Nonparametric Tests

The most commonly used methods for inference about the means of quantitative response variables assume that the variables in question have Normal distributions in the population or populations from which we draw our data. In practice, of course, no distribution is exactly Normal. Fortunately, our usual methods for inference about population means (the one-sample and two-sample $t$ procedures and analysis of variance) are quite robust. That is, the results of inference are not very sensitive to moderate lack of Normality, especially when the samples are reasonably large. Practical guidelines for taking advantage of the robustness of these methods appear in Chapters 16, 17, and 22.

What can we do if plots suggest that the data are clearly not Normal, especially when we have only a few observations? This is not a simple question. Here are the basic options:

1. If there are extreme outliers in a small data set, any inference method may be suspect. An outlier is an observation that may not come from the same population as the others. To decide what to do, you must find the cause of the outlier. Equipment failure that produced a bad measurement, for example, entitles you to remove the outlier and analyze the remaining data. If the outlier appears to be "real data," it is risky to draw any conclusion from just a few observations. This is the advice we gave to the child development researcher in Example 5.6 (text page 118).

2. Sometimes we can transform our data so that their distribution is more nearly Normal. Transformations such as the logarithm that pull in the...
long tail of right-skewed distributions are particularly helpful. We used
the logarithm transformation in Example 4.5 (text page 84) and in
Exercises I.45 and I.46 (text page 170).

3. In some settings, other standard distributions replace the Normal
distributions as models for the overall pattern in the population. The
lifetimes in service of equipment or the survival times of cancer patients
after treatment usually have right-skewed distributions. Statistical
studies in these areas use families of right-skewed distributions rather
than Normal distributions. There are inference procedures for the parameters
of these distributions that replace the \( t \) procedures.

4. Finally, there are inference procedures that do not assume any specific
form for the distribution of the population. These are called
nonparametric methods. They are the subject of this chapter.

This chapter concerns one type of nonparametric procedure: tests that can
replace the \( t \) tests and one-way analysis of variance when the Normality condi-
tions for those tests are not met. The most useful nonparametric tests are rank
tests based on the rank (place in order) of each observation in the set of all the
data.

Figure 23.1 presents an outline of the standard tests (based on Normal distri-
butions) and the rank tests that compete with them. All of these tests require
that the population or populations have continuous distributions. That is, each
distribution must be described by a density curve that allows observations to
take any value in some interval of outcomes. The Normal curves are one shape
density curve. Rank tests allow curves of any shape.

The rank tests we will study concern the center of a population or popu-
lations. When a population has at least roughly a Normal distribution, we
describe its center by the mean. The “Normal tests” in Figure 23.1 all test
hypotheses about population means. When distributions are strongly skewed,
we often prefer the median to the mean as a measure of center. In simplest form,
the hypotheses for rank tests just replace mean by median.

We begin by describing the most common rank test, for comparing two sam-
ples. In this setting we also explain ideas common to all rank tests: the big
idea of using ranks, the conditions required by rank tests, the nature of the
hypotheses tested, and the contrast between exact distributions for use with
small samples and Normal approximations for use with larger samples.

<table>
<thead>
<tr>
<th>Setting</th>
<th>Normal Test</th>
<th>Rank Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>One sample</td>
<td>One-sample ( t ) test</td>
<td>Wilcoxon signed-rank test</td>
</tr>
<tr>
<td>Matched pairs</td>
<td>Apply one-sample test to differences within pairs</td>
<td>Wilcoxon rank sum test</td>
</tr>
<tr>
<td>Two independent samples</td>
<td>Two-sample ( t ) test</td>
<td>Chapter 17</td>
</tr>
<tr>
<td>Several independent samples</td>
<td>One-way ANOVA ( F ) test</td>
<td>Chapter 22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Kruskal-Wallis test</td>
</tr>
</tbody>
</table>
Comparing two samples: the Wilcoxon rank sum test

Two-sample problems (see Chapter 17) are among the most common in statistics. The most useful nonparametric significance test compares two distributions. Here is an example of this setting.

**EXAMPLE 23.1  Weeds among the corn**

Does the presence of small numbers of weeds reduce the yield of corn? Lamb’s-quarter is a common weed in corn fields. A researcher planted corn at the same rate in 8 small plots of ground, then weeded the corn rows by hand to allow no weeds in 4 randomly selected plots and exactly 3 lamb’s-quarter plants per meter of row in the other 4 plots. Here are the yields of corn (bushels per acre) in each of the plots.1

<table>
<thead>
<tr>
<th>Weeds per meter</th>
<th>Yield (bu/acre)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>166.7 172.2 165.0 176.9</td>
</tr>
<tr>
<td>3</td>
<td>158.6 176.4 153.1 156.0</td>
</tr>
</tbody>
</table>

A back-to-back stemplot (Figure 23.2) suggests that yields may be lower when weeds are present. There is one outlier; because it is correct data, it cannot be removed. The samples are too small to rely on the robustness of the two-sample t test. We may prefer to use a test that does not require Normality.

We first rank all 8 observations together. To do this, arrange them in order from smallest to largest:

153.1 156.0 158.6 165.0 166.7 172.2 176.4 176.9

The boldface entries in the list are the yields with no weeds present. We see that four of the five highest yields come from that group, suggesting that yields are higher with no weeds. The idea of rank tests is to look just at position in this ordered list. To do this, replace each observation by its order, from 1 (smallest) to 8 (largest). These numbers are the ranks:

<table>
<thead>
<tr>
<th>Yield</th>
<th>153.1</th>
<th>156.0</th>
<th>158.6</th>
<th>165.0</th>
<th>166.7</th>
<th>172.2</th>
<th>176.4</th>
<th>176.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>

**Figure 23.2  Back-to-back stemplot of corn yields from plots with no weeds and with 3 weeds per meter of row. We split the stems, with leaves 0 to 4 on the first stem and leaves 5 to 9 on the second stem.**

<table>
<thead>
<tr>
<th>0 weeds/meter</th>
<th>3 weeds/meter</th>
</tr>
</thead>
<tbody>
<tr>
<td>151</td>
<td>3</td>
</tr>
<tr>
<td>155</td>
<td>69</td>
</tr>
<tr>
<td>161</td>
<td></td>
</tr>
<tr>
<td>163</td>
<td></td>
</tr>
<tr>
<td>164</td>
<td></td>
</tr>
<tr>
<td>198</td>
<td></td>
</tr>
<tr>
<td>217</td>
<td></td>
</tr>
<tr>
<td>227</td>
<td></td>
</tr>
<tr>
<td>237</td>
<td></td>
</tr>
<tr>
<td>247</td>
<td></td>
</tr>
<tr>
<td>7136</td>
<td>6</td>
</tr>
</tbody>
</table>
CHAPTER 23 • Nonparametric Tests

RANKS

To rank observations, first arrange them in order from smallest to largest. The rank of each observation is its position in this ordered list, starting with rank 1 for the smallest observation.

Moving from the original observations to their ranks retains only the ordering of the observations and makes no other use of their numerical values. Working with ranks allows us to dispense with specific assumptions about the shape of the distribution, such as Normality.

If the presence of weeds reduces corn yields, we expect the ranks of the yields from plots with weeds to be smaller as a group than the ranks from plots without weeds. We might compare the sums of the ranks from the two treatments:

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Sum of ranks</th>
</tr>
</thead>
<tbody>
<tr>
<td>No weeds</td>
<td>23</td>
</tr>
<tr>
<td>Weeds</td>
<td>13</td>
</tr>
</tbody>
</table>

These sums measure how much the ranks of the weed-free plots as a group exceed those of the weedy plots. In fact, the sum of the ranks from 1 to 8 is always equal to 36, so it is enough to report the sum for one of the two groups. If the sum of the ranks for the weed-free group is 23, the ranks for the other group must add to 13 because $23 + 13 = 36$. If the weeds have no effect, we would expect the sum of the ranks in either group to be 18 (half of 36). Here are the facts we need in a more general form that takes account of the fact that our two samples need not be the same size.

THE WILCOXON RANK SUM TEST

Draw an SRS of size $n_1$ from one population and draw an independent SRS of size $n_2$ from a second population. There are $N$ observations in all, where $N = n_1 + n_2$. Rank all $N$ observations. The sum $W$ of the ranks for the first sample is the Wilcoxon rank sum statistic. If the two populations have the same continuous distribution, then $W$ has mean

$$\mu_W = \frac{n_1(N + 1)}{2}$$

and standard deviation

$$\sigma_W = \sqrt{\frac{n_1n_2(N + 1)}{12}}$$

The Wilcoxon rank sum test rejects the hypothesis that the two populations have identical distributions when the rank sum $W$ is far from its mean.
Comparing two samples: the Wilcoxon rank sum test

In the corn yield study of Example 23.1, we want to test

\[ H_0: \text{no difference in distribution of yields} \]

against the one-sided alternative

\[ H_1: \text{yields are systematically higher in weed-free plots} \]

Our test statistic is the rank sum \( W = 23 \) for the weed-free plots.

**Example 23.2 Weeds among the corn**

In Example 23.1, \( n_1 = 4 \), \( n_2 = 4 \), and there are \( N = 8 \) observations in all. The sum of ranks for the weed-free plots has mean

\[
\mu_W = \frac{n_1(N+1)}{2} = \frac{(4)(9)}{2} = 18
\]

and standard deviation

\[
\sigma_W = \sqrt{\frac{n_1n_2(N+1)}{12}} = \sqrt{\frac{(4)(4)(9)}{12}} = \sqrt{12} = 3.464
\]

Although the observed rank sum \( W = 23 \) is higher than the mean, it is only about 1.4 standard deviations high. We now suspect that the data do not give strong evidence that yields are higher in the population of weed-free corn.

