



CHAPTER 25

Ariel Skelly/CORBIS

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 17, 18, and 24.

robustness

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 lack of Normality is due to outliers, it may be legitimate to **remove outliers** if you have reason to think that they do not come from the same population as the other observations. Equipment failure that produced a bad measurement, for example, entitles you to remove the outlier and analyze the remaining data. *But if an outlier appears to be “real data,” you should not arbitrarily remove it.*



2. In some settings, **other standard distributions** replace the Normal distributions as models for the overall pattern in the population. The lifetimes in

IN THIS CHAPTER WE COVER...

- Comparing two samples: the Wilcoxon rank sum test
- The Normal approximation for W
- Using technology
- What hypotheses does Wilcoxon test?
- Dealing with ties in rank tests
- Matched pairs: the Wilcoxon signed rank test
- The Normal approximation for W^+
- Dealing with ties in the signed rank test
- Comparing several samples: the Kruskal-Wallis test
- Hypotheses and conditions for the Kruskal-Wallis test
- The Kruskal-Wallis test statistic

25-1

25-2 CHAPTER 25 • Nonparametric Tests

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.

3. Modern **bootstrap methods** and **permutation tests** use heavy computing to avoid requiring Normality or any other specific form of sampling distribution. We recommend these methods unless the sample is so small that it may not represent the population well. For an introduction, see Companion Chapter 16 of the somewhat more advanced text *Introduction to the Practice of Statistics*, available online at www.whfreeman.com/ips.
4. Finally, there are other **nonparametric methods**, which do not assume any specific form for the distribution of the population. Unlike bootstrap and permutation methods, common nonparametric methods do not make use of the actual values of the observations.

rank tests

This chapter concerns one type of nonparametric procedure: tests that can replace the t tests and one-way analysis of variance when the Normality conditions 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 25.1 presents an outline of the standard tests (based on Normal distributions) and the rank tests that compete with them. The rank tests require that the population or populations have *continuous distributions*. That is, each distribution must be described by a *density curve* (Chapter 3, page 69) that allows observations to take any value in some interval of outcomes. The Normal curves are one shape of density curve. Rank tests allow curves of any shape.

The rank tests we will study concern the *center* of a population or populations. When a population has at least roughly a Normal distribution, we describe its center by the mean. The “Normal tests” in Figure 25.1 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.

Setting	Normal test	Rank test
One sample	One-sample t test Chapter 17	Wilcoxon signed rank test
Matched pairs	Apply one-sample test to differences within pairs	
Two independent samples	Two-sample t test Chapter 18	Wilcoxon rank sum test
Several independent samples	One-way ANOVA F test Chapter 24	Kruskal-Wallis test

FIGURE 25.1

Comparison of tests based on Normal distributions with rank tests for similar settings.

We begin by describing the most common rank test, for comparing two samples. 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.

Comparing two samples: the Wilcoxon rank sum test

Two-sample problems (see Chapter 18) are among the most common in statistics. The most useful nonparametric significance test compares two distributions. Here is an example of this setting.

EXAMPLE 25.1 Weeds among the corn

STATE: 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

0 weeds per meter	166.7	172.2	165.0	176.9
3 weeds per meter	158.6	176.4	153.1	156.0

PLAN: Make a graph to compare the two sets of yields. Test the hypothesis that there is no difference against the one-sided alternative that yields are higher when no weeds are present.

SOLVE (first steps): A back-to-back stemplot (Figure 25.2) suggests that yields may be higher when there are no weeds. There is one outlier; because it is correct data, we cannot remove it. The samples are too small to rely on the robustness of the two-sample t test. We will now develop a test that does not require Normality. ■

0 weeds/meter		3 weeds/meter
	15	3
	15	6 9
	16	
7 5	16	
2	17	
7	17	6

First, arrange all 8 observations from both samples in order from smallest to largest:

153.1 156.0 158.6 165.0 166.7 172.2 176.4 176.9



FIGURE 25.2

Back-to-back stemplot of corn yields from plots with no weeds and with 3 weeds per meter of row, for Example 25.1. Notice the split stems, with leaves 0 to 4 on the first stem and leaves 5 to 9 on the second stem.

25-4 CHAPTER 25 • Nonparametric Tests

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*:

Yield	153.1	156.0	158.6	165.0	166.7	172.2	176.4	176.9
Rank	1	2	3	4	5	6	7	8

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 conditions on 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 without weeds to be larger as a group than the ranks from plots with weeds. Let's compare the *sums* of the ranks from the two treatments:

Treatment	Sum of ranks
No weeds	23
Weeds	13

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

$$\sigma_W = \sqrt{\frac{n_1 n_2 (N + 1)}{12}}$$

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.

In the corn yield study of Example 25.1, we want to test the hypotheses

H_0 : no difference in distribution of yields

H_a : yields are systematically higher in weed-free plots

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

EXAMPLE 25.2 Weeds among the corn: inference



SOLVE: First note that the conditions for the Wilcoxon test are met: the data come from a randomized comparative experiment and the yield of corn in bushels per acre has a continuous distribution.

There are $N = 8$ observations in all, with $n_1 = 4$ and $n_2 = 4$. The sum of ranks for the weed-free plots has mean

$$\begin{aligned}\mu_W &= \frac{n_1(N + 1)}{2} \\ &= \frac{(4)(9)}{2} = 18\end{aligned}$$

and standard deviation

$$\begin{aligned}\sigma_W &= \sqrt{\frac{n_1 n_2 (N + 1)}{12}} \\ &= \sqrt{\frac{(4)(4)(9)}{12}} = \sqrt{12} = 3.464\end{aligned}$$

Although the observed rank sum $W = 23$ is higher than the mean, it is only about 1.4 standard deviations higher. 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. Software tells us that this probability is $P = 0.1$.

CONCLUDE: The data provide some evidence ($P = 0.1$) that corn yields are lower when weeds are present. There are only 4 observations in each group, so even quite large effects can fail to reach the levels of significance usually considered convincing, such as $P < 0.05$. A larger experiment might clarify the effect of weeds on corn yield. ■

APPLY YOUR KNOWLEDGE

25.1 Daily activity and obesity. Our lead example for the two-sample t procedures in Chapter 18 concerned a study comparing the level of physical activity of lean and mildly obese people who don't exercise. Here are the minutes per day that the subjects spent standing or walking over a 10-day period:

Lean subjects		Obese subjects	
511.100	543.388	260.244	416.531
607.925	677.188	464.756	358.650
319.212	555.656	367.138	267.344
584.644	374.831	413.667	410.631
578.869	504.700	347.375	426.356

The data are a bit irregular but not distinctly non-Normal. Let's use the Wilcoxon test for comparison with the two-sample t test.

