



Nigel Cattlin/Science Source

## CHAPTER

# 28

## 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 20, 21, and 27.

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 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



### In this chapter, we cover ...

- 28.1 Comparing two samples: The Wilcoxon rank sum test
- 28.2 The Normal approximation for  $W$
- 28.3 Examples of technology
- 28.4 What hypotheses does Wilcoxon test?
- 28.5 Dealing with ties in rank tests
- 28.6 Matched pairs: The Wilcoxon signed rank test
- 28.7 The Normal approximation for  $W^+$
- 28.8 Dealing with ties in the signed rank test
- 28.9 Comparing several samples: The Kruskal–Wallis test
- 28.10 Hypotheses and conditions for the Kruskal–Wallis test
- 28.11 The Kruskal–Wallis test statistic

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 the online Chapter 32.
- 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 28.1 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 77) 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.

**FIGURE 28.1**  
Comparison of tests based on Normal distributions with rank tests for similar settings.

Setting	Normal test	Rank test
One sample	One-sample <i>t</i> test Chapter 20	Wilcoxon signed rank test
Matched pairs	Apply one-sample test to differences within pairs	
Two independent samples	Two-sample <i>t</i> test Chapter 21	Wilcoxon rank sum test
Several independent samples	One-way ANOVA <i>F</i> test Chapter 27	Kruskal–Wallis test

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 28.1 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 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.

**28.1 Comparing Two Samples: The Wilcoxon Rank Sum Test**

Two-sample problems (see Chapters 21 and 23) are among the most common in statistics. The most useful nonparametric significance test compares two distributions. Here is an example of this setting.

**EXAMPLE 28.1 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 eight small plots of ground, then weeded the corn rows by hand to allow no weeds in four randomly selected plots and exactly three lamb's-quarter plants per meter of row in the other four plots. Here are the yields of corn (bushels per acre) in each of the plots:<sup>1</sup>

Zero weeds per meter	166.7	172.2	165.0	176.9
Three 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 28.2) 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



WEEDS3

**FIGURE 28.2**

Back-to-back stemplot of corn yields from plots with no weeds and with three weeds per meter of row, for Example 28.1. Notice the split stems, with leaves 0 to 4 on the first stem and leaves 5 to 9 on the second stem.

First, arrange all eight 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

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	<b>165.0</b>	<b>166.7</b>	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 28.1, 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 28.2 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 four 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

**28.1 Daily Activity and Obesity.** Our lead example for the two-sample  $t$  procedures in Chapter 21 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:  STANDTME

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?

**28.2 Does Playing Video Games Make Better Surgeons?** In laparoscopic surgery, a video camera and several thin instruments are inserted into the patient's abdominal cavity. The surgeon uses the image from the video camera positioned inside the patient's body to perform the procedure by manipulating the instruments that have been inserted. The Top Gun Laparoscopic Skills and Suturing Program was developed to help surgeons develop the skill set

necessary for laparoscopic surgery. Because of the similarity in many of the skills involved in video games and laparoscopic surgery, it was hypothesized that surgeons with greater prior video game experience might acquire the skills required in laparoscopic surgery more easily. Sixteen surgeons with previous video game experience participated in the study and were classified into the two categories, under three hours and more than three hours—depending on the number of hours they played video games at the height of their video game use. They also performed Top Gun drills and received a score based on the time to complete the drill and the number of errors made, with lower scores indicating better performance. Here are the Top Gun scores and video game categories for the 16 participants:<sup>2</sup>



TOPGUN2

Under three hours:	5540	6259	5163	6149	4398	3968	7367	4217	5716
Three or more hours:	7288	4010	4859	4432	4845	5394	2703	5797	3758

- Arrange the Top Gun scores in order and find their ranks.
- Find the Wilcoxon statistic  $W$  for the “Under three hours” group, along with its mean and standard deviation under the null hypothesis (no difference between the groups).
- Does comparing  $W$  with the mean and standard deviation suggest that surgeons with greater video game experience scored better on the Top Gun drills?

## 28.2 The Normal Approximation for $W$

To calculate the  $P$ -value  $P(W \geq 23)$  for Example 28.2, 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 28.3 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 of the Appendix; page 696)

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

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.

## APPLY YOUR KNOWLEDGE

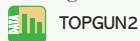
**28.3 Daily Activity and Obesity, continued.** In Exercise 28.1, you found the Wilcoxon rank sum  $W$  and its mean and standard deviation. We want to test the null hypothesis that the two groups don't differ in activity against the alternative hypothesis that the lean subjects spend more time standing and walking.



STANDTIME

- 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?

**28.4 Does Playing Video Games Make Better Surgeons? continued.** Use your values of  $W$ ,  $\mu_W$ , and  $\sigma_W$  from Exercise 28.2 to see whether surgeons with greater video game experience score better on the Top Gun Drills.



TOPGUN2

- 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?

**28.5 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:<sup>3</sup>



STORY2



Tyler Olson/Shutterstock

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 28.1 and 28.2.

## 28.3 Examples of Technology

For samples as small as those in the corn yield study of Example 28.1, 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 28.4 Weeds among the Corn: Software Output

Figure 28.3 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 28.3, is quite accurate. There are several differences between the CrunchIt! output and our work in Example 28.3. 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 statistics are related by simple algebra.

The second panel in Figure 28.3 is the two-sample  $t$  test from Chapter 21, which does not assume that the two populations have the same standard deviation. It gives

#### Mann–Whitney test

**FIGURE 28.3**

Output from CrunchIt!, for the data of Example 28.1. The output compares the results of three tests that could be used to compare yields for the two groups of corn plots.