The \( P \)-value for our one-sided alternative is \( P(W \geq 23) \), the probability that \( W \) is at least as large as the value for our data when \( H_0 \) is true.

**Apply Your Knowledge**

23.1 Attracting beetles. To detect the presence of harmful insects in farm fields, we can put up boards covered with a sticky material and examine the insects trapped on the boards. Which colors attract insects best? Experimenters placed boards of several colors at random locations in a field of oats. Here are the counts of cereal leaf beetles trapped by boards colored blue and green:

<table>
<thead>
<tr>
<th></th>
<th>Blue</th>
<th>Green</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>16</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>32</td>
</tr>
</tbody>
</table>

Because the samples are small, it is difficult to verify approximate Normality. We might use the Wilcoxon rank sum test.

(a) Arrange the 12 observations in order and find the ranks.
(b) Take \( W \) to be the sum of the ranks for green boards. What is the value of \( W \)?
23.2 Is red wine better than white wine? Exercise 17.7 (text page 448) describes an experiment that compared the change in the level of polyphenols in the blood after two weeks of drinking either red wine or white wine. (Polyphenols may reduce the risk of a heart attack.) Here are the data:

<table>
<thead>
<tr>
<th>Red wine</th>
<th>3.5</th>
<th>8.1</th>
<th>7.4</th>
<th>4.0</th>
<th>0.7</th>
<th>4.9</th>
<th>8.4</th>
<th>7.0</th>
<th>5.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>White wine</td>
<td>3.1</td>
<td>0.5</td>
<td>-3.8</td>
<td>4.1</td>
<td>-0.6</td>
<td>2.7</td>
<td>1.9</td>
<td>-5.9</td>
<td>0.1</td>
</tr>
</tbody>
</table>

(a) Arrange the observations in order and find their ranks.
(b) Take $W$ to be the rank sum for red wine. What is the value of $W$?
(c) The study hoped to show that red wine raises polyphenols more than white wine. The null hypothesis is “no difference.” What are the mean $\mu_W$ and standard deviation $\sigma_W$ if the null hypothesis is true? Does comparing $W$ with $\mu_W$ and $\sigma_W$ suggest that red wine does raise polyphenols more than white wine?

The Normal approximation for $W$

To calculate the $P$-value $P(W \geq 23)$ for Example 23.2, we need to know the sampling distribution of the rank sum $W$ when the null hypothesis is true. This distribution depends on the two sample sizes $n_1$ and $n_2$. Tables are therefore unwieldy, though you can find them in handbooks of statistical tables. Most statistical software will give you $P$-values, as well as carry out the ranking and calculate $W$. However, many software packages give only approximate $P$-values. You must learn what your software offers.

With or without software, $P$-values for the Wilcoxon test are often based on the fact that the rank sum statistic $W$ becomes approximately Normal as the two sample sizes increase. We can then form yet another $z$ statistic by standardizing $W$:

$$z = \frac{W - \mu_W}{\sigma_W} = \frac{W - n_1(N + 1)/2}{\sqrt{n_1n_2(N + 1)/12}}$$

**continuity correction**

Use standard Normal probability calculations to find $P$-values for this statistic. Because $W$ takes only whole-number values, we use a trick called the continuity correction to improve the accuracy of the approximation. To apply the continuity correction when a variable takes only whole-number values, act as
The Normal approximation for \( W \)

if each whole number occupies the entire interval from 0.5 below the number to 0.5 above it.

**EXAMPLE 23.3** Using the Normal approximation

The standardized rank sum statistic \( W \) in our corn yield example is

\[
z = \frac{W - \mu_W}{\sigma_W} = \frac{23 - 18.7}{3.464} = 1.44
\]

We expect \( W \) to be larger when the alternative hypothesis is true, so the approximate \( P \)-value is

\[
P(Z \geq 1.44) = 0.0749
\]

We can improve this approximation by using the continuity correction. To do this, act as if the whole number 23 occupies the entire interval from 22.5 to 23.5. Calculate the \( P \)-value \( P(W \geq 23) \) as \( P(W \geq 22.5) \) because the value 23 is included in the range whose probability we want. Here is the calculation:

\[
P(W \geq 22.5) = P\left(\frac{W - \mu_W}{\sigma_W} \geq \frac{22.5 - 18.7}{3.464}\right) = P(Z \geq 1.30) = 0.0968
\]

We recommend always using either the exact distribution (from software or tables) or the continuity correction for the rank sum statistic \( W \).

**APPLY YOUR KNOWLEDGE**

Use the Normal approximation with continuity correction in these exercises.

23.3 Attracting beetles, continued. In Exercise 23.1, you found the Wilcoxon rank sum \( W \) and its mean and standard deviation. We want to test the null hypothesis that the two colors don’t differ against the alternative hypothesis that green boards will attract more beetles.

(a) What is the probability expression for the \( P \)-value of \( W \) if we use the continuity correction?

(b) Find the \( P \)-value. What do you conclude?

23.4 Red wine versus white wine, continued. Use your values of \( W, \mu_W \), and \( \sigma_W \) from Exercise 23.2 to see whether red wine raises blood polyphenols significantly more than white wine.

(a) The \( P \)-value is \( P(W \geq ?) \). Using the continuity correction, what number replaces the \( ? \) in this probability?

(b) Find the \( P \)-value. What do you conclude about red wine versus white wine?

23.5 Tell me a story. A study of early childhood education asked kindergarten students to tell fairy tales that had been read to them earlier in the week. The 10 children in the study included
5 high-progress readers and 5 low-progress readers. Each child told two stories. Story 1 had been read to them; Story 2 had been read and also illustrated with pictures. An expert listened to a recording of the children and assigned a score for certain uses of language. Here are the data:

<table>
<thead>
<tr>
<th>Child</th>
<th>Progress</th>
<th>Story 1 score</th>
<th>Story 2 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>high</td>
<td>0.55</td>
<td>0.80</td>
</tr>
<tr>
<td>2</td>
<td>high</td>
<td>0.57</td>
<td>0.82</td>
</tr>
<tr>
<td>3</td>
<td>high</td>
<td>0.72</td>
<td>0.54</td>
</tr>
<tr>
<td>4</td>
<td>high</td>
<td>0.70</td>
<td>0.79</td>
</tr>
<tr>
<td>5</td>
<td>high</td>
<td>0.84</td>
<td>0.89</td>
</tr>
<tr>
<td>6</td>
<td>low</td>
<td>0.40</td>
<td>0.77</td>
</tr>
<tr>
<td>7</td>
<td>low</td>
<td>0.72</td>
<td>0.49</td>
</tr>
<tr>
<td>8</td>
<td>low</td>
<td>0.00</td>
<td>0.66</td>
</tr>
<tr>
<td>9</td>
<td>low</td>
<td>0.36</td>
<td>0.28</td>
</tr>
<tr>
<td>10</td>
<td>low</td>
<td>0.55</td>
<td>0.38</td>
</tr>
</tbody>
</table>

Use the data for children telling Story 2 to carry out by hand the steps in the Wilcoxon rank sum test.

(a) Arrange the 10 observations in order and assign ranks.

(b) Find the rank sum $W$ for the 5 high-progress readers. What are the mean and standard deviation of $W$ under the null hypothesis that low-progress and high-progress readers do not differ?

(c) Standarize $W$ to obtain a $z$ statistic. Do a Normal probability calculation with the continuity correction to obtain a one-sided $P$-value.

Using technology

Neither the Excel spreadsheet nor the TI-83 calculator has menu entries for rank tests, though either can be programmed to find ranks, calculate statistics, and use Normal approximations for $P$-values. Figure 23.3 displays Minitab output for the corn yield data. Minitab offers only the Normal approximation, and it gives the results of the Mann-Whitney test. This is an alternative form of the Wilcoxon rank sum test. The one-sided $P$-value is $P = 0.0970$. This agrees (up to roundoff error) with our result in Example 23.3. You see that Minitab does use the continuity correction.

Figure 23.3 also displays output from a professional statistical software program, S-PLUS. This output compares four different ways of testing the null hypothesis that weed counts do not change yield. Here they are, in the order in which the results appear in the figure:

- **Exact Wilcoxon.** The rank sum is $W = 23$ and the $P$-value using the exact distribution of $W$ is $P = 0.100$. 

Mann-Whitney test
Using technology

Mann-Whitney Test and CI: weeds0, weeds3

weeds0  N =  4     Median =     169.45
weeds3  N =  4     Median =     157.30
Point estimate for ETA1-ETA2 is     11.30
W = 23.0
Test of ETA1 = ETA2 vs ETA1 > ETA2 is significant at 0.0970
Cannot reject at alpha = 0.05

Wilcoxon rank-sum test
data:  x: weeds0 in eg23.01 , and y: weeds3 in eg23.01
rank-sum statistic W = 23, n = 4, m = 4, p-value = 0.1
alternative hypothesis: true mu is greater than 0

Welch Modified Two-Sample t-Test
data:  x: weeds0 in eg23.01 , and y: weeds3 in eg23.01
t = 1.5536, df = 4.495, p-value = 0.0937
alternative hypothesis: true difference in means is greater than 0

Standard Two-Sample t-Test
data:  x: weeds0 in eg23.01 , and y: weeds3 in eg23.01
t = 1.5536, df = 6, p-value = 0.0856
alternative hypothesis: true difference in means is greater than 0

Figure 23.3  Output from Minitab and S-PLUS for the data in Example 23.1. The S-PLUS output compares the results of four tests that could be used to compare yields for the two groups of plots.