- Find the median minutes spent standing or walking for each group. Which group appears more active?
- Arrange all 20 observations in order and find the ranks.
- Take W to be the sum of the ranks for the lean group. What is the value of W ? If the null hypothesis (no difference between the groups) is true, what are the mean and standard deviation of W ?
- Does comparing W with the mean and standard deviation suggest that the lean subjects are more active than the obese subjects?

25.2 How strong are durable press fabrics? Exercise 18.38 (text page 496) describes an experiment comparing the strengths of cotton fabric treated with two "durable press" processes. Here are the breaking strengths in pounds:

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

There is a mild outlier in the Permafresh group. Perhaps we should use the Wilcoxon test.

- Arrange the breaking strengths in order and find their ranks.
- Find the Wilcoxon statistic W for the Permafresh group, along with its mean and standard deviation under the null hypothesis (no difference between the groups).
- Is W far enough from the mean to suggest that there may be a difference between the groups?

The Normal approximation for W

To calculate the P -value $P(W \geq 23)$ for Example 25.2, 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. 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 :

$$\begin{aligned} z &= \frac{W - \mu_W}{\sigma_W} \\ &= \frac{W - n_1(N+1)/2}{\sqrt{n_1 n_2 (N+1)/12}} \end{aligned}$$

Use standard Normal probability calculations to find P -values for this statistic. Because W takes only whole-number values, an idea called the *continuity correction* improves the accuracy of the approximation.

CONTINUITY CORRECTION

To apply the **continuity correction** in a Normal approximation for a variable that takes only whole-number values, act as if each whole number occupies the entire interval from 0.5 below the number to 0.5 above it.

EXAMPLE 25.3 Weeds among the corn: Normal approximation

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

$$z = \frac{W - \mu_W}{\sigma_W} = \frac{23 - 18}{3.464} = 1.44$$

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

$$P(Z \geq 1.44) = 0.0749$$

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:

$$\begin{aligned} P(W \geq 22.5) &= P\left(\frac{W - \mu_W}{\sigma_W} \geq \frac{22.5 - 18}{3.464}\right) \\ &= P(Z \geq 1.30) \\ &= 0.0968 \end{aligned}$$

This is close to the software value, $P = 0.1$. If you do not use the exact distribution of W (from software or tables), you should always use the continuity correction in calculating P -values. ■

25-8 CHAPTER 25 • Nonparametric Tests

APPLY YOUR KNOWLEDGE

25.3 Daily activity and obesity, continued. In Exercise 25.1, you found the Wilcoxon rank sum W and its mean and standard deviation. We want to test the null hypothesis that the two groups don't differ in activity against the alternative hypothesis that the lean subjects spend more time standing and walking.

- What is the probability expression for the P -value of W if we use the continuity correction?
- Find the P -value. What do you conclude?

25.4 Strength of durable press fabrics, continued. Use your values of W , μ_W , and σ_W from Exercise 25.2 to see whether fabrics treated with the two processes differ in breaking strength.

- The two-sided P -value is $2P(W \geq ?)$. Using the continuity correction, what number replaces the $?$ in this probability?
- Find the P -value. What do you conclude?



25.5 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:²



Ariel Skelly/CORBIS

Child	Progress	Story 1 score	Story 2 score
1	high	0.55	0.80
2	high	0.57	0.82
3	high	0.72	0.54
4	high	0.70	0.79
5	high	0.84	0.89
6	low	0.40	0.77
7	low	0.72	0.49
8	low	0.00	0.66
9	low	0.36	0.28
10	low	0.55	0.38

Look only at the data for Story 2. Is there good evidence that high-progress readers score higher than low-progress readers? Follow the four-step process as illustrated in Examples 25.1 and 25.2.

Using technology

For samples as small as those in the corn yield study of Example 25.1, we prefer software that gives the exact P -value for the Wilcoxon test rather than the Normal approximation. Neither the Excel spreadsheet nor TI graphing calculators have menu entries for rank tests. Minitab offers only the Normal approximation.

EXAMPLE 25.4 Weeds among the corn: software output

Figure 25.3 displays output from CrunchIt! for the corn yield data. The top panel reports the exact Wilcoxon P -value as $P = 0.1$. The Normal approximation with continuity correction, $P = 0.0968$ in Example 25.3, is quite accurate. There are several differences between the CrunchIt! output and our work in Example 25.3. The most important is that CrunchIt! carries out the **Mann-Whitney test** rather than the Wilcoxon test. The two tests always have the same P -value because the two test statistics are related by simple algebra.

Mann-Whitney test

The second panel in Figure 25.3 is the two-sample t test from Chapter 18, which does not assume that the two populations have the same standard deviation. It gives $P = 0.0937$, close to the Wilcoxon value. Because the t test is quite robust, it is somewhat unusual for P -values from t and W to differ greatly.

The bottom panel shows the result of the “pooled” version of t , now outdated, that assumes equal population standard deviations. You see that its P is a bit different from the others, another reminder that you should never use this test. ■

APPLY YOUR KNOWLEDGE

- 25.6 Strength of durable press fabrics: software.** Use your software to repeat the Wilcoxon test you did in Exercise 25.4. By comparing the results, state how your software finds P -values for W : exact distribution, Normal approximation with continuity correction, or Normal approximation without continuity correction.
- 25.7 Daily activity and obesity: software.** Use your software to carry out the one-sided Wilcoxon rank sum test that you did by hand in Exercise 25.3. Use the exact distribution if your software will do it. Compare the software result with your result in Exercise 25.3.
- 25.8 Weeds among the corn.** The corn yield study of Example 25.1 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

162.8 142.4 162.7 162.4

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 t procedures because \bar{x} and s are not resistant.

- 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 25.1 to answer this question.
- Compare the results from (a) with those from the two-sample t test for these data.
- Now remove the low outlier 142.4 from the data with 9 weeds per meter. Repeat both the Wilcoxon and t analyses. By how much did the outlier reduce the mean yield in its group? By how much did it increase the standard deviation? Did it have a practically important impact on your conclusions?

25-10 CHAPTER 25 • Nonparametric Tests

Mann-Whitney						
Hypothesis test results:						
m1 = median of weeds0						
m2 = median of weeds3						
Parameter : m1 - m2						
H_0 : Parameter = 0						
H_A : Parameter > 0						
Difference	n1	n2	Diff. Est.	Test Stat.	P-value	Method
m1 - m2	4	4	11.3	23	0.1	Exact

Two sample T statistics					
Hypothesis test results:					
μ_1 : mean of weeds0					
μ_2 : mean of weeds3					
$\mu_1 - \mu_2$: mean difference					
H_0 : $\mu_1 - \mu_2 = 0$					
H_A : $\mu_1 - \mu_2 > 0$					
(without pooled variances)					
Difference	Sample Mean	Std. Err.	DF	T-Stat	P-value
$\mu_1 - \mu_2$	9.175	5.9055586	4.4951386	1.5536209	0.0937

Two sample T statistics					
Hypothesis test results:					
μ_1 : mean of weeds0					
μ_2 : mean of weeds3					
$\mu_1 - \mu_2$: mean difference					
H_0 : $\mu_1 - \mu_2 = 0$					
H_A : $\mu_1 - \mu_2 > 0$					
(with pooled variances)					
Difference	Sample Mean	Std. Err.	DF	T-Stat	P-value
$\mu_1 - \mu_2$	9.175	5.9055586	6	1.5536209	0.0856

FIGURE 25.3

Output from CrunchIt! for the data of Example 25.1. The output compares the results of three tests that could be used to compare yields for the two groups of corn plots.