#### CrunchIt!

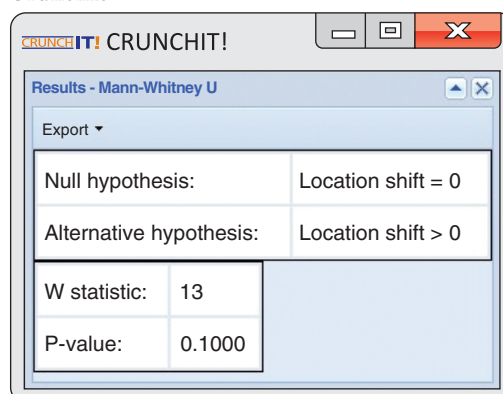
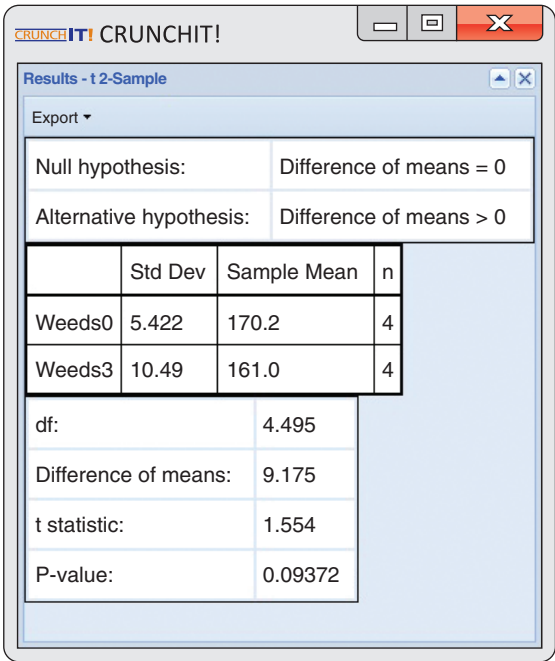
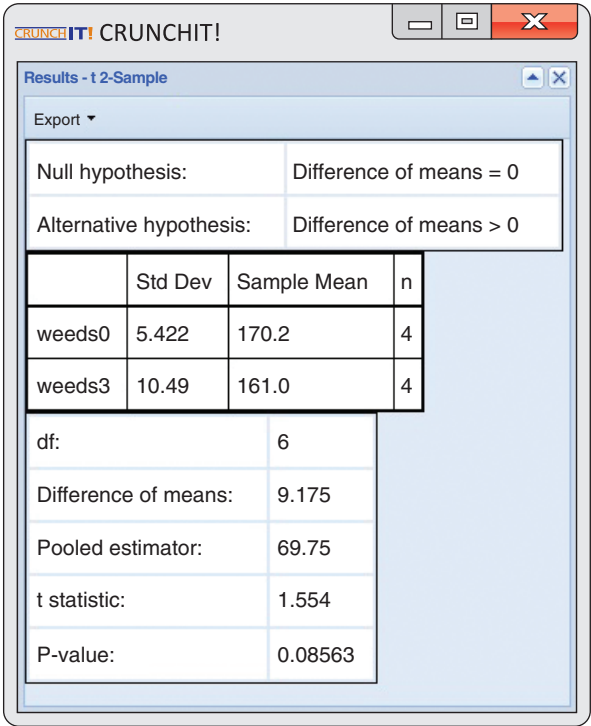




FIGURE 28.3  
(Continued)




CrunchIt!





$P = 0.0937$ , 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  $P$ -value is a bit different from the others. We do not recommend its use in general, despite the reasonable agreement in this example.

### APPLY YOUR KNOWLEDGE

**28.6 Does Playing Video Games Make Better Surgeons? Software.** Use your software to repeat the Wilcoxon test you did in Exercise 28.4. 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.  TOPGUN2

**28.7 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 28.3. Use the exact distribution if your software will do it. Compare the software result with your result in Exercise 28.3.  STANDTME

**28.8 Weeds among the Corn.** The corn yield study of Example 28.1 also examined yields in four plots having nine lamb's-quarter plants per meter of row. The yields (bushels per acre) in these plots were  WEEDS9

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 nine weeds per meter reduces corn yields when compared with weed-free corn? Use the Wilcoxon rank sum test with the preceding data and part of the data from Example 28.1 to answer this question.
- Compare the results from part (a) with those from the two-sample  $t$  test for these data.
- Now remove the low outlier 142.4 from the data with nine 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?

## 28.4 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

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 > \mu_2$$

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 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 three 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 28.3 states the hypotheses in terms of a location shift, the difference in the 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

$H_0$ : the two distributions are the same

$H_a$ : 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.<sup>4</sup> 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.

## APPLY YOUR KNOWLEDGE

**28.9 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  $W$ .

**28.10 Does Playing Video Games Make Better Surgeons? Hypotheses.** We are interested in whether surgeons with greater video experience score better “on the average” than those with less experience.

- 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?

## 28.5 Dealing with Ties in Rank Tests

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 nine observations:

*average ranks*

Observation	153	155	158	158	158	161	164	164	168
Rank	1	2	4	4	4	6	7.5	7.5	9

The first set of tied observations at 158 occupy the third, fourth, and fifth places in the ordered list, so they share the average rank of  $(3 + 4 + 5)/3 = 4$ . The second set of tied observations at 164 occupy the seventh and eighth places in the ordered list, so they share the average rank of  $(7 + 8)/2 = 7.5$ .

The exact distribution we have been using 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. Most statistical software will detect ties and make the necessary adjustment when using the Normal approximation. *Although there is an exact distribution when the data contain tied values, it is more complex and requires specialized software to compute. In practice, be careful using rank tests for very small sample sizes when ties are present.*



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 28.5**   Food Safety at Fairs



Donny Lehman/Corbis/VCG/Getty Images

**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.<sup>5</sup> 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 data file. 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):

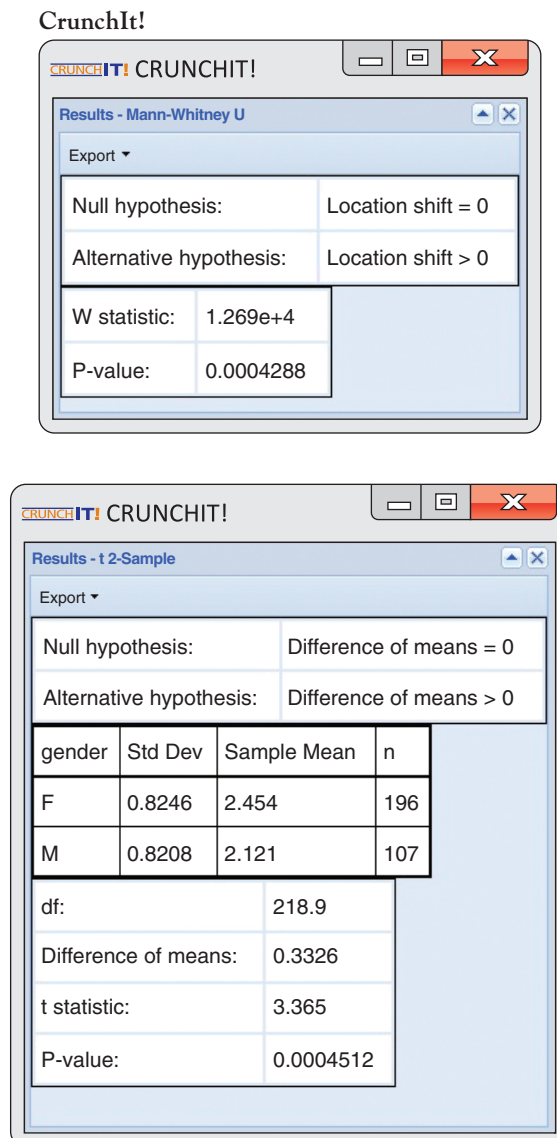
	Response					Total
	1	2	3	4	5	
Percent of females	6.6	55.1	25.5	11.7	1.0	100
Percent of males	20.6	53.3	20.6	4.7	1.0	100

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.