- Wilcoxon with Normal approximation. The approximate P-value is \( P = 0.097 \). This agrees with Minitab. The Normal approximation with continuity correction is close to the exact value \( P = 0.100 \).
- The two-sample t test. This is the test from Chapter 17, which does not assume that the two populations have the same standard deviation. S-PLUS calls this the “Welch modified two-sample t test.” It gives
CHAPTER 23  •  Nonparametric Tests

\[ P = 0.0937, \text{ close to the Wilcoxon value. Because the } t \text{ test is quite robust, it is somewhat unusual for } P \text{-values from } t \text{ and } W \text{ to differ greatly.} \]

* The standard \textit{t} test. This is the version of \textit{t}, now outdated, that assumes equal population standard deviations. You see that its \textit{P}-value is a bit different from all the others, another reminder that you should never use this test.

**APPLY YOUR KNOWLEDGE**

23.6 Attracting beetles: software. Use your software to carry out the one-sided Wilcoxon rank sum test that you did by hand in Exercise 23.3. Use the exact distribution if your software will do it. Compare the software result with your result in Exercise 23.3.

23.7 Red wine versus white wine: software. Use your software to repeat the Wilcoxon test you did in Exercise 23.4. By comparing the results, state how your software finds \textit{P}-values for \textit{W}: exact distribution, Normal approximation with continuity correction, or Normal approximation without continuity correction.

23.8 Student study time: effects of an outlier. Exercise I.3 (text page 156) contains data from a sample of first-year college students who were asked how long (in minutes) they study on a typical school night. We ask if there is a significant difference between men and women.

(a) Carry out the two-sample \textit{t} test. What are \textit{t} and its two-sided \textit{P}-value?

(b) Carry out the Wilcoxon rank sum test. What are \textit{W} and its two-sided \textit{P}-value? Do \textit{t} and \textit{W} lead to the same practical conclusion?

(c) One male student in the sample claimed to study 30,000 minutes per night. This is clearly a joke, so this student was replaced by another who claimed 180 minutes. Replace the final 180 in the male data by 30,000. Recalculate \textit{t} and \textit{W} and their \textit{P}-values. Compare the effect of the outlier on the two tests.

23.9 Weeds among the corn. The corn yield study of Example 23.1 also examined yields in four plots having 9 lamb’s-quarter plants per meter of row. The yields (bushels per acre) in these plots were

\[ 162.8 \quad 142.4 \quad 162.7 \quad 162.4 \]

There is a clear outlier, but rechecking the results found that this is the correct yield for this plot. The outlier makes us hesitant to use \textit{t} procedures because \textit{t} and \textit{s} are not resistant.

(a) Is there evidence that 9 weeds per meter reduces corn yields when compared with weed-free corn? Use the Wilcoxon rank sum test with the data above and part of the data from Example 23.1 to answer this question.
What hypotheses does Wilcoxon test?

Our null hypothesis is that weeds do not affect yield. The alternative hypothesis is that yields are lower when weeds are present. If we are willing to assume that yields are Normally distributed, or if we have reasonably large samples, we use the two-sample $t$ test for means. Our hypotheses then become

$$H_0: \mu_1 = \mu_2$$
$$H_a: \mu_1 > \mu_2$$

When the distributions may not be Normal, we might restate the hypotheses in terms of population medians rather than means:

$$H_0: \text{median}_1 = \text{median}_2$$
$$H_a: \text{median}_1 > \text{median}_2$$

The Wilcoxon rank sum test provides a significance test for these hypotheses, but only if an additional condition is met: both populations must have distributions of the same shape. That is, the density curve for corn yields with 3 weeds per meter looks exactly like that for no weeds except that it may slide to a different location on the scale of yields. The Minitab output in Figure 23.3 states the hypotheses in terms of population medians, which it calls “eta.” Minitab will also give a confidence interval for the difference between the two population medians.

The same-shape condition is too strict to be reasonable in practice. Fortunately, the Wilcoxon test also applies in a much more general and more useful setting. It tests hypotheses that we can state in words as

$$H_0: \text{two distributions are the same}$$
$$H_a: \text{one has values that are systematically larger}$$

A more exact statement of the “systematically larger” alternative hypothesis is a bit tricky, so we won’t try to give it here. These hypotheses really are “non-parametric” because they do not involve any specific parameter such as the mean or median. If the two distributions do have the same shape, the general hypotheses reduce to comparing medians. Many texts and computer outputs state the hypotheses in terms of medians, sometimes ignoring the same-shape requirement. We recommend that you express the hypotheses in words rather than symbols. “Yields are systematically higher in weed-free plots” is easy to understand and is a good statement of the effect that the Wilcoxon test looks for.
Chapter 23 - Nonparametric Tests

23.10 Attracting beetles: hypotheses. We could use either two-sample t or the Wilcoxon rank sum to test the null hypothesis that blue and green boards don't differ in their ability to attract beetles against the alternative that green attracts more beetles. Explain carefully what \( H_0 \) and \( H_a \) are for \( t \) and for \( W \).

23.11 Red wine versus white wine: hypotheses. We are interested in whether red wine raises blood polyphenols more than white wine “on the average.”

(a) State null and alternative hypotheses in terms of population means. What test would we typically use for these hypotheses? What conditions does this test require?

(b) State null and alternative hypotheses in terms of population medians. What test would we typically use for these hypotheses? What conditions does this test require?

Dealing with ties in rank tests

The exact distribution for the Wilcoxon rank sum is obtained assuming that all observations in both samples take different values. This allows us to rank them all. In practice, however, we often find observations tied at the same value. What shall we do? The usual practice is to assign all tied values the average of the ranks they occupy. Here is an example with 6 observations:

<table>
<thead>
<tr>
<th>Observation</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>153</td>
<td>1</td>
</tr>
<tr>
<td>155</td>
<td>2</td>
</tr>
<tr>
<td>158</td>
<td>3.5</td>
</tr>
<tr>
<td>158</td>
<td>3.5</td>
</tr>
<tr>
<td>161</td>
<td>5</td>
</tr>
<tr>
<td>164</td>
<td>6</td>
</tr>
</tbody>
</table>

The tied observations occupy the third and fourth places in the ordered list, so they share rank 3.5.

The exact distribution for the Wilcoxon rank sum \( W \) applies only to data without ties. Moreover, the standard deviation \( \sigma_W \) must be adjusted if ties are present. The Normal approximation can be used after the standard deviation is adjusted. Statistical software will detect ties, make the necessary adjustment, and switch to the Normal approximation. In practice, software is required if you want to use rank tests when the data contain tied values.

It is sometimes useful to use rank tests on data that have very many ties because the scale of measurement has only a few values. Here is an example.

Example 23.4 Food safety at fairs

Food sold at outdoor fairs and festivals may be less safe than food sold in restaurants because it is prepared in temporary locations and often by volunteer help. What do people who attend fairs think about the safety of the food served? One study asked this question of people at a number of fairs in the Midwest:
Dealing with ties in rank tests

How often do you think people become sick because of food they consume prepared at outdoor fairs and festivals?

The possible responses were
1 = very rarely
2 = once in a while
3 = often
4 = more often than not
5 = always

In all, 303 people answered the question. Of these, 196 were women and 107 were men. We suspect that women are more concerned than men about food safety. Is there good evidence for this conclusion?

We should first ask if the subjects in Example 23.4 are a random sample of people who attend fairs, at least in the Midwest. The researcher visited 11 different fairs. She stood near the entrance and stopped every 25th adult who passed. Because no personal choice was involved in choosing the subjects, we can reasonably treat the data as coming from a random sample. (As usual, there was some nonresponse, which could create bias.)

Here are the data, presented as a two-way table of counts:

<table>
<thead>
<tr>
<th>Response</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>13</td>
<td>108</td>
<td>50</td>
<td>23</td>
<td>2</td>
<td>196</td>
</tr>
<tr>
<td>Male</td>
<td>22</td>
<td>57</td>
<td>22</td>
<td>5</td>
<td>1</td>
<td>107</td>
</tr>
<tr>
<td>Total</td>
<td>35</td>
<td>165</td>
<td>72</td>
<td>28</td>
<td>3</td>
<td>303</td>
</tr>
</tbody>
</table>

Comparing row percentages shows that the women in the sample do tend to give higher responses (showing more concern):

<table>
<thead>
<tr>
<th>Response</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>6.6%</td>
<td>55.1%</td>
<td>25.5%</td>
<td>11.7%</td>
<td>1.0%</td>
<td>100%</td>
</tr>
<tr>
<td>Male</td>
<td>20.6%</td>
<td>53.3%</td>
<td>20.6%</td>
<td>4.7%</td>
<td>1.0%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Is the difference between the genders statistically significant? We might apply the chi-square test (Chapter 20), which asks if there is a relationship between gender and safety response. Although the chi-square test answers this general question, it ignores the ordering of the responses. We would really like to know whether women are more concerned than men about the safety of the food served. This question depends on the ordering of responses

We might apply the sign test (Chapter 21) to test if the distribution of responses differs between genders. Although the sign test answers this specific question, it ignores which responses are tied. In this data set, 28% of the responses are tied. We would really like to know whether women are more concerned than men about the safety of the food served. This question depends on the ordering of responses.
from least concerned to most concerned. We can use the Wilcoxon test for the hypotheses:

\[ H_0: \text{men and women do not differ in their responses} \]
\[ H_a: \text{women give systematically higher responses than men} \]

The alternative hypothesis is one-sided. Because the responses can take only five values, there are very many ties. All 35 people who chose “very rarely” are tied at 1, and all 165 who chose “once in a while” are tied at 2.

