Nonparametric Tests

CHAPTER 15

15.1 The Wilcoxon Rank Sum Test
15.2 The Wilcoxon Signed Rank Test
15.3 The Kruskal-Wallis Test

Introduction

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. Some practical guidelines for taking advantage of the **robustness** of these methods appear in Chapter 7.

robustness

What can we do if plots suggest that the population distribution is 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 the outliers. An outlier is an observation that may 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. If the outlier appears to be "real data," you can base inference on statistics that are more resistant than $\bar{x}$ and s. Options 4 and 5 allow this.

outlier

LOOK BACK
transformations, p. 93

2. Sometimes we can **transform** our data so that their distribution is more nearly Normal. Transformations such as the logarithm that pull in the long tail of right-skewed distributions are particularly helpful. Example 7.10 (page 436) illustrates use of the logarithm.

other standard distributions

3. In some settings, **other standard distributions** replace the Normal distributions as models for the overall pattern in the population. We mentioned in Chapter 5 (page 315) that the Weibull distributions are common models for the lifetimes in service of equipment in statistical studies of reliability. Also, we studied the exponential distributions (page 309) and the Poisson distributions (page 339) in Chapter 5. There are inference procedures for the parameters of these distributions that replace the t procedures when we use specific non-Normal models.

bootstrap methods
permutation tests

4. Modern **bootstrap methods** and **permutation tests** do not require Normality or any other specific form of sampling distribution. Moreover, you can base inference on resistant statistics such as the trimmed mean. We recommend these methods unless the sample is so small that it may not represent the population well. Chapter 16 gives a full discussion.

nonparametric methods

5. Finally, there are other **nonparametric methods** that do not require 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. The *sign test* (page 438) works with *counts* of observations. This chapter presents **rank tests** based on the *rank* (place in order) of each observation in the set of all the data.

rank tests

This chapter concerns rank tests that are designed to replace the t tests and one-way analysis of variance when the Normality conditions for those tests are not met. Figure 15.1 presents an outline of the standard tests (based on Normal distributions) and the rank tests that compete with them.

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 15.1 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.

FIGURE 15.1 Comparison of tests based on Normal distributions with nonparametric tests for similar settings.

Setting	Normal test	Rank test
One sample	One-sample t test Section 7.1	Wilcoxon signed rank test Section 15.2
Matched pairs	Apply one-sample test to differences within pairs	
Two independent samples	Two-sample t test Section 7.2	Wilcoxon rank sum test Section 15.1
Several independent samples	One-way ANOVA F test Chapter 12	Kruskal-Wallis test Section 15.3

We devote a section of this chapter to each of the rank procedures. Section 15.1, which discusses the most common of these tests, also contains general information about rank tests. The kind of assumptions required, the nature of the hypotheses tested, the big idea of using ranks, and the contrast between exact distributions for use with small samples and approximations for use with larger samples are common to all rank tests. Sections 15.2 and 15.3 more briefly describe other rank tests.

15.1 The Wilcoxon Rank Sum Test

When you complete this section, you will be able to

- **Find the rank transformation for a set of data.**
- **Compute the Wilcoxon rank sum statistic for the comparison of two populations.**
- **State the null and alternative hypotheses that are used for the analysis of data using the Wilcoxon rank sum test.**
- **Use the two sample sizes to find the mean and the standard deviation of the sampling distribution of the Wilcoxon rank sum statistic under the null hypothesis.**
- **Find the *P*-value for the Wilcoxon rank sum significance test using the Normal approximation with the continuity correction.**
- **For the Wilcoxon rank sum test, use computer output to determine the results of the significance test.**

LOOK BACK
two sample, p. 447

Two-sample problems (see Section 7.2) are among the most common in statistics. The most useful nonparametric significance test compares two distributions. Here is an example of this setting.

EXAMPLE

HITS

15.1 Does the American League get more hits? In 1973, the American League adopted the designated-hitter rule, which allows a substitute player to take the place of the pitcher when it is the pitcher's turn to bat. Since pitchers typically do not hit as well as other players, it was expected that the rule would produce more hits and therefore more excitement for the fans. The National League has not adopted this rule. Let's look at some data to see if we can detect a difference in hits between the American League and the National League. Here are the number of hits for eight games played on the same spring day, four from each league.

League	Hits			
American	21	18	24	20
National	19	7	11	13

The samples are too small to assess Normality adequately or to rely on the robustness of the t test. We prefer to use a test that does not require Normality.

The rank transformation

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

7 11 13 **18** 19 **20** **21** **24**

The boldface entries in the list are the hits for the American League. 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:*

Runs	7	11	13	**18**	19	**20**	**21**	**24**
Rank	1	2	3	**4**	5	**6**	**7**	**8**

It would not be surprising if we had sampled a day where more than one game had the same number of hits. We will discuss how to handle ties later in this section.

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 is a transformation of the data, like moving from the observations to their logarithms. The rank transformation retains only the ordering of the observations and makes no other use of their numerical values. Working with ranks allows us to dispense with specific assumptions about the shape of the distribution, such as Normality.

USE YOUR KNOWLEDGE

SPAS

SPAS2

15.1 Numbers of rooms in top spas. A report of a readers' poll in *Condé Nast Traveler* magazine ranked 100 top resort spas.[1] Let Group A be the 25 top-ranked spas, and let Group B be the spas ranked 26 to 50. A simple random sample of size 5 was taken from each group, and the number of rooms in each selected spa was recorded. Here are the data:

Group A	106	145	312	60	49
Group B	190	500	1293	161	225

Rank all the observations together and make a list of the ranks for Group A and Group B.

15.2 The effect of Animal Kingdom on the result. Refer to the previous exercise. Disney's Animal Kingdom in Lake Buena Vista, Florida, with 1293 rooms, was the third spa selected in Group B. Suppose, instead, a different spa, with 540 rooms, had been selected. Replace the observation 1293 in Group B by 540. Use the modified data to make a list of the ranks for Groups A and B combined. What changes?

The Wilcoxon rank sum test

If the American League games tend to have more hits than the National League, we expect the ranks of the American League games to be higher than those for the National League games. Let's compare the *sums* of the ranks from the two treatments:

League	Sum of ranks
American	25
National	11

These sums compare the hits of the American League with those of the National League. 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 American League is 25, then the ranks for the National League must be 11 because $25 + 11 = 36$. If there was no difference between the leagues, we would expect the sum of the ranks for each league 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.* This test is also called the **Mann-Whitney test.**

For the baseball question of Example 15.1, we want to test

H_0: no difference in number of hits

against the one-sided alternative

H_a: more hits are made in American League games than in National League games

Our test statistic is the rank sum $W = 25$ for the American League games.

*This test was invented by Frank Wilcoxon (1892–1965) in 1945. Wilcoxon was a chemist who encountered statistical problems in his work at the research laboratories of American Cyanamid Company.

USE YOUR KNOWLEDGE

SPAS

15.3 Hypotheses and test statistic for top spas. Refer to Exercise 15.1. State appropriate null and alternative hypotheses for this setting and calculate the value of W, the test statistic.

SPAS2

15.4 Effect of Animal Kingdom on the test statistic. Refer to Exercise 15.2. Using the altered data, state appropriate null and alternative hypotheses and calculate the value of W, the test statistic.

EXAMPLE

15.2 Perform the significance test. In Example 15.1, $n_1 = 4$, $n_2 = 4$, and there are $N = 8$ observations in all. The sum of ranks for the American League games has mean

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

$$= \frac{(4)(9)}{2} = 18$$

and standard deviation

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

$$= \sqrt{\frac{(4)(4)(9)}{12}} = \sqrt{12} = 3.464$$

The observed sum of the ranks, $W = 25$, is higher than the mean, about 2 standard deviations higher ([25 −18]/3.464). It appears that the data support our idea that American League games have more hits than National League games. The P-value for our one-sided alternative is $P(W \geq 25)$, the probability that W is at least as large as the value for our data when H_0 is true.

To calculate the P-value $P(W \geq 25)$, 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 a bit unwieldy, though you can find them in handbooks of statistical tables. Most statistical software will give you P-values, as well as carry out the ranking and calculate W. However, some software gives only approximate P-values. You must learn what your software offers.

EXAMPLE

15.3 Software output. Figure 15.2 shows the output from software that calculates the exact sampling distribution of W. We see that the sum of the ranks (called scores in the output) for the American League is $W = 25$, with P-value $P = 0.0286$ against the one-sided alternative that American League games have more hits than the National League games.

FIGURE 15.2 Output from SAS for the baseball hit data, for Example 15.3.

SAS

The SAS System

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable Hits Classified by Variable League

League	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
American	4	25.0	18.0	3.464102	6.250
National	4	11.0	18.0	3.464102	2.750

Wilcoxon Two-Sample Test	
Statistic (S)	25.0000
Normal Approximation	
Z	1.8764
One-Sided Pr > Z	0.0303
Two-Sided Pr > |Z|	0.0606
t Approximation	
One-Sided Pr > Z	0.0514
Two-Sided Pr > |Z|	0.1027
Exact Test	
One-Sided Pr ≥ S	0.0286
Two-Sided Pr ≥ |S – Mean|	0.0571
Z includes a continuity correction of 0.5.	

Done

LOOK BACK
two-sample *t* test, p. 454

It is worth noting that the two-sample t test for the one-sided alternative gives essentially the same result as the Wilcoxon test in Example 15.3 ($t = 2.95$, $P = 0.016$).