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 can use the two-sample t test for means. Our hypotheses then have the form

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ H_a &: \mu_1 > \mu_2 \end{aligned}$$

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

$$\begin{aligned} H_0 &: \text{median}_1 = \text{median}_2 \\ H_a &: \text{median}_1 > \text{median}_2 \end{aligned}$$

The Wilcoxon rank sum test provides a test of 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 CrunchIt! output in the top panel of Figure 25.3 states the hypotheses in terms of population medians. CrunchIt! 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 more useful setting. It compares any two continuous distributions, whether or not they have the same shape, by testing hypotheses that we can state in words as

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

A more exact statement of the “systematically larger” alternative hypothesis is a bit tricky, so we won’t try to give it here.³ These hypotheses really are “nonparametric” 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 condition.* 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.

Why don’t we discuss the confidence intervals for the difference in population medians that software such as CrunchIt! offers? These intervals require the unrealistic same-shape condition. The more general “systematically larger” hypothesis does not involve a specific parameter, so there is no accompanying confidence interval.



APPLY YOUR KNOWLEDGE

- 25.9 Daily activity and obesity: hypotheses.** We could use either two-sample t or the Wilcoxon rank sum to test the null hypothesis that lean and mildly obese people don't differ in the time they spend standing and walking against the alternative hypothesis that lean people generally spend more time in these activities. Explain carefully what H_0 and H_a are for t and for W .
- 25.10 Strength of durable press fabrics: hypotheses.** We are interested in whether fabrics treated with the Permafresh and Hylite processes have the same breaking strength "on the average."
- 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?
 - 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

average ranks

We have chosen our examples and exercises to this point rather carefully: they all involve data in which *no two values are the same*. This allowed us to rank all the values. 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:

Observation	153	155	158	158	161	164
Rank	1	2	3.5	3.5	5	6

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 σ_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 to use rank tests when the data contain tied values.*



Some data have many ties because the scale of measurement has only a few values. Rank tests are often used for such data. Here is an example.



EXAMPLE 25.5 Food safety at fairs

STATE: 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: "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.⁴ We suspect that women are more concerned than men about food safety. Is there good evidence for this conclusion?

PLAN: Do data analysis to understand the difference between women and men. Check the conditions required by the Wilcoxon test. If the conditions are met, use the Wilcoxon test for the hypotheses

- H_0 : men and women do not differ in their responses
- H_a : women give systematically higher responses than men

SOLVE: The responses for the 303 subjects appear in the file `eg25-05.dat` on the text CD and Web site. We can summarize them in a two-way table of counts:

	Response					Total
	1	2	3	4	5	
Female	13	108	50	23	2	196
Male	22	57	22	5	1	107
Total	35	165	72	28	3	303

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

	Response					Total
	1	2	3	4	5	
Percent of females	6.6	55.1	25.5	11.7	1.0	100
Percent of males	20.6	53.3	20.6	4.7	1.0	100

Are these differences between women and men statistically significant?

The most important condition for inference is that the subjects 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.) The Wilcoxon test also requires that responses have *continuous distributions*. We think that the subjects really have a continuous distribution of opinions about how often people become sick from food at fairs. The questionnaire asks them to round off their opinions to the nearest value in the five-point scale. So we are willing to use the Wilcoxon test.

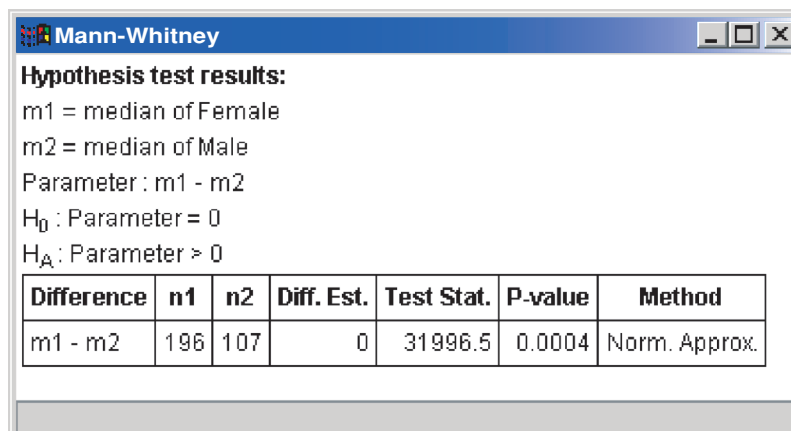


Danny Lehman/CORBIS

25-14 CHAPTER 25 • Nonparametric Tests

FIGURE 25.4

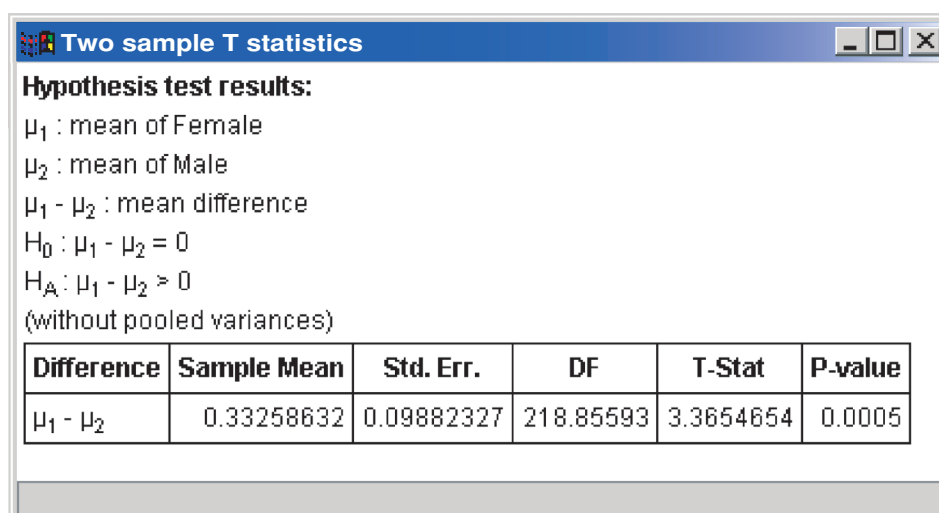
Output from CrunchIt! for the data of Example 25.5. The Wilcoxon rank sum test and the two-sample t test give similar results.