**EXAMPLE 23.5** Food safety: male-female differences

Figure 23.4 gives output from S-PLUS for three tests. The chi-square test is highly significant (\( \chi^2 = 16.1205, P = 0.0029 \)). There is strong evidence that opinions about food safety are related to gender.

The Wilcoxon test for the one-sided alternative that women are more concerned about food safety at fairs is also highly significant (\( z = 3.3335, P = 0.0004 \)). The statistic here is the standardized version of the Wilcoxon \( W \) with adjustment for the many ties. There is very strong evidence of a difference. Women are more concerned than men about the safety of food served at fairs.

With more than 100 observations in each group and no outliers, we might use the two-sample \( t \) test even though responses take only five values. Figure 23.4 shows that \( t = 3.3655 \) with \( P = 0.0005 \). The one-sided \( P \)-value for the two-sample \( t \) test is essentially the same as that for the Wilcoxon test.

As is often the case, \( t \) and \( W \) agree closely in Example 23.5. There is, however, another reason to prefer the rank test in this example. The \( t \) statistic treats the response values 1 through 5 as meaningful numbers. In particular, the possible responses are treated as though they are equally spaced. The difference between “very rarely” and “once in a while” is the same as the difference...
Dealing with ties in rank tests

between “once in a while” and “often.” This may not make sense. The rank test, on the other hand, uses only the order of the responses, not their actual values. The responses are arranged in order from least to most concerned about safety, so the rank test makes sense. Some statisticians avoid using t procedures when there is not a fully meaningful scale of measurement.

**APPLY YOUR KNOWLEDGE**

Software is required to adequately carry out the Wilcoxon rank sum test in the presence of ties. All of the following exercises concern data with ties.

### 23.12 Does polyester decay?

In Example 17.2 (text page 441), we compared the breaking strength of polyester strips buried for 16 weeks with that of strips buried for 2 weeks. The breaking strengths in pounds were

<table>
<thead>
<tr>
<th></th>
<th>2 weeks</th>
<th></th>
<th>16 weeks</th>
</tr>
</thead>
<tbody>
<tr>
<td>weeks</td>
<td>118</td>
<td>126</td>
<td>126</td>
</tr>
</tbody>
</table>

(a) There are two pairs of tied observations. What ranks do you assign to each observation, using average ranks for ties?

(b) Apply the Wilcoxon rank sum test to these data and compare your result with the \( P = 0.1857 \) obtained from the two-sample t test in Chapter 17.

(c) What are the null and alternative hypotheses for the t test? For the Wilcoxon test?

### 23.13 Do birds learn to time their breeding?

Exercises 17.32 to 17.34 (text page 464) concern a study of whether supplementing the diet of blue tits with extra caterpillars will prevent them from adjusting their breeding date the following year in search of a better supply of food. Here are the data (days after the caterpillar peak):

<table>
<thead>
<tr>
<th></th>
<th>4.6</th>
<th>2.3</th>
<th>7.7</th>
<th>6.0</th>
<th>4.6</th>
<th>1.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supplemented</td>
<td>15.5</td>
<td>11.3</td>
<td>5.4</td>
<td>16.5</td>
<td>11.3</td>
<td>11.4</td>
</tr>
</tbody>
</table>

The null hypothesis is no difference in timing; the alternative hypothesis is that the supplemented birds miss the peak by more days because they don’t adjust their breeding date.

(a) There are 3 sets of ties, at 4.6, 7.7, and 11.3. Arrange the observations in order and assign average ranks to each tied observation.

(b) Take \( W \) to be the rank sum for the supplemented group. What is the value of \( W \)?

(c) Find the \( P \)-value of the Wilcoxon test and state your conclusion.
23.14 Tell me a story, continued. The data in Exercise 23.5 for a story told without pictures (Story 1) have tied observations. Is there good evidence that high-progress readers score higher than low-progress readers when they retell a story they have heard without pictures?
(a) Make a back-to-back stemplot of the 5 responses in each group. Are any major deviations from Normality apparent?
(b) Carry out a two-sample t test. State hypotheses and give the two sample means, the t statistic and its P-value, and your conclusion.
(c) Carry out the Wilcoxon rank sum test. State hypotheses and give the rank sum W for high-progress readers, its P-value, and your conclusion. Do the t and Wilcoxon tests lead you to different conclusions?

23.15 Each day I am getting better in math. Table 17.1 (text page 463) gives the pretest and posttest scores for two groups of students taking a program to improve their basic mathematics skills. Did the treatment group show significantly greater improvement than the control group?
(a) Apply the Wilcoxon rank sum test to the posttest versus pretest differences. Note that there are some ties. What do you conclude? Compare your findings with those from the two-sample t test in Exercise 17.28.
(b) What are the null and alternative hypotheses for each of the two tests we have applied to these data?
(c) What must we assume about the data to apply each of the tests?

23.16 Food safety in restaurants. Example 23.4 describes a study of the attitudes of people attending outdoor fairs about the safety of the food served at such locations. The full data set is stored on the CD and online as the file ex23-16.dat. It contains the responses of 303 people to several questions. The variables in this data set are (in order)

subject hfair sfair sfast srest gender

The variable “sfair” contains the responses described in the example concerning safety of food served at outdoor fairs and festivals. The variable “srest” contains responses to the same question asked about food served in restaurants. The variable "gender" contains 1 if the respondent is a woman, 2 if he is a man. We saw that women are more concerned than men about the safety of food served at fairs. Is this also true for restaurants?

23.17 Shopping in secondhand stores. To study customers’ attitudes toward secondhand stores, researchers interviewed samples of shoppers at two secondhand stores of the same chain in two cities. Here are data on the incomes of shoppers at the two stores, presented as a two-way table of counts.
Matched pairs: the Wilcoxon signed rank test

Income code | Income          | City 1 | City 2 |
-------------|----------------|-------|-------|
1            | Under $10,000  | 70    | 62    |
2            | $10,000 to $19,999 | 52    | 63    |
3            | $20,000 to $24,999 | 69    | 50    |
4            | $25,000 to $34,999 | 22    | 19    |
5            | $35,000 or more  | 28    | 24    |

(a) Is there a relationship between city and income? Use the chi-square test to answer this question.

(b) The chi-square test ignores the ordering of the income categories. The data file ex23-17.dat contains data on the 459 shoppers in this study. The first variable is the city (City1 or City2) and the second is the income code as it appears in the table above (1 to 5). Is there good evidence that shoppers in one city have systematically higher incomes than in the other?

23.18 More on food safety. The data file used in Example 23.4 and Exercise 23.16 contains 303 rows, one for each of the 303 respondents. Each row contains the responses of one person to several questions. We wonder if people are more concerned about safety of food served at fairs than they are about the safety of food served at restaurants. Explain carefully why we cannot answer this question by applying the Wilcoxon rank sum test to the variables “sfair” and “srest.”

Matched pairs: the Wilcoxon signed rank test

We use the one-sample t procedures for inference about the mean of one population or for inference about the mean difference in a matched pairs setting. The matched pairs setting is more important because good studies are generally comparative. We will now meet a rank test for this setting.

EXAMPLE 23.6 Tell me a story

A study of early childhood education asked kindergarten students to tell fairy tales that had been read to them earlier in the week. Each child told two stories. The first had been read to them and the second had been read but also illustrated with pictures. An expert listened to a recording of the children and assigned a score for certain uses of language. Here are the data for five low-progress readers in a pilot study:

<table>
<thead>
<tr>
<th>Child</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Story 2</td>
<td>0.77</td>
<td>0.49</td>
<td>0.66</td>
<td>0.28</td>
<td>0.18</td>
</tr>
<tr>
<td>Story 1</td>
<td>0.40</td>
<td>0.72</td>
<td>0.00</td>
<td>0.36</td>
<td>0.55</td>
</tr>
<tr>
<td>Difference</td>
<td>0.37</td>
<td>−0.23</td>
<td>0.66</td>
<td>−0.08</td>
<td>−0.17</td>
</tr>
</tbody>
</table>
CHAPTER 23 • Nonparametric Tests

We wonder if illustrations improve how the children retell a story. We would like to test the hypotheses

\[ H_0: \text{scores have the same distribution for both stories} \]

\[ H_1: \text{scores are systematically higher for Story 2} \]

Because this is a matched pairs design, we base our inference on the differences. The matched pairs \( t \) test gives \( t = 0.635 \) with one-sided \( P \)-value \( P = 0.280 \). We cannot assess Normality from so few observations. We would therefore like to use a rank test.

Positive differences in Example 23.6 indicate that the child performed better telling Story 2. If scores are generally higher with illustrations, the positive differences should be farther from zero in the positive direction than the negative differences are in the negative direction. We therefore compare the absolute values of the differences, that is, their magnitudes without a sign. Here they are, with boldface indicating the positive values:

\[
0.37 \quad 0.23 \quad 0.66 \quad 0.08 \quad 0.17
\]

Arrange these in increasing order and assign ranks, keeping track of which values were originally positive. Tied values receive the average of their ranks. If there are zero differences, discard them before ranking.

