ALEX SEGRE/ALAMY

CHAPTER 16

Nonparametric Tests

Introduction

The most commonly used statistical methods for inference ultimately rely on certain distributional assumptions. In regression, it is assumed that the residual variation is Normal. In testing of means, if data are not Normal, there is a presumption that sample sizes are large enough to allow for appropriate application of *t*-tests and *F*-tests. In addition, it is assumed that the values taken on by the response variables have clear numerical interpretation.

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. But, with this said, there are applications where the adequacy of standard methods can be seriously challenged:

- A health care payer, such as Humana or Blue Cross, wants to determine which of five potential benefit plans will appeal to its 10 largest employers.
- Toyota wants to compare the quality of service at a various service locations based on customer surveys using the following scale and coded responses: 1 = very poor, 2 = poor, 3 = fair, 4 = good, and 5 = excellent.
- Generac is a leading manufacturer of residential and commercial standby generators. Reliability of its generators is essential to protect consumers from loss and inconvenience during power outages. The number of repairs of generators during the warranty period is typically quite small. Generac wishes to use frequency of repair data across different generator product models to see if there is evidence of different levels of reliability.

CHAPTER OUTLINE

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

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. Exercises 7.112 (page 413) and 11.8 (page 538) illustrate the use of the logarithm transformation.

3. In some settings, **other standard distributions** replace the Normal distributions as models for the overall pattern in the population. There are inference procedures for the parameters of these distributions that replace the *t* procedures when we use specific non-Normal models.

REMINDER
sign test, p. 407

4. Finally, there are inference procedures that do not require any specific form for the distribution of the population. These are called **nonparametric methods**. The *sign test* and the *runs test* are examples of nonparametric tests that have been earlier covered.

REMINDER
runs test, p. 648

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 16.1 presents an outline of the standard tests (based on Normal distributions) and the rank tests that compete with them.

The rank tests we 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 16.1 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.

We devote a section of this chapter to each of the rank procedures. Section 16.1, 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

FIGURE 16.1 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 16.2
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 16.1
Several independent samples	One-way ANOVA *F* test Chapter 14	Kruskal-Wallis test Section 16.3

distributions for use with small samples and approximations for use with larger samples are common to all rank tests. Sections 16.2 and 16.3 more briefly describe other rank tests.

16.1 The Wilcoxon Rank Sum Test

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.

CASE 16.1

DATA USED

Price Discrimination? A Midwestern automobile dealership, which we call simply Midwest Auto to preserve confidentiality, has been accused of price discrimination in its selling of used cars. Midwest Auto sells used cars on two different lots. One lot is the standard lot where both new and used cars are sold. This lot is marketed under the name of Midwest Auto. Its other lot exclusively sells used cars. The exclusively used cars lot is marketed under another name than Midwest Auto, which we will call Used Auto.

Customers approaching Midwest Auto who are interested in purchasing a used car are subjected to a credit check. If the dealership assesses the customer as "risky," then that customer is given only the option to purchase a used car from its subsidiary Used Auto.

Given the customer base for Used Auto is credit risky, car loans offered to these customers come at a higher interest rate. However, the legal matter brought forth to the legal system is that Used Car sells used cars at a higher prices than would have been sold at its Midwest Auto lot. The implication is that there is unfair pricing for a class of consumers. A particular Used Car consumer retained a law firm to pursue a possible class action suit against Midwest Auto.

In the pretrial stage, the Court would not grant opening all records of sales at the two lots. Instead, the Court granted access of data for used cars sold at the two lots during the week of the sale to the plaintiff.

For each used car sold, selling price was compared to the car's valuation based on the National Automobile Dealers Association (NADA) and converted to a percent. NADA valuations are similar to valuations found in the Black Book or Kelley Blue Book. Data can be viewed as selling price markups relative to NADA valuations. For the week in question, there were 12 used cars sold at Midwest Auto and eight cars sold at Used Auto. Here are the data.

Lot	Markup %											
Used Auto	54.7	18.3	24.3	12.8	26.9	6.7	10.9	34.4	35.3	5.4	38.4	30.3
Midwest Auto	13.4	24.0	19.4	8.1	3.5	5.8	7.4	5.1				

Even though the markup percents are all positive here, it is quite possible to get a negative value, which simply means that the consumer purchased the car for less-than book value.

Figure 16.2 is a back-to-back stemplot that compares the markup percents of the 12 Used Auto consumers and the eight Midwest Auto consumers in our sample. The Used Auto distribution has a high outlier, and the Midwest Auto distribution is

FIGURE 16.2 Back-to-back stemplot of the markup percents for Midwest Auto consumers and Used Auto consumers.

Midwest		Used
4	0	
8765	0	5 7
3	1	1 3
9	1	8
4	2	4
	2	7
	3	0 4
	3	5 8
	4	
	4	
	5	
	5	5

strongly skewed to the right. If data show neither outliers nor skewness, then two-sample t procedures are fairly robust, even for a total combined sample size of 20 as we have here. However, it is not so clear we can rely on the robustness of the two-sample t test for the data of Case 16.1. We prefer to consider an alternative test that is robust to departures in Normality.

REMINDER
robustness of the two-sample t procedures, p. 384

The rank transformation

We first rank all 20 observations together. To do this, arrange them in order from smallest to largest:

3.5	5.1	**5.4**	5.8	**6.7**	7.4	8.1	**10.9**	**12.8**	13.4
18.3	19.4	24.0	**24.3**	**26.9**	**30.3**	**34.4**	**35.3**	**38.4**	**54.7**

The boldface entries in the list are the sales at the Used Auto lot. 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 20 (largest). These numbers are the *ranks*:

Markup %	3.5	5.1	**5.4**	5.8	**6.7**	7.4	8.1	**10.9**	**12.8**	13.4
Rank	1	2	**3**	4	**5**	6	7	**8**	**9**	10
Markup %	**18.3**	19.4	24.0	**24.3**	**26.9**	**30.3**	**34.4**	**35.3**	**38.4**	**54.7**
Rank	**11**	12	13	**14**	**15**	**16**	**17**	**18**	**19**	**20**

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.

REMINDER
log transformation, p. 68

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.

APPLY YOUR KNOWLEDGE

SPAS

16.1 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.

SPAS2

16.2 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 Used Car consumers tend to have a higher markup percent relative to Midwest Car consumers, we expect the ranks of the Used Car consumers to be larger than the ranks of the Midwest consumers. Let's compare the *sums* of the ranks from the two groups:

Lot	Sum of ranks
Used Auto	155
Midwest Auto	55

These sums measure how much the ranks of the Used Auto consumers as a group exceed those of the Midwest Auto consumers. In fact, the sum of the ranks from 1 to 20 is always equal to 210, so it is enough to report the sum for one of the two groups. If the sum of the ranks for the Used Auto consumers is 155, the ranks for the Midwest Auto consumers must add to 55 because $155 + 55 = 210$.

Because there are more Used Auto consumers than Midwest Auto consumers, we would expect the sum of the Used Auto ranks to be greater than the sum of the Midwest Auto ranks if there are no systematic price discrimination differences. But how much greater? Here are the facts we need stated in general form.

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.[2]

In the used car consumer study of Case 16.1, we want to test

H_0: no difference in distribution of the markup percent for consumers from Midwest Auto and Used Auto lots.

against the one-sided alternative

H_a: Used Auto consumers are systematically paying a higher markup percent.

Our test statistic is the rank sum $W = 155$ for the Used Auto consumers.

APPLY YOUR KNOWLEDGE

SPAS

16.3 Hypotheses and test statistic for top spas. Refer to Exercise 16.1. State appropriate null and alternative hypotheses for this setting and calculate the value of W, the test statistic.

SPAS2

16.4 Effect of Animal Kingdom on the test statistic. Refer to Exercise 16.2. Using the altered data, state appropriate null and alternative hypotheses and calculate the value of W, the test statistic.

EXAMPLE 16.1 Perform the Significance Test

CASE 16.1 In Case 16.1, $n_1 = 12$, $n_2 = 8$, and there are $N = 20$ observations in all. The sum of ranks for the 12 Used Auto consumers has mean

$$\mu_W = \frac{n_1(N+1)}{2} = \frac{(12)(21)}{2} = 126$$

and standard deviation

$$\sigma_W = \sqrt{\frac{n_1 n_2 (N+1)}{12}} = \sqrt{\frac{(12)(8)(21)}{12}} = \sqrt{168} = 12.961$$

The observed sum of the ranks, $W = 155$, is higher than the mean by a bit more than two standard deviations since $(155 - 126)/12.961 = 2.24$. It appears that the data support the suspicion that Used Auto consumers are being discriminated against regarding price. The P-value for our one-sided alternative is $P(W \geq 155)$, 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, $P(W \geq 155)$, 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 16.2 The *P*-value

DATA USED

CASE 16.1 Figure 16.3 shows the output from JMP that calculates the exact sampling distribution of W. We see that the sum of the ranks in the first group (Used Auto) is $W = 155$. The reported one-sided P-value is 0.0126. It should be noted that the software computed the P-value based on the sum of ranks for the second group which is 55. It is reporting the probability that the sum of ranks for Midwest Auto consumers is at most as large as 55. This is equivalent to the probability that the sum of ranks for Used Auto consumers is at least as large as 155.