The Normal approximation

The rank sum statistic W becomes approximately Normal as the two sample sizes increase. We can then form yet another z statistic by standardizing W:

$$z = \frac{W - \mu_W}{\sigma_W}$$

$$= \frac{W - n_1(N + 1)/2}{\sqrt{n_1 n_2 (N + 1)/12}}$$

LOOK BACK
continuity correction, p. 335

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

EXAMPLE

15.4 The continuity correction. The standardized rank sum statistic W in our baseball example is

$$z = \frac{W - \mu_W}{\sigma_W} = \frac{25 - 18}{3.464} = 2.02$$

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

$$P(Z \geq 2.02) = 0.0217$$

The continuity correction acts as if the whole number 25 occupies the entire interval from 24.5 to 25.5. We calculate the P-value $P(W \geq 25)$ as $P(W \geq 24.5)$ because the value 25 is included in the range whose probability we want. Here is the calculation:

$$\begin{aligned} P(W \geq 24.5) &= P\left(\frac{W - \mu_W}{\sigma_W} \geq \frac{24.5 - 18}{3.464}\right) \\ &= P(Z \geq 1.876) \\ &= 0.303 \end{aligned}$$

The continuity correction gives a result closer to the exact value $P = 0.0286$ (see Figure 15.2).

USE YOUR KNOWLEDGE

SPAS

15.5 The *P*-value for top spas. Refer to Exercises 15.1 and 15.3 (pages 15-4 and 15-6). Find μ_W, σ_W, and the standardized rank sum statistic. Then give an approximate P-value using the Normal approximation. What do you conclude?

SPAS2

15.6 The effect of Animal Kingdom on the *P*-value. Refer to Exercises 15.2 and 15.4 (pages 15-4 and 15-6). Repeat the analysis in Exercise 15.5 using the altered data.

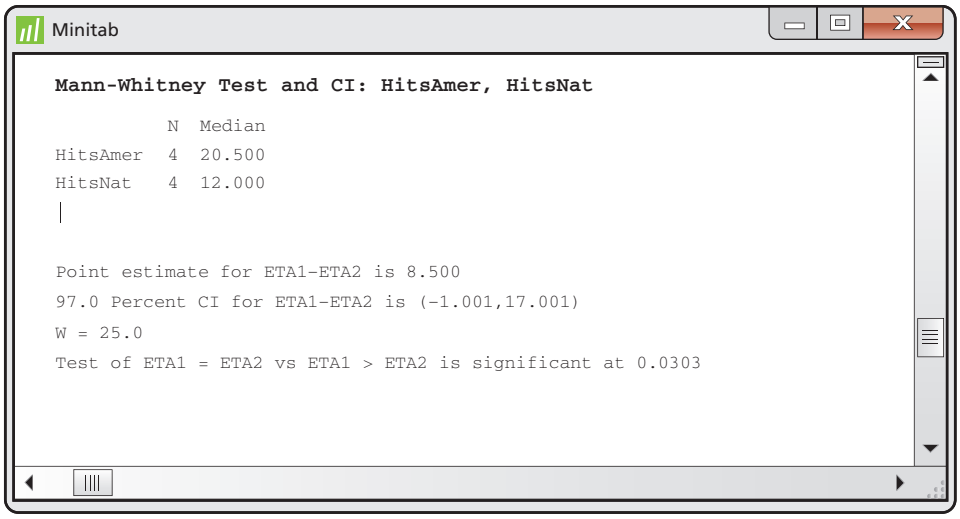
We recommend always using either the exact distribution (from software or tables) or the continuity correction for the rank sum statistic W. The exact distribution is safer for small samples. As Example 15.4 illustrates, however, the Normal approximation with the continuity correction is often adequate.

EXAMPLE

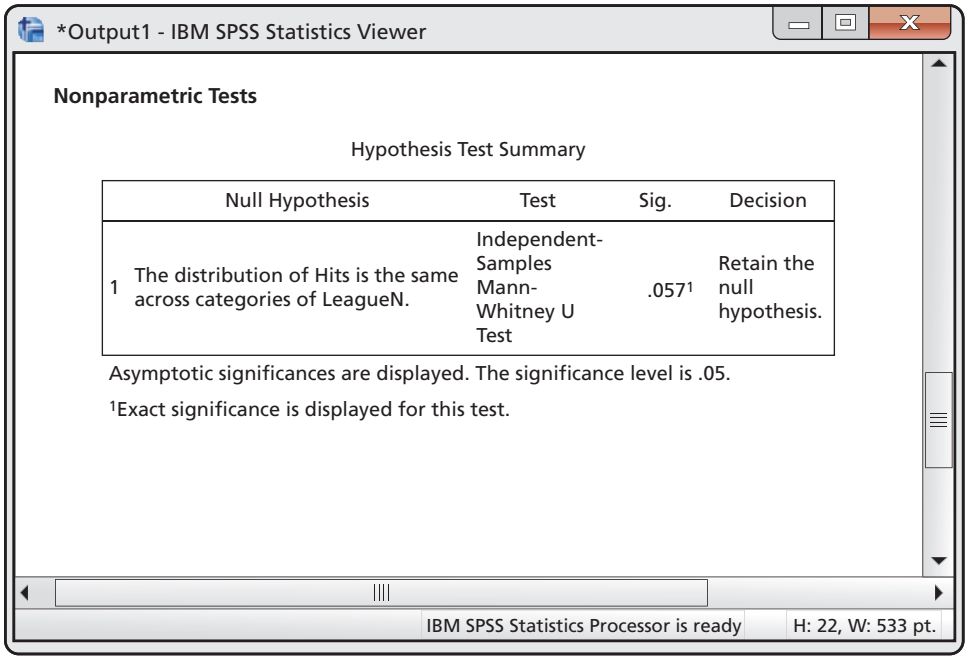
Mann-Whitney test

15.5 Software output. Figure 15.3 shows the output for our data from two additional statistical programs. Minitab gives the Normal approximation, and it refers to the **Mann-Whitney test.** This is an alternative form of the Wilcoxon rank sum test. SPSS uses the exact calculation for the P-value here but tests the null hypothesis only against the two-sided alternative.

FIGURE 15.3 Output from the Minitab and SPSS statistical software for the data in Example 15.1. (a) Minitab uses the Normal approximation for the distribution of W. (b) SPSS gives the exact value for the two-sided alternative.



(a) Minitab



(b) SPSS

What hypotheses does Wilcoxon test?

Our null hypothesis is that the distribution of hits is the same in the two leagues. Our alternative hypothesis is that there are more hits in the American League than in the National League. If we are willing to assume that hits are Normally distributed, or if we have reasonably large samples, we use the two-sample t test for means. Our hypotheses then become

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 > \mu_2$$

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

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

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

The Wilcoxon rank sum test does test hypotheses about population medians, but only if an additional assumption is met: both populations must have distributions of the same shape. That is, the density curve for hits in the American League must look exactly like that for the National League except that it may be shifted to the left or to the right. The Minitab output in Figure 15.3(a) states the hypotheses in terms of population medians (which it calls "ETA") and also gives a confidence interval for the difference between the two population medians.

The same-shape assumption is too strict to be reasonable in practice. Recall that our preferred version of the two-sample t test does not require that the two populations have the same standard deviation—that is, it does not make a same-shape assumption. Fortunately, the Wilcoxon test also applies in a much more general and more useful setting. It tests hypotheses that we can state in words as

H_0: The two distributions are the same.

H_a: One distribution has values that are systematically larger.

systematically larger

Here is a more exact statement of the **systematically larger** alternative hypothesis. Take X_1 to be hits in the American League and X_2 to be hits in the National League. These hits are random variables. That is, for each game in the American League, the number of hits is a value of the variable X_1. The probability that the number of hits is more than 15 is $P(X_1 > 15)$. Similarly, $P(X_2 > 15)$ is the corresponding probability for the National League. If the number of American League hits is "systematically larger" than the number of National League hits, getting more hits than 15 should be more likely in the American League. That is, we should have

$$P(X_1 > 15) > P(X_2 > 15)$$

The alternative hypothesis says that this inequality holds not just for 15 hits but for *any* number of hits.[2]

This exact statement of the hypotheses we are testing is a bit awkward. The 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 requirement. We recommend that you express the hypotheses in words rather than symbols. "The number of American League hits is systematically higher than the number of National League hits" is easy to understand and is a good statement of the effect that the Wilcoxon test looks for.

Ties

The exact distribution for the Wilcoxon rank sum is obtained assuming that all observations in both samples take different values. This allows us to rank them all. In practice, however, we often find observations tied at the same value.

average ranks

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 six observations:

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

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

The exact distribution for the Wilcoxon rank sum W changes if the data contain ties. Moreover, the standard deviation σ_W must be adjusted if ties are present. The Normal approximation can be used after the standard deviation is adjusted. Statistical software will detect ties, make the necessary adjustment, and switch to the Normal approximation. In practice, software is required if you want to use rank tests when the data contain tied values.

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

DATA EXERG

EXAMPLE

15.6 Exergaming in Canada. Exergames are active video games such as rhythmic dancing games, virtual bicycles, balance board simulators, and virtual sports simulators that require a screen and a console. A study of exergaming in students from grades 10 and 11 in Montreal, Canada, examined many factors related to participation in exergaming.[3] In Exercise 14.23 (page 14-22) we used logistic regression to examine the relationship between exergaming and time spent viewing television. Here are the data displayed in a two-way table of counts:

	TV time (hours per day)		
Exergamer	None	Some but less than 2 hours	2 hours or more
Yes	6	160	115
No	48	616	255

USE YOUR KNOWLEDGE

DATA EXERG

15.7 Analyze as a two-way table. Analyze the exergaming data in Example 15.6 as a two-way table.

(a) Compute the percents in the three categories of TV watching for the exergamers. Do the same for those who are not exergamers. Display the percents graphically and summarize the differences in the two distributions.