Mann-Whitney

Hypothesis test results:
 m1 = median of Female
 m2 = median of Male
 Parameter : m1 - m2
 H_0 : Parameter = 0
 H_A : Parameter > 0

Difference	n1	n2	Diff. Est.	Test Stat.	P-value	Method
m1 - m2	196	107	0	31996.5	0.0004	Norm. Approx.



Two sample T statistics

Hypothesis test results:
 μ_1 : mean of Female
 μ_2 : mean of Male
 $\mu_1 - \mu_2$: mean difference
 H_0 : $\mu_1 - \mu_2 = 0$
 H_A : $\mu_1 - \mu_2 > 0$
 (without pooled variances)

Difference	Sample Mean	Std. Err.	DF	T-Stat	P-value
$\mu_1 - \mu_2$	0.33258632	0.09882327	218.85593	3.3654654	0.0005

Because the responses can take only five values, there are 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. Figure 25.4 gives output from CrunchIt! The Wilcoxon (reported as Mann-Whitney) test for the one-sided alternative that women are more concerned about food safety at fairs is highly significant ($P = 0.0004$).

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 25.4 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.

CONCLUDE: There is very strong evidence ($P = 0.0004$) that women are more concerned than men about the safety of food served at fairs. ■

As is often the case, t and W for the data in Example 25.5 agree closely. 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 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.*



Because we have a two-way table, we might have applied the chi-square test (Chapter 22), which asks if there is a significant relationship of *any kind* between gender and response. The chi-square test ignores the ordering of the responses and so doesn't tell us 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.

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.

25.11 Does polyester decay? Exercise 18.8 (text page 482) compares the breaking strength of polyester strips buried for 16 weeks with that of strips buried for 2 weeks. The breaking strengths in pounds are

2 weeks	118	126	126	120	129
16 weeks	124	98	110	140	110

- What are the null and alternative hypotheses for the Wilcoxon test? For the two-sample t test?
- There are two pairs of tied observations. What ranks do you assign to each observation, using average ranks for ties?
- Apply the Wilcoxon rank sum test to these data. Compare your result with the $P = 0.1857$ obtained from the two-sample t test in Figure 18.5.

25.12 Do birds learn to time their breeding? Exercises 18.42 to 18.44 (text pages 497–498) concern a study of whether supplementing the diet of blue titmice with extra caterpillars will prevent them from adjusting their breeding date the following year in search of a better food supply. Here are the data (days after the caterpillar peak):

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

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

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

25-16 CHAPTER 25 • Nonparametric Tests

- (b) Take W to be the rank sum for the supplemented group. What is the value of W ?
- (c) Use software: find the P -value of the Wilcoxon test and state your conclusion.

25.13 Tell me a story, continued. The data in Exercise 25.5 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?



25.14 Do good smells bring good business? Exercise 18.9 (text page 483) describes an experiment that asked whether background aromas in a restaurant encourage customers to stay longer and spend more. The data on amount spent (in euros) are as follows:

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

Examine the data and comment on departures from Normality. Is there significant evidence that the lavender odor encourages customers to spend more? Follow the four-step process.



25.15 Cicadas as fertilizer? Exercise 7.41 (text page 193) gives data from an experiment in which some bellflower plants in a forest were “fertilized” with dead cicadas and other plants were not disturbed. The data record the mass of seeds produced by 39 cicada plants and 33 undisturbed (control) plants. Do the data show that dead cicadas increase seed mass? Do data analysis to compare the two groups, explain why you would be reluctant to use the two-sample t test, and apply the Wilcoxon test. Follow the four-step process in your report.



25.16 Food safety in restaurants. Example 25.5 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 *ex25-16.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 F if the respondent is a woman, M 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? Follow the four-step process in your answer.

25.17 More on food safety. The data file used in Exercise 25.16 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 (Chapter 17) 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 25.6 Tell me a story

STATE: 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:



	Child				
	1	2	3	4	5
Story 2	0.77	0.49	0.66	0.28	0.38
Story 1	0.40	0.72	0.00	0.36	0.55
Difference	0.37	-0.23	0.66	-0.08	-0.17

We wonder if illustrations improve how the children retell a story.

PLAN: We would like to test the hypotheses

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

$$H_a: \text{scores are systematically higher for Story 2}$$

SOLVE (first steps): 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. ■

25-18 CHAPTER 25 • Nonparametric Tests

absolute value

Positive differences in Example 25.6 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 0.23 **0.66** 0.08 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.

Absolute value	0.08	0.17	0.23	0.37	0.66
Rank	1	2	3	4	5

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 25.7 Tell me a story, continued

SOLVE: In the storytelling study of Example 25.6, $n = 5$. 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

$$\begin{aligned}\sigma_{W^+} &= \sqrt{\frac{n(n+1)(2n+1)}{24}} \\ &= \sqrt{\frac{(5)(6)(11)}{24}} \\ &= \sqrt{13.75} = 3.708\end{aligned}$$

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

The P -value for our one-sided alternative is $P(W^+ \geq 9)$, calculated using the distribution of W^+ when the null hypothesis is true. Software gives the P -value $P = 0.4063$.

CONCLUDE: The data give no evidence ($P = 0.4$) that scores are higher for Story 2. The data do show an effect, but it fails to be significant because the sample is very small. ■

APPLY YOUR KNOWLEDGE

25.18 Growing trees faster. Exercise 17.37 (text page 465) 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:

Pair	Control plot	Treated plot
1	9.752	10.587
2	7.263	9.244
3	5.742	8.675

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.

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

25.19 Fighting cancer. Lymphocytes (white blood cells) play an important role in defending our bodies against tumors and infections. Can lymphocytes be genetically modified to recognize and destroy cancer cells? In one study of this idea, modified cells were infused into 11 patients with metastatic melanoma (serious skin cancer) that had not responded to existing treatments. Here are data for an "ELISA" test for the presence of cells that trigger an immune response, in counts per 100,000 cells before and after infusion.⁵ High counts suggest that infusion had a beneficial effect.