FIGURE 16.3 Output from JMP for the data in Case 16.1. JMP provides the option of reporting the *P*-value for the exact distribution of *W*.

JMP

Wilcoxon / Kruskal-Wallis Tests (Rank Sums)

Level	Count	Score Sum	Expected Score	Score Mean	(Mean-Mean0)/Std0
Used	12	155.000	126.000	12.9167	2.199
Midwest	8	55.000	84.000	6.8750	-2.199

2-Sample: Exact Test

S	Prob≤S	Prob≥|S-Mean|
55	0.0126*	0.0252*

The two-sample t test gives a somewhat more significant result than the Wilcoxon test in Example 16.2 ($t = 2.80$, $P = 0.006$). We hesitate to trust the t test because of the skewness of one sample and the outlier in the other sample along with the fact that the sample sizes are not large.

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:

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

REMINDER continuity correction, p. 260

Use standard Normal probability calculations to find P-values for this statistic. Because W takes only whole-number values, the **continuity correction** can improve the accuracy of the approximation. The idea of continuity correction was first introduced with binomial probability calculations.

EXAMPLE 16.3 The Normal Approximation

CASE 16.1 In our used car consumer example, we are interested in approximating $P(W \geq 155)$. Continuity correction acts as if the whole number 155 occupies the entire interval from 154.5 to 155.5. We calculate the P-value $P(W \geq 155)$ as $P(W \geq 154.5)$ because the value 155 is included in the range whose probability we want. Here is the calculation:

$$\begin{aligned} P(W \geq 154.5) &= P\left(\frac{W - \mu_W}{\sigma_W} \geq \frac{154.5 - 126}{12.961}\right) \\ &= P(Z \geq 2.1989) \\ &= 0.0139 \end{aligned}$$

The continuity correction gives a result close to the exact value $P = 0.0126$.

We recommend 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 16.3 illustrates, however, the Normal approximation with the continuity correction is often adequate.

DATA
USED
Mann-Whitney test

EXAMPLE 16.4 Reading Software Output

CASE 16.1 Figure 16.4 shows the output for our data from Minitab. Minitab offers only the Normal approximation, and it refers to the **Mann-Whitney test**.This is an alternative form of the Wilcoxon rank sum test. Minitab gives the approximate one-sided P-value as 0.0139 which agrees with our result in Example 16.3.

FIGURE 16.4 Output from Minitab for the data in Case 16.1. Minitab uses the Normal approximation for the distribution of W.

Minitab

Mann-Whitney Test and CI: Used, Midwest

```
           N  Median
Used      12   25.60
Midwest    8    7.75


Point estimate for η1 - η2 is 13.35
95.1 Percent CI for η1 - η2 is (1.60,26.80)
W = 155.0
Test of η1 = η2 vs η1 > η2 is significant at 0.0139
```

APPLY YOUR KNOWLEDGE

DATA
SPAS

16.5 The *P*-value for top spas. Refer to Exercises 16.1 and 16.3 (pages 16-5 and 16-6). 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?

DATA
SPAS2

16.6 The effect of Animal Kingdom on the *P*-value. Refer to Exercises 16.2 and 16.4 (pages 16-5 and 16-6). Repeat the analysis in Exercise 16.5 using the altered data.

What hypotheses do the Wilcoxon test?

Our null hypothesis is that the markup percents of Used Auto consumers and Midwest Auto consumers do not differ systematically. Our alternative hypothesis is that Used Auto consumers' markup percents are higher. If we are willing to assume that markup percents 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:

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

CAUTION

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 markup percents for Used Auto consumers must look exactly like that for markup percents for Midwest Auto consumers, except

that it may slide to a different location on the scale of markup percents. The Minitab output in Figure 16.4 states the hypotheses in terms of population medians (which it denotes as η) 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 the markup percent paid by a randomly chosen Used Auto consumer and X_2 to be the markup percent paid by a randomly chosen Midwest Auto consumer. These markup percents are random variables. That is, every time we choose a Used Auto consumer at random, the consumer's markup percent is a value of the variable X_1. The probability that a Used Auto consumer's markup percent is more than 10% is $P(X_1 > 10)$. Similarly, $P(X_2 > 10)$ is the corresponding probability for a randomly chosen Midwest Auto consumer. If the markup percents for Used Auto consumers are "systematically larger" than those of Midwest consumers, getting getting a markup percent greater than 10% should be more likely for Used Auto consumers. That is, we should have

$$P(X_1 > 10) > P(X_2 > 10)$$

The alternative hypothesis says that this inequality holds not just for 10% markup but for *any* markup percent.[3]

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 software 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. "Used Auto consumers pay a systematically higher markup percent than Midwest Auto consumers" 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. In theory, with continuous distributions, the probability is 0 that we encounter observations of *exactly* the same value. Having different values allows us to rank them all. In practice, however, we can encounter 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 six observations:

average ranks

Observation	10.8	13.2	23.1	23.1	29.7	30.4
Rank	1	2	3.5	3.5	5	6

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

The exact distribution for the Wilcoxon rank sum W applies only to data without ties. Moreover, the standard deviation σ_W must be adjusted if ties are present.

The Normal approximation can be used after the standard deviation is adjusted. Statistical software will detect ties, make the necessary adjustment, and switch to the Normal approximation. In practice, software is required 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.

CASE 16.2

FSAFETY

Consumer Perceptions of Food Safety Vendors of prepared food are very sensitive to the public's perception of the safety of the food they sell. 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. Is there good evidence that men and women differ in their perceptions about food safety at fairs?[4]

We should first ask if the subjects in Case 16.2 are a random sample of people who attend fairs, at least in the Midwest. The researcher visited 11 different fairs. She stood near the entrance and stopped every 25th adult who passed. Because no personal choice was involved in choosing the subjects, we can reasonably treat the data as coming from a random sample. (As usual, there was some nonresponse, which could create bias.)

Here are the data (variable "sfair" in the associated data file), presented as a two-way table of counts:

	Response					
	1	**2**	**3**	**4**	**5**	**Total**
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 are more concerned about food safety than the men:

	Response					
	1	**2**	**3**	**4**	**5**	**Total**
Female	6.6%	55.1%	25.5%	11.7%	1.0%	100%
Male	20.6%	53.3%	20.6%	4.7%	1.0%	100%

Is the difference between the genders statistically significant?

We might apply the chi-square test (Chapter 9). It is highly significant ($X^2 = 16.120$, df = 4, $P = 0.0029$). Although the chi-square test answers our general question, it ignores the ordering of the responses and so does not use all of

the available information. We would really like to know whether men or women are more concerned about the safety of the food served. This question depends on the ordering of responses from least concerned to most concerned. We can use the Wilcoxon test for the hypotheses:

H_0: men and women do not differ in their responses

H_a: one gender gives systematically larger responses than the other

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

EXAMPLE 16.5 Attitudes Toward Food Sold at Fairs

FSAFETY

CASE 16.2 Figure 16.5 gives JMP output for the Wilcoxon test. The rank sum for men (using average ranks for ties) is $W = 14{,}059.5$. The standardized value is $z = -3.33$, with two-sided P-value $P = 0.0009$. There is very strong evidence of a difference. Women are more concerned than men about the safety of food served at fairs.

FIGURE 16.5 Output from JMP for the food safety study of Case 16.2. The approximate two-sided P-value is 0.0009.

JMP

Wilcoxon / Kruskal-Wallis Tests (Rank Sums)

Level	Count	Score Sum	Expected Score	Score Mean	(Mean-Mean0)/Std0
Man	107	14059.5	16264.0	131.397	-3.334
Woman	196	31996.5	29792.0	163.247	3.334

2-Sample Test, Normal Approximation

S	Z	Prob>|Z|
14059.5	-3.33353	0.0009*

With more than 100 observations in each group and no outliers, we might use the two-sample t even though responses take only five values. In fact, the results for Example 16.5 are $t = -3.3655$ with $P = 0.0009$. The P-value for two-sample t is the same as that for the Wilcoxon test. There is, however, another reason to prefer the rank test in this example. The t statistic treats the response values 1 through 5 as meaningful numbers. In particular, the possible responses are treated as though they are equally spaced. The difference between "very rarely" and "once in a while" is the same as the difference between "once in a while" and "often." This may not make sense. The rank test, on the other hand, uses only the order of the responses, not their actual values. The responses are arranged in order from least to most concerned about safety, so the rank test makes sense. Some statisticians avoid using t procedures when there is not a fully meaningful scale of measurement.