(b) Perform the chi-square test for the counts in the two-way table. Report the test statistic, the degrees of freedom, and the P-value. Give a brief summary of what you can conclude from this significance test.

LOOK BACK
chi-square test, p. 539

How do we approach the analysis of these data using the Wilcoxon test? We start with the hypotheses. We have two distributions of TV viewing, one for the exergamers and one for those who are not exergamers. The null hypothesis states that these two distributions are the same. The alternative hypothesis uses the fact that the responses are ordered from no TV to 2 hours or more per day. It states that one of the exerciser groups watches more TV than the other.

H_0: The amount of time spent viewing TV is the same for students who are exergamers and students who are not.

H_a: One of the two groups views more TV than the other.

The alternative hypothesis is two-sided. Because the responses can take only three values, there are very many ties. All 54 students who watch no TV are tied. Similarly, all students in each of the other two columns of the table are tied. The graphical display that you prepared in Exercise 15.7 suggests that the exergamers watch more TV than those who are not exergamers. Is this difference statistically significant?

EXAMPLE

EXERG

15.7 Software output. Look at Figure 15.4, which gives SAS output for the Wilcoxon test. The rank sum for the exergamers (using average ranks for ties) is $W = 187{,}747.5$. The expected rank sum under the null hypothesis is 168,740.5, so the exergamers have a higher rank sum than we would expect. The Normal approximation test statistic is $z = 4.47$ and the two-sided P-value is reported as $P < 0.0001$. There is very strong evidence of a difference. Exergamers watch more TV than the students who are not exergamers.

We can use our framework of "systematically larger" (page 15-10) to summarize these data. For the exergamers, 98% watch some TV and 41% watch two or more hours per day. The corresponding percents for the students who are not exergamers are 95% and 28%.

In our discussion of TV viewing and exergaming, we have expressed results in terms of the amount of TV watched. In fact, we do not have the actual hours of TV watched by each student in the study. Only data with the hours classified into three groups are available. Many government surveys summarize quantitative data categorized into ranges of values. *When summarizing the analysis of data, it is very important to explain clearly how the data are recorded.* In this setting, we have chosen to use phrases such as "watch more TV" because they express the findings based on the data available.

Note that the two-sample t test would not be appropriate in this setting. If we coded the TV-watching categories as 1, 2, and 3, the average of these coded values would not be meaningful.

On the other hand, we frequently encounter variables measured in scales such as "strongly agree," "agree," "neither agree nor disagree," "disagree," and "strongly disagree." In these circumstances, many would code the responses with the integers 1 to 5 and then use standard methods such as a t test or ANOVA. Whether to do this or not is a matter of judgment. Rank tests avoid

FIGURE 15.4 Output from SAS for the exergaming data, for Example 15.7.

SAS

The SAS System

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable TVN Classified by Variable Exergame

Exergame	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Yes	281	187747.50	168740.50	4253.97554	668.140569
No	919	532852.50	551859.50	4253.97554	579.817737

Average scores were used for ties.

Wilcoxon Two-Sample Test	
Statistic (S)	187747.5000
Normal Approximation	
Z	4.4679
One-Sided Pr > Z	<.0001
Two-Sided Pr > |Z|	<.0001
t Approximation	
One-Sided Pr > Z	<.0001
Two-Sided Pr > |Z|	<.0001
Exact Test	
One-Sided Pr ≥ S	4.899E-06
Two-Sided Pr ≥ |S – Mean|	7.713E-06
Z includes a continuity correction of 0.5.	

Done

the dilemma because they use only the order of the responses, not their actual values. *Some statisticians use t procedures when there is not a fully meaningful scale of measurement, but others avoid them.*

Rank, *t*, and permutation tests

The two-sample *t* procedures are the most common method for comparing the centers of two populations based on random samples from each. The Wilcoxon rank sum test is a competing procedure that does not start from the condition that the populations have Normal distributions. Permutation tests (Chapter 16) also avoid the need for Normality. Tests based on Normality, rank tests, and permutation tests apply in many other settings as well. How do these three approaches compare in general?

First, let's consider rank tests versus traditional tests based on Normal distributions. Both are available in almost all statistical software.

• Moving from the actual data values to their ranks allows us to find an exact sampling distribution for rank statistics such as the Wilcoxon rank sum W when the null hypothesis is true. (Most software will do this only if there are no ties and if the samples are quite small.) When our samples are small, are truly random samples from the populations, and show non-Normal distributions of the same shape, the Wilcoxon test is more reliable than the two-sample t test. In practice, the robustness of t procedures implies that we rarely encounter data that require nonparametric procedures to obtain reasonably accurate P-values. The t and W tests gave very similar results for the baseball hit data in Example 15.1, but we would not use a t procedure for the exergame data in Example 15.6.

• Normal tests compare means and are accompanied by simple confidence intervals for means or differences between means. When we use rank tests to compare medians, we can also give confidence intervals for medians. However, the usefulness of rank tests is clearest in settings when they do not simply compare medians—see the discussion "What Hypotheses Does Wilcoxon Test?" (page 15-9). Rank methods focus on significance tests, not confidence intervals.

• Inference based on ranks is largely restricted to simple settings. Normal inference extends to methods for use with complex experimental designs and multiple regression, but nonparametric tests do not. We stress Normal inference in part because it leads to more advanced statistics.

If you read Chapter 16 and use software that makes permutation tests available to you, you will also want to compare rank tests with resampling methods.

• Both rank and permutation tests are nonparametric. That is, they require no assumptions about the shape of the population distribution. A two-sample permutation test has the same null hypothesis as the Wilcoxon rank sum test: that the two population distributions are identical. Calculation of the sampling distribution under the null hypothesis is similar for both tests but is simpler for rank tests because it depends only on the sizes of the samples. As a result, software often gives exact P-values for rank tests but not for permutation tests.

• Permutation tests have the advantage of flexibility. They allow wide choice of the statistic used to compare two samples, an advantage over both the t and Wilcoxon tests. In fact, we could apply the permutation test method to sample means (imitating t) or to rank sums (imitating Wilcoxon), as well as to other statistics such as the trimmed mean that we used in Exercise 1.99. Permutation tests are not available in some settings, such as testing hypotheses about a single population, though bootstrap confidence intervals do allow resampling tests in these settings. Permutation tests are available for multiple regression and some other quite elaborate settings.

LOOK BACK
trimmed mean, p. 53

• An important advantage of resampling methods over both Normal and rank procedures is that we can get bootstrap confidence intervals for the parameter corresponding to whatever statistic we choose for the

permutation test. If the samples are very small, however, bootstrap confidence intervals may be unreliable because the samples don't represent the population well enough to provide a good basis for bootstrapping.

In general, both Normal distribution methods and resampling methods are more useful than rank tests. *If you are familiar with resampling, we recommend rank tests only for very small samples, and even then only if your software gives exact P-values for rank tests but not for permutation tests.*

SECTION 15.1 Summary

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

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

The **Wilcoxon rank sum test** compares two distributions to assess whether one has systematically larger values than the other. The Wilcoxon test is based on the **Wilcoxon rank sum statistic *W***, which is the sum of the ranks of one of the samples. The Wilcoxon test can replace the **two-sample *t* test.**

***P*-values for the Wilcoxon test** are based on the sampling distribution of the rank sum statistic W when the null hypothesis (no difference in distributions) is true. You can find P-values from special tables, software, or a Normal approximation (with continuity correction).

SECTION 15.1 Exercises

For Exercises 15.1 and 15.2, see page 15-4; for Exercises 15.3 and 15.4, see page 15-6; for Exercises 15.5 and 15.6, see page 15-8; and for Exercise 15.7, see page 15-11.

15.8 Time spent studying. Students in a large first-year college class were asked how much time they spent studying on a typical weeknight. Here are the responses, in minutes, for five female students in the class: STUDYT

120 360 115 60 170

Find the ranks for these data.

15.9 Find the rank sum statistic. Refer to the previous exercise. Here are the data for six men in the class: STUDYT

0 300 75 90 30 130

Compute the value of the Wilcoxon statistic. Take the first sample to be the women.

15.10 State the hypotheses. Refer to the previous exercise. State appropriate null and alternative hypotheses for this setting. STUDYT

15.11 Find the mean and standard deviation of the distribution of the statistic. The statistic W that you calculated in Exercise 15.9 is a random variable with a sampling distribution. What are the mean and the standard deviation of this sampling distribution under the null hypothesis? STUDYT

15.12 Find the *P*-value. Refer to Exercises 15.8 to 15.11. Find the P-value using the Normal approximation with the continuity correction and interpret the result of the significance test. STUDYT

15.13 Is civic engagement related to education? A Pew Internet Poll of adults aged 18 and older examined factors related to civic engagement. Participants were asked whether or not they had participated in a civic group or activity in the preceding 12 months. One analysis looked at the relationship between this variable and education. Here are the data:[4] CIVIC

	Education			
Civic participation	No high school	High school	Some college	College
Civic	76	294	295	428
No civic	155	424	273	298

FIGURE 15.5 Output from SAS for the civic participation data, for Exercise 15.13.

SAS

The SAS System

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable EdN Classified by Variable Group

Group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Civic	1093	1351159.50	1226346.0	14672.9591	1236.19350
NoCivic	1150	1165486.50	1290300.0	14672.9591	1013.46652

Average scores were used for ties.