<table>
<thead>
<tr>
<th>Absolute value</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.08</td>
<td>1</td>
</tr>
<tr>
<td>0.17</td>
<td>2</td>
</tr>
<tr>
<td>0.23</td>
<td>3</td>
</tr>
<tr>
<td>0.37</td>
<td>4</td>
</tr>
<tr>
<td>0.66</td>
<td>5</td>
</tr>
</tbody>
</table>

The test statistic is the sum of the ranks of the positive differences. (We could equally well use the sum of the ranks of the negative differences.) This is the Wilcoxon signed rank statistic. Its value here is \( W^+ = 9 \).

THE WILCOXON SIGNED RANK TEST FOR MATCHED PAIRS

Draw an SRS of size \( n \) from a population for a matched pairs study and take the differences in responses within pairs. Rank the absolute values of these differences. The sum \( W^+ \) of the ranks for the positive differences is the Wilcoxon signed rank statistic. If the distribution of the responses is not affected by the different treatments within pairs, then \( W^+ \) has mean

\[
\mu_{W^+} = \frac{n(n + 1)}{4}
\]

and standard deviation

\[
\sigma_{W^+} = \sqrt{\frac{n(n + 1)(2n + 1)}{24}}
\]

The Wilcoxon signed rank test rejects the hypothesis that there are no systematic differences within pairs when the rank sum \( W^+ \) is far from its mean.
EXAMPLE 23.7  Tell me a story: mean and standard deviation

In the storytelling study of Example 23.6, \( n = 5 \). If the null hypothesis (no systematic effect of illustrations) is true, the mean of the signed rank statistic is

\[
\mu_{W^+} = \frac{n(n + 1)}{4} = \frac{5(6)}{4} = 7.5
\]

The standard deviation of \( W^+ \) under the null hypothesis is

\[
\sigma_{W^+} = \sqrt{\frac{n(n + 1)(2n + 1)}{24}} = \sqrt{\frac{5(6)(11)}{24}} = \sqrt{\frac{330}{24}} = \sqrt{13.75} = 3.708
\]

The observed value \( W^+ = 9 \) is only slightly larger than the mean. We now expect that the data are not statistically significant.

APPLY YOUR KNOWLEDGE

23.19 Growing trees faster. Exercise 16.14 (text page 428) describes an experiment in which extra carbon dioxide was piped to some plots in a pine forest. Each plot was paired with a nearby control plot left in its natural state. Do trees grow faster with extra carbon dioxide? Here are the average percent increases in base area for trees in the plots:

<table>
<thead>
<tr>
<th>Pair</th>
<th>Control plot</th>
<th>Treated plot</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9.752</td>
<td>10.587</td>
</tr>
<tr>
<td>2</td>
<td>7.263</td>
<td>9.244</td>
</tr>
<tr>
<td>3</td>
<td>5.742</td>
<td>8.675</td>
</tr>
</tbody>
</table>

The investigators used the matched pairs \( t \) test. With only 3 pairs, we can’t verify Normality. We will try the Wilcoxon signed rank test.

(a) Find the differences within pairs, arrange them in order, and rank the absolute values. What is the signed rank statistic \( W^+ \)?

(b) If the null hypothesis (no difference in growth) is true, what are the mean and standard deviation of \( W^+ \)? Does comparing \( W^+ \) to this mean lead to a tentative conclusion?

23.20 Floral scents and learning. Table 16.1 (text page 423) gives matched pairs data for 21 subjects. The response variable is time to complete a maze, both wearing a scented mask and wearing an identical mask that is unscented. Does the scent improve performance (that is, shorten the time needed to complete the maze)? The matched pairs \( t \) test (Example 16.3) works well, and gives \( P = 0.365 \). Let’s compare the Wilcoxon signed rank test.
(a) What are the ranks for the absolute values of the differences in Table 16.1? What is the value of $W^+$?

(b) What would be the mean and standard deviation of $W^+$ if the null hypothesis (scent makes no difference) were true? Compare $W^+$ with this mean (in standard deviation units) to reach a tentative conclusion about significance.

The Normal approximation for $W^+$

The distribution of the signed rank statistic when the null hypothesis (no difference) is true becomes approximately Normal as the sample size becomes large. We can then use Normal probability calculations (with the continuity correction) to obtain approximate $P$-values for $W^+$. Let’s see how this works in the storytelling example, even though $n = 5$ is certainly not a large sample.

**EXAMPLE 23.8** Tell me a story: the Normal approximation

For $n = 5$ observations, we saw in Example 23.7 that $\mu_{W^+} = 7.5$ and that $\sigma_{W^+} = 3.708$. We observed $W = 9$, so the one-sided $P$-value is $P(W^+ \geq 9)$. The continuity correction calculates this as $P(W^+ \geq 8.5)$, treating the value $W^+ = 9$ as occupying the interval from 8.5 to 9.5. We find the Normal approximation for the $P$-value by standardizing and using the standard Normal table:

$$P(W^+ \geq 8.5) = P\left( \frac{W^+ - 7.5}{3.708} \geq \frac{8.5 - 7.5}{3.708} \right) = P(Z \geq \frac{8.5 - 7.5}{3.708}) = P(Z \geq 0.27) = 0.394$$

Figure 23.5 displays the output of two statistical programs. Minitab uses the Normal approximation and agrees with our calculation $P = 0.394$. We asked S-PLUS to do three analyses: using the exact distribution of $W^+$, using the Normal approximation, and using the matched pairs $t$ test. The exact one-sided $P$-value for the Wilcoxon signed rank test is $P = 0.4062$. The Normal approximation is quite close to this. The $t$ test result is a bit different, $P = 0.280$, but all three tests tell us that this very small sample gives no evidence that seeing illustrations improves the storytelling of low-progress readers.

**APPLY YOUR KNOWLEDGE**

23.21 Growing trees faster: Normal approximation. Continue your work from Exercise 23.19. Use the Normal approximation (with continuity correction) to find the $P$-value for the signed rank test against the one-sided alternative that trees grow faster with added carbon dioxide. What do you conclude?

23.22 $W^+$ versus $t$. Find the one-sided $P$-value for the matched pairs $t$ test applied to the tree growth data in Exercise 23.19. The smaller $P$-value of $t$ relative to $W^+$ means that $t$ gives stronger evidence of the effect...
The Normal approximation for \( W \)

The Wilcoxon Signed Rank Test: Diff

Test of median = 0.000000 versus median > 0.000000

<table>
<thead>
<tr>
<th>N for Wilcoxon</th>
<th>Estimated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diff</td>
<td>N</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

Exact Wilcoxon signed-rank test

data: Diff in eg23.06
signed-rank statistic \( V = 9, n = 5 \)
p-value = 0.4042
alternative hypothesis: true mu is greater than 0

Wilcoxon signed-rank test

data: Diff in eg23.06
signed-rank normal statistic with correction \( Z = 0.2697 \)
p-value = 0.3937
alternative hypothesis: true mu is greater than 0

One-sample t-Test

data: Diff in eg23.06
t = 0.635, df = 4, p-value = 0.28
alternative hypothesis: true mean is greater than 0

S-Plus

Figure 23.5 Output from Minitab and S-PLUS for the storytelling data of Example 23.6. The S-PLUS output compares the Wilcoxon signed rank test with the exact distribution and with the Normal approximation and the matched pairs \( t \) test.

of carbon dioxide on growth. The \( t \) test takes advantage of assuming that the data are Normal, a considerable advantage for these very small samples.

23.23 Floral scents and learning: Normal approximation. Use the Normal approximation with continuity correction to find the \( P \)-value for the test in Exercise 23.20. Does the Wilcoxon signed rank test lead to essentially the same result as the \( t \) test?

23.24 Ancient air. Exercise 16.7 (text page 416) reports the following data on the percent of nitrogen in bubbles of ancient air trapped in amber:

- 63.4 65.0 64.4 63.3 54.8 64.5 60.8 49.1 51.0

We wonder if ancient air differs significantly from the present atmosphere, which is 78.1% nitrogen.
(a) Graph the data, and comment on skewness and outliers. A rank test is appropriate.

(b) We would like to test hypotheses about the median percent of nitrogen in ancient air (the population):

\[ H_0: \text{median} = 78.1 \]
\[ H_1: \text{median} \neq 78.1 \]

To do this, apply the Wilcoxon signed rank statistic to the differences between the observations and 78.1. (This is the one-sample version of the test.) What do you conclude?

### Dealing with ties in the signed rank test

Ties among the absolute differences are handled by assigning average ranks. A tie within a pair creates a difference of zero. Because these are neither positive nor negative, we drop such pairs from our sample. Ties within pairs simply reduce the number of observations, but ties among the absolute differences complicate finding a \( P \)-value. There is no longer a usable exact distribution for the signed rank statistic \( W^+ \), and the standard deviation \( \sigma_{W^+} \) must be adjusted for the ties before we can use the Normal approximation. Software will do this. Here is an example.