APPLY YOUR KNOWLEDGE

GERCARS

16.7 CO_2 Emissions. Consider data on CO_2 emissions for 2014 compact cars manufactured by the German makers of Mercedes Benz and BMW. For each car, the data are expressed in units of grams of carbon dioxide per kilometer driven. There are many ties among the observations. Arrange the readings in order and assign ranks, assigning all tied values the average of the ranks they occupy.

GERCARS

16.8 Mercedes Benz versus BMW. Using your ranks from the previous exercise, what is the rank sum W for the Mercedes Benz cars? Using software, is there a significant difference between the CO_2 emissions of Mercedes Benz and BMW compact cars?

Rank versus *t* tests

The two-sample t procedures are the most common methods 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. Both are available in almost all statistical software. How do these two approaches compare in general?

- 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. 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 most other situations in practice, the robustness of t procedures allows us to obtain reasonably accurate P-values.
- 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 do the Wilcoxon test?" (page 16-8). 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.

SECTION 16.1 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 16.1 Exercises

For Exercises 16.1 and 16.2, see page 16-5; for 16.3 and 16.4, see page 16-6; for 16.5 and 16.6, see page 16-8; for 16.7, see page 16-11; for 16.8, see page 16-12.

CASE 16.1 **16.9 Price Discrimination?** In Examples 16.1 through 16.4, the testing of significance was based on the rank sum of the Used Auto consumers being the first group. Suppose we associate the first group with the Midwest Auto consumers. As noted in the section (page 16-5), the rank sum for Midwest Auto consumers is 55. USED

(a) Find the mean and standard deviation of the rank sum for Midwest Auto consumers under the null hypothesis that markup percents do not differ significantly.
(b) Refer to Example 16.1 where the mean of the rank sum for Used Auto consumers was found. What is the relationship between the two means?
(c) In Example 16.3, the one-sided P-value was determined using the rank sum of Used Auto consumers. Show how the same reported P-value can be found using the rank sum of Midwest Auto consumers.

16.10 Wheat prices. Example 7.13 (pages 385–386) reports the results of a small survey that asked separate samples of 5 wheat producers in each of January and July what price they received for wheat sold that month. Here are the data: WHEAT

Month	Price of wheat ($/bushel)				
January	$6.6125	$6.4775	$6.3500	$6.7525	$6.7625
July	$6.7350	$6.9000	$6.6475	$7.2025	$7.0550

The stemplot on page 386 shows a large difference between months. We cannot assess Normality from such small samples. Carry out by hand the steps in the Wilcoxon rank sum test for comparing prices in January and July.
(a) Arrange the 10 observations in order and assign ranks. There are no ties.
(b) Find the rank sum W for July. What are the mean and standard deviation of W under the null hypothesis that prices in January and July do not differ systematically?
(c) Standardize W to obtain a z statistic. Do a Normal probability calculation with the continuity correction to obtain a two-sided P-value.

16.11 Online discussion posting. Students in a fully online MBA statistics course are required as to post relevant learning contributions in the course's discussion forum throughout the semester. These posts serve as a form of online participation and factor into their grades. Below find the number of posts during the semester by the six female students in the class.[5] POSTS

23 27 20 16 10 31

Find the ranks for these data.

16.12 Find the rank sum statistic. Refer to the previous exercise. Here are the data for the 10 men in the class. POSTS

18 9 14 15 13 17 4 12 5 3

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

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

16.14 Find the mean and standard deviation of the distribution of the statistic. The statistic W that you calculated in Exercise 16.12 is a random variable with a sampling distribution. What are the mean and the standard deviation of this sampling distribution under the null hypothesis?

16.15 Find the P-value. Refer to Exercises 16.11 through 16.14. Find the P-value using the Normal approximation with the continuity correction and interpret the result of the significance test.

16.16 Counts of seeds in one-pound scoops. Exercise 7.55 (page 395) discusses a study of two different packaging plants in terms of the packaging of seeds. An SRS of 50 one-pound scoops of seeds was collected from Plant 1746, and an SRS of 19 one-pound scoops of seeds was collected from Plant 1748. The number of seeds found in each scoop was recorded. Histograms and Normal quantile plots suggest the data arise from non-Normal distributions. Using software, is there a significant difference in the number of seeds per pound between the two plants based on the Wilcoxon test? SEEDCNT2

16.17 Polyester fabrics in landfills. How quickly do synthetic fabrics such as polyester decay in landfills? A researcher buried polyester strips in the soil for different lengths of time, then dug up the strips and measured the force required to break them. Breaking strength is easy to measure and is a good indicator of decay. Lower strength means the fabric has decayed. Part of the study involved burying 10 polyester strips in well-drained soil in the summer. Five of the strips, chosen at random, were dug up after two weeks; the other five were dug up after 16 weeks. Here are the breaking strengths in pounds:[6] POLY

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

(a) Make side-by-side stemplots for the two groups. Does it appear reasonable to assume that the two distributions have the same shape?
(b) Is there evidence that breaking strengths are lower for strips buried longer?

16.18 Economic growth. The most commonly used measure of economic growth is the rate of growth in a country's total output of goods and services gauged by the gross domestic product (GDP) adjusted for inflation. The level of a country's GDP growth reflects

on the growth of businesses, jobs, and personal income. Here are World Bank data on the average growth of GDP (percent per year) for the period 2010 to 2013 in developing countries of Europe:[7] REGGDP

Country	Growth	Country	Growth
Albania	2.3	Macedonia, FYR	2.1
Armenia	4.4	Moldova	5.5
Azerbaijan	3.2	Montenegro	1.7
Belarus	4.0	Romania	1.3
Bosnia and Herzegovina	0.4	Serbia	0.9
Bulgaria	0.9	Turkey	6.0
Georgia	5.6	Ukraine	2.9
Kosovo	3.4		

Here are the data for nearby developing countries of the Central Asia:

Country	Growth	Country	Growth
Uzbekistan	8.2	Kyrgyz Republic	4.0
Turkmenistan	11.3	Kazakhstan	6.5
Tajikistan	7.2		

(a) Arrange the 20 observations in order and assign ranks. Be aware of ties.
(b) What is the rank sum W for the Central Asia region?
(c) Perform the two-sided test at the $\alpha = 0.05$ level, making sure to report the test statistic, and P-value. What is your conclusion?

16.19 It's your choice. Exercise 16.18 asks for the rank sum W for Central Asia. REGGDP
(a) What is the rank sum for European developing countries? The ranks of 20 observations always add to 210. Do your two sums add to 210?
(b) Repeat the previous exercise using the the rank sum for European countries. Show that you obtain exactly the same P-value. That is, your choice between the two possible W's does not affect the results of the Wilcoxon test.

16.20 ERP implementation. Companies worldwide are investing in enterprise resource planning (ERP) systems. ERP is an integrated business management software system that allows companies to share common data across all functional business areas. By linking areas of a company with a single system, the premise is that companies will reduce costs and improve efficiencies corporate-wide. In a study, researchers investigated if ERP implementation had positive impact on facility management (FM) services. A survey was conducted on companies with and without ERP systems. Data were collected by company on the number of FM-related areas that have had productivity improvements in the given calendar year. The researchers were interested in testing the alternative hypothesis that the number of FM-related productivity improvements was greater for ERP companies than non-ERP companies.[8] ERP
(a) Examine the data from each group. Explain why a two-sample t test may not be the best choice for conducting a test between the two groups.
(b) There are many ties among the observations. Arrange the observations in order and assign ranks, assigning all tied values the average of the ranks they occupy.
(c) What is the rank sum W for the ERP group?
(d) Perform the appropriate one-sided test at the $\alpha = 0.05$ level, making sure to report the test statistic and P-value. What is your conclusion?

16.21 The influence of subliminal messages. Can "subliminal" messages that are below our threshold of awareness nonetheless influence us? Advertisers, among others, want to know. One study asked if 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 eight 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.[9] 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?

16.22 Fitness and ego. Exercise 7.63 describes a study of fitness and personality. In particular, 28 middle-aged college faculty were evenly divided into low-fitness and high-fitness groups. The subjects then took the Cattell Sixteen Personality Factor Questionnaire. The provided data are the measurements of "ego strength." EGO
(a) Arrange the observations in order and assign ranks. Assign any tied values the average of the ranks they occupy.
(b) What is the rank sum W for the high-fitness group?
(c) Perform the significance test at the $\alpha = 0.05$ level, making sure to report the test statistic, and P-value. What is your conclusion?