Because the responses can take only five values, there are many ties. All 35 people who chose “very rarely” are tied with a score of 1 and receive the average rank of  $(1 + 2 + \cdots + 34 + 35)/35 = 18$ , all 165 who chose “once in a while” are tied with a score of 2 and receive the average rank of  $(36 + 37 + \cdots + 199 + 200)/135 = 118$ , and so forth. Figure 28.4 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 ( $P = 0.000429$ ).



**FIGURE 28.4**

Output from CrunchIt!, for the data of Example 28.5. The Wilcoxon rank sum test and the two-sample  $t$  test give similar results.

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 28.4 shows that  $t = 3.365$  with  $P = 0.000451$ . 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 ( $P = 0.0004$  for the Wilcoxon test) 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 28.5 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 no fully meaningful scale of measurement.*




Because we have a two-way table, we might have applied the chi-square test (Chapter 25), 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.

### Macmillan Learning Online Resources

- The StatBoards video, *The Wilcoxon Rank Sum Test*, illustrates the use of ranks in the presence of ties for the comparison of two populations. The example discusses the appropriate hypotheses, the assignment of ranks to the observations, the computation of the test statistic, and the use of the normal approximation to draw conclusions.

### 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.

**28.11 Perception of Life Expectancy.** Exercise 21.8 (page 490) compares the perceived life expectancies of men and women. A researcher asked a sample of men and women to indicate their life expectancy. This was compared with values from actuarial tables, and the relative percent difference was computed (perceived life expectancy minus life expectancy from actuarial tables was divided by life expectancy from actuarial tables and converted to a percent). Here are the relative percent differences for all men and women over the age of 70 in the sample:  LIFEEXP

Men	-28	-23	-20	-19	-14	-13	
Women	-20	-19	-15	-12	-10	-8	-5


- 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?

- (c) Apply the Wilcoxon rank sum test to these data. Compare your result with the  $P = 0.0528$  obtained from the two-sample  $t$  test in Figure 21.5.

**28.12 Does the Wall Street Culture Corrupt Bankers?** Bank employees from a large international bank were recruited, with 67 assigned at random to a control group and the remaining 61 assigned to a treatment group. All subjects first completed a short online survey. After answering some general filler questions, the treatment group were asked seven questions about their professional background such as, “At which bank are you currently employed?” or “What is your function at this bank?” These are referred to as “identity priming” questions. The control group were asked seven innocuous questions unrelated to their profession such as, “How many hours a week on average do you watch television?” After the survey, all subjects performed a coin-tossing task that required tossing any coin 10 times and reporting the results online. They were told they would win \$20 for each head tossed for a maximum payoff of \$200. Subjects were unobserved during the task, making it impossible to tell if a particular subject cheated. If the banking culture favors dishonest behavior, it was conjectured that it should be possible to trigger this behavior by reminding subjects of their profession.<sup>6</sup> Here are the results for the 67 subjects in the control group, with the first line giving the possible number of heads on 10 tosses followed by the number of subjects that reported tossing this number of heads for the control and treatment groups, respectively:  **BANKERS**

	Number of Heads											Total
	0	1	2	3	4	5	6	7	8	9	10	
Control group	0	0	1	8	16	17	14	6	2	1	2	67
Treatment group	0	0	2	4	8	14	15	7	6	0	5	61

- (a) There are three observations tied at 2. What is the average rank assigned to these three observations? What is the average rank assigned to the 12 observations tied at 3?
- (b) What are the null and alternative hypotheses for the Wilcoxon test?
- (c) Use software: find the  $P$ -value of the Wilcoxon test and state your conclusion.

**28.13 Tell Me a Story, continued.** The data in Exercise 28.5 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?  **STORY1**

- (a) Make a back-to-back stemplot of the five 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?

**28.14 Shared Pain and Bonding.** Although painful experiences are involved in social rituals in many parts of the world, little is known about the social effects of pain. Will sharing painful experiences in a small group lead to greater bonding of group members than sharing a similar nonpainful experience? Fifty-four University students in South Wales were divided at random into a pain group containing 27 students, with the remaining students in the no-pain group. Pain was induced by two tasks. In the first task, students submerged their hands in freezing water for as long as possible, moving metal balls at the bottom of the

vessel into a submerged container, and in the second task, students performed a standing wall squat with back straight and knees at 90 degrees for as long as possible. The no-pain group completed the first task using room temperature water for 90 seconds and the second task by balancing on one foot for 60 seconds, changing feet if necessary. In both the pain and no-pain groups, the students completed the tasks in small groups which typically consisted of four students and contained similar levels of group interaction. Afterward, each student completed a questionnaire to create a bonding score based on answers to questions such as “I feel the participants in this study have a lot in common” or “I feel I can trust the other participants.” Here are the bonding scores for the two groups:<sup>7</sup>



BONDING

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

- Do the data show that sharing a painful experience in a small group leads to higher bonding scores for group members than sharing a similar nonpainful experience? 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.
- 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.
- What features of the data account for the differences in the results of the  $t$  test and the Wilcoxon test? Explain briefly.



**28.15 Food Safety in Restaurants.** Example 28.5 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 contains the responses of 303 people to several questions. The variables in this data set are (in order)



FOODSAFE

subject hfair sfair sfair sfair srest 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, and M if the respondent 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.



**28.16 Shared Pain and Bonding, continued.** The subjects in the experiment of Exercise 28.14 consist of both male and female students. Is there a difference in the level of bonding experienced by male and female students when having a painful experience in a small group? The “Pain” group contains 9 males and 18 females. Here are the bonding scores for the 9 males and 18 females in the “Pain” group:



BONDGEN

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



Examine the data and comment on departures from Normality and/or outliers. Is there significant evidence that sharing a painful experience in a small group leads to higher bonding scores for either males or females? Follow the four-step process.

**28.17 More on Food Safety.** The data file used in Exercise 28.15 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.”

## 28.6 Matched Pairs: The Wilcoxon Signed Rank Test

We use the one-sample  $t$  procedures (Chapter 20) 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 28.6 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

$H_0$ : scores have the same distribution for both stories

$H_a$ : 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.