#### EXAMPLE 23.9 Golf scores

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

<table>
<thead>
<tr>
<th>Player</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Round 2</td>
<td>94</td>
<td>85</td>
<td>89</td>
<td>89</td>
<td>81</td>
<td>107</td>
<td>89</td>
<td>87</td>
<td>91</td>
<td>88</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>Round 1</td>
<td>89</td>
<td>90</td>
<td>87</td>
<td>95</td>
<td>86</td>
<td>102</td>
<td>105</td>
<td>83</td>
<td>88</td>
<td>91</td>
<td>79</td>
<td></td>
</tr>
<tr>
<td>Difference</td>
<td>-5</td>
<td>-5</td>
<td>2</td>
<td>-6</td>
<td>-5</td>
<td>-5</td>
<td>5</td>
<td>-16</td>
<td>4</td>
<td>3</td>
<td>-3</td>
<td>1</td>
</tr>
</tbody>
</table>

Negative differences indicate better (lower) scores on the second round. We see that 6 of the 12 golfers improved their scores. We would like to test the hypotheses that in a large population of collegiate woman golfers

\[ H_0: \text{scores have the same distribution in rounds 1 and 2} \]
\[ H_1: \text{scores are systematically lower or higher in round 2} \]

A stemplot plot of the differences (Figure 23.6) shows some irregularity and a low outlier. We will use the Wilcoxon signed rank test.

The absolute values of the differences, with boldface indicating those that were negative, are

\[ 5 \ 5 \ 2 \ 6 \ 5 \ 5 \ 16 \ 4 \ 3 \ 3 \ 1 \]
Arrange these in increasing order and assign ranks, keeping track of which values were originally negative. Tied values receive the average of their ranks.

<table>
<thead>
<tr>
<th>Absolute value</th>
<th>1 2 3 4 5 5 5 6 16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank</td>
<td>1 2 3.5 3.5 5 8 8 8 11 12</td>
</tr>
</tbody>
</table>

The Wilcoxon signed rank statistic is the sum of the ranks of the negative differences. (We could equally well use the sum for the ranks of the positive differences.) Its value is $W^+ = 50.5$.

**EXAMPLE 23.10 Golf scores: computer output**

Figure 23.7 displays Minitab and S-PLUS output for the golf score data. The two programs give slightly different two-sided $P$-values, $P = 0.388$ from Minitab and $P = 0.384$ from S-PLUS. Both lead to the same practical conclusion: these data give no evidence for a systematic change in scores between rounds. The results differ because the programs use slightly different versions of the approximate calculations needed when ties are present.

**Minitab**

Wilcoxon Signed Rank Test: Diff

<table>
<thead>
<tr>
<th>Test of median = 0.000000 versus median not = 0.000000</th>
</tr>
</thead>
<tbody>
<tr>
<td>N for Wilcoxon                  Estimated</td>
</tr>
<tr>
<td>12                  12              57.5 0.388 -1.000</td>
</tr>
</tbody>
</table>

**S-PLUS**

Wilcoxon signed-rank test

data:Diff in eg23.09
signed-rank normal statistic with correction Z = -0.47, p-value = 0.6443
alternative hypothesis: true mu is not equal to 0

One-sample t-Test

data: Diff in eg23.09
t = -0.9414, df = 11, p-value = 0.3716
alternative hypothesis: true mean is not equal to 0

Figure 23.7 Output from Minitab and S-PLUS for the golf scores data of Example 23.9. Because there are ties, a Normal approximation must be used.
Nonparametric Tests

The S-PLUS display also includes the matched pairs t test: \( t = -0.9314 \) with \( P = 0.3716 \). Once again, t and \( W^+ \) lead to the same conclusion.

**APPLY YOUR KNOWLEDGE**

23.25 Stepping up your heart rate. The EESEE story “Stepping Up Your Heart Rate” describes a student project that asked subjects to step up and down for three minutes and measured their heart rates before and after the exercise. Here are data for five subjects and two treatments: stepping at a low rate (14 steps per minute) and at a medium rate (21 steps per minute). For each subject, we give the resting heart rate (beats per minute) and the heart rate at the end of the exercise. (The data are slightly modified.)

<table>
<thead>
<tr>
<th>Subject</th>
<th>Resting Rate</th>
<th>Final Rate</th>
<th>Resting Rate</th>
<th>Final Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>60</td>
<td>75</td>
<td>63</td>
<td>84</td>
</tr>
<tr>
<td>2</td>
<td>90</td>
<td>98</td>
<td>69</td>
<td>93</td>
</tr>
<tr>
<td>3</td>
<td>87</td>
<td>93</td>
<td>81</td>
<td>96</td>
</tr>
<tr>
<td>4</td>
<td>78</td>
<td>87</td>
<td>75</td>
<td>90</td>
</tr>
<tr>
<td>5</td>
<td>84</td>
<td>54</td>
<td>90</td>
<td>108</td>
</tr>
</tbody>
</table>

Does exercise at the low rate raise heart rate significantly?

(a) State hypotheses in terms of the median increase in heart rate.

(b) Find the differences and assign ranks to the absolute values. Remember that zeros are dropped from the data before ranking, so that \( n \) is the number of nonzero differences within pairs. What is the value of \( W^+ \)?

(c) Because there are no ties in the absolute values, you can use the usual Normal approximation (with the reduced \( n \)). Find the \( P \)-value and state your conclusion.

23.26 Stepping up your heart rate, continued. Do the data in the previous exercise give good evidence that stepping at the medium rate raises heart rates? Now there are ties in the absolute values of the differences, so software is required for an accurate analysis.

23.27 Stepping up your heart rate: medium versus low. Do the data from Exercise 23.25 give good reason to think that stepping at the medium rate increases heart rates more than stepping at the low rate?

(a) State hypotheses in terms of comparing the median increases for the two treatments. What is the proper rank test for these hypotheses?

(b) Carry out your test and state a conclusion.
23.28 Does nature heal best? Table 16.3 (text page 435) gives data on the healing rate (micrometers per hour) of the skin of newts under two conditions. This is a matched pairs design, with the body’s natural electric field for one limb (control) and half the natural value for another limb of the same newt (experimental). We want to know if the healing rates are systematically different under the two conditions. You decide to use a rank test.

(a) There are several ties among the absolute differences. Find the ranks and give the value of the signed rank statistic $W^+$.

(b) Use software to find the $P$-value. Give a conclusion. Be sure to include a description of what the data show in addition to the test results.

23.29 Sweetening colas. Cola makers test new recipes for loss of sweetness during storage. Trained tasters rate the sweetness before and after storage. Here are the sweetness losses (sweetness before storage minus sweetness after storage) found by 10 tasters for one new cola recipe:

2.0 0.4 0.7 2.0 $-0.4$ 2.2 $-1.3$ 1.2 1.1 2.3

Are these data good evidence that the cola lost sweetness?

(a) These data are the differences from a matched pairs design. State hypotheses in terms of the median difference in the population of all tasters, carry out a test, and give your conclusion.

(b) In Example 16.2 we found that the one-sample $t$ test had $P$-value $P = 0.0122$ for these data. How does this compare with your result from (a)? What are the hypotheses for the $t$ test? What assumptions must we make for each of the $t$ and Wilcoxon tests?

Comparing several samples: the Kruskal-Wallis test

We have now considered alternatives to the paired-sample and two-sample $t$ tests for comparing the magnitude of responses to two treatments. To compare more than two treatments, we use one-way analysis of variance (ANOVA) if the distributions of the responses to each treatment are at least roughly Normal and have similar spreads. What can we do when these distribution requirements are violated?

EXAMPLE 23.11 Weeds among the corn

Lamb’s-quarter is a common weed that interferes with the growth of corn. A researcher planted corn at the same rate in 16 small plots of ground, then randomly assigned the plots to four groups. He weeded the plots by hand to allow a fixed number of lamb’s-quarter plants to grow in each meter of corn row. These numbers were 0, 1, 3, and 9 in the four groups of plots. No other weeds were allowed to grow, and all plots received identical treatment except for the weeds. Here are the yields of corn (bushels per acre) in each of the plots:
CHAPTER 23  Nonparametric Tests

The summary statistics are

<table>
<thead>
<tr>
<th>Weeds per meter</th>
<th>Corn yield</th>
<th>Weeds per meter</th>
<th>Corn yield</th>
<th>Weeds per meter</th>
<th>Corn yield</th>
<th>Weeds per meter</th>
<th>Corn yield</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>166.7</td>
<td>1</td>
<td>166.2</td>
<td>3</td>
<td>156.6</td>
<td>9</td>
<td>162.8</td>
</tr>
<tr>
<td>0</td>
<td>172.2</td>
<td>1</td>
<td>157.3</td>
<td>3</td>
<td>176.4</td>
<td>9</td>
<td>142.4</td>
</tr>
<tr>
<td>0</td>
<td>165.0</td>
<td>1</td>
<td>166.7</td>
<td>3</td>
<td>153.1</td>
<td>9</td>
<td>162.7</td>
</tr>
<tr>
<td>0</td>
<td>176.9</td>
<td>1</td>
<td>161.1</td>
<td>3</td>
<td>156.0</td>
<td>9</td>
<td>162.4</td>
</tr>
</tbody>
</table>

The sample standard deviations do not quite satisfy our rule of thumb that for safe use of ANOVA the largest should not exceed twice the smallest. Moreover, we see that outliers are present in the yields for 3 and 9 weeds per meter. These are the correct yields for their plots, so we have no justification for removing them. We may want to use a nonparametric test.

Hypotheses and conditions for the Kruskal-Wallis test

The ANOVA $F$ test concerns the means of the several populations represented by our samples. For Example 23.11, the ANOVA hypotheses are

$H_0: \mu_0 = \mu_1 = \mu_3 = \mu_9$

$H_a: \text{not all four means are equal}$

For example, $\mu_0$ is the mean yield in the population of all corn planted under the conditions of the experiment with no weeds present. The data should consist of four independent random samples from the four populations, all Normally distributed with the same standard deviation.