CASE 16.2 **16.23 Safety of restaurant food.** Case 16.2 (page 16-10) describes a study of the attitudes of people attending outdoor fairs about the safety of the food served at such locations. In the associated data file, you will find the responses of 303 people to several questions. The variables in this data set are (in order)

subject hfair sfair sfast srest gender

The variable "sfair" contains the responses described in the example concerning safety of food served at outdoor fairs and festivals. The variable "srest" contains responses to the same question asked about food served in restaurants. We saw that women are more concerned than men about the safety of food served at fairs. Is this also true for restaurants? FSAFETY

CASE 16.2 **16.24 Food safety at fairs and in restaurants.** The data file used in Example 16.5 (page 16-11) and Exercise 16.23 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 the 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."

16.25 Sadness and spending. Exercise 7.52 (page 394) studies the effect of sadness on a person's spending judgment. In the exercise, we find 17 participants in the sad group and 14 participants in the neutral group. SADNESS
(a) Arrange the 31 observations in order and assign ranks. Be aware of ties.
(b) What is the rank sum W for the sad group?
(c) Using software, is there evidence of a sadness effect in that people in a sad mood systematically spend more than people not in a sad mood?

16.26 Patient satisfaction. A Wisconsin health care provider has a network of hospitals serving communities throughout eastern Wisconsin. In an attempt to continually improve its services, this provider conducts patient and employee satisfaction surveys. To measure overall rating of a given hospital, patients are asked, "Would you recommend this hospital to your friends and family?" Answer choices are definitely no, probably no, probably yes, definitely yes. The responses are coded numerically from 1 to 4. Here are quarterly survey data for one of its urban-based hospitals and for one of its suburban-based hospitals. HSURVEY

Response code	Response	Urban	Suburb
1	definitely no	12	6
2	probably no	33	9
3	probably yes	47	33
4	definitely yes	58	52

(a) Is there a relationship between location and rating? Use the chi-square test to answer this question.
(b) The chi-square test ignores the ordering of the rating categories. The provided data file contains data on the 250 patients surveyed. The first variable is the location (Urban or Suburb) and the second is the rating code as it appears in the table (1 to 4). Is there good evidence that patients in one location have systematically higher satisfaction ratings than in the other?

16.2 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. We now meet a rank test for matched pairs and single samples. The matched pairs setting is more important because good studies are generally comparative.

EXAMPLE 16.6 Loss of Product Value

DATA
VITC

Food products are often enriched with vitamins and other supplements. Does the level of a supplement decline over time so that the user receives less than the manufacturer intended? Here are data on the vitamin C levels (milligrams per 100 grams) in wheat soy blend, a flour-like product supplied by international aid programs mainly for feeding children. The same nine bags of blend were measured at the factory and five months later in Haiti.[10]

Bag	1	2	3	4	5	6	7	8	9
Factory	45	32	47	40	38	41	37	52	37
Haiti	38	40	35	38	34	35	38	38	40
Difference	7	−8	12	2	4	6	−1	14	−3

We suspect that vitamin C levels are generally higher at the factory than they are five months later. We would like to test the hypotheses

H_0: vitamin C has the same distribution at both times

H_a: vitamin C is systematically higher at the factory

Because these are matched pairs data, we base our inference on the differences.

absolute value

Positive differences in Example 16.6 indicate that the vitamin C level of a bag was higher at the factory than in Haiti. If factory values are generally higher, the positive differences should be farther from zero in the positive direction than the negative differences are in the negative direction. Therefore, we compare the **absolute values** of the differences—that is, their magnitudes without a sign. Here they are, with boldface indicating the positive values:

7 8 **12** **2** **4** **6** 1 **14** 3

Arrange the absolute values 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. In our example, there are no zeros and no ties.

Absolute value	1	**2**	3	**4**	**6**	**7**	8	**12**	**14**
Rank	1	**2**	3	**4**	**5**	**6**	7	**8**	**9**

The test statistic is the sum of the ranks of the positive differences. This is the *Wilcoxon signed rank statistic*. Its value here is $W^+ = 34$. (We could equally well use the sum of the ranks of the negative differences, which is 11.)

The Wilcoxon Signed Rank Test for Matched Pairs

Draw an SRS from a population for a matched pairs study and take the differences in responses within pairs. *Remove all zero differences*, so that n nonzero differences remain. 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.

APPLY YOUR KNOWLEDGE

SPAS3

16.27 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 16.1 also reported scores on service and on food. Here are the scores for a random sample of seven 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

16.28 Scores for the next 25 spas. Refer to the previous exercise. Here are the scores for a random sample of seven 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 16.7 Loss of Product Value: Rank Test

VITC

In the vitamin loss study of Example 16.6, $n = 9$. If the null hypothesis (no systematic loss of vitamin C) is true, the mean of the signed rank statistic is

$$\mu_{W^+} = \frac{n(n+1)}{4} = \frac{(9)(10)}{4} = 22.5$$

Our observed value $W^+ = 34$ is somewhat larger than this mean. The one-sided P-value is $P(W^+ \geq 34)$.

Figure 16.6 displays the output of two statistical programs. We see from Figure 16.6(a) that the one-sided P-value is $P = 0.1016$. JMP reports a statistic S as being 11.5. This is simply the difference between W^+ and μ_{W^+}. This small sample does not give convincing evidence of vitamin loss.

In fact, the Normal quantile plot in Figure 16.7 shows that the differences are reasonably Normal. We could use the paired-sample t to get a similar conclusion ($t = 1.5595$, df $= 8$, $P = 0.0787$). The t test has a slightly lower P-value because it is somewhat more powerful than the rank test when the data are actually Normal.

FIGURE 16.6 Output from (a) JMP and (b) Minitab for the loss of product value study of Example 16.6. JMP reports the exact one-sided *P*-value, $P = 0.1016$. Minitab uses the Normal approximation with the continuity correction and so gives an approximate one-sided *P*-value, $P = 0.096$.

(a)

JMP

Wilcoxon Signed Rank

	Factory-Haiti
Test Statistic S	11.500
Prob>|S|	0.2031
Prob>S	0.1016
Prob<S	0.8984

(b)