Wilcoxon Two-Sample Test	
Statistic (S)	1351159.5000
Normal Approximation	
Z	8.5063
One-Sided Pr > Z	<.0001
Two-Sided Pr > |Z|	<.0001
t Approximation	
One-Sided Pr > Z	<.0001
Two-Sided Pr > |Z|	<.0001
Exact Test	
One-Sided Pr ≥ S	6.557E-18
Two-Sided Pr ≥ |S – Mean|	1.316E-17

Z includes a continuity correction of 0.5.

Figure 15.5 gives the SAS output for analyzing these data using the Wilcoxon rank sum procedure.

(a) Describe the relevant parts of the output and write a short summary of the results.

(b) Apply the "systematically larger" framework that we used in Example 15.7 (page 15-12) to these data. Is this a useful way to describe the results of this analysis? Give reasons for your answer.

15.14 Do women talk more? Conventional wisdom suggests that women are more talkative than men. One study designed to examine this stereotype collected data on the speech of 10 men and 10 women in the United States.[5] The variable recorded is the number of words per day. Here are the data: TALK10

Men				Women			
23,871	5,180	9,951	12,460	10,592	24,608	13,739	22,376
17,155	10,344	9,811	12,387	9,351	7,694	16,812	21,066
29,920	21,791			32,291	12,320		

(a) Summarize the data for the two groups using w numerical and graphical methods. Describe the two distributions.

(b) Compare the words per day spoken by the men with the words per day spoken by the women using the Wilcoxon rank sum test. Summarize your results and conclusion in a short paragraph.

15.15 More data for women and men talking. The data in the previous exercise were a sample of the data collected in a larger study of 42 men and 37 women. Use the larger data set to answer the questions in the previous exercise. Discuss the advisability of using the Wilcoxon test versus the t test for this exercise and for the previous one. TALK

15.16 Learning math through subliminal messages. A "subliminal" message is below our threshold of awareness but may nonetheless influence us. Can subliminal messages help students learn math? A group of students who had failed the mathematics part of the City University of New York Skills Assessment Test agreed to participate in a study to find out. All received a daily subliminal message, flashed on a screen too rapidly to be consciously read. The treatment group of 10 students was exposed to "Each day I am getting better in math." The control group of 8 students was exposed to a neutral message, "People are walking on the street." All students participated in a summer program designed to raise their math skills, and all took the assessment test again at the end of the program. Here are data on the subjects' scores before and after the program:[6] SUBLIM

Treatment Group		Control Group	
Pretest	Posttest	Pretest	Posttest
18	24	18	29
18	25	24	29
21	33	20	24
18	29	18	26
18	33	24	38
20	36	22	27
23	34	15	22
23	36	19	31
21	34		
17	27		

(a) The study design was a randomized comparative experiment. Outline this design.

(b) Compare the gain in scores in the two groups, using a graph and numerical descriptions. Does it appear that the treatment group's scores rose more than the scores for the control group?

(c) Apply the Wilcoxon rank sum test to the posttest versus pretest differences. Note that there are some ties. What do you conclude?

15.17 Storytelling and the use of language. A study of early childhood education asked kindergarten students to retell two 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 each child and assigned a score for certain uses of language. Here are the data:[7] STORY

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

Is there evidence that the scores of high-progress readers are higher than those of low-progress readers when they retell a story they have heard without pictures (Story 1)?

(a) Make Normal quantile plots for the 5 responses in each group. Are any major deviations from Normality apparent?

(b) Carry out a two-sample t test. State hypotheses and give the two sample means, the t statistic and its P-value, and your conclusion.

(c) Carry out the Wilcoxon rank sum test. State hypotheses and give the rank sum W for high-progress readers, its P-value, and your conclusion. Do the t and Wilcoxon tests lead you to different conclusions?

15.18 Repeat the analysis for Story 2. Repeat the analysis of Exercise 15.17 for the scores when children retell a story they have heard and seen illustrated with pictures (Story 2). STORY

15.19 Do the calculations by hand. Use the data in Exercise 15.17 for children telling Story 2 to carry out by hand the steps in the Wilcoxon rank sum test. STORY

(a) Arrange the 10 observations in order and assign ranks. There are no ties.

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

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

(d) The data for Story 1 contain tied observations. What ranks would you assign to the 10 scores for Story 1?

15.2 The Wilcoxon Signed Rank Test

When you complete this section, you will be able to

- **For a set of paired sample data, take the differences between the pairs, take the absolute values of the differences, put the absolute values of the differences in order, from smallest to largest with an indication of which absolute differences were from positive differences.**
- **Compute the Wilcoxon signed rank statistic W^+ from an ordered list of differences with an indication of which absolute differences were from positive differences.**
- **State the null and alternative hypotheses that are used for the analysis of data using the Wilcoxon signed rank test.**
- **Using the sample size (that is, the number of pairs), find the mean and the standard deviation of the sampling distribution of the Wilcoxon signed rank statistic under the null hypothesis.**
- **Find the *P*-value for the Wilcoxon signed rank significance test using the Normal approximation with the continuity correction.**
- **For the Wilcoxon signed rank test, use computer output to determine the results of the significance test.**
- **Test a hypothesis about the median of a distribution using the Wilcoxon signed rank test.**

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

EXAMPLE

15.8 Storytelling and reading. A study of early childhood education asked kindergarten students to retell two fairy tales that had been read to them earlier in the week. The first (Story 1) had been read to them, and the second (Story 2) had been read but also illustrated with pictures. An expert listened to recordings of the children retelling each story and assigned a score for certain uses of language. Here are the data for five "low-progress" readers in a pilot study:[8] Higher scores are better.

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. 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.

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 $P = 0.280$. Displays of the data (Figure 15.6) suggest some lack of Normality. We would therefore like to use a rank test.

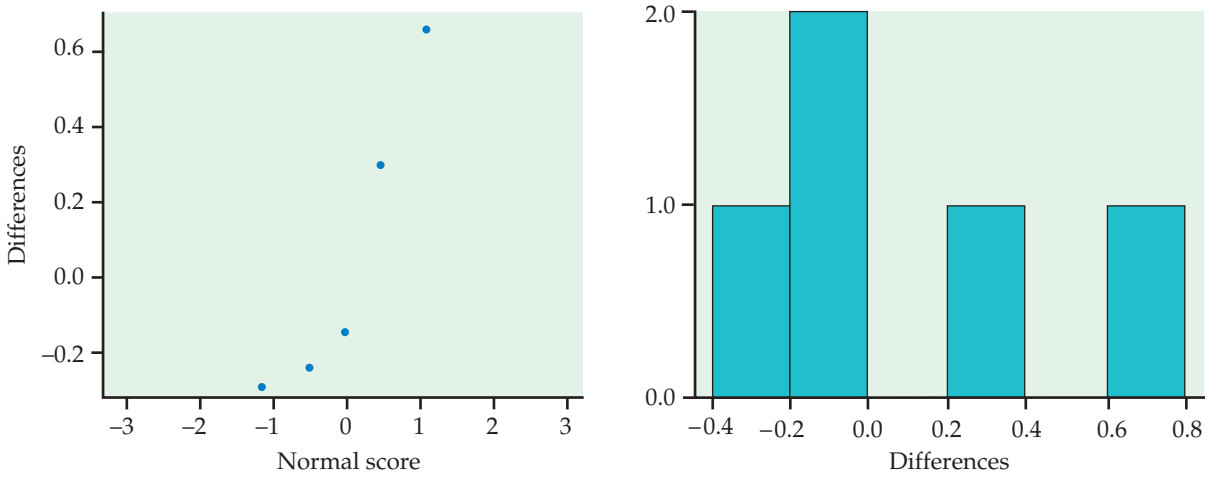


FIGURE 15.6 Normal quantile plot and histogram for the five differences in story scores, for Example 15.8.

absolute value

Positive differences in Example 15.8 indicate that the child performed better telling Story 2. If scores are generally higher with illustrations, the positive differences should be farther from zero in the positive direction than the negative differences are in the negative direction. We therefore compare the **absolute values** of the differences, that is, their magnitudes without a sign. Here they are, with boldface indicating the positive values:

0.37 0.23 **0.66** 0.08 0.17

Arrange these in increasing order and assign ranks, keeping track of which values were originally positive. Tied values receive the average of their ranks. If there are cases with 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.

USE YOUR KNOWLEDGE

SPAS3

15.20 Service and food provided by top 25 spas. The readers' poll in *Condé Nast Traveler* magazine that ranked 100 top resort spas and that was described in Exercise 15.1 also reported scores on service and on food. Here are the scores for a random sample of 7 spas that ranked in the top 25:

Spa	1	2	3	4	5	6	7
Service	89.6	89.8	87.3	94.2	95.8	87.9	91.0
Food	83.1	88.1	85.8	92.9	95.7	80.7	83.6

Is service more important than food for a top ranking? Formulate this question in terms of null and alternative hypotheses. Then compute the differences and find the value of the Wilcoxon signed rank statistic, W^+.

SPAS4

15.21 Scores for the next 25 spas. Refer to the previous exercise. Here are the scores for a random sample of 7 spas that ranked between 26 and 50:

Spa	1	2	3	4	5	6	7
Service	90.6	87.2	95.0	88.4	91.5	88.2	91.2
Food	86.6	74.4	89.1	81.0	85.7	83.2	93.1

Answer the questions from the previous exercise for this setting.