25-20 CHAPTER 25 • Nonparametric Tests

Patient	1	2	3	4	5	6	7	8	9	10	11
Pre	14	0	1	0	0	0	0	20	1	6	0
Post	41	7	1	215	20	700	13	530	35	92	108

- Examine the differences (post minus pre). Why can't we use the matched pairs t test to see if infusion raised the ELISA counts?
- We will apply the Wilcoxon signed rank test. What are the ranks for the absolute values of the differences in counts? What is the value of W^+ ?
- What would be the mean and standard deviation of W^+ if the null hypothesis (infusion 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 25.8 Tell me a story: Normal approximation

For $n = 5$ observations, we saw in Example 25.7 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 either from software or by standardizing and using the standard Normal table:

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

Figure 25.5 displays the output of two statistical programs. Minitab uses the Normal approximation and agrees with our calculation $P = 0.394$. We asked CrunchIt! to do two analyses: using the exact distribution of W^+ and using the matched pairs t test. The exact one-sided P -value for the Wilcoxon signed rank test is $P = 0.4063$, as we reported in Example 25.7. The Normal approximation is quite close to this. The t test result is a bit different, $P = 0.28$, but all three tests tell us that this very small sample gives no evidence that seeing illustrations improves the storytelling of low-progress readers.

● The Normal approximation for W^+ 25-21

Minitab

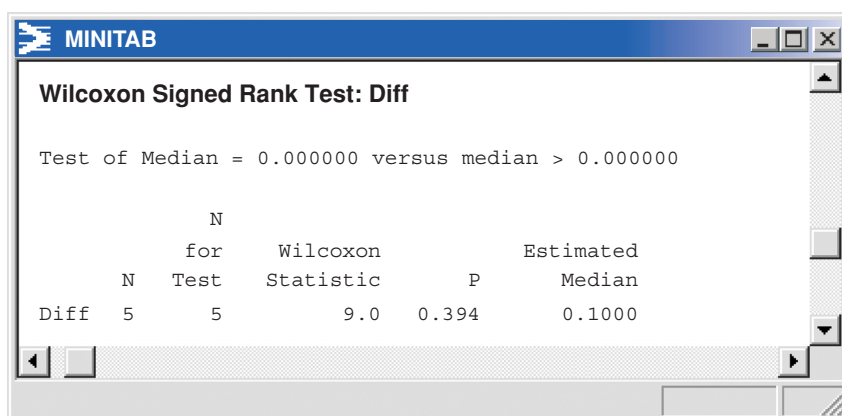
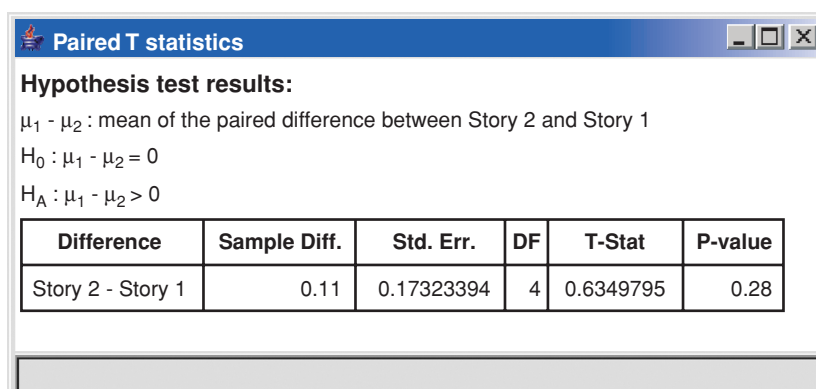
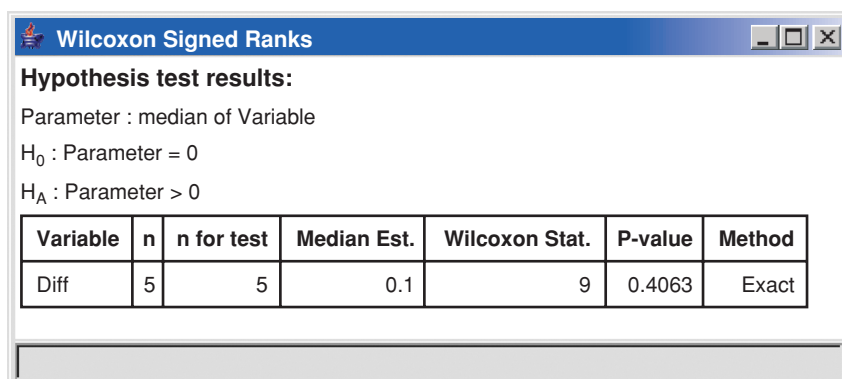


FIGURE 25.5

Output from Minitab and CrunchIt! for the storytelling data of Example 25.6. The CrunchIt! output compares the Wilcoxon signed rank test (with the exact distribution) and the matched pairs t test.

CrunchIt!



APPLY YOUR KNOWLEDGE

25.20 Growing trees faster: Normal approximation. Continue your work from Exercise 25.18. 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?

25-22 CHAPTER 25 • Nonparametric Tests

25.21 W^+ versus t . Find the one-sided P -value for the matched pairs t test applied to the tree growth data in Exercise 25.18. The smaller P -value of t relative to W^+ means that t gives stronger evidence of the effect 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.

25.22 Fighting cancer: Normal approximation. Use the Normal approximation with continuity correction to find the P -value for the test in Exercise 25.19. What do you conclude about the effect of infusing modified cells on the ELISA count?

25.23 Ancient air. Exercise 17.7 (text page 449) 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.

- Graph the data, and comment on skewness and outliers. A rank test is appropriate.
- We would like to test hypotheses about the median percent of nitrogen in ancient air (the population):

$$H_0 : \text{median} = 78.1$$

$$H_a : \text{median} \neq 78.1$$

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?



David Sanger Photography/Alamy

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 σ_{W^+} must be adjusted for the ties before we can use the Normal approximation. Software will do this. Here is an example.



John Cumming/Digital Vision/Getty Images

EXAMPLE 25.9 Golf scores

STATE: 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.)



● Dealing with ties in the signed rank test 25-23

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

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

PLAN: We would like to test the hypotheses that in a tournament play

H_0 : scores have the same distribution in Rounds 1 and 2

H_a : scores are systematically lower or higher in Round 2

SOLVE: A stemplot of the differences (Figure 25.6) shows some irregularity and a low outlier. We will use the Wilcoxon signed rank test.

Figure 25.7 displays CrunchIt! output for the golf score data. The Wilcoxon statistic is $W^+ = 50.5$ with two-sided P -value $P = 0.3843$. The output also includes the matched pairs t test, for which $P = 0.3716$. The two P -values are once again similar.

CONCLUDE: These data give no evidence for a systematic change in scores between rounds. ■

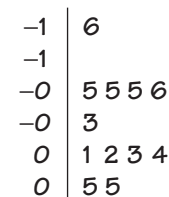


FIGURE 25.6

Stemplot (with split stems) of the differences in scores for two rounds of a golf tournament, for Example 25.9.

Wilcoxon Signed Ranks						
Hypothesis test results:						
Parameter : median of Variable						
H_0 : Parameter = 0						
H_A : Parameter \neq 0						
Variable	n	n for test	Median Est.	Wilcoxon Stat.	P-value	Method
Round 2-Round 1	12	12	1	50.5	0.3843	Norm. Approx.