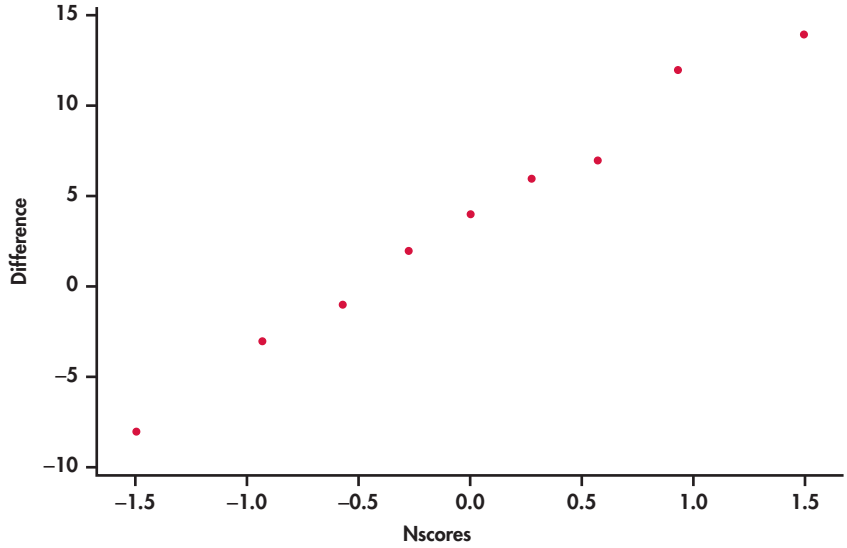
Minitab

Wilcoxon Signed Rank Test: Difference

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

                N for   Wilcoxon          Estimated
           N    Test  Statistic      P     Median
Difference 9       9       34.0  0.096      4.000
```

FIGURE 16.7 Normal quantile plot of the differences in loss of product value, Example 16.7.



Although we emphasize the matched pairs setting, W^+ can also be applied to a single sample. It then tests the hypothesis that the population median is zero. For matched pairs, we are testing that the median of the differences is zero. To test the hypothesis that the population median has a specific value m, apply the test to the differences $X_i - m$.

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 loss of product value example, even though $n = 9$ is certainly not a large sample.

EXAMPLE 16.8 Loss of Product Value: Normal Approximation

DATA VITC

For $n = 9$ observations, we saw in Example 16.7 that $\mu_{W^+} = 22.5$. The standard deviation of W^+ under the null hypothesis is

$$\sigma_{W^+} = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$
$$= \sqrt{\frac{(9)(10)(19)}{24}}$$
$$= \sqrt{71.25} = 8.441$$

The continuity correction calculates the P-value $P(W^+ \geq 34)$ as $P(W^+ \geq 33.5)$, treating the value $W^+ = 34$ as occupying the interval from 33.5 to 34.5. We find the Normal approximation for the P-value by standardizing and using the standard Normal table:

$$P(W^+ \geq 33.5) = P\left(\frac{W^+ - 22.5}{8.441} \geq \frac{33.5 - 22.5}{8.441}\right)$$
$$= P(Z \geq 1.303)$$
$$= 0.0968$$

Despite the small sample size, the Normal approximation gives a result quite close to the exact value $P = 0.1016$. The Minitab output in Figure 16.6(b) gives $P = 0.096$ based on a Normal calculation rather than the table.

APPLY YOUR KNOWLEDGE

DATA SPAS3

16.29 Significance test for top-ranked spas. Refer to Exercise 16.27 (page 16-17). Find μ_{W^+}, σ_{W^+}, and the Normal approximation for the P-value for the Wilcoxon signed rank test.

DATA SPAS4

16.30 Significance test for lower-ranked spas. Refer to Exercise 16.28 (page 16-17). Find μ_{W^+}, σ_{W^+}, and the Normal approximation for the P-value for the Wilcoxon signed rank test.

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, we drop such pairs from our sample. As in the case of the Wilcoxon rank sum, ties complicate finding a P-value. There is no longer a usable exact distribution for the signed rank statistic W^+, and the standard deviation σ_{W^+} must be adjusted for the ties before we can use the Normal approximation. Software will do this. Here is an example.

EXAMPLE 16.9 Two Rounds of Golf Scores

DATA GOLF

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

H_0: 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 16.8) shows some irregularity and a low outlier. We will use the Wilcoxon signed rank test.

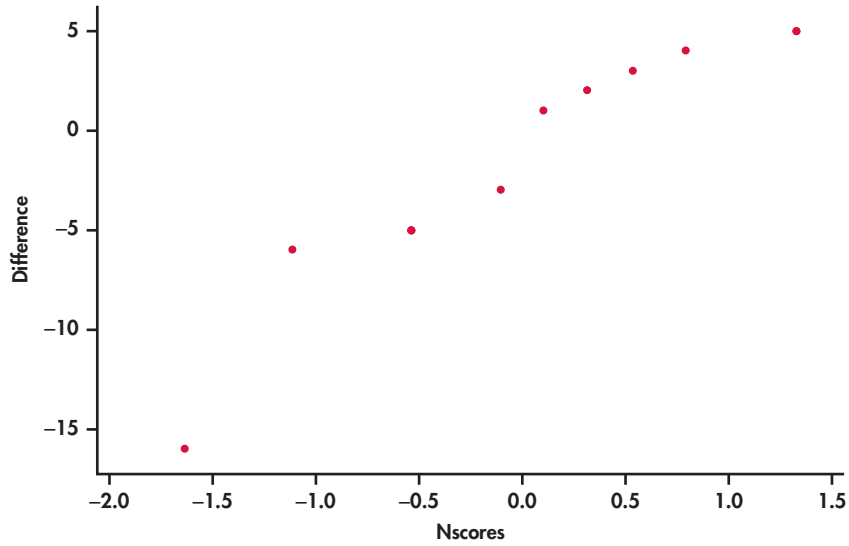


FIGURE 16.8 Normal quantile plot of the differences in scores for two rounds of a golf tournament, Example 16.9.

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 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 16.10 Software Results for Golf Scores

GOLF

Here are the two-sided P-values for the Wilcoxon signed rank test for the golf score data from several statistical programs:

Program	P-value
Minitab	$P = 0.388$
JMP	$P = 0.388$
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-values reported differ a bit from program to program. The reason for the variations is that the programs use slightly different versions of the approximate calculations needed when ties are present. The exact result depends on which of these variations the programmer chooses to use.

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

SECTION 16.2 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.
- The Wilcoxon signed rank test can also be used to test hypotheses about a population median by applying it to a single sample.
- ***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 16.2 Exercises

For Exercises 16.27 and 16.28, see page 16-17; for 16.29 and 16.30, see page 16-19.

16.31 Online bookstore prices. For more than 80 years, the *New York Times* has been publishing its weekly list of best-selling books in the United States. For certain book subcategories, the best-selling lists are done monthly. On the *New York Times* best-selling book website, there is a "buy" button next to each listed book. The button directs the viewer to one of two online bookstore options: Amazon or Barnes & Noble. Consider the following table of the online prices of the best-selling business books for September 2014. Note that older titles are paperback versus paperback comparisons, while newer titles are hardback versus hardback comparisons.[11]


NYTBOOK

Book title	Barnes & Noble price ($)	Amazon price ($)
Outliers	10.39	10.19
The Path Between the Seas	13.44	13.18
Thinking, Fast and Slow	9.99	9.90
The Power of Habit	9.97	9.78
#GIRLBOSS	16.25	16.17
The Organized Mind	18.03	17.68
Capital in the Twenty-First Century	24.08	23.97
Think Like a Freak	17.73	17.73
Business Adventures	10.25	10.25
Lean In	16.11	16.03

Is there is a systematic difference in book prices between the two online booksellers? 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^+.

16.32 Significance test for online bookstore prices. Refer to the previous exercise. Find μ_{W^+}, σ_{W^+}, and the Normal approximation for the P-value for the Wilcoxon signed rank test. What do you conclude?

16.33 Potential insurance fraud? Insurance adjusters are concerned about the high estimates they are receiving from Jocko's Garage. To see if the estimates are unreasonably high, each of the 10 damaged cars was taken to Jocko's and to another garage, and the estimates (in dollars) were recorded. Here are the results. FRAUD

Car	1	2	3	4	5
Jocko's	1410	1550	1250	1300	900
Other	1250	1300	1250	1200	950
Car	6	7	8	9	10
Jocko's	1520	1750	3600	2250	2840
Other	1575	1600	3380	2125	2600

Test the null hypothesis that there is no difference between the two garages. Remember that zeros are dropped from the data before ranking, so that n is the number of nonzero differences within pairs.

16.34 It's all in the thumbs. New mobile phone software is advertised as resulting in faster text message creation (measured in characters per second) when compared with traditional mobile phone software. The traditional software offers suggested word completions based on a list of commonly used words, while the new software uses a smartword completion algorithm that takes into account the context of the text message. Ten individuals are involved in an experiment to test the marketing claim. The following data are the differences in their text message creation speeds: a positive value indicates more characters per second with the new mobile phone software. TXTMES

(new minus old) 5 3 0 1 −1 7 2 −2 4 0

(a) Plot these data on a Normal quantile plot. Would a t procedure be appropriate? Explain your answer.
(b) Is there statistical evidence to support the marketing claim? Remember that zeros are dropped from the data before ranking, so that n is the number of nonzero differences within pairs. Clearly state the relevant details from the signed rank test.

16.35 Oil-free frying comparison. Exercise 7.10 (page 371) describes an experiment in which a group of food experts to compare foods made with hot oil and their oil-free fryer. Here are the data. FRYER

Expert	1	2	3	4	5
Hot Oil	78	83	61	71	63
Oil Free	75	85	67	75	66

Using the signed rank test, is there a significant difference in taste? State appropriate hypotheses and carry out the test using $\alpha = 0.05$.

16.36 Loss of product value. In Example 16.6 (page 16-16), we considered the alternative hypothesis of vitamin C being systematically higher at the factory versus five months later when measured in Haiti. For the data provided in the example, the Wilcoxon signed rank statistic W^+ was found to be 34. On page 16-16, we noted that we could have equally used the sum of the ranks of the negative differences, which is 11. VITC
(a) If we were to use this sum instead, what would be the implied alternative hypothesis?
(b) Show how continuity correction would be used in conjunction with the value of 11 to arrive at the same P-value reported in Example 16.8 (page 16-19).

16.37 Marketing a health aid. Exercise that helps health and fitness should raise our heart rate for some period of time. A firm that markets a "Step Up to Health" apparatus consisting of a step and handrails for users to hold must tell buyers how to use their new device. The firm has subjects use the step at several stepping rates and measures their heart rates before and after stepping. Here are data for five subjects and two

treatments: low rate (14 steps per minute) and medium rate (21 steps per minute). For each subject, we give the resting heart rate (beats per minutes) and the heart rate at the end of the exercise.[12] HEART

	Low rate		Medium rate	
Subject	Resting	Final	Resting	Final
1	60	75	63	84
2	90	99	69	93
3	87	93	81	96
4	78	87	75	90
5	84	84	90	108

Does exercise at the low rate raise heart rate significantly? State hypotheses in terms of the median increase in heart rate and apply the Wilcoxon signed rank test. What do you conclude?

16.38 Marketing a health aid, continued. Do the data from the previous exercise give good reason to think that stepping at the medium rate increases heart rates more than stepping at the low rate? HEART
(a) State hypotheses in terms of comparing the median increases for the two treatments. What is the proper rank test for these hypotheses?
(b) Carry out your test and state a conclusion.

16.39 Executives learn Spanish. A matched pairs study of the effect of a language institute on the ability of executives to comprehend spoken Spanish had these improvements in scores between the pretest and the posttest for 20 executives: SPNISH1

−1	2	1	4	−4	−5	−3	3	5	5
2	−1	3	3	−2	7	2	4	1	3

(Exercise 7.34, page 377, shows the actual pretest and posttest scores.) Show the assignment of ranks and the calculation of the signed rank statistic W^+ for these data.

CASE 16.2 **16.40 Consumer perceptions of food safety.** Case 16.2 (page 16-10) describes a study of the attitudes of people attending outdoor fairs about the safety of the food served at such locations. In the associated data file, you will find the responses of 303 people to several questions. The variables in this data set are (in order)

subject hfair sfair sfast srest gender

The variable "sfair" contains responses to the safety question described in Case 16.2. The variable "srest" contains responses to the same question asked about food served in restaurants. We suspect that restaurant food will appear safer than food served outdoors at a fair. Do the data give good evidence for this suspicion? (Give descriptive measures, a test statistic and its P-value, and your conclusion.) FSAFETY

CASE 16.2 **16.41 Consumer perceptions of food safety.** The food safety survey data described in Case 16.2 (page 16-10) and Exercise 16.40 also contain the responses of the 303 subjects to the same question asked about food served at fast-food restaurants. These responses are the values of the variable "sfast." Is there a systematic difference between the level of concern about food safety at outdoor fairs and at fast-food restaurants? FSAFETY

16.42 Home 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.[13] 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?

16.43 The Platinum Gasaver. Platinum Gasaver is a device its maker says "may increase gas mileage by 22%." An advertisement reports the results of a matched pairs study with 15 identical vehicles. The claimed percent changes in gas mileage with the Gasaver are GASAVER

48.3	46.9	46.8	44.6	40.2	38.5	34.6	33.7
28.7	28.7	24.8	10.8	10.4	6.9	−12.4	

Is there good evidence that the Gasaver improves median gas mileage by 22% or more? (Apply the Wilcoxon signed rank test to the differences between the sample percents and the claimed 22%.)

16.44 Home radon detectors, continued. Some software (Minitab, for example) calculates a confidence interval for the population median as part of the Wilcoxon signed rank test. Using software and the data in Exercise 16.42, give a 95% confidence interval for the median reading of home radon detectors when exposed to 105 picocuries per liter of radon. RADON

16.45 Service call time. Some software (Minitab, for example) calculates a confidence interval for the population median as part of the Wilcoxon signed rank test. Exercise 1.37 (page 26) gives the service times for 80 calls to a customer service center. Using software, find a 95% confidence interval for the median service call time. CC80

CASE 1.2 **16.46 Time to start a business.** Case 1.2 (page 23) provides World Bank data for 24 countries on the time, in days, to complete all the procedures required to start a business in the country. Some software (Minitab, for example) calculates a confidence interval for the population median as part of the Wilcoxon signed rank test. Using software, find a 95% confidence interval for the median time required to start a business. TTS24

16.3 The Kruskal-Wallis Test

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

EXAMPLE 16.11 Weeds and Corn Yield

DATA WEEDS

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 zero, one, three, and nine 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	Standard deviation
0	4	170.200	5.422
1	4	162.825	4.469
3	4	161.025	10.493
9	4	157.575	10.118

REMINDER rule for standard deviations in ANOVA, p. 720

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. 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 16.11, 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.

Kruskal-Wallis test

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

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

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 16.12 Perform the Significance Test

WEEDS

In Example 16.11, there are $I = 4$ populations and $N = 16$ observations. The sample sizes are equal, $n_i = 4$. The 16 observations arranged in increasing order, with their ranks, are

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,

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

Referring to the table of chi-square critical points (Table 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 16.9 displays the output from Minitab and JMP for the analysis of the data in Example 16.12. Minitab gives the H statistic adjusted for ties as $H = 5.57$ with 3 degrees of freedom and $P = 0.134$. JMP 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 16.9 Output from (a) Minitab and (b) JMP for the Kruskal-Wallis test applied to the data in Example 16.11.

(a)

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)
```