EXAMPLE

15.9 Software output. In the storytelling study of Example 15.8, $n = 5$. If the null hypothesis (no systematic effect of illustrations) is true, the mean of the signed rank statistic is

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

Our observed value $W^+ = 9$ is only slightly larger than this mean. The one-sided P-value is $P(W^+ \geq 9)$.

Most statistical software uses the differences between the two variables, *with the signs,* as input. Alternatively, the differences can sometimes be calculated within the software. Figure 15.7 displays the output from three statistical programs. Each does things a little differently. The Minitab output in Figure 15.7(a) gives $P = 0.394$ for the one-sided Wilcoxon signed rank test with $n = 5$ observations and $W^+ = 9$. In Figure 15.7(b), the SPSS output gives $P = 0.686$ for testing the two-sided alternative. The results from SAS in Figure 15.7(c) are part of the usual output for the analysis of a single

FIGURE 15.7 Output from (a) Minitab, (b) SPSS, and (c) SAS for the storytelling data, for Example 15.9.

Minitab

```
Test of median = 0.000000 versus median > 0.000000

                  N for    Wilcoxon          Estimated
          N  N*    test   Statistic      P     Median
DiffLow   5   5       5         9.0  0.394     0.1000
```

(a) Minitab

*Output1 - IBM SPSS Statistics Viewer

Nonparametric Tests

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The median of DiffLow equals 0.00.	One-Sample Wilcoxon Signed Rank Test	.686	Retain the null hypothesis.

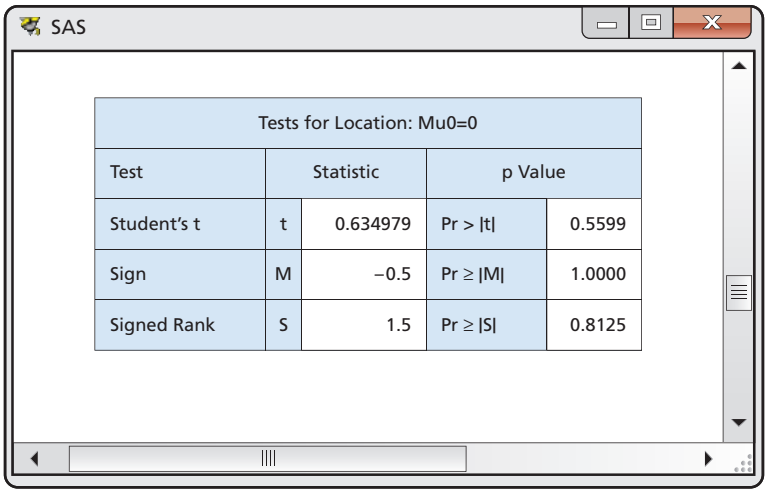
Asymptotic significances are displayed. The significance level is .05.

IBM SPSS Statistics Processor is ready

(b) SPSS

(*Continued*)

FIGURE 15.7 (*Continued*)



(c) SAS

variable. The two-sided alternative is used. The test statistic for the signed rank test is given as $S = 1.5$. This quantity is W^+ minus its expected value $\mu_{W^+} = 7.5$, $S = W^+ - \mu_{W^+}$. The P-value is given as $P = 0.8125$.

Results reported in the three outputs lead us to the same qualitative conclusion: the data do not provide evidence to support the idea that the Story 2 scores are higher than (or not equal to) the Story 1 scores. Different methods and approximations are used to compute the P-values. With larger sample sizes, we would not expect so much variation in the P-values. Note that the t test results reported in SAS also give the same conclusion, $P = 0.5599$.

When the sampling distribution of a test statistic is symmetric, we can use output that gives a P-value for a two-sided alternative to compute a P-value for a one-sided alternative. Check that the effect is in the direction specified by the one-sided alternative and then divide the P-value by 2.

The Normal approximation

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

15.10 The Normal approximation. For $n = 5$ observations, we saw in Example 15.9 that $\mu_{W^+} = 7.5$. The standard deviation of W^+ under the null hypothesis is

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

$$= \sqrt{\frac{(5)(6)(11)}{24}}$$

$$= \sqrt{13.75} = 3.708$$

The continuity correction calculates the P-value $P(W^+ \geq 9)$ as $P(W^+ \geq 8.5)$, treating the value $W^+ = 9$ as occupying the interval from 8.5 to 9.5. We find the Normal approximation for the P-value by standardizing and using the standard Normal table:

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

Despite the small sample size, the Normal approximation gives a result quite close to the exact value $P = 0.4062$. Figure 15.7(b) shows that the approximation is much less accurate without the continuity correction. *This output reminds us not to trust software unless we know exactly what it does.*

USE YOUR KNOWLEDGE

15.22 Significance test for top-ranked spas. Refer to Exercise 15.20 (page 15-20). Find μ_{W^+}, σ_{W^+}, and the Normal approximation for the P-value for the Wilcoxon signed rank test.

SPAS3

15.23 Significance test for lower-ranked spas. Refer to Exercise 15.21 (page 15-20). Find μ_{W^+}, σ_{W^+}, and the Normal approximation for the P-value for the Wilcoxon signed rank test.

SPAS4

Ties

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, the usual procedure simply drops such pairs from the sample. *This amounts to dropping observations that favor the null hypothesis (no difference). If there are many ties, the test may be biased in favor of the alternative hypothesis.* As in the case of the Wilcoxon rank sum, ties complicate finding a P-value. Most software no longer provides an exact distribution for the signed rank statistic W^+, and the standard deviation σ_{W^+} must be adjusted for the ties before we can use the Normal approximation. Software will do this. Here is an example.

EXAMPLE

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

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. We see that 6 of the 12 golfers improved their scores. We would like to test the hypotheses that in a large population of collegiate women golfers

H_0: Scores have the same distribution in Rounds 1 and 2.

H_a: Scores are systematically lower or higher in Round 2.

A Normal quantile plot of the differences (Figure 15.8) shows some irregularity and a low outlier. We will use the Wilcoxon signed rank test.

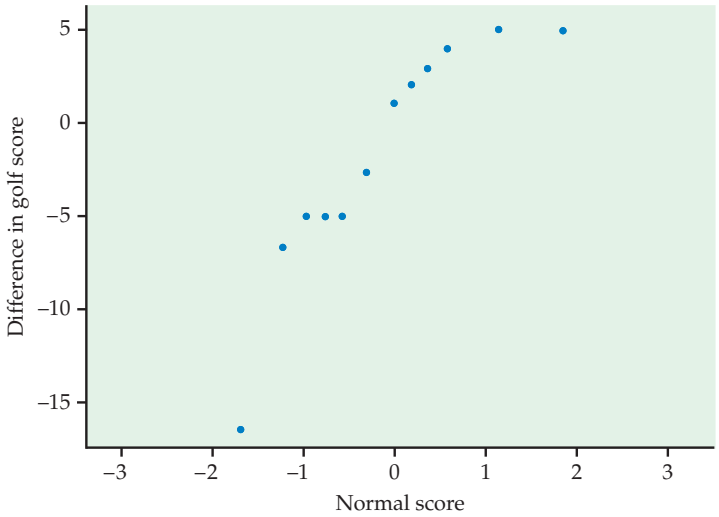


FIGURE 15.8 Normal quantile plot of the difference in scores for two rounds of a golf tournament, for Example 15.11.

The absolute values of the differences, with boldface indicating those that are 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 of the ranks of the negative differences. (We could equally well use the sum of the ranks of the positive differences.) Its value is $W^+ = 50.5$.

EXAMPLE

15.12 Software output. Here are the two-sided *P*-values for the Wilcoxon signed rank test for the golf score data from three statistical programs:

Program	*P*-value
Minitab	$P = 0.388$
SAS	$P = 0.388$
SPSS	$P = 0.363$

All lead to the same practical conclusion: these data give no evidence for a systematic change in scores between rounds. However, the *P*-value reported by SPSS differs a bit from the other two. The reason for the variation is that the programs use slightly different versions of the approximate calculations needed when ties are present. The exact result depends on which version the software programmer chose to use.

For the golf data, the matched pairs t test gives $t = 0.9314$ with $P = 0.3716$. Once again, t and W^+ lead to the same conclusion.

Testing a hypothesis about the median of a distribution

Let's take another look at how the Wilcoxon signed rank test works. We have data for a pair of variables measured on the same individuals. The analysis starts with the differences between the two variables. These differences are what we input to statistical software.

At this stage we can think of our data as consisting of a single variable. The Wilcoxon signed rank test tests the null hypothesis that the population median of the differences is zero. The alternative is that the median is not zero.

Think about starting the analysis at the stage where we have a single variable and we are interested in testing a hypothesis about the median. The null hypothesis does not necessarily need to be zero. If it is some other value, we simply subtract that value from each observation before we start the analysis. Exercise 15.35 (page 15-27) leads you through the steps needed for this analysis.

SECTION 15.2 Summary

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** and the **sign test** are alternative tests 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).

SECTION 15.2 Exercises

For Exercises 15.20 and 15.21, see page 15-20; and for Exercises 15.22 and 15.23, see page 15-23.

15.24 Fuel efficiency. Computers in some vehicles calculate various quantities related to performance. One of these is the fuel efficiency, or gas mileage, usually expressed as miles per gallon (mpg). For one vehicle equipped in this way, the mpg were recorded each time the gas tank was filled, and the computer was then reset. In addition to the computer calculating mpg, the driver also recorded the mpg by dividing the miles driven by the number of gallons at fill-up.[9] The driver wants to determine if these calculations are different. MPG8

Fill-up	1	2	3	4	5	6	7	8
Computer	41.5	50.7	36.6	37.3	34.2	45.0	48.0	43.2
Driver	36.5	44.2	37.2	35.6	30.5	40.5	40.0	41.0

(a) For each of the eight fill-ups find the difference between the computer mpg and the driver mpg.