A positive difference in Example 28.6 indicates 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 difference—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



STORIES

**absolute values**

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 28.7 Tell Me a Story, continued



**SOLVE:** In the storytelling study of Example 28.6,  $n = 5$ . 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

**28.18 Growing Trees Faster.** Exercise 20.39 (text page 474) 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:  TREES

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 three 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  $W^+$  with this mean lead to a tentative conclusion?

**28.19 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.<sup>8</sup> High counts suggest that infusion had a beneficial effect.  ELISACT

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.

## 28.7 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.

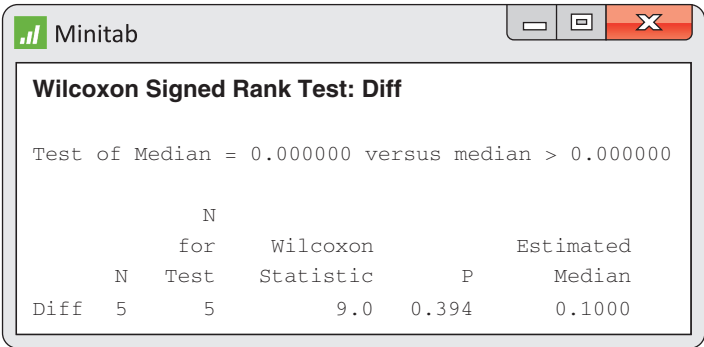
**EXAMPLE 28.8** Tell Me a Story: Normal Approximation

For  $n = 5$  observations, we saw in Example 28.7 that  $\mu_{W^+} = 7.5$  and  $\sigma_{W^+} = 3.708$ . 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:

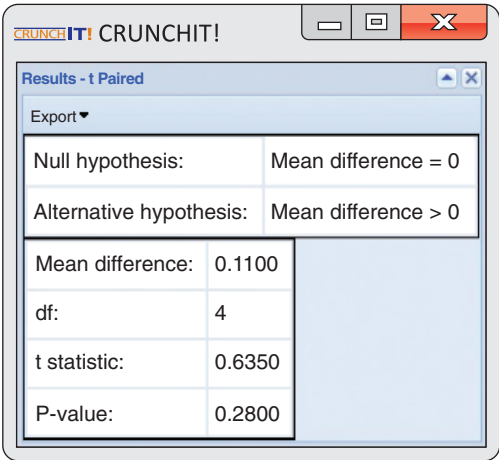
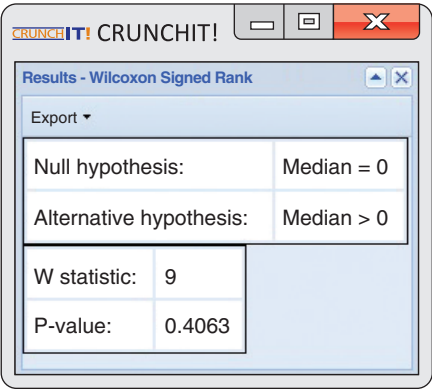
$$\begin{aligned} P(W^+ \geq 8.5) &= P\left(\frac{W^+ - 7.5}{3.708} \geq \frac{8.5 - 7.5}{3.708}\right) \\ &= P(Z \geq 0.27) \\ &= 0.394 \end{aligned}$$

Figure 28.5 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 28.7. 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.

**Minitab**




**CrunchIt!**





**FIGURE 28.5** Output from Minitab and CrunchIt!, for the storytelling data of Example 28.6. The CrunchIt! output compares the Wilcoxon signed rank test (with the exact distribution) and the matched pairs  $t$  test.




## APPLY YOUR KNOWLEDGE

**28.20 Growing Trees Faster: Normal Approximation.** Continue your work from Exercise 28.18. 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?  TREES

**28.21  $W^+$  versus  $t$ .** Find the one-sided  $P$ -value for the matched pairs  $t$  test applied to the tree growth data in Exercise 28.18. 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.  TREES

**28.22 Fighting Cancer: Normal Approximation.** Use the Normal approximation with continuity correction to find the  $P$ -value for the test in Exercise 28.19. What do you conclude about the effect of infusing modified cells on the ELISA count?  ELISACT

**28.23 Ancient Air.** The composition of the earth's atmosphere may have changed over time. To try to discover the nature of the atmosphere long ago, we can examine the gas in bubbles inside ancient amber. Amber is tree resin that has hardened and been trapped in rocks. The gas in bubbles within amber should be a sample of the atmosphere at the time the amber was formed. Measurements on specimens of amber from the late Cretaceous era (75 million to 95 million years ago) give these percents of nitrogen:<sup>9</sup>  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. Assume (this is not yet agreed on by experts) that these observations are an SRS from the late Cretaceous atmosphere.

- 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

## 28.8 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. Special software is required to use the 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.



GOLF



```

-1 | 6
-1 |
-0 | 6 5 5 5
-0 | 3
 0 | 1 2 3 4
 0 | 5 5

```

**FIGURE 28.6**

Stemplot (with split stems) of the differences in scores for two rounds of a golf tournament, for Example 28.9.

**EXAMPLE 28.9 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 low scores are better.)

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

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

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

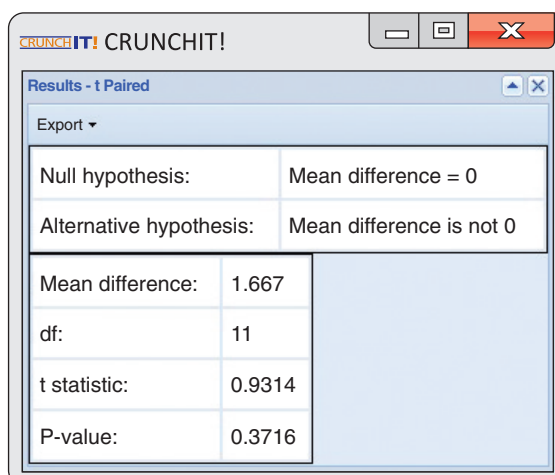
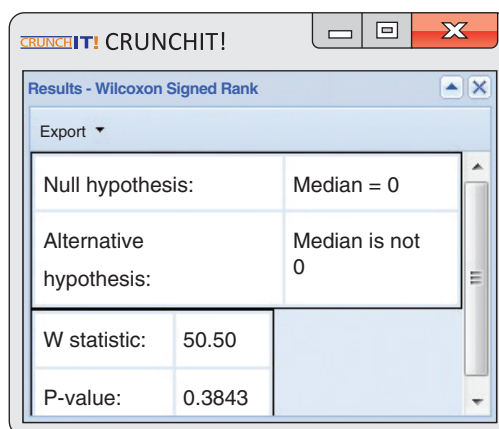
$H_0$ : scores have the same distribution in Rounds 1 and 2

$H_a$ : scores are systematically lower or higher in Round 2

**SOLVE:** A stemplot of the differences (Figure 28.6) shows some irregularity and a low outlier. We will use the Wilcoxon signed rank test.