(b)

JMP

Wilcoxon / Kruskal-Wallis Tests (Rank Sums)

Level	Count	Score Sum	Expected Score	Score Mean	(Mean-Mean0)/Std0
0	4	52.500	34.000	13.1250	2.184
1	4	33.500	34.000	8.3750	0.000
3	4	25.000	34.000	6.2500	-1.032
9	4	25.000	34.000	6.2500	-1.032

1-way Test, ChiSquare Approximation

ChiSquare	DF	Prob>ChiSq
5.5725	3	0.1344

SECTION 16.3 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 16.3 Exercises

16.47 French tourism economy. Ski resort activities make up nearly 20% of the French tourism economy. The French ski resort economy is under pressure to remain competitive in light of new entrants to the ski resort market, which are less expensive (for example, Slovenia and Montenegro). A research study was conducted to assess the productivity and efficiency of French ski resorts.[15] The study examined the productivity of 64 French ski resorts with the Luenberger Productivity Indicator (LPI) over a two-year time frame. LPI as an

overall measure of productivity is commonly used by economists because it can be decomposed into the usual constituents of productivity growth: technological change and efficiency change. A positive LPI indicates an increase in productivity, while a negative LPI indicates a decrease in productivity. The researcher of this study wished to investigate the relationship between the ski resorts' size and productivity. Ski resorts in the study were classified as being "large" (level 1), "medium" (level 2), or "small" (level 3). SKI

(a) What are the null hypothesis and the alternative hypothesis? Explain why a nonparametric procedure would be appropriate in this setting.

(b) Use the Kruskal-Wallis test to compare LPI across the three size classifications of ski resorts. Write a brief statement of your findings.

16.48 Evaluating an educational product. Case 14.2 (pages 733–734) considers the evaluation of a new educational product designed to improve children's reading comprehension. Three methods (Basal, SRTA, and Strat) are evaluated on three groups of 22 children. Your company markets educational materials aimed at parents of young children. The response variable is a measure of reading comprehension called COMP that was obtained by a test taken after the instruction was completed. Use the Kruskal-Wallis test to compare the three methods. EDUPROD

16.49 Loss of vitamin C 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 one, three, five, or seven days. The 10 observations are from 10 different loaves of bread.[16] 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 two 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, then write a brief summary of what the data show.

(b) Because there are only two 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 the SAS software) is $P = 0.0011$. Compare this with your P-value from (a). Is the difference large enough to affect your conclusion?

16.50 Exercise and bone density. Many studies have suggested that there is a link between exercise and healthy bones. Exercise stresses the bones and this causes them to get stronger. One study examined the effect of jumping on the bone density of growing rats.[17] 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. Here are the bone densities (in milligrams per cubic centimeter) after eight weeks of 10 jumps per day. BONE

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.

16.51 Decay of polyester fabric. In Exercise 16.17 (page 16-13), the breaking strengths (in pounds) of strips of polyester fabric buried in the ground were considered for two time points. Breaking strength is a good measure of the extent to which the fabric has decayed. Here are the breaking strengths for several lengths of time.[18] POLY2

Time	Breaking strength				
2 weeks	118	126	126	120	129
4 weeks	130	120	114	126	128
8 weeks	122	136	128	146	131
16 weeks	124	98	110	140	110

(a) Find the standard deviations of the four samples. They do not meet our rule of thumb for applying ANOVA. In addition, the sample buried for 16 weeks contains an outlier. We will use the Kruskal-Wallis test.

(b) Find the medians of the four samples. What are the hypotheses for the Kruskal-Wallis test, expressed in terms of medians?
(c) Carry out the test and report your conclusion.

CASE 16.2 **16.52 Food safety: Fairs, fast food, restaurants.** Case 16.2 (page 16-10) describes a study of the attitudes of people attending outdoor fairs about the safety of the food served at such locations. It contains the responses of 303 people to several questions. The variables in this data set are (in order): FSAFETY

subject hfair sfair sfast srest gender

The variable "sfair" contains responses to the safety question described in Case 16.2. 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 16 Review Exercises

16.53 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.