The Kruskal-Wallis test is a rank test that can replace the ANOVA $F$ test. The condition about data production (independent random samples from each population) remains important, but we can relax the Normality condition. We assume only that the response has a continuous distribution in each population.

The hypotheses tested in our example are

$H_0: \text{yields have the same distribution in all groups}$

$H_a: \text{yields are systematically higher in some groups than in others}$

If all of the population distributions have the same shape (Normal or not), these hypotheses take a simpler form. The null hypothesis is that all four populations have the same median yield. The alternative hypothesis is that not all four median yields are equal.
The Kruskal-Wallis test statistic

Recall the analysis of variance idea: we write the total observed variation in the responses as the sum of two parts, one measuring variation among the groups (sum of squares for groups, SSG) and one measuring variation among individual observations within the same group (sum of squares for error, SSE). The ANOVA $F$ test rejects the null hypothesis that the mean responses are equal in all groups if SSG is large relative to SSE.

The idea of the Kruskal-Wallis rank test is to rank all the responses from all groups together and then apply one-way ANOVA to the ranks rather than to the original observations. If there are $N$ observations in all, the ranks are always the whole numbers from 1 to $N$. The total sum of squares for the ranks is therefore a fixed number no matter what the data are. So we do not need to look at both SSG and SSE. Although it isn’t obvious without some unpleasant algebra, the Kruskal-Wallis test statistic is essentially just SSG for the ranks. We give the formula, but you should rely on software to do the arithmetic.

When SSG is large, that is evidence that the groups differ.

**THE KRUSKAL-WALLIS TEST**

Draw independent SRSs of sizes $n_1, n_2, \ldots, n_I$ from $I$ populations. There are $N$ observations in all. Rank all $N$ observations and let $R_i$ be the sum of the ranks for the $i$th sample. The Kruskal-Wallis statistic is

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

When the sample sizes $n_i$ are large and all $I$ populations have the same continuous distribution, $H$ has approximately the chi-square distribution with $I - 1$ degrees of freedom.

The Kruskal-Wallis test rejects the null hypothesis that all populations have the same distribution when $H$ is large.

We now see that, like the Wilcoxon rank sum statistic, the Kruskal-Wallis statistic is based on the sums of the ranks for the groups we are comparing. The more different these sums are, the stronger is the evidence that responses are systematically larger in some groups than in others.

The exact distribution of the Kruskal-Wallis statistic $H$ under the null hypothesis depends on all the sample sizes $n_1$ to $n_I$, so tables are awkward. The calculation of the exact distribution is so time-consuming for all but the smallest problems that even most statistical software uses the chi-square approximation to obtain $P$-values. As usual, there is no usable exact distribution when there are ties among the responses. We again assign average ranks to tied observations.
EXAMPLE 23.12. Weeds among the corn, continued.

In Example 23.11, there are $I = 4$ populations and $N = 16$ observations. The sample sizes are equal, $n_i = 4$. The 16 observations arranged in increasing order, with their ranks, are:

<table>
<thead>
<tr>
<th>Yield</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>142.4</td>
<td>1</td>
</tr>
<tr>
<td>153.1</td>
<td>2</td>
</tr>
<tr>
<td>156.0</td>
<td>3</td>
</tr>
<tr>
<td>157.3</td>
<td>4</td>
</tr>
<tr>
<td>158.6</td>
<td>5</td>
</tr>
<tr>
<td>161.1</td>
<td>6</td>
</tr>
<tr>
<td>162.4</td>
<td>7</td>
</tr>
<tr>
<td>162.7</td>
<td>8</td>
</tr>
<tr>
<td>162.8</td>
<td>9</td>
</tr>
<tr>
<td>165.0</td>
<td>10</td>
</tr>
<tr>
<td>166.2</td>
<td>11</td>
</tr>
<tr>
<td>166.7</td>
<td>12</td>
</tr>
<tr>
<td>166.7</td>
<td>12.5</td>
</tr>
<tr>
<td>172.2</td>
<td>14</td>
</tr>
<tr>
<td>176.4</td>
<td>15</td>
</tr>
<tr>
<td>176.9</td>
<td>16</td>
</tr>
</tbody>
</table>

There is one pair of tied observations. The ranks for each of the four treatments are:

<table>
<thead>
<tr>
<th>Weeds</th>
<th>Ranks</th>
<th>Sum of ranks</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10</td>
<td>12.5 14 16</td>
</tr>
<tr>
<td>1</td>
<td>4 6</td>
<td>11 12.5   33.5</td>
</tr>
<tr>
<td>3</td>
<td>2 3</td>
<td>5 15     25.0</td>
</tr>
<tr>
<td>9</td>
<td>1 7</td>
<td>8 9      25.0</td>
</tr>
</tbody>
</table>

The Kruskal-Wallis statistic is therefore:

$$H = \frac{12}{N(N+1)} \sum R_i^2 - 3(N+1)$$

$$= \frac{12}{(16)(17)} \left( \frac{52.5^2}{4} + \frac{33.5^2}{4} + \frac{25^2}{4} + \frac{25^2}{4} \right) - 3(17)$$

$$= \frac{12}{272} \left(1282.125 \right) - 51$$

$$= 5.56$$

Referring to the table of chi-square critical points (Table E) with df = 3, we find that the $P$-value lies in the interval $0.10 < P < 0.15$. This small experiment suggests that more weeds decrease yield but does not provide convincing evidence that weeds have an effect.

Figure 23.8 displays the Minitab output for both ANOVA and the Kruskal-Wallis test. Minitab agrees that $H = 5.56$ and gives $P = 0.135$. Minitab also gives the results of making an adjustment that makes the chi-square approximation more accurate when there are ties. For these data, the adjustment has no practical effect. It would be important if there were many ties. A calculation that takes several hours of computer time shows that the exact $P$-value is $P = 0.1299$. The chi-square approximation is quite accurate.

The ANOVA $F$ test gives $F = 1.73$ with $P = 0.213$. Although the practical conclusion is the same, ANOVA and Kruskal-Wallis do not agree closely in this example. The rank test is more reliable for these small samples with outliers.
The Kruskal-Wallis test statistic

### Kruskal-Wallis Test on Yield

**Kruskal-Wallis Test: Yield versus Weeds**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weeds</td>
<td>3</td>
<td>340.7</td>
<td>113.6</td>
<td>1.73</td>
<td>0.213</td>
</tr>
<tr>
<td>Error</td>
<td>12</td>
<td>785.5</td>
<td>65.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>15</td>
<td>1126.2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Median

<table>
<thead>
<tr>
<th>Level</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>169.5</td>
</tr>
<tr>
<td>1</td>
<td>163.6</td>
</tr>
<tr>
<td>3</td>
<td>157.3</td>
</tr>
<tr>
<td>9</td>
<td>162.6</td>
</tr>
</tbody>
</table>

Ave Rank

<table>
<thead>
<tr>
<th>Level</th>
<th>Ave Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>13.1</td>
</tr>
<tr>
<td>1</td>
<td>8.4</td>
</tr>
<tr>
<td>3</td>
<td>6.3</td>
</tr>
<tr>
<td>9</td>
<td>8.5</td>
</tr>
</tbody>
</table>

N

<table>
<thead>
<tr>
<th>Level</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
</tr>
</tbody>
</table>

16

Z

<table>
<thead>
<tr>
<th>Level</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2.24</td>
</tr>
<tr>
<td>1</td>
<td>-0.06</td>
</tr>
<tr>
<td>3</td>
<td>-1.09</td>
</tr>
<tr>
<td>9</td>
<td>-1.09</td>
</tr>
</tbody>
</table>

H = 5.56  DF = 3  P = 0.135
H = 5.57  DF = 3  P = 0.134 (adjusted for ties)

*NOTE* One or more small samples

### One-way ANOVA: Yield versus Weeds

Analysis of Variance for Yield

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>ME</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error</td>
<td>12</td>
<td>785.5</td>
<td>65.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>15</td>
<td>1126.2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Level

<table>
<thead>
<tr>
<th>Color</th>
<th>Mean</th>
<th>StDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>170.20</td>
<td>5.42</td>
</tr>
<tr>
<td>1</td>
<td>161.03</td>
<td>10.49</td>
</tr>
<tr>
<td>3</td>
<td>157.57</td>
<td>10.49</td>
</tr>
<tr>
<td>9</td>
<td>157.57</td>
<td>10.49</td>
</tr>
</tbody>
</table>

Pooled StDev = 8.09

**Figure 23.8** Minitab output for the corn yield data of Example 23.11. For comparison, both the Kruskal-Wallis test and one-way ANOVA are shown.

### APPLY YOUR KNOWLEDGE

**23.30 Which color attracts beetles best?** Example 22.6 used ANOVA to analyze the results of a study to see which of four colors best attracts cereal leaf beetles. Here are the data:

<table>
<thead>
<tr>
<th>Color</th>
<th>Insects trapped</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue</td>
<td>16 11 20 21 14 7</td>
</tr>
<tr>
<td>Green</td>
<td>37 32 15 25 39 41</td>
</tr>
<tr>
<td>White</td>
<td>21 12 14 17 33 41</td>
</tr>
<tr>
<td>Yellow</td>
<td>45 59 48 46 38 47</td>
</tr>
</tbody>
</table>

Because the samples are small, we will apply a nonparametric test. (a) Find the median number of beetles trapped by boards of each color. Which colors appear more effective?
(b) What hypotheses does ANOVA test? What hypotheses does Kruskal-Wallis test?
(c) What are \( I \), the \( n_i \), and \( N \)? Arrange the counts in order and assign ranks. Be careful about ties.
(d) Calculate the Kruskal-Wallis statistic \( H \). How many degrees of freedom should you use for the chi-square approximation to its null distribution? Use the chi-square table to give an approximate \( P \)-value. What does the test lead you to conclude?