(b) Find the absolute values of the differences you found in part (a).

(c) Order the absolute values of the differences that you found in part (b) from smallest to largest, and underline those absolute differences that came from positive differences in part (a).

15.25 Find the Wilcoxon signed rank statistic. Using the work that you performed in the previous exercise, find the value of the Wilcoxon signed rank statistic W^+.

15.26 State the hypotheses. Refer to Exercise 15.24. State the null hypothesis and the alternative hypothesis for this setting.

15.27 Find the mean and the standard deviation. Refer to Exercise 15.24. Use the sample size to find the mean and the standard deviation of the sampling distribution of the Wilcoxon signed rank statistic W^+ under the null hypothesis.

15.28 Find the *P*-value. Refer to Exercises 15.24 to 15.27. Find the *P*-value for the Wilcoxon signed rank statistic using the Normal approximation with the continuity correction.

15.29 Read the output. The data in Exercise 15.24 are a subset of a larger set of data. Figure 15.9 gives Minitab output for the analysis of this larger set of data. MPGCOMP

(a) How many pairs of observations are in the larger data set?

(b) What is the value of the Wilcoxon signed rank statistic W^+?

(c) Report the *P*-value for the significance test and give a brief statement of your conclusion.

Minitab

```
Wilcoxon Signed Rank Test: Diff

Test of median = 0.000000 versus median not = 0.000000

             N for    Wilcoxon          Estimated
        N     Test   Statistic      P     Median
Diff   20       20       192.0  0.001      2.925
```

FIGURE 15.9 Minitab output for the fuel efficiency data, for Exercise 15.29.

(d) The output reports an estimated median. Explain how this statistic is calculated from the data.

15.30 Number of friends on Facebook. Facebook recently examined all active Facebook users (more than 10% of the global population) and determined that the average user has 190 friends. This distribution takes only integer values, so it is certainly not Normal. It is also highly skewed to the right, with a median of 100 friends.[10] Consider the following SRS of $n = 30$ Facebook users from your large university. FACEFR

594	60	417	120	132	176	516	319	734	8
31	325	52	63	537	27	368	11	12	190
85	165	288	65	57	81	257	24	297	148

(a) Use the Wilcoxon signed rank procedure to test the null hypothesis that the median number of Facebook friends for Facebook users at your university is 190. Describe the steps in the procedure and summarize the results.

(b) Exercise 7.26 (page 442) asked you to analyze these data using the *t* procedure. Perform this analysis and compare the results with those that you found in part (a).

15.31 The full moon and behavior. Can the full moon influence behavior? A study observed 15 nursing-home patients with dementia. The number of incidents of aggressive behavior was recorded each day for 12 weeks. Call a day a "moon day" if it is the day of a full moon or the day before or after a full moon. Here are the average numbers of aggressive incidents for moon days and other days for each subject:[11] MOON

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

The matched pairs t test (Example 7.7, page 429) gives $P < 0.000015$, and a permutation test (Example 16.14, page 16-50) gives $P = 0.0001$. Does the Wilcoxon signed rank test, based on ranks rather than means, agree that there is strong evidence that there are more aggressive incidents on moon days?

15.32 Comparison of two energy drinks. Consider the following study to compare two popular energy drinks. For each subject, a coin was flipped to determine which drink to rate first. Each drink was rated on a 0 to 100 scale, with 100 being the highest rating. ENERDR6

	Subject					
Drink	1	2	3	4	5	6
A	43	83	66	87	78	67
B	45	78	64	79	71	62

(a) Inspect the data. Is there a tendency for these subjects to prefer one of the two energy drinks?

(b) Use the matched pairs t test of Chapter 7 (page 429) to compare the two drinks.

(c) Use the Wilcoxon signed rank test to compare the two drinks.

(d) Write a summary of your results and explain why the two tests give different conclusions.

15.33 Comparison of two energy drinks with an additional subject. Refer to the previous exercise. Let's suppose that there is an additional subject who expresses a strong preference for energy drink "A." Here is the new data set: ENERDR7

	Subject						
Drink	1	2	3	4	5	6	7
A	43	83	66	87	78	67	90
B	45	78	64	79	71	62	60

Answer the questions given in the previous exercise. Write a summary comparing this exercise with the previous one. Include a discussion of what you have learned regarding the choice of the t test versus the Wilcoxon signed rank test for different sets of data.

15.34 A summer language institute for teachers. A matched pairs study of the effect of a summer language institute on the ability of teachers to comprehend spoken French had these improvements in scores between the pretest and the posttest for 20 teachers: SUMLANG

2	0	6	6	3	3	2	3	−6	6
6	6	3	0	1	1	0	2	3	3

(Exercise 7.45, page 446, applies the t test to these data; Exercise 16.59, page 16-49, applies a permutation test based on the means.) Show the assignment of ranks and the calculation of the signed rank statistic W^+ for these data. Remember that zeros are dropped from the data before ranking, so that n is the number of nonzero differences within pairs.

15.35 Radon detectors. How accurate are radon detectors of a type sold to homeowners? To answer this question, university researchers placed 12 detectors in a chamber that exposed them to 105 picocuries per liter (pCi/l) of radon.[12] The detector readings are as follows: RADON

91.9	97.8	111.4	122.3	105.4	95.0
103.8	99.6	96.6	119.3	104.8	101.7

We wonder if the median reading differs significantly from the true value 105.

(a) Graph the data, and comment on skewness and outliers. A rank test is appropriate.

(b) We would like to test hypotheses about the median reading from home radon detectors:

$$H_0: \text{median} = 105$$

$$H_a: \text{median} \neq 105$$

To do this, apply the Wilcoxon signed rank statistic to the differences between the observations and 105. (This is the one-sample version of the test.) What do you conclude?

15.36 Vitamin C in wheat-soy blend. The U.S. Agency for International Development provides large quantities of wheat-soy blend (WSB) for development programs and emergency relief in countries throughout the world. One study collected data on the vitamin C content of 5 bags of WSB at the factory and five months later in Haiti.[13] Here are the data: WSBVITC

Sample	1	2	3	4	5
Before	73	79	86	88	78
After	20	27	29	36	17

We want to know if vitamin C has been lost during transportation and storage. Describe what the data show about this question. Then use a rank test to see whether there has been a significant loss.

15.3 The Kruskal-Wallis Test*

When you complete this section, you will be able to

- **Describe the setting where the Kruskal-Wallis test can be used.**
- **Specify the null and alternative hypotheses for the Kruskal-Wallis test.**
- **For the Kruskal-Wallace test, use computer output to determine the results of the significance test.**

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

EXAMPLE

WEEDS

15.13 Weeds and corn yield. 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 and then randomly assigned the plots to four groups. He weeded the plots by hand to allow a fixed number of lamb's-quarter plants to grow in each meter of corn row. These numbers were 0, 1, 3, and 9 in the four groups of plots. No other weeds were allowed to grow, and all plots received identical treatment except for the weeds. Here are the yields of corn (bushels per acre) in each of the plots:[14]

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

The summary statistics are

Weeds	n	Mean	Std. dev.
0	4	170.200	5.422
1	4	162.825	4.469
3	4	161.025	10.493
9	4	157.575	10.118

*Because this test is an alternative to the one-way analysis of variance F test, you should first read Chapter 12.

The sample standard deviations do not satisfy our rule of thumb that for safe use of ANOVA the largest should not exceed twice the smallest. A careful look at the data suggests that there may be some outliers in the 3 and 9 weeds per meter groups. These are the correct yields for their plots, so we have no justification for removing them. Let's use a rank test that is not sensitive to outliers.

Hypotheses and assumptions

The ANOVA F test concerns the means of the several populations represented by our samples. For Example 15.13, the ANOVA hypotheses are

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

H_a: not all four means are equal

Here, μ_0 is the mean yield in the population of all corn planted under the conditions of the experiment with no weeds present. The data should consist of four independent random samples from the four populations, all Normally distributed with the same standard deviation.

The **Kruskal-Wallis test** is a rank test that can replace the ANOVA F test. The assumption about data production (independent random samples from each population) remains important, but we can relax the Normality assumption. 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 Kruskal-Wallis test

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

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

THE KRUSKAL-WALLIS TEST

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

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

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

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

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

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

EXAMPLE

WEEDS

15.14 Perform the significance test. In Example 15.13, 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

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

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

Weeds	Ranks				Rank sums
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

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

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

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

$$= 5.56$$

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

Figure 15.10 displays the output from Minitab, SPSS, and SAS for the analysis of the data in Example 15.14. Minitab gives the H statistic adjusted for ties as $H = 5.57$ with 3 degrees of freedom and $P = 0.134$. SPSS reports the same P-value. SAS reports a chi-square statistic with 3 degrees of freedom and $P = 0.1344$. All agree that there is not sufficient evidence in the data to reject the null hypothesis that the number of weeds per meter has no effect on the yield.

FIGURE 15.10 Output from (a) Minitab, (b) SPSS, and (c) SAS for the Kruskal-Wallis test applied to the weed data, for Example 15.14.