16.54 Design of controls. Exercise 7.36 (page 378) contains data from a student project that investigated whether right-handed people can turn a knob faster clockwise than they can counterclockwise. Describe what the data show, then state hypotheses and do a test that does not require Normality. Report your conclusions carefully. CNTROLS

16.55 Retail sales. Exercise 7.97 (page 411) gives monthly sales of 64GB flash drives at a sample of 50 retail stores. That exercise reports 95% confidence intervals for the mean sales in all stores obtained by the bootstrap method, which does not require Normality. Find a 95% confidence interval for the median sales and compare your results with those in Exercise 7.97. RETAIL

16.56 Drive-thru customer service. Exercise 7.57 (page 395) considered data gathered by **QSRMagazine.com** in the assessment of drive-thru customer service for the fast-food chains Taco Bell and McDonald's. Consider in this exercise the comparison between two competing burger fast-food chains: Burger King and Wendy's.[19] Responses ranged from "rude (1)" to "very friendly (5)." The following table breaks down the responses according to two of the chains studied. BURGER

	Rating				
Chain	**1**	**2**	**3**	**4**	**5**
Burger King	11	28	92	117	51
Wendy's	6	17	70	134	98

Is there evidence that one restaurant chain has systematically higher satisfaction ratings than the other?

16.57 Calories in hot dog brands. Table 16.1 presents data on the calorie and sodium content of selected brands of beef, meat, and poultry hot dogs.[20] We regard these brands as random samples from all brands available in food stores. HOTDOG
(a) Make stemplots of the calorie contents side by side, using the same stems for easy comparison. Give the five-number summaries for the three types of hot dog. What do the data suggest about the calorie content of different types of hot dog?
(b) Are any of the three distributions clearly not Normal? Which ones, and why?

TABLE 16.1 Calories and sodium in three types of hot dogs

Beef hot dogs		Meat hot dogs		Poultry hot dogs	
Calories	Sodium	Calories	Sodium	Calories	Sodium
186	495	173	458	129	430
181	477	191	506	132	375
176	425	182	473	102	396
149	322	190	545	106	383
184	482	172	496	94	387
190	587	147	360	102	542
158	370	146	387	87	359
139	322	139	386	99	357
175	479	175	507	170	529
148	375	136	393	113	513
152	330	179	405	135	426
111	300	153	372	142	513
141	386	107	144	86	358
153	401	195	511	143	581
190	645	135	405	152	588
157	440	140	428	146	522
131	317	138	339	144	545
149	319				
135	298				
132	253				

(c) Carry out a nonparametric test. Report your conclusions carefully.

16.58 Comparative study of U.S. and Indian firms. Are managerial efficiencies of similar U.S. and Indian firms comparable? A study was conducted to evaluate the relative managerial efficiencies of 14 match-paired U.S. and Indian firms.[21] Each pair of firms chosen in the study consists of one U.S. firm and one Indian firm producing similar product and having approximately the same size. Here are the data on the cumulative return on assets (ROA) for each firm over a five-year span. USIN

Paired firms	1	2	3	4	5	6	7
U.S.	34.32	9.61	11.73	20.74	23.58	10.38	40.24
Indian	59.87	39.25	13.13	32.79	18.13	28.14	81.45
Paired firms	8	9	10	11	12	13	14
U.S.	84.55	9.95	16.3	0.50	7.42	9.38	43.53
Indian	58.27	37.14	9.73	6.61	3.22	0.18	43.74

Test the null hypothesis that there is no difference between the two countries in terms of managerial efficiency as measured by ROA.

16.59 Sodium in hot dog brands. Repeat the analysis of Exercise 16.57 for the sodium content of hot dogs. HOTDOG

CASE 11.3 **16.60 House prices.** Case 11.3 (page 566) provides data the selling prices of 504 houses in West Lafayette, Indiana. We wonder if there is a difference between the average prices of houses with three bathrooms and houses with more than three bathrooms in this community. In the provided data file, three bathroom houses are coded by "0" and houses with more than three bathrooms are coded by "1." BATHS

(a) Make a Normal quantile plot of the prices of three bathroom houses. What kind of deviation from Normality do you see?

(b) The t tests are quite robust. State the hypotheses for the proper t test, carry out the test, and present your results including appropriate data summaries.

(c) Carry out a nonparametric test. Once more, state the hypotheses tested and present your results for both the test and the data summaries that should go with it.

(d) Compare the nonparametric result of part (c) with the t test result of part (b).

16.61 French tourism economy. Refer to Exercise 16.47 (page 16-27). The study also considered two other components of ski resort efficiency: technological change (TECH) and technical efficiency change (EFFCH). SKI

(a) With relationship to the TECH measure, the researcher states: "There is no reason to believe that size differences affect technological change." Is this statement consistent with your results? Explain.

(b) With relationship to the EFFCH measure, the researcher states: "A relationship between size and technical efficiency change can be made." Is this statement consistent with your results? Explain.

(c) With respect to the EFFCH measure, the researcher concludes: "These observations permit the conclusion that large ski resorts are better organized and ensure better management than medium and small ski resorts." Explain how the data support this conclusion.

16.62 Does the type of cooking pot affect iron content? Iron-deficiency anemia is the most common form of malnutrition in developing countries, affecting about 50% of children and women and 25% of men. Iron pots for cooking foods had traditionally been used in many of these countries, but they have been largely replaced by aluminum pots, which are cheaper and lighter. Some research has suggested that food cooked in iron pots will contain more iron than food cooked in other types of pots. One study designed to investigate this issue compared the iron content of some Ethiopian foods cooked in aluminum, clay, and iron pots.[22] One

of the foods was *yesiga wet'*, beef cut into small pieces and prepared with several Ethiopian spices. The iron content of four samples of yesiga wet' cooked in each of the three types of pots is given here. The units are milligrams of iron per 100 grams of cooked food.

Type of pot	Iron content (mg/100 g food)			
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

We want to know if the dish varies in iron content when cooked in aluminum, clay, and iron pots. COOK
(a) Check the requirements for one-way ANOVA. Which requirements are a bit dubious in this setting?
(b) Instead of ANOVA, do a nonparametric test. Summarize your conclusions about the effect of pot material on iron content, including both descriptive measures and your test result.

NOTES AND DATA SOURCES

1. *Condé Nast Traveler* readers poll data for 2013, from **cntraveler.com**.
2. This test was invented by Frank Wilcoxon (1892–1965) in 1945. Wilcoxon was a chemist who met statistical problems in his work at the research laboratories of American Cyanimid Company.
3. For purists, here is the precise definition: X_1 is *stochastically larger* than X_2 if

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

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.

4. Data from Huey Chern Boo, "Consumers' perceptions and concerns about safety and healthfulness of food served at fairs and festivals," M.S. thesis, Purdue University, 1997.
5. Discussion forum count taken from University of Wisconsin–Milwaukee MBA course titled "Business Analytics for Managers."
6. From Sapna Aneja, "Biodeterioration of textile fibers in soil," M.S. thesis, Purdue University, 1994.
7. Data obtained from **data.worldbank.org/indicator.**
8. Data loosely based on Alexander Redlein and Michael Zobl, "ERP systems within facility management," *Advanced Research in Scientific Areas Proceedings*, December 2013, pp. 153–155.
9. Data provided by Warren Page, New York City Technical College, from a study done by John Hudesman.
10. Data 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.
11. Data from **www.nytimes.com/best-sellers-books.**
12. Simplified from the EESEE story "Stepping Up Your Heart Rate."
13. Data provided by Diana Schellenberg, Purdue University School of Health Sciences.
14. Data provided by Sam Phillips, Purdue University.
15. Data from Olga Goncalves, "Efficiency and productivity of French ski resorts," *Tourism Management* 36 (2013), pp. 650–657.
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. See Note 6.
19. The 2013 study can be found at **www.qsrmagazine.com/content/drive-thru-performance-study-customer-service.**
20. *Consumer Reports*, June 1986, pp. 366–367.
21. Data from Parviz Asheghian, "The managerial efficiencies of Indian firms as compared to American firms," *International Journal of Economics and Management Sciences* 6 (2012), pp. 45–55.
22. 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.

ANSWERS TO ODD-NUMBERED EXERCISES

16.1 Group A ranks are 1, 2, 3, 4, 8. Group B ranks are 5, 6, 7, 9, 10.

16.3 H_0: No difference in distribution of the number of rooms among the 25 top-ranked spas and the spas ranked 26 to 50. H_α: There is a systematic difference in distribution of the number of rooms among the 25 top-ranked spas and the spas ranked 26 to 50. $W = 18$.