FIGURE 25.7

Output from CrunchIt! for the golf scores data of Example 25.9. Because there are ties, a Normal approximation must be used for the Wilcoxon signed rank test.

Paired T statistics					
Hypothesis test results:					
$\mu_1 - \mu_2$: mean of the paired difference between Round 2 and Round 1					
H_0 : $\mu_1 - \mu_2 = 0$					
H_A : $\mu_1 - \mu_2 \neq 0$					
Difference	Sample Diff.	Std. Err.	DF	T-Stat	P-value
Round 2 - Round 1	1.66666666	1.7894189	11	0.931401	0.3716

25-24 CHAPTER 25 • Nonparametric Tests

Let's see where the value $W^+ = 50.5$ came from. The absolute values of the differences, with boldface indicating those that were negative, are

5 5 2 6 5 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.

Absolute value	1	2	3	3	4	5	5	5	5	5	6	16
Rank	1	2	3.5	3.5	5	8	8	8	8	8	11	12

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

APPLY YOUR KNOWLEDGE

25.24 Does nature heal best? Exercise 17.33 (text page 464) gives these data on the healing rate (micrometers per hour) for cuts in the hind limbs of 12 newts:

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

The electrical field in the experimental limbs was reduced to zero by applying a voltage. The control limbs were not treated, so that they had their natural electrical field. The paired differences include an outlier, so we may choose to use the Wilcoxon signed rank test.

- Find the ranks and give the value of the test statistic W^+ .
- 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.

25.25 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?

- 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.
- The output in Figure 17.6 (text page 454) showed that the one-sample t test had P -value $P = 0.0123$ for these data. How does this compare with your result from (a)? What are the hypotheses for the t test? What conditions must be met for each of the t and Wilcoxon tests?

● Comparing several samples: the Kruskal-Wallis test 25-25

25.26 Fungus in the air. The air in poultry-processing plants often contains fungus spores. Inadequate ventilation can damage the health of the workers. The problem is most serious during the summer. To measure the presence of spores, air samples are pumped to an agar plate, and “colony-forming units (CFUs)” are counted after an incubation period. Here are data from two locations in a plant that processes 37,000 turkeys per day, taken on four days in the summer. The units are CFUs per cubic meter of air.⁶



	Day			
	1	2	3	4
Kill room	3175	2526	1763	1090
Processing	529	141	362	224

Spore counts are clearly much higher in the kill room, but with only 4 pairs of observations, the difference may not be statistically significant. Apply a rank test.

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 mean responses for 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 25.10 Weeds among the corn

STATE: 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:⁷

Weeds per meter	Corn yield	Weeds per meter	Corn yield	Weeds per meter	Corn yield	Weeds per meter	Corn yield
0	166.7	1	166.2	3	158.6	9	162.8
0	172.2	1	157.3	3	176.4	9	142.4
0	165.0	1	166.7	3	153.1	9	162.7
0	176.9	1	161.1	3	156.0	9	162.4

Do yields change as the presence of weeds changes?

PLAN: Do data analysis to see how the yields change. Test the null hypothesis “no difference in the distribution of yields” against the alternative that the groups do differ.



SOLVE (first steps): The summary statistics are

Weeds	<i>n</i>	Median	Mean	Std. dev.
0	4	169.45	170.200	5.422
1	4	163.65	162.825	4.469
3	4	157.30	161.025	10.493
9	4	162.55	157.575	10.118

The mean yields do go down as more weeds are added. ANOVA tests whether the differences are statistically significant. Can we safely use ANOVA? Outliers are present in the yields for 3 and 9 weeds per meter. The outliers explain the differences between the means and the medians. They are the correct yields for their plots, so we cannot remove them. Moreover, the sample standard deviations do not quite satisfy our rule of thumb for ANOVA that the largest should not exceed twice the smallest. We may prefer 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 25.10, 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, μ_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 different standard deviations suggest that the four distributions in Example 25.10 do *not* all have the same shape.

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, \dots, 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 25.11 Weeds among the corn, continued

SOLVE (inference): In Example 25.10, 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



25-28 CHAPTER 25 • Nonparametric Tests

Yield	142.4	153.1	156.0	157.3	158.6	161.1	162.4	162.7
Rank	1	2	3	4	5	6	7	8
Yield	162.8	165.0	166.2	166.7	166.7	172.2	176.4	176.9
Rank	9	10	11	12.5	12.5	14	15	16

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

Weeds	Ranks					Sum of ranks
0	10	12.5	14	16		52.5
1	4	6	11	12.5		33.5
3	2	3	5	15		25.0
9	1	7	8	9		25.0

The Kruskal-Wallis statistic is therefore

$$\begin{aligned}
 H &= \frac{12}{N(N+1)} \sum \frac{R_i^2}{n_i} - 3(N+1) \\
 &= \frac{12}{(16)(17)} \left(\frac{52.5^2}{4} + \frac{33.5^2}{4} + \frac{25^2}{4} + \frac{25^2}{4} \right) - (3)(17) \\
 &= \frac{12}{272}(1282.125) - 51 \\
 &= 5.56
 \end{aligned}$$

Referring to the table of chi-square critical points (Table D) with $df = 3$, we see that the P -value lies in the interval $0.10 < P < 0.15$.

CONCLUDE: Although this small experiment suggests that more weeds decrease yield, it does not provide convincing evidence that weeds have an effect. ■

Figure 25.8 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 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 very lengthy computer calculation 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.

APPLY YOUR KNOWLEDGE

25.27 More rain for California? Exercise 24.30 describes an experiment that examines the effect on plant biomass in plots of California grassland randomly assigned to receive added water in the winter, added water in the spring, or no added water.