Minitab

Kruskal-Wallis Test: Yield versus Weeds

Kruskal-Wallis Test on Yield

Weeds	N	Median	Ave Rank	Z
0	4	169.4	13.1	2.24
1	4	163.6	8.4	-0.06
3	4	157.3	6.3	-1.09
9	4	162.6	6.3	-1.09
Overall	16		8.5	

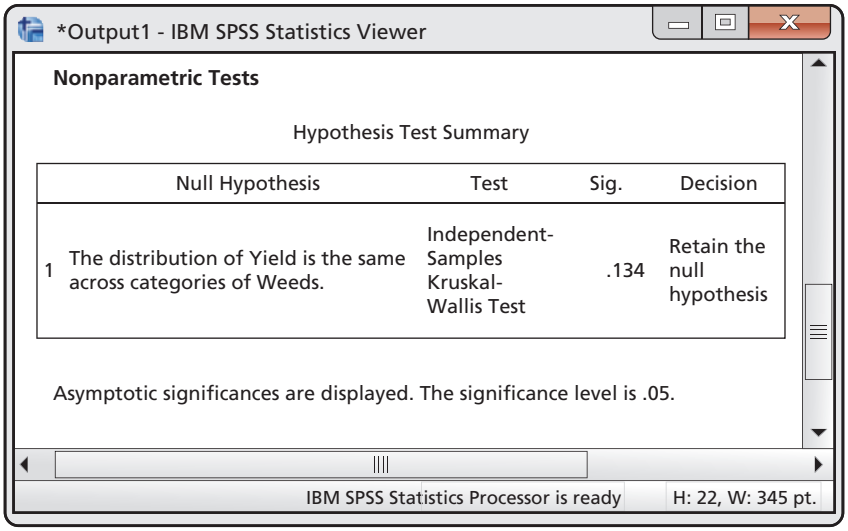
H = 5.56 DF = 3 P = 0.135
H = 5.57 DF = 3 P = 0.134 (adjusted for ties)

* NOTE * One or more small samples

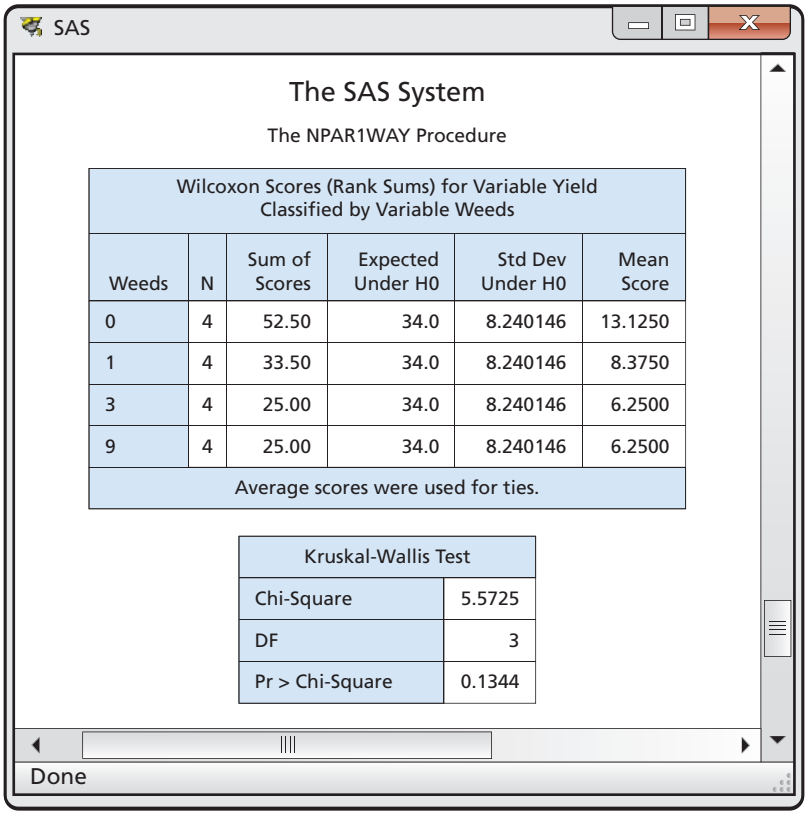
(a) Minitab

(*Continued*)

FIGURE 15.10 (*Continued*)



(b) SPSS



(c) SAS

SECTION 15.3 Summary

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, H for comparing I populations has approximately the chi-square distribution with $I - 1$ degrees of freedom. We use this approximate distribution to obtain P-values.

SECTION 15.3 Exercises

15.37 Number of Facebook friends. An experiment was run to examine the relationship between the number of Facebook friends and the user's perceived social attractiveness.[15] A total of 134 undergraduate participants were randomly assigned to observe one of five Facebook profiles. Everything about the profile was the same except the number of friends, which appeared on the profile as 102, 302, 502, 702, or 902. After viewing the profile, each participant was asked to fill out a questionnaire on the physical and social attractiveness of the profile user. Each attractiveness score is an average of several seven-point questionnaire items, ranging from 1 (strongly disagree) to 7 (strongly agree). In Example 12.3 (page 648), we analyzed these data using a one-way ANOVA. Explain the setting for this problem. Include the number of groups to be compared, assumptions about independence, and the distribution of the distributions. FRIENDS

15.38 What are the hypotheses? Refer to the previous exercise. What are the null hypothesis and the alternative hypothesis? Explain why a nonparametric procedure is appropriate in this setting.

15.39 Read the output. Figure 15.11 gives the Minitab output for the analysis of the data described in Exercise 15.37. Describe the results given in the output and write a short summary of your conclusions from the analysis.

FIGURE 15.11 Output from Minitab for the Kruskal-Wallis test applied to the Facebook data, for Exercise 15.39.

Minitab

```
Kruskal-Wallis Test: Score versus Friends

Kruskal-Wallis Test on Score

Friends      N    Median   Ave Rank           Z
102         24     3.600       46.3       -2.96
302         33     5.000       84.9        2.97
502         26     4.700       72.3        0.70
702         30     4.600       70.7        0.51
902         21     4.200       54.0       -1.74
Overall    134                 67.5

H = 16.98   DF = 4    P = 0.002
H = 17.05   DF = 4    P = 0.002    (adjusted for ties)
```

15.40 Do we experience emotions differently? In Exercise 12.37 (page 684) you analyzed data related to the way people from different cultures experience emotions. The study subjects were 410 college students from five different cultures. They were asked to record, on a 1 (never) to 7 (always) scale, how much of the time they typically felt eight specific emotions. These were averaged to produce the global emotion score for each participant. Analyze the data using the Kruskal-Wallis test and write a summary of your analysis and conclusions. Be sure to include your assumptions, hypotheses, and the results of the significance test. EMOTION

15.41 Do isoflavones increase bone mineral density? In Exercise 12.45 (page 686) you investigated the effects of isoflavones from kudzu on bone mineral density (BMD). The experiment randomized rats to three diets: control, low isoflavones, and high isoflavones. Here are the data: KUDZU

Treatment	BMD (g/cm^2)							
Control	0.228	0.207	0.234	0.220	0.217	0.228	0.209	0.221
	0.204	0.220	0.203	0.219	0.218	0.245	0.210	
Low dose	0.211	0.220	0.211	0.233	0.219	0.233	0.226	0.228
	0.216	0.225	0.200	0.208	0.198	0.208	0.203	
High dose	0.250	0.237	0.217	0.206	0.247	0.228	0.245	0.232
	0.267	0.261	0.221	0.219	0.232	0.209	0.255	

(a) Use the Kruskal-Wallace test to compare the three diets.

(b) How do these results compare with what you find using the ANOVA F statistic?

15.42 Vitamins in bread. Does bread lose its vitamins when stored? Here are data on the vitamin C content (milligrams per 100 grams of flour) in bread baked from the same recipe and stored for 1, 3, 5, or 7 days.[16] The 10 observations are from 10 different loaves of bread. BREAD

Condition	Vitamin C (mg/100 g)	
Immediately after baking	47.62	49.79
One day after baking	40.45	43.46
Three days after baking	21.25	22.34
Five days after baking	13.18	11.65
Seven days after baking	8.51	8.13

The loss of vitamin C over time is clear, but with only 2 loaves of bread for each storage time we wonder if the differences among the groups are significant.

(a) Use the Kruskal-Wallis test to assess significance and then write a brief summary of what the data show.

(b) Because there are only 2 observations per group, we suspect that the common chi-square approximation to the distribution of the Kruskal-Wallis statistic may not be accurate. The exact P-value (from SAS software) is $P = 0.0011$. Compare this with your P-value from part (a). Is the difference large enough to affect your conclusion?

15.43 Jumping and strong bones. In Exercise 12.47 (page 687) you studied the effects of jumping on the bones of rats. Ten rats were assigned to each of three treatments: a 60-centimeter "high jump," a 30-centimeter "low jump," and a control group with no jumping.[17] Here are the bone densities (in milligrams per cubic centimeter) after eight weeks of 10 jumps per day: JUMP

Group	Bone density (mg/cm^3)				
Control	611	621	614	593	593
	653	600	554	603	569
Low jump	635	605	638	594	599
	632	631	588	607	596
High jump	650	622	626	626	631
	622	643	674	643	650

(a) The study was a randomized comparative experiment. Outline the design of this experiment.

(b) Make side-by-side stemplots for the three groups, with the stems lined up for easy comparison. The distributions are a bit irregular but not strongly non-Normal. We would usually use analysis of variance to assess the significance of the difference in group means.

(c) Do the Kruskal-Wallis test. Explain the distinction between the hypotheses tested by Kruskal-Wallis and ANOVA.

(d) Write a brief statement of your findings. Include a numerical comparison of the groups as well as your test result.