16.5 $\mu_W = 27.5$, $\sigma_W = 4.7871$. We found $W = 18$, so using continuity correction, we get $z = -1.88$, P-value $= 0.0301$. There is evidence to show a systematic difference in the number of rooms among the 25 top-ranked spas and the spas ranked 26 to 50 at the 5% significance level.

16.7 There are many tied values; the table below shows the corresponding ranks.

Type	**BMW**	**Mercedes**	**BMW**	**BMW**
CO_2	151	152	161	166
Rank	1	2	3	5
#	1	1	1	3
Type	**BMW**	**BMW**	**Mercedes**	**Mercedes**
CO_2	172	179	182	186
Rank	8.5	11.5	13	14
#	4	2	1	1
Type	**BMW**	**Mercedes**	**BMW**	**BMW**
CO_2	189	197	198	200
Rank	15.5	17	18.5	20.5
#	2	1	2	2
Type	**BMW**	**BMW**	**Mercedes**	**Mercedes**
CO_2	202	205	207	209
Rank	22	23.5	25	26.5
#	1	2	1	2
Type	**Mercedes**	**BMW**	**BMW**	**Mercedes**
CO_2	232	246	255	264
Rank	28	29	30	32
#	1	1	1	3
Type	**Mercedes**	**Mercedes**		
CO_2	271	306		
Rank	34	35		
#	1	1		

16.9 **(a)** $\mu_W = 84$, $\sigma_W = 12.961$. **(b)** $84 + 126 = 210$, which is the sum of the ranks. **(c)** $z = -2.20$, P-value $= 0.0139$, which is the same as in Example 16.3.

16.11 There are no tied values; the following table shows the corresponding ranks.

Sex	**M**	**M**	**M**	**M**	**F**	**M**	**M**	**M**
Posts	3	4	5	9	10	12	13	14
Rank	1	2	3	4	5	6	7	8
Sex	**M**	**F**	**M**	**M**	**F**	**F**	**F**	**F**
Posts	15	16	17	18	20	23	27	31
Rank	9	10	11	12	13	14	15	16

16.13 H_0: There is no difference in distribution of the number of posts between females and males. H_α: One gender has a systematically higher number of posts than the other.

16.15 $z = 2.33$, P-value $= 0.0198$. There is evidence of a systematic difference in the number of posts by gender.

16.17 **(a)** The distributions are very different and not the same shape. **(b)** From software, $z = -1.05$, P-value $= 0.1467$. The data do not show a systematic difference in breaking strengths.

16.19 **(a)** $W = 124.5$. Yes, $124.5 + 85.5 = 210$. **(b)** $z = 2.839$, P-value $= 0.0045$.

16.21 **(b)** The histograms show that the differences for the Treatment group appear higher than those for the control group. **(c)** $z = 1.6506$, P-value $= 0.0494$. The treatment group has systematically higher differences than the control group; the subliminal messages appear to work.

16.23 $W = 32267.5$, $z = 3.8133$, P-value < 0.0001. Women are systematically more concerned than men about the safety of food served at restaurants.

16.25 **(b)** $W = 357$. **(c)** $z = 3.4008$, P-value $= 0.0003$. People in a sad mood systematically spend more than people not in a sad mood.

16.27 H_0: There is no difference in the distribution between service and food scores. H_α: Service scores are systematically higher than food scores. $W^+ = 28$.

16.29 $\mu_{W+} = 14$, $\sigma_{W+} = 5.9161$. $z = 2.28$, P-value $= 0.0113$.

16.31 H_0: There is no difference in the distribution of price between Barnes & Noble and Amazon. H_α: One online retailer has systematically higher prices than the other. $W^+ = 36$.

16.33 H_0: There is no difference in the distribution of estimates between Jocko's garage and the other garage. H_α: Jocko's estimates are systematically higher than the other garage's estimates. $W^+ = 42$. $\mu_{W+} = 22.5$, $\sigma_{W+} = 8.441$. $z = 2.26$, P-value $= 0.0119$. Jocko's estimates are systematically higher than the other garage's estimates.

16.35 H_0: There is no difference in the distribution of taste between food made with hot oil and food made oil free. H_α: One oil preparation method has systematically higher taste ratings than the other. $W^+ = 2.5$. $\mu_{W+} = 7.5$, $\sigma_{W+} = 3.7081$. $z = -1.21$, P-value $= 0.2262$. The data do not show a systematic difference in food taste between food made with hot oil and food made oil free.

16.37 H_0: median increase$_{low} = 0$, H_α: median increase$_{low} > 0$. $W^+ = 10$. $\mu_{W+} = 5$, $\sigma_{W+} = 2.7386$. $z = 1.64$, P-value $= 0.0505$. The data do not show a systematic difference in heart rate.

16.39 $W^+ = 154.5$.

16.41 From software: P-value $= 0.206$. The data do not show a systematic difference between the level of concern about food safety at outdoor fairs and at fast-food restaurants.

16.43 From software: $W^+ = 88$, P-value $= 0.059$. The data do not show a systematic difference between the sample percents and the claimed 22%.

16.45 (95.5, 161.0).

16.47 **(a)** H_0: LPIs have the same distribution in all groups. H_α: LPIs are systematically higher in some groups than in others. The standard deviations for the three sizes are very different. **(b)** $H = 6.8205$, P-value $= 0.0330$. There are significant differences in LPI across different resort sizes.

16.49 H_0: Vitamin C values have the same distribution in all groups. H_α: Vitamin C values are systematically higher in some groups than in others. $H = 8.7273$, P-value $= 0.0683$. The data do not show differences in vitamin C across different conditions. Although the test is not significant, looking at the data suggests that there is vitamin C lost over time. **(b)** With the new P-value from SAS, we would reject the null hypothesis and conclude there are systematically higher vitamin C values in some groups than in others, showing the loss of vitamin C over time.

16.51 **(a)** The standard deviations are 4.60, 6.54, 9.04, 16.09. **(b)** The medians are 126, 126, 131, 110. H_0: The median breaking strength for all groups are the same. H_α: Some medians are higher in some groups than in others. **(c)** $H = 5.3797$, P-value $= 0.1460$. The data do not show median breaking strength differences across different times; there appears to be no decay.

16.53 **(a)** The CLEC customers generally have to wait longer than the Verizon customers for repairs. **(b)** The data are count data and far from normally distributed. $W = 786.5$. $z = 3.2399$, P-value $= 0.0006$. The CLEC customers have systematically higher repair wait times than the Verizon customers. **(c)** The CLEC group only has two values, 1 and 5. The Verizon group is strongly right skewed and has a large outlier.

16.55 (25, 32). This is quite close to the confidence interval for the mean we found in Exercise 7.97, (26.06, 33.18).

16.57 **(a)** It looks like beef and meat hotdogs have more calories than poultry hotdogs.

Analysis Variable : Calories					
Type	Minimum	Lower Quartile	Median	Upper Quartile	Maximum
Beef	111	140	152.5	178.5	190
Meat	107	139	153	179	195
Poultry	86	102	129	143	170

(b) None of the distributions look very normal. The beef and meat hotdogs have similar distributions: somewhat uniform with a gap in the 160s, creating a bimodal distribution. There also may be one low outlier in each. The poultry hotdogs also has a somewhat bimodal distribution with a low point around 120; it additionally may have a high outlier. **(c)** H_0: Calories have the same distribution for all types of hotdogs. H_α: Calories are systematically higher for some types of hotdogs than in others. $H = 15.8994$, P-value $= 0.0004$. The data show that some hotdog types have systematically higher calories than others.

16.59 **(a)** It looks like poultry hotdogs have a slightly higher sodium content, followed by meat hotdogs;

beef hotdogs have the lowest sodium content. **(b)** The beef hotdogs have right-skewed distribution with maybe two potential high outliers with large sodium contents. The meat hotdogs category is roughly normal but has one extreme low outlier, with less than half the sodium of all other hotdogs. The poultry hotdogs have a bimodal distribution with two main groups, one with low sodium amounts and one with high sodium amounts. **(c)** H_0: Sodium contents have the same distribution for all types of hotdogs. H_α: Sodium contents are systematically higher for some types of hotdogs than in others. $H = 4.7128$, P-value $= 0.0948$. The data do not show systematic differences among sodium content for different types of hotdogs.

16.61 **(a)** This statement is correct. H_0: TECH values have the same distribution in all groups. H_α: TECH values are systematically higher in some groups than in others. $H = 3.6386$, P-value $= 0.1621$. The data do not show systematic differences in TECH values across resort sizes. **(b)** This statement is correct. H_0: EFFCH values have the same distribution in all groups. H_α: EFFCH values are systematically higher in some groups than in others. $H = 7.0163$, P-value $= 0.0300$. The data show systematic differences in EFFCH values across resort sizes. **(c)** The rank sum for the large resort size is much larger than expected under H_0.