● The Kruskal-Wallis test statistic 25-29

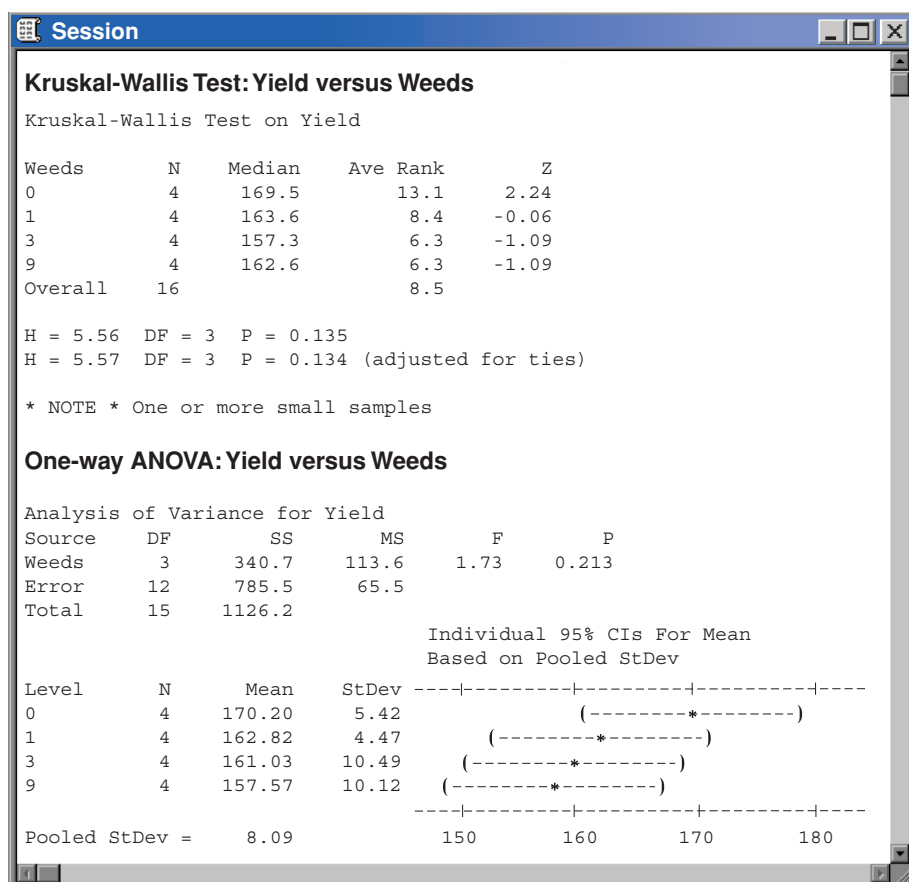


FIGURE 25.8

Minitab output for the corn yield data of Example 25.10. For comparison, both the Kruskal-Wallis test and one-way ANOVA are shown.

The experiment continued for several years. Here are data for 2004 (mass in grams per square meter):

	Winter	Spring	Control
	254.6453	517.6650	178.9988
	233.8155	342.2825	205.5165
	253.4506	270.5785	242.6795
	228.5882	212.5324	231.7639
	158.6675	213.9879	134.9847
	212.3232	240.1927	212.4862

The sample sizes are small and the data contain some possible outliers. We will apply a nonparametric test.

- Examine the data. Show that the conditions for ANOVA (text page 644) are not met. What appear to be the effects of extra rain in winter or spring?
- What hypotheses does ANOVA test? What hypotheses does Kruskal-Wallis test?
- What are I , the n_i , and N ? Arrange the counts in order and assign ranks.

25-30 CHAPTER 25 • Nonparametric Tests

- (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?

25.28 Logging in the rain forest: species richness. Table 24.2 (text page 640) contains data comparing the number of trees and number of tree species in plots of land in a tropical rain forest 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 and 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.



25.29 Does polyester decay? Here are the breaking strengths (in pounds) of strips of polyester fabric buried in the ground for several lengths of time:⁸

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. Do a complete analysis that compares the four groups. Give the Kruskal-Wallis test along with a statement in words of the null and alternative hypotheses.



25.30 Compressing soil. Farmers know that driving heavy equipment on wet soil compresses the soil and injures future crops. Table 2.5 (text page 65) gives data on the “penetrability” of the same soil at three levels of compression. Penetrability is a measure of how much resistance plant roots will meet when they try to grow through the soil. Does penetrability systematically change with the degree of compression? Do a complete analysis that includes a test of significance. Include a statement in words of your null and alternative hypotheses.

25.31 Food safety. Example 25.5 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 *ex25-16.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 25.5. 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 25 SUMMARY

- **Nonparametric tests** do not require any specific form for the distributions of the populations 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. Use rank tests when the data come from random samples or randomized comparative experiments and the populations have continuous distributions.
- 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**. Software may perform the **Mann-Whitney test**, another form of the Wilcoxon 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.
- 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, the Kruskal-Wallis test statistic 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.

S T A T I S T I C S I N S U M M A R Y

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

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 for H .

C H E C K Y O U R S K I L L S

25.32 A study of “road rage” gives randomly selected drivers a test that measures “angry/threatening driving.” You wonder if the scores go down with age. You compare the scores for three age groups: less than 30 years, 30 to 55 years, and over 55 years. You use the

- (a) Wilcoxon rank sum test.
- (b) Wilcoxon signed rank test.
- (c) Kruskal-Wallis test.

25.33 You interview college students who have done community service and another group of students who have not. To compare the scores of the two groups on a test of attitude toward people of other races, you use the

- (a) Wilcoxon rank sum test.
- (b) Wilcoxon signed rank test.
- (c) Kruskal-Wallis test.

Check Your Skills 25-33

- 25.34** You interview 75 students in their freshman year and again in their senior year. Each interview includes a test of knowledge of world affairs. To assess whether there has been a significant change from freshman to senior year, you use the
- Wilcoxon rank sum test.
 - Wilcoxon signed rank test.
 - Kruskal-Wallis test.

- 25.35** When some plants are attacked by leaf-eating insects, they release chemical compounds that repel the insects. Here are data on emissions of one compound by plants attacked by leaf bugs and by plants in an undamaged control group:

Control group	14.4	15.2	12.6	11.9	5.1	8.0
Attacked group	10.6	15.3	25.2	19.8	17.1	14.6

The rank sum W for the control group is

- 21.
 - 26.
 - 52.
- 25.36** If there is no difference in emissions between the attacked group and the control group, the mean of W in the previous exercise is
- 39.
 - 78.
 - 6.2.
- 25.37** Suppose that the 12 observations in Exercise 25.35 were

Control group	14.4	15.2	12.6	11.9	5.1	8.0
Attacked group	12.6	15.3	25.2	19.8	17.1	14.4

The rank sum for the control group is now

- 21.
 - 25.
 - 26.
- 25.38** Interview 10 young married couples, wife and husband separately. One question asks how important the attractiveness of their spouse is to them on a scale of 1 to 10. Here are the responses:

	Couple									
	1	2	3	4	5	6	7	8	9	10
Husband	7	7	7	3	9	5	10	6	6	7
Wife	4	2	5	2	2	2	4	7	1	5

The Wilcoxon signed rank statistic W^+ (based on husband's score minus wife's score) is

- 51.
 - 53.5.
 - 54.
- 25.39** If husbands and wives don't differ in how important the attractiveness of their spouse is, the mean of W^+ in the previous exercise is
- 27.5.
 - 55.
 - 105.

25-34 CHAPTER 25 • Nonparametric Tests

25.40 Suppose that the responses in Exercise 25.38 are

	Couple									
	1	2	3	4	5	6	7	8	9	10
Husband	7	7	7	3	9	5	10	6	6	5
Wife	4	2	5	3	2	2	4	7	1	5

The Wilcoxon signed rank statistic W^+ (based on husband's score minus wife's score) is now

- (a) 35. (b) 36. (c) 52.