Figure 28.7 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.

CrunchIt!

**FIGURE 28.7**

Output from CrunchIt!, for the golf scores data of Example 28.9. Because there are ties, a Normal approximation must be used for the Wilcoxon signed rank test.

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

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 of the ranks of the positive differences.)

### Macmillan Learning Online Resources

- The StatBoards video, *The Wilcoxon Signed Rank Test*, illustrates the use of ranks in the presence of ties for the comparison of two treatments when the data are paired observations. The example discusses the appropriate hypotheses, the assignment of ranks to the observations, the computation of the test statistic and the use of the normal approximation to draw conclusions.


## APPLY YOUR KNOWLEDGE

**28.24 Food Safety at Fairs and Fast-Food Restaurants.** Example 28.5 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 contains the responses of 303 people to several questions. The variables in this data set are (in order)  FOODSAFE

subject    hfair    sfair    sfast    srest    gender

The variable “sfair” contains responses to the safety question described in Example 28.5. The variable “sfast” contains responses to the same question asked about food served in fast-food restaurants. Is there a systematic difference between the level of concern about food safety at outdoor fairs and at fast-food restaurants? You will use the Wilcoxon signed rank statistic  $W^+$  (based on “sfair” response minus “sfast” response) to answer this question.

- In how many of the observations is the difference, “sfair” response minus “sfast” response, equal to zero? How are these observations used when computing the Wilcoxon signed rank statistic  $W^+$ ?
- Each of the variables “sfair” and “sfast” take the values 1, 2, 3, 4, or 5. What are the possible values of the absolute differences of “sfair” response minus “sfast” response? What does this say about the quantity of ties that will be found when ranking the absolute values of the differences?
- Do the data give evidence that there is a systematic difference between the level of concern about food safety at outdoor fairs and at fast-food restaurants? Use the Wilcoxon signed rank statistic  $W^+$  and software to answer this question. Given your work in part (b), make sure that your software adjusts for ties when computing the  $P$ -value.

**28.25 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:  COLA

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 one-sample matched pairs  $t$  test had  $P$ -value  $P = 0.0123$  for these data. How does this compare with your result from part (a)? What are the hypotheses for the  $t$  test? What conditions must be met for each of the  $t$  and Wilcoxon tests?

**28.26 Comparing Insect Repellents.** West Nile virus, Chikungunya, Rocky Mountain spotted fever, and Lyme disease are becoming increasingly common insect-borne diseases in North America. Insect repellents can provide protection from bites of insects that carry these diseases, but which are the most effective? To investigate, we compare two insect repellents. The active ingredient in one is 15% Deet. The active ingredient in the other is oil of lemon eucalyptus. Repellents are tested on four volunteers. For each volunteer, the left arm is sprayed with one of the repellents and the right arm with the other. Which arm receives which repellent is determined randomly. Beginning 30 minutes after applying the repellents, once every hour volunteers put each arm in separate 8-cubic-foot cages containing 200 disease-free female mosquitoes in need of a blood meal to lay their eggs. During each hour, volunteers leave their arms in the cage for five minutes. The repellent is considered to have failed if a volunteer was bitten two or more times in a five-minute session or bitten once in the five-minute sessions for two consecutive hours. The response is the number of five-minute sessions until a repellent fails.<sup>10</sup> Here are the number of hours until failure for the two repellents.

 REPEL

Subject	1	2	3	4
Deet	5	7	4	4
Oil of lemon eucalyptus	8	7	6	7
Difference (Deet minus oil of lemon eucalyptus)	-3	0	-2	-3

Is either repellent more effective?

- The tie *within* the pair for subject 2 creates a difference of zero. Because these are neither positive nor negative, this subject is dropped from our sample and the computations are done as if there were only three subjects, reducing the sample size in our calculations. 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.
- Although oil of lemon eucalyptus is superior for the remaining three subjects, the  $P$ -value is not that small. Explain briefly why this is the case.

## 28.9 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 28.10 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 zero, one, three, and nine weeds 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:<sup>11</sup>



WEEDFULL

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	$n$	Median	Mean	Standard Deviation
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 three and nine 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.



## 28.10 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 28.10, the ANOVA hypotheses are

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

$H_a$ : 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

$H_0$ : yields have the same distribution in all groups

$H_a$ : yields are systematically higher in some groups than in others

If all 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 28.10 do *not* all have the same shape.

## 28.11 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, the exact distribution when there are ties among the responses requires special software. We again assign average ranks to tied observations.

### EXAMPLE 28.11 Weeds among the Corn, continued

**SOLVE (inference):** In Example 28.10, 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 as follows:

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



WEEDFULL

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,

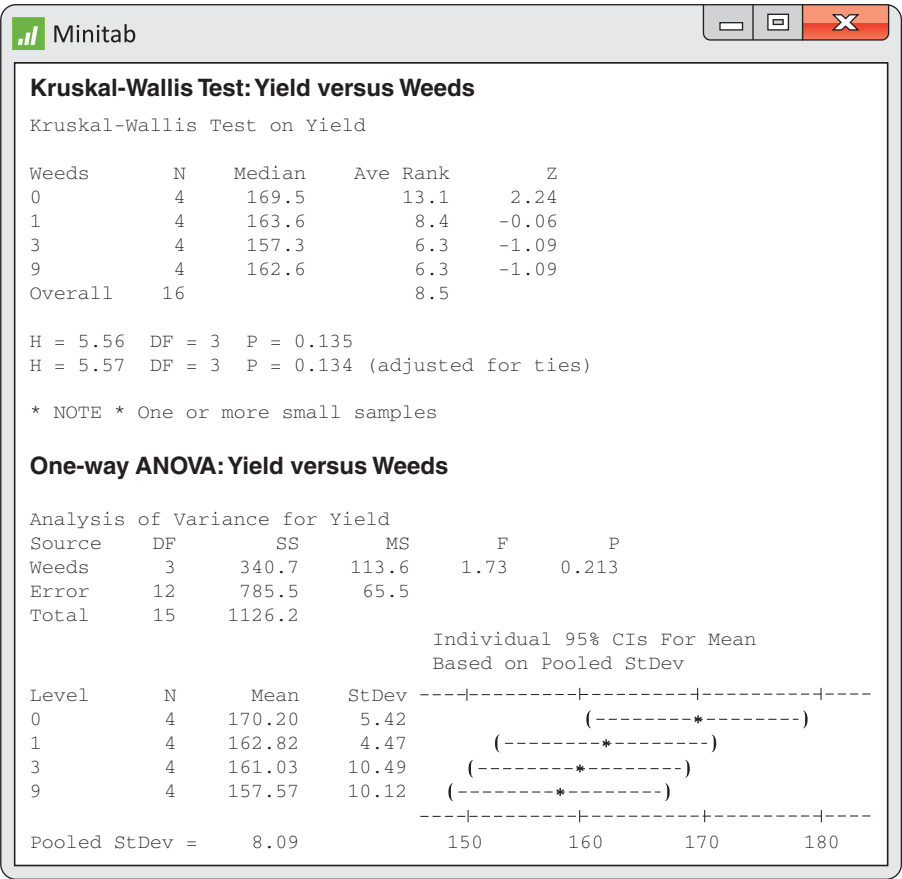
$$\begin{aligned}
 H &= \frac{12}{N(N+1)} \sum \frac{R_i^2}{n_i} - 3(N+1) \\
 &= \frac{12}{(16)(17)} \left( \frac{52.5^2}{4} + \frac{33.5^2}{4} + \frac{25^2}{4} + \frac{25^2}{4} \right) - (3)(17) \\
 &= \frac{12}{272} (1282.125) - 51 \\
 &= 5.56
 \end{aligned}$$

