely eliminated by the ranking procedure.)

**Key Terms**

- Parametric tests (Introduction)
- Distribution-free tests (Introduction)
- Resampling procedures (Introduction)
- Sampling with replacement (Introduction)
- Permutation tests (Introduction)
- Randomization tests (Introduction)
- Sampling without replacement (Introduction)
- Wilcoxon rank-sum test (18.6)
- Rank-randomization tests (18.6)
- Mann–Whitney \( U \) test (18.6)
- Wilcoxon matched-pairs signed-ranks test (18.7)
- Sign test (18.8)
- Kruskal–Wallis one-way analysis of variance (18.9)
- Friedman’s rank test for \( k \) correlated samples (18.10)

**Exercises**

18.1 McConaughy (1980) has argued that younger children organize stories in terms of simple descriptive (“and then . . .”) models, whereas older children incorporate causal statements and social inferences. Suppose that we asked two groups of children differing in age to summarize a story they just read. We then counted the number of statements in the summary that can be classed as inferences. The data follow:

Younger Children: 0 1 0 3 2 5 2
Older Children: 4 7 6 4 8 7
a. Analyze these data using the two-tailed rank-sum test.

b. What can you conclude?

c. How would you go about analyzing these data if you had access to a program that would do resampling for you?

18.2 Kapp, Fryzsinger, Gallagher, and Hazelton (1979) have demonstrated that lesions in the amygdala can reduce certain responses commonly associated with fear (e.g., decreases in heart rate). If fear is really reduced, then it should be more difficult to train an avoidance response in lesioned animals because the aversiveness of the stimulus will be reduced. Assume two groups of rabbits: One group has lesions in the amygdala, and the other is an untreated control group. The following data represent the number of trials to learn an avoidance response for each animal:

<table>
<thead>
<tr>
<th>Group with Lesions:</th>
<th>15</th>
<th>14</th>
<th>15</th>
<th>8</th>
<th>7</th>
<th>22</th>
<th>36</th>
<th>19</th>
<th>14</th>
<th>18</th>
<th>17</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control Group:</td>
<td>9</td>
<td>4</td>
<td>9</td>
<td>10</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>5</td>
<td>9</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Analyze the data using the Wilcoxon rank-sum test (two-tailed).

b. What can you conclude?

18.3 Repeat the analysis in Exercise 18.2 using the normal approximation.

18.4 Repeat the analysis in Exercise 18.2 using the appropriate one-tailed test.

18.5 Nurcombe and Fitzhenry-Coor (1979) have argued that training in diagnostic techniques should lead a clinician to generate (and test) more hypotheses in coming to a decision about a case. Suppose we take 10 psychiatric residents who are just beginning their residency and ask them to watch a videotape of an interview and to record their thoughts on the case every few minutes. We then count the number of hypotheses each resident includes in his or her written remarks. The experiment is repeated at the end of the residency with a comparable videotape. The data follow:

| Subject: | 12345 678910 | Before: | 84224 83139 | After: | 79363106787 |

a. Analyze the data using Wilcoxon’s matched-pairs signed-ranks test.

b. What can you conclude?

18.6 Refer to Exercise 18.5.

a. Repeat the analysis using the normal approximation.

b. How well do the two answers (18.5a and 18.6a) agree? Why do they not agree exactly?

18.7 How would you go about applying a resampling procedure to test the difference between Before and After scores in Exercise 18.6?

18.8 It has been argued that first-born children tend to be more independent than later-born children. Suppose we develop a 25-point scale of independence and rate each of 20 first-born children and their second-born siblings using our scale. We do this when both siblings are adults, thus eliminating obvious age effects. The data on independence are as follows (a higher score means that the person is more independent):

<table>
<thead>
<tr>
<th>Sibling Pair:</th>
<th>1 2 3 4 5 6 7 8 9 10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>11 12 13 14 15 16 17 18 19 20</td>
</tr>
<tr>
<td>First Born:</td>
<td>12 18 13 17 8 15 16 5 8 12</td>
</tr>
<tr>
<td></td>
<td>13 5 14 20 19 17 2 5 15 18</td>
</tr>
<tr>
<td>Second Born:</td>
<td>10 12 15 13 9 12 13 8 10 8</td>
</tr>
<tr>
<td></td>
<td>8 9 8 10 14 11 7 7 13 12</td>
</tr>
</tbody>
</table>

Exercises 687
a. Analyze the data using Wilcoxon’s matched-pairs signed-ranks test.
b. What can you conclude?

18.9 Rerun the analysis in Exercise 18.8 using the normal approximation.

18.10 How would we run a standard resampling test for the data in Exercise 18.8?

18.11 The results in Exercise 18.8 are not quite as clear-cut as we might like. Plot the differences as a function of the first-born’s score. What does this figure suggest?