25.41 You compare the incomes of 4 college freshmen, 5 sophomores, 6 juniors, and 7 seniors. If the four income distributions are the same, the Kruskal-Wallis statistic H has approximately a chi-square distribution. The degrees of freedom are

- (a) 3. (b) 4. (c) 18.

CHAPTER 25 EXERCISES

One of the rank tests discussed in this chapter is appropriate for each of the following exercises. Follow the **Plan**, **Solve**, and **Conclude** parts of the four-step process in your answers.



25.42 **Each day I am getting better in math.** Table 18.3 (text page 499) 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?



25.43 **Which blue is most blue?** The color of a fabric depends on the dye used and also on how the dye is applied. This matters to clothing manufacturers, who want the color of the fabric to be just right. Dye fabric made of ramie with the same “procion blue” die applied in four different ways. Then use a colorimeter to measure the lightness of the color on a scale in which black is 0 and white is 100. Here are the data for 8 pieces of fabric dyed in each way:⁹

Method A	41.72	41.83	42.05	41.44	41.27	42.27	41.12	41.49
Method B	40.98	40.88	41.30	41.28	41.66	41.50	41.39	41.27
Method C	42.30	42.20	42.65	42.43	42.50	42.28	43.13	42.45
Method D	41.68	41.65	42.30	42.04	42.25	41.99	41.72	41.97

Do the methods differ in color lightness?



25.44 **Right versus left.** Table 17.5 (text page 469) contains data from a student project that investigated whether right-handed people can turn a knob faster clockwise than they can counterclockwise. We expect that right-handed people work more quickly when they turn the knob clockwise.



25.45 **Logging in the rain forest.** Investigators compared the number of tree species in unlogged plots in the rain forest of Borneo with the number of species in plots logged 8 years earlier. Here are the data:¹⁰

Unlogged	22	18	22	20	15	21	13	13	19	13	19	15
Logged	17	4	18	14	18	15	15	10	12			

Does logging significantly reduce the number of species in a plot after 8 years?

25.46 Food safety at fairs and restaurants. Example 25.5 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 *ex25-16.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 25.5. 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?

25.47 Food safety at fairs and fast-food restaurants. The food safety survey data described in Example 25.5 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?

25.48 Nematodes and plant growth. A botanist prepares 16 identical planting pots and then introduces different numbers of nematodes (microscopic worms) 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:¹¹

Nematodes	Seedling growth			
0	10.8	9.1	13.5	9.2
1,000	11.1	11.1	8.2	11.3
5,000	5.4	4.6	7.4	5.0
10,000	5.8	5.3	3.2	7.5

Do nematodes in soil affect plant growth?

25.49 Mutual fund performance. Mutual funds often compare their performance with a benchmark provided by an “index” that describes the performance of the class of assets in which the funds invest. For example, the Vanguard International Growth Fund benchmarks its performance against the EAFE (Europe, Australasia, Far East) index. Table 17.4 (text page 468) gives the annual returns (percent) for the fund and the index. Does the fund’s performance differ significantly from that of its benchmark?

How does the meeting of large rivers influence the diversity of fish? A study of the Amazon and 13 of its major tributaries concentrated on electric fish, which are common in South America. The researchers trawled in more than 1000 locations in the Amazon above and below each tributary and in the lower part of the tributaries themselves. In all, they found 43 species of electric fish. These distinctive fish can “stand in” for fish in general, which are too numerous to count easily. The researchers concluded that the number of fish species increases when a tributary joins the Amazon, but that the effect is local: there is no steady increase in diversity as we move downstream. Table 25.1 gives the estimated number of electric fish species in the Amazon upstream and downstream from each tributary and in the tributaries themselves just before they flow into the Amazon.¹² The researchers used nonparametric tests to assess the statistical significance of their results. Exercises 25.50 to 25.52 quote conclusions from the study.



TABLE 25.1 Electric fish species in the Amazon

Tributary	Species Counts		
	Upstream	Tributary	Downstream
Içá	14	23	19
Jutaí	11	15	18
Juruá	8	13	8
Japurá	9	16	11
Coari	5	7	7
Purus	10	23	16
Manacapuru	5	8	6
Negro	23	26	24
Madeira	29	24	30
Trombetas	19	20	16
Tapajós	16	5	20
Xingu	25	24	21
Tocantins	10	12	12

25.50 Downstream versus upstream. “We identified a significant positive effect of tributaries on Amazon mainstem species richness in two respects. First, we found that sample stations downstream of each tributary contained more species than did their respective upstream stations.” Do a test to confirm the statistical significance of this effect and report your conclusion.

25.51 Tributary versus upstream. “Second, we found that species richness within tributaries exceeded that within their adjacent upstream mainstem stations.” Again, do a test to confirm significance and report your finding.

25.52 Tributary versus downstream. Species richness “was comparable between tributaries and their adjacent downstream mainstem stations.” Verify this conclusion by comparing tributary and downstream species counts.

NOTES AND DATA SOURCES

1. Data provided by Samuel Phillips, Purdue University.
2. Data provided by Susan Stadler, Purdue University.
3. The precise meaning of “yields are systematically larger in plots with no weeds” is that for every fixed value a , the probability that the yield with no weeds is larger than a is at least as great as the same probability for the yield with weeds.
4. Huey Chern Boo, “Consumers’ perceptions and concerns about safety and healthfulness of food served at fairs and festivals,” MS thesis, Purdue University, 1997.
5. Richard A. Morgan et al., “Cancer regression in patients after transfer of genetically engineered lymphocytes,” *Science*, 314 (2006), pp. 126–129. The data appear in the Online Supplementary Material.

Notes and Data Sources 25-37

6. Michael W. Peugh, "Field investigation of ventilation and air quality in duck and turkey slaughter plants," MS thesis, Purdue University, 1996.
7. See Note 1.
8. Sapna Aneja, "Biodeterioration of textile fibers in soil," MS thesis, Purdue University, 1994.
9. Yvan R. Germain, "The dyeing of ramie with fiber reactive dyes using the cold pad-batch method," MS thesis, Purdue University, 1988.
10. I thank Charles Cannon of Duke University for providing the data. The study report is C. H. Cannon, D. R. Peart, and M. Leighton, "Tree species diversity in commercially logged Bornean rainforest," *Science*, 281 (1998), pp. 1366–1367.
11. Data provided by Matthew Moore.
12. Cristina Cox Fernandes, Jeffrey Podos, and John G. Lundberg, "Amazonian ecology: tributaries enhance the diversity of electric fishes," *Science*, 305 (2004), pp. 1960–1962.

P1: OSO
FREE013-25 FREE013-Moore

August 25, 2008 16:4