15.44 Do poets die young? In Exercise 12.46 (page 686) you analyzed the age at death for female writers. They were classified as novelists, poets, and nonfiction writers. The data are given in Table 12.1 (page 686). POETS

(a) Use the Kruskal-Wallace test to compare the three groups of female writers.

(b) Compare these results with what you find using the ANOVA F statistic.

CHAPTER 15 Exercises

15.45 Plants and hummingbirds. Different varieties of the tropical flower *Heliconia* are fertilized by different species of hummingbirds. Over time, the lengths of the flowers and the forms of the hummingbirds' beaks have evolved to match each other. Here are data on the lengths in millimeters of three varieties of these flowers on the island of Dominica:[18] HBIRDS

H. bihai					
47.12	46.75	46.81	47.12	46.67	47.43
46.44	46.64	48.07	48.34	48.15	50.26
50.12	46.34	46.94	48.36		
***H. caribaea* red**					
41.90	42.01	41.93	43.09	41.47	41.69
39.78	40.57	39.63	42.18	40.66	37.87
39.16	37.40	38.20	38.07	38.10	37.97
38.79	38.23	38.87	37.78	38.01	
***H. caribaea* yellow**					
36.78	37.02	36.52	36.11	36.03	35.45
38.13	37.10	35.17	36.82	36.66	35.68
36.03	34.57	34.63			

Do a complete analysis that includes description of the data and a rank test for the significance of the differences in lengths among the three species.

15.46 Time spent studying. In Exercise 1.173 (page 50) you compared the time spent studying by men and women. The students in a large first-year college class were asked how many minutes they studied on a typical weeknight. Here are the responses of random samples of 30 women and 30 men from the class: STIME

Women					Men				
170	120	180	360	240	80	120	30	90	200
120	180	120	240	170	90	45	30	120	75
150	120	180	180	150	150	120	60	240	300
200	150	180	150	180	240	60	120	60	30
120	60	120	180	180	30	230	120	95	150
90	240	180	115	120	0	200	120	120	180

(a) Summarize the data numerically and graphically.

(b) Use the Wilcoxon rank sum test to compare the men and women. Write a short summary of your results.

(c) Use a two-sample t test to compare the men and women. Write a short summary of your results.

(d) Which procedure is more appropriate for these data? Give reasons for your answer.

15.47 Response times for telephone repair calls. A study examined the time required for the telephone company Verizon to respond to repair calls from its own customers and from customers of a CLEC, another phone company that pays Verizon to use its local lines. Here are the data, which are rounded to the nearest hour: TREPAIR

Verizon											
1	1	1	1	2	2	1	1	1	1	2	2
1	1	1	1	2	2	1	1	1	1	2	3
1	1	1	1	2	3	1	1	1	1	2	3
1	1	1	1	2	3	1	1	1	1	2	3
1	1	1	1	2	3	1	1	1	1	2	4
1	1	1	1	2	5	1	1	1	1	2	5
1	1	1	1	2	6	1	1	1	1	2	8
1	1	1	1	2	15	1	1	1	2	2	
CLEC											
1	1	5	5	5	1	5	5	5	5		

(a) Does Verizon appear to give CLEC customers the same level of service as its own customers? Compare the data using graphs and descriptive measures and express your opinion.

(b) We would like to see if times are significantly longer for CLEC customers than for Verizon customers. Why would you hesitate to use a t test for this purpose? Carry out a rank test. What can you conclude?

(c) Explain why a nonparametric procedure is appropriate in this setting.

Iron-deficiency anemia is the most common form of malnutrition in developing countries. Does the type of cooking pot affect the iron content of food? We have data from a study in Ethiopia that measured the iron content (milligrams per 100 grams of food) for three types of food cooked in each of three types of pots:[19] COOK

Type of Pot	Iron Content			
	Meat			
Aluminum	1.77	2.36	1.96	2.14
Clay	2.27	1.28	2.48	2.68
Iron	5.27	5.17	4.06	4.22
	Legumes			
Aluminum	2.40	2.17	2.41	2.34
Clay	2.41	2.43	2.57	2.48
Iron	3.69	3.43	3.84	3.72
	Vegetables			
Aluminum	1.03	1.53	1.07	1.30
Clay	1.55	0.79	1.68	1.82
Iron	2.45	2.99	2.80	2.92

Exercises 15.48 to 15.50 use these data.

15.48 Cooking vegetables in different pots. Does the vegetable dish vary in iron content when cooked in aluminum, clay, and iron pots? COOK

(a) What do the data appear to show? Check the conditions for one-way ANOVA. Which requirements are a bit dubious in this setting?

(b) Instead of ANOVA, do a rank test. Summarize your conclusions about the effect of pot material on the iron content of the vegetable dish.

15.49 Cooking meat and legumes in aluminum and clay pots. There appears to be little difference between the iron content of food cooked in aluminum pots and food cooked in clay pots. Is there a significant difference between the iron content of meat cooked in aluminum and clay? Is the difference between aluminum and clay significant for legumes? Use rank tests. COOK

15.50 Iron in food cooked in iron pots. The data show that food cooked in iron pots has the highest iron content. They also suggest that the three types of food differ in iron content. Is there significant evidence that the three types of food differ in iron content when all are cooked in iron pots? COOK

15.51 Multiple comparisons for plants and hummingbirds. As in ANOVA, we often want to carry out a **multiple-comparisons** procedure following a Kruskal-Wallis test to tell us *which* groups differ significantly.[20] The Bonferroni method (page 670) is a simple method: If we carry out k tests at fixed significance level $0.05/k$, the probability of *any* false rejection among the k tests is always no greater than 0.05. That is, to get overall significance level 0.05 for all of k comparisons, do each individual comparison at the $0.05/k$ level. In Exercise 15.45 you found a significant difference among the lengths of three varieties of the flower *Heliconia*. Now we will explore multiple comparisons. HBIRDS

(a) Write down all the pairwise comparisons we can make, for example, *bihai* versus *caribaea* red. There are three possible pairwise comparisons.

(b) Carry out three Wilcoxon rank sum tests, one for each of the three pairs of flower varieties. What are the three two-sided P-values?

(c) For purposes of multiple comparisons, any of these three tests is significant if its P-value is no greater than $0.05/3 = 0.0167$. Which pairs differ significantly at the overall 0.05 level?

15.52 Multiple comparisons for cooking pots. The previous exercise outlines how to use the Wilcoxon rank sum test several times for multiple comparisons with overall significance level 0.05 for all comparisons together. Apply this procedure to the data used in each of Exercises 15.48 to 15.50. COOK

CHAPTER 15 Notes and Data Sources

1. *Condé Nast Traveler* readers poll data for 2013, from **cntraveler.com/spas/2013/03/best-spas-united-states-caribbean-mexico-cruise-ships.**

2. For purists, here is the precise definition: X_1 is *stochastically larger* than X_2 if

$$P(X_1 > a) \geq P(X_2 > a)$$

for all a, with strict inequality for at least one a. The Wilcoxon rank sum test is effective against this alternative in the sense that the power of the test approaches 1 (that is, the test becomes more certain to reject the null hypothesis) as the number of observations increases.

3. Erin K. O'Loughlin et al., "Prevalence and correlates of exergaming in youth," *Pediatrics,* 130 (2012), pp. 806–814.

4. From the PEW Internet and American Life website, **pewinternet.org/Reports/2013/Civic-Engagement.aspx.**

5. From Matthias R. Mehl et al., "Are women really more talkative than men?," *Science,* 317, No. 5834 (2007), p. 82. The raw data were provided by Matthias Mehl.

6. Data provided by Warren Page, New York City Technical College, from a study done by John Hudesman.

7. Data provided by Susan Stadler, Purdue University.

8. Ibid.

9. The vehicle is a 2002 Toyota Prius owned by the third author.

10. Statistics regarding Facebook usage can be found at **facebook.com/notes/facebook-data-team/anatomy-of-facebook/10150388519243859.**

11. These data were collected as part of a larger study of dementia patients conducted by Nancy Edwards, School of Nursing, and Alan Beck, School of Veterinary Medicine, Purdue University.

12. Data provided by Diana Schellenberg, Purdue University School of Health Sciences.

13. These data are from "Results report on the vitamin C pilot program," prepared by SUSTAIN (Sharing

United States Technology to Aid in the Improvement of Nutrition) for the U.S. Agency for International Development. The report was used by the Committee on International Nutrition of the National Academy of Sciences/Institute of Medicine to make recommendations on whether or not the vitamin C content of food commodities used in U.S. food aid programs should be increased. The program was directed by Peter Ranum and Françoise Chomé. The second author was a member of the committee.

14. Data provided by Sam Phillips, Purdue University.

15. See Note 10.

16. Data provided by Helen Park. See H. Park et al., "Fortifying bread with each of three antioxidants," *Cereal Chemistry,* 74 (1997), pp. 202–206.

17. Data provided by Jo Welch, Purdue University Department of Foods and Nutrition.

18. We thank Ethan J. Temeles of Amherst College for providing the data. His work is described in Ethan J. Temeles and W. John Kress, "Adaptation in a plant-hummingbird association," *Science,* 300 (2003), pp. 630–633.

19. Based on A. A. Adish et al., "Effect of consumption of food cooked in iron pots on iron status and growth of young children: A randomised trial," *The Lancet,* 353 (1999), pp. 712–716.

20. For more details on multiple comparisons, see M. Hollander and D. A. Wolfe, *Nonparametric Statistical Methods,* 2nd ed., Wiley, 1999. This book is a useful reference on applied aspects of nonparametric inference in general.