18.12 What is the difference between the null hypothesis tested by Wilcoxon’s rank-sum test and the corresponding \( t \) test?

18.13 What is the difference between the null hypothesis tested by Wilcoxon’s matched-pairs signed-ranks test and the corresponding \( t \) test?

18.14 One of the arguments put forth in favor of nonparametric tests is that they are more appropriate for ordinal-scale data. This issue was addressed earlier in the book in a different context. Give a reason why this argument is not a good one.

18.15 Why is rejection of the null hypothesis using a \( t \) test a more specific statement than rejection of the null hypothesis using the appropriate nonparametric test?

18.16 Three rival professors teaching English I all claim the honor of having the best students. To settle the issue, eight students are randomly drawn from each class and are given the same exam, which is graded by a neutral professor who does not know from which class the students came.

The data follow:

\[
\begin{align*}
\text{Professor A:} & \quad 82 & 71 & 56 & 58 & 63 & 64 & 62 & 53 \\
\text{Professor B:} & \quad 55 & 88 & 85 & 83 & 71 & 70 & 68 & 72 \\
\text{Professor C:} & \quad 65 & 54 & 66 & 68 & 72 & 78 & 65 & 73
\end{align*}
\]

Run the appropriate test and draw the appropriate conclusions.

18.17 A psychologist operating a group home for delinquent adolescents needs to show that it is successful at reducing delinquency. He samples nine adolescents living in their parents’ home whom the police have identified as having problems, nine similar adolescents living in foster homes, and nine adolescents living in the group home. As an indicator variable, he uses truancy (number of days truant in the past semester), which is readily obtained from school records. On the basis of the following data, draw the appropriate conclusions.

\[
\begin{align*}
\text{Natural Home:} & \quad 15 & 18 & 19 & 14 & 5 & 8 & 12 & 13 & 7 \\
\text{Foster Home:} & \quad 16 & 14 & 20 & 22 & 19 & 5 & 17 & 18 & 12 \\
\text{Group Home:} & \quad 10 & 13 & 14 & 11 & 7 & 3 & 4 & 18 & 2
\end{align*}
\]

18.18 As an alternative method of evaluating a group home, suppose that we take 12 adolescents who have been declared delinquent. We take the number of days truant (1) during the month before they are placed in the home, (2) during the month they live in the home, and (3) during the month after they leave the home.

The data follow:

\[
\begin{align*}
\text{Adolescent:} & \quad 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\
\text{Before:} & \quad 10 & 12 & 12 & 19 & 5 & 13 & 20 & 8 & 12 & 10 & 8 & 18 \\
\text{During:} & \quad 5 & 8 & 13 & 10 & 10 & 8 & 16 & 4 & 14 & 3 & 3 & 16 \\
\text{After:} & \quad 8 & 7 & 10 & 12 & 8 & 7 & 12 & 5 & 9 & 5 & 3 & 2
\end{align*}
\]

Apply Friedman’s test. What do you conclude?

18.19 I did not discuss randomization tests on the evaluation of data that are laid out like a one-way analysis of variance (as in Exercise 18.17), but you should be able to suggest an analysis that would be appropriate if we had the software to carry out the calculations. How would you outline that test?
18.20 The test referred to in Exercise 18.19 is available on my Web site. Run that program on the data for Exercise 18.18 and report the results. (There is a “read-me” file on the disk that will tell you how to run the resampling program.

18.21 What advantage does the study described in Exercise 18.18 have over the study described in Exercise 18.17?

18.22 It would be possible to apply Friedman’s test to the data in Exercise 18.5. What would we lose if we did?

18.23 For the data in Exercise 18.5, we could say that 3 out of 10 residents used fewer hypotheses the second time and 7 used more. We could test this with $\chi^2$. How would this differ from Friedman’s test applied to those data?

18.24 The history of statistical hypothesis testing really began with a tea-tasting experiment (Fisher, 1935), so it seems fitting for this book to end with one. The owner of a small tea-room does not think that people really can tell the difference between the first cup made with a given tea bag and the second and third cups made with the same bag (perhaps that is why it is still a small tearoom). He chooses eight different brands of tea bags, makes three cups of tea with each, reusing the same tea bag, and then has a group of customers rate each cup on a 20-point scale (without knowing which cup is which). The data are shown here, with higher ratings indicating better tea.

<table>
<thead>
<tr>
<th>Tea Brands</th>
<th>First Cup</th>
<th>Second Cup</th>
<th>Third Cup</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>14</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>16</td>
<td>17</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>10</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>12</td>
<td>10</td>
<td>2</td>
</tr>
</tbody>
</table>

Using Friedman’s test, draw the appropriate conclusions.
This page intentionally left blank