Referring to the table of chi-square critical points (Table D of the Appendix; page 700) with  $df = 3$ , we see that the  $P$ -value lies in the interval  $0.10 < P < 0.15$ .

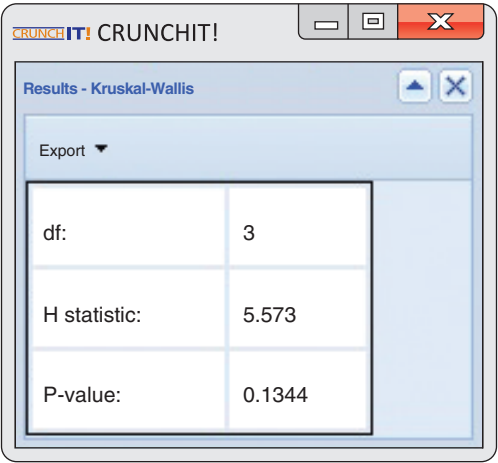
**CONCLUDE:** Although this small experiment *suggests* that more weeds decrease yield, it does not provide *convincing* evidence that weeds have an effect.

Figure 28.8 displays the Minitab output for both ANOVA and the Kruskal-Wallis test and the CrunchIt! output for 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. CrunchIt! automatically makes a correction for ties as well. For these data, the

**Minitab**



**CrunchIt!**



**FIGURE 28.8**  
Minitab output, for the corn yield data of Example 28.10. For comparison, both the Kruskal-Wallis test and one-way ANOVA are shown.


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.

### Macmillan Learning Online Resources

- The StatBoards video, *The Kruskal–Wallis Test*, provides a detailed example of a nonparametric one-way ANOVA using the Kruskal–Wallis test.


### APPLY YOUR KNOWLEDGE

**28.27 More Rain for California?** Exercise 27.35 (page 667) 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 experiment continued for several years. Here are data for 2004 (mass in grams per square meter):  MASS2004

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 (page 646) 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  $I$ ,  $n_i$ , and  $N$ ? Arrange the counts in order and assign ranks.
- 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?

**28.28 Logging in the Rain Forest: Species Richness.** Table 27.2 (page 642) 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 one year earlier and eight 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.  LOGFULL

- 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.



**28.29 Does Playing Video Games Make Better Surgeons?** In laparoscopic surgery, a video camera and several thin instruments are inserted into the patient's abdominal cavity. The surgeon uses the image from the video camera positioned inside the patient's body to perform the procedure by manipulating the instruments that have been inserted. The Top Gun Laparoscopic Skills and Suturing Program was developed to help surgeons develop the skill set necessary for laparoscopic surgery. Because of the similarity in many of the skills involved in video games and laparoscopic surgery, it was hypothesized that surgeons with greater prior video game experience might acquire the skills required in laparoscopic surgery more easily. Thirty-three surgeons participated in the study and were classified into the three categories, never used, under three hours, and three or more hours—depending on the number of hours they played video games at the height of their video game use. They also performed Top Gun drills and received a score based on the time to complete the drill and the number of errors made, with lower scores indicating better performance. Here are the Top Gun scores and video game categories for the 33 participants:<sup>12</sup> **TOPGUN**

Never played:	9379	8302	5489	5334	4605	4789	9185	7216	9930
	4828	5655	4623	7778	8837	5947			
Under three hours:	5540	6259	5163	6149	4398	3968	7367	4217	5716
Three or more hours:	7288	4010	4859	4432	4845	5394	2703	5797	3758


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



**28.30 Good Weather and Tipping.** Favorable weather has been shown to be associated with increased tipping. Exercise 27.46 (page 670) describes a study to investigate whether just the belief that future weather will be favorable can lead to higher tips. The researchers gave 60 index cards to a waitress at an Italian restaurant in New Jersey. Before delivering the bill to each customer, the waitress randomly selected a card and wrote on the bill the same message that was printed on the index card. Twenty of the cards had the message, “The weather is supposed to be really good tomorrow. I hope you enjoy the day!” Another 20 cards contained the message, “The weather is supposed to be not so good tomorrow. I hope you enjoy the day anyway!” The remaining 20 cards were blank, indicating that the waitress was not supposed to write any message. Choosing a card at random ensured that there was a random assignment of the diners to the three experimental conditions. Here are the tips as a percentage of the total bill for the three messages:<sup>13</sup> **TIPPING**

Good weather report	20.8	18.7	19.9	20.6	22.0	23.4	22.8	24.9	22.2	20.3
	24.9	22.3	27.0	20.4	22.2	24.0	21.2	22.1	22.0	22.7
Bad weather report	18.0	19.0	19.2	18.8	18.4	19.0	18.5	16.1	16.8	14.0
	17.0	13.6	17.5	19.9	20.2	18.8	18.0	23.2	18.2	19.4
No weather report	19.9	16.0	15.0	20.1	19.3	19.2	18.0	19.2	21.2	18.8
	18.5	19.3	19.3	19.4	10.8	19.1	19.7	19.8	21.3	20.6

Do a complete analysis that includes a test of significance. Include a statement in words of your null and alternative hypotheses.