23.31 Logging in the rainforest: species richness. Table 22.2 (text page 604) contains data comparing the number of trees and number of tree species in plots of land in a tropical rainforest that had never been logged with similar plots nearby that had been logged 1 year earlier and 8 years earlier. The third response variable is species richness, the number of tree species divided by the number of trees. There are low outliers in the data, and a histogram of the ANOVA residuals shows outliers as well. Because of lack of Normality in small samples, we may prefer the Kruskal-Wallis test.

(a) Make a graph to compare the distributions of richness for the three groups of plots. Also give the median richness for the three groups.
(b) Use the Kruskal-Wallis test to compare the distributions of richness. State hypotheses, the test statistic and its \( P \)-value, and your conclusions.

23.32 Does polyester decay? Here are the breaking strengths (in pounds) of strips of polyester fabric buried in the ground for several lengths of time:

```
2 weeks  118 126 126 120 129
4 weeks  130 120 114 126 128
8 weeks  122 136 128 146 140
16 weeks 124  98 110 140 110
```

Breaking strength is a good measure of the extent to which the fabric has decayed.
(a) Find the standard deviations of the 4 samples. They do not meet our rule of thumb for applying ANOVA. In addition, the sample buried for 16 weeks contains an outlier. We will use the Kruskal-Wallis test.
(b) Find the medians of the four samples. What are the hypotheses for the Kruskal-Wallis test, expressed in terms of medians?
(c) Carry out the test and report your conclusion.

23.33 Food safety. Example 23.4 describes a study of the attitudes of people attending outdoor fairs about the safety of the food served at such locations. The full data set is stored on the CD and online as the file
**Chapter 23 Summary**

*Nonparametric tests* do not require any specific form for the distribution of the population from which our samples come.

*Rank tests* are nonparametric tests based on the *ranks* of observations, their positions in the list ordered from smallest (rank 1) to largest. Tied observations receive the average of their ranks.

The *Wilcoxon rank sum test* compares two distributions to assess whether one has systematically larger values than the other. The Wilcoxon test is based on the *Wilcoxon rank sum statistic* $W$, which is the sum of the ranks of one of the samples. The Wilcoxon test can replace the *two-sample t test*. $P$-values for the Wilcoxon test are based on the sampling distribution of the rank sum statistic $W$ when the null hypothesis (no difference in distributions) is true. You can find $P$-values from special tables, software, or a Normal approximation (with continuity correction).

The *Wilcoxon signed rank test* applies to matched pairs studies. It tests the null hypothesis that there is no systematic difference within pairs against alternatives that assert a systematic difference (either one-sided or two-sided). The test is based on the *Wilcoxon signed rank statistic* $W^+$, which is the sum of the ranks of the positive (or negative) differences when we rank the absolute values of the differences. The *matched pairs t test* is an alternative test in this setting. $P$-values for the signed rank test are based on the sampling distribution of $W^+$ when the null hypothesis is true. You can find $P$-values from special tables, software, or a Normal approximation (with continuity correction).

The *Kruskal-Wallis test* compares several populations on the basis of independent random samples from each population. This is the *one-way analysis of variance* setting. The null hypothesis for the Kruskal-Wallis test is that the distribution of the response variable is the same in all the populations. The alternative hypothesis is that responses are systematically larger in some populations than in others.

*ex23-16.dat*. It contains the responses of 303 people to several questions. The variables in this data set are (in order):

- subject
- hfair
- sfair
- sfast
- srest
- gender

The variable “sfair” contains responses to the safety question described in Example 23.4. The variables “srest” and “sfast” contain responses to the same question asked about food served in restaurants and in fast-food chains. Explain carefully why we cannot use the Kruskal-Wallis test to see if there are systematic differences in perceptions of food safety in these three locations.
CHAPTER 23 • Nonparametric Tests

The Kruskal-Wallis statistic $H$ can be viewed in two ways. It is essentially the result of applying one-way ANOVA to the ranks of the observations. It is also a comparison of the sums of the ranks for the several samples.

When the sample sizes are not too small and the null hypothesis is true, $H$ for comparing $I$ populations has approximately the chi-square distribution with $I - 1$ degrees of freedom. We use this approximate distribution to obtain $P$-values.

**STATISTICS IN SUMMARY**

After studying this chapter, you should be able to do the following.

**A. RANKS**
1. Assign ranks to a moderate number of observations. Use average ranks if there are ties among the observations.
2. From the ranks, calculate the rank sums when the observations come from two or several samples.

**B. RANK TEST STATISTICS**
1. Determine which of the rank sum tests is appropriate in a specific problem setting.
2. Calculate the Wilcoxon rank sum $W$ from ranks for two samples, the Wilcoxon signed rank sum $W^+$ for matched pairs, and the Kruskal-Wallis statistic $H$ for two or more samples.
3. State the hypotheses tested by each of these statistics in specific problem settings.
4. Determine when it is appropriate to state the hypotheses for $W$ and $H$ in terms of population medians.

**C. RANK TESTS**
1. Use software to carry out any of the rank tests. Combine the test with data description and give a clear statement of findings in specific problem settings.
2. Use the Normal approximation with continuity correction to find approximate $P$-values for $W$ and $W^+$. Use a table of chi-square critical values to approximate the $P$-value of $H$.

**Chapter 23 EXERCISES**

23.34 DDT poisoning. Exercise 17.13 (text page 453) reports the results of a study of the effect of the pesticide DDT on nerve activity in rats. The data for the DDT group are

\[
12.207 \quad 16.869 \quad 25.050 \quad 22.429 \quad 8.456 \quad 20.589
\]
The control group data are

It is difficult to assess Normality from such small samples, so we might use a nonparametric test to assess whether DDT affects nerve response. Describe the data to compare responses in the two groups. State hypotheses, carry out a rank test, and give your conclusions.

23.35 Right versus left. Table 16.2 (text page 425) contains data from a student project that investigated whether right-handed people can turn a knob faster clockwise than they can counterclockwise. Describe what the data show, then state hypotheses and do a test that does not require Normality. Report your conclusions carefully.

23.36 Logging in the rainforest. Exercise 17.8 (text page 448) compared the number of tree species in unlogged plots in the rainforest of Borneo with the number of species in plots logged 8 years earlier. Here are the data:

<table>
<thead>
<tr>
<th>Unlogged</th>
<th>22 18 22 20 15 21 13 13 19 13 19 15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logged</td>
<td>17 4 18 14 18 15 10 12</td>
</tr>
</tbody>
</table>

(a) Make a back-to-back stemplot of the data. Does there appear to be a difference in species counts for logged and unlogged plots?
(b) Does logging significantly reduce the mean number of species in a plot after 8 years? State the hypotheses, do a rank test, and state your conclusion.

23.37 Food safety at fairs and restaurants. Example 23.4 describes a study of the attitudes of people attending outdoor fairs about the safety of the food served at such locations. The full data set is stored on the CD and online as the file ex23-16.dat. It contains the responses of 303 people to several questions. The variables in this data set are (in order):

subject hfair sfair sfast srest gender

The variable “sfair” contains responses to the safety question described in Example 23.4. The variable “srest” contains responses to the same question asked about food served in restaurants. We suspect that restaurant food will appear safer than food served outdoors at a fair. Do the data give good evidence for this suspicion? (Give descriptive measures, a test statistic and its P-value, and your conclusion.)

23.38 Food safety at fairs and fast-food restaurants. The food safety survey data described in Example 23.4 also contain the responses of the 303 subjects to the same question asked about food served at fast-food restaurants. These responses are the values of the variable “sfast.” Is there a systematic difference between the level of concern about food safety at outdoor fairs and at fast-food restaurants?
23.39 Nematodes and plant growth. How do nematodes (microscopic worms) affect plant growth? A botanist prepares 16 identical planting pots and then introduces different numbers of nematodes into the pots. A tomato seedling is transplanted into each pot. Here are data on the increase in height of the seedlings (in centimeters) 16 days after planting:

<table>
<thead>
<tr>
<th>Nematodes</th>
<th>Seedling growth</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10.8 9.1 13.5 9.2</td>
</tr>
<tr>
<td>1,000</td>
<td>11.1 11.1 8.2 11.3</td>
</tr>
<tr>
<td>5,000</td>
<td>5.4 4.6 7.4 5.0</td>
</tr>
<tr>
<td>10,000</td>
<td>5.8 5.3 3.2 7.5</td>
</tr>
</tbody>
</table>

We applied ANOVA to these data in Exercise 22.27. Because the samples are very small, it is difficult to assess Normality.

(a) What hypotheses does ANOVA test? What hypotheses does Kruskal-Wallis test?

(b) Find the median growth in each group. Do nematodes appear to retard growth? Apply the Kruskal-Wallis test. What do you conclude?

Notes and Data Sources

1. Data provided by Samuel Phillips, Purdue University.
2. Data provided by Susan Stadler, Purdue University.
5. See Note 1.
6. Data provided by Matthew Moore.