**28.31 Food Safety.** Example 28.5 describes a study of the attitudes of people attending outdoor fairs about the safety of the food served at such locations. The data set contains the responses of 303 people to several questions. The variables in this data set are (in order)  FOODSAFE

subject hfair sfair sfast srest gender

The variable “sfair” contains responses to the safety question described in Example 28.5. 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 28 SUMMARY

### Chapter Specifics

- **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.



## Statistics in Summary

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  $H$ .

## Link It

The statistical methods in Chapters 20, 21, and 27 were developed for Normal distributions, but are fairly robust against a failure in this assumption. However, when the sample sizes are very small or there are outlying observations, it is important to have alternative techniques. The nonparametric methods described in this chapter provide such alternatives. Nonparametric methods often involve replacing the actual data by their ranks, which makes these methods fairly insensitive to outlying observations and also allows us to use them with minimal assumptions regarding the distributions of the data.

For the paired-data problem, the Wilcoxon signed rank test can be used as an alternative to the paired  $t$  test of Chapter 20, whereas for the two-sample problem the rank sum test provides an alternative to the two-sample  $t$  test described in Chapter 21. The Kruskal–Wallis test can be used in place of the  $F$  test for one-way ANOVA. Although we have concentrated on hypothesis testing in this chapter, these nonparametric methods also provide confidence intervals, and in one-way ANOVA there are associated multiple comparisons for determining which of the treatments differ. As with the other supplemental chapters, this chapter is intended to provide only an introduction to a more advanced topic in statistical inference. More advanced courses in statistics will provide additional details for these methods.

## Macmillan Learning Online Resources

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

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

## CHECK YOUR SKILLS

**28.32** A study of the effects of exercise used rats bred to have high or low capacity for exercise. There were eight high-capacity and eight low-capacity rats. To compare the blood pressures of the groups, you use the

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

**28.33** 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.

**28.34** You interview both the husband and the wife in 64 married couples and give each a test that measures marital satisfaction. To assess whether there is a difference in level of marital satisfaction between husbands and wives, you use the

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

**28.35** 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	15.2	12.6	11.9	5.1	
Attacked group	10.6	15.3	25.2	17.1	14.6

The rank sum  $W$  for the control group is

- (a) 10.
- (b) 14.
- (c) 17.

**28.36** If there is no difference in emissions between the attacked group and the control group, the mean of  $W$  in Exercise 28.35 is

- (a) 18.
- (b) 20.
- (c) 25.

**28.37** Suppose that the nine observations in Exercise 28.35 were

Control group	15.2	12.6	11.9	5.1	
Attacked group	10.6	15.2	25.2	15.2	14.6

The rank sum for the control group is now

- (a) 12.
- (b) 14.
- (c) 15.

**28.38** Interview eight young married couples, speaking with 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
Husband	7	7	8	5	9	6	8	6
Wife	4	3	6	6	2	2	10	2

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

- (a) 30.
- (b) 32.5.
- (c) 34.5.

**28.39** If husbands and wives don't differ in how important the attractiveness of their spouse is, the mean of  $W^+$  in Exercise 28.38 is

- (a) 18.
- (b) 36.
- (c) 68.

**28.40** Suppose that the responses in Exercise 28.38 are

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

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

- (a) 18.5.
- (b) 20.
- (c) 21.5.

**28.41** You compare the starting salaries of seven graduates who majored in accounting, nine who majored in finance, five who majored in marketing, and six who majored in business administration. If the four starting-salary distributions are the same, the Kruskal-Wallis statistic  $H$  has approximately a chi-square distribution. The degrees of freedom are

- (a) 1.
- (b) 2.
- (c) 3.


## CHAPTER 28 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.

**28.42 Right versus left.** The design of controls and instruments affects how easily people can use them. Timothy Sturm investigated this effect in a course project, asking




25 right-handed students to turn a knob (with their right hands) that moved an indicator by screw action. There were two identical instruments: one with a right-hand thread (the knob turns clockwise) and the other with a left-hand thread (the knob turns counterclockwise). Each of the 25 students used both instruments with the order

randomized separately for each student. Table 20.4 (page 476) gives the times in seconds each subject took to move the indicator a fixed distance. The project hoped to show that right-handed people find right-hand threads easier to use. Do an analysis that leads to a conclusion about this issue.  RTLFT


**28.43 College students need better sleep!** A random sample of 898 students between the ages of 20 and 24 at a large midwestern university completed a survey including questions about their sleep quality, moods, academic performance, physical health, and psychoactive drug use. Sleep quality was measured using the Pittsburgh Sleep Quality Index (PSQI), with students scoring less than or equal to 5 on the index classified as optimal sleepers, those scoring a 6 or 7 classified as borderline, and those scoring over 7 classified as poor sleepers. The depression subscale of the Profile of Moods State (POMS) was used to assess how severely students experienced depression on a typical day, with high scores indicating greater levels of depression. The full data set is too large to print here, but here are the first seven individuals:  DEPRESSED

Quality of sleep:	poor	poor	border	poor	poor	optimal	border
Depression score:	5	8	5	11	7	7	10

We want to know if there is a significant difference in depression scores among the three classifications of sleep.<sup>14</sup>

**28.44 The brain responds to sound.** Table 7.1 (page 191) contains data from a study comparing the brain's response to "pure tones" and recognizable sounds. Researchers anesthetized macaque monkeys and fed pure tones and also monkey calls directly to their brains by inserting electrodes. Response to the stimulus was measured by the firing rate (electrical spikes per second) of neurons in various areas of the brain. You may assume the 37 responses are independent. Researchers suspected that the response to monkey calls would be stronger than the response to a pure tone. Do the data support this idea?  BRESPN

**28.45 Adolescent obesity.** Adolescent obesity is a serious health risk affecting more than 5 million young people in the United States alone. Laparoscopic adjustable gastric banding has the potential to provide a safe and effective treatment. Fifty adolescents between 14 and 18 years old with a body mass index higher than 35 were recruited from the Melbourne, Australia, community for the study.<sup>15</sup> Twenty-five were randomly selected to undergo gastric banding, and the remaining 25 were assigned to a supervised lifestyle intervention program involving diet, exercise, and be-


havior modification. All subjects were followed for two years. Here are the weight losses in kilograms for the subjects who completed the study. In the gastric banding group:  GASTRIC

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

In the lifestyle intervention group:

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


Does gastric banding result in significantly greater weight loss than a supervised lifestyle intervention program?

**28.46 Protective equipment and risk taking.** Studies have shown that people who are using safety equipment when engaging in an activity tend to take increased risks. Will risk taking increase when people are not aware they are wearing protective equipment and are engaged in an activity that cannot be made safer by this equipment? Participants in the study were falsely told they were taking part in an eye-tracking experiment for which they needed to wear an eye-tracking device. Eighty subjects were divided at random into two groups of 40 each, with one group wearing the tracking device mounted on a baseball cap and the other group wearing it mounted on a bicycle helmet. Subjects were told that the helmet or cap was just being used to mount the eye tracker. All subjects watched an animated balloon on a video screen and pressed a button to inflate it. The balloon was programmed to burst at a random point, but until that point, each press of the button inflated the balloon further and increased the amount of fictional currency a subject would earn. Subjects were free to stop pumping at any point and keep their earnings, knowing that if the balloon burst they would lose all earnings for that round. The score was the average number of pumps on the trials, with lower scores corresponding to less risk taking and more conservative play. Here are the first 10 observations from each group:<sup>16</sup>  HELMET

Helmet:	3.67	36.50	29.28	30.50	24.08	32.10	50.67	26.26	41.05	20.56
Baseball cap:	29.38	42.50	41.57	47.77	32.45	30.65	7.04	2.68	22.04	25.86


Does wearing of a helmet rather than a baseball cap lead to greater risk taking in this experiment?

**28.47 Food safety at fairs and restaurants.** Example 28.5 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 contains the responses of 303 people to several questions. The variables in this data set are (in order)  FOODSAFE


subject hfair sfair sfast srest gender

The variable “sfair” contains responses to the safety question described in Example 28.5. 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?

**28.48 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:<sup>17</sup>  NEMATODE

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?


**28.49 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 Spliced International index. Table 20.3 (page 476) gives the annual returns (percent) for the fund and the index. Does the fund’s performance differ significantly from that of its benchmark?  MFUND


*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 28.1*


**TABLE 28.1 Electric fish species in the Amazon**

Tributary	Species Counts		
	Upstream	Tributary	Downstream
Içá	14	23	19
Jutai	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

*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.<sup>18</sup> The researchers used nonparametric tests to assess the statistical significance of their results. Exercises 28.50 through 28.52 quote conclusions from the study.*

**28.50 Downstream versus upstream.** “We identified a significant positive effect of tributaries on Amazon mainstream 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.  AMAZON

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

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

## EXPLORING THE WEB

**28.53 Confidence in the Congress.** The General Social Survey (GSS) is a sociological survey used to collect data on demographic characteristics and attitudes of residents of the United States. The survey is conducted by the National Opinion Research Center of the University of Chicago, which interviews face-to-face a

randomly selected sample of adults (18 and older). SDA (Survey Documentation and Analysis) is a set of programs that allows you to analyze survey data and includes the GSS as part of its archive. In Exercises 27.54 and 27.55 (found within the “Web Exercises” resource at [www.macmillanlearning.com/bps8e](http://www.macmillanlearning.com/bps8e)), you used data from the GSS to study the relationship between average respondent age and confidence in the U.S. Congress. A one-way ANOVA showed a statistically significant difference between the age of the respondent and his or her confidence in Congress. Download the data file following the directions in Exercise 27.55, and do the Kruskal–Wallis test to see if the ages are systematically higher for some levels of confidence in the Congress than others. Are your results similar to the one-way ANOVA?

**28.54 Who’s More Liberal?** The American National Election Studies (ANES) is the leading academically run national survey of voters in the United States and is conducted before and after every presidential election. SDA (Survey Documentation and Analysis) is a set of programs that allows you to analyze survey data and includes the ANES survey as part of its archive. Go to the website [sda.berkeley.edu/](http://sda.berkeley.edu/) and click on the Archive link at the top of the page. Under the American National Election Studies, go to the most recent ANES survey (at the time of writing, this was the 2012 ANES survey).

- (a) You want to see if either gender rates itself as more liberal. The variable gender is “GEND\_GENDOBS” and the variable for liberal/conservative self-placement on a seven-point scale is “LIBCPRE\_SELF.” You need to download these two variables using the same procedure as in Exercise 28.53. After downloading the data, you first need to eliminate all observations for which either variable is missing. For gender, any value other than 1 or 2 is a missing value code, and for the liberal/conservative self-placement scale any value other than 1, 2, . . . , 7 is a missing value code. Once these observations have been eliminated, use your statistical software to perform the Wilcoxon rank sum test, adjusting for ties (the two groups are males and females and the response is the self-placement score). What is your conclusion?
- (b) Can the two-sample  $t$  test be used to answer this question? Explain.

## Notes and Data Sources

1. Data provided by Samuel Phillips, Purdue University.
2. James C. Rosser et al., “The impact of video games on training surgeons in the 21st century,” *Archives of Surgery*, 142 (2007), pp. 181–186. We thank Douglas Gentile for providing the data.
3. Data provided by Susan Stadler, Purdue University.
4. 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.
5. Huey Chern Boo, “Consumers’ perceptions and concerns about safety and healthfulness of food served at fairs and festivals,” MS thesis, Purdue University, 1997.
6. Alain Cohen et al., “Business culture and dishonesty in the banking industry,” *Nature*, 516 (2014), pp. 86–89.
7. Brock Bastian et al., “Pain as social glue: Shared pain increases cooperation,” *Psychological Science*, 25 (2014), pp. 2079–2085.
8. 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.
9. R. A. Berner and G.P. Landis, “Gas bubbles in fossil amber as possible indicators for the major gas composition of ancient air,” *Science*, 239 (1988), pp. 1406–1409.

10. The description of this experiment is based on the methodology described in “How to win the battle of the bugs,” *Consumer Reports*, July 2015, pp. 34–37. Although artificial, the data given in the example are consistent with the findings of *Consumer Reports*.
11. See note 1.
12. See note 2.
13. Bruce Rind and David Strohmetz, “Effects of beliefs about future weather conditions on restaurant tipping,” *Journal of Applied Social Psychology*, 31 (2001), pp. 2160–2164. We would like to thank the authors for supplying the original data.
14. Hannah Lund et al., “Sleep patterns and predictors of disturbed sleep in a large population of college students,” *Journal of Adolescent Health*, 46 (2010), pp. 124–132. We would like to thank the authors for supplying the data.
15. Paul E. O’Brien et al., “Laparoscopic adjustable gastric banding in severely obese adolescents,” *Journal of the American Medical Association*, 303 (2010), pp. 519–526. We thank the authors for providing the data.
16. Tim Gamble and Ian Walker, “Wearing a bicycle helmet can increase risk taking and sensation seeking in adults,” *Psychological Science*, 27 (2016), pp. 289–294.
17. Data provided by Matthew Moore.
18. Cristina Cox Fernandes, Jeffrey Podos, and John G. Lundberg, “Amazonian ecology: Tributaries enhance the diversity of electric fishes,” *Science*, 305 (2004), pp. 1960–1962.