CHAPTER 16

© Alan Thorton/Getty Images

Nonparametric (Distribution-Free) Statistical Methods

Many of the inferential techniques presented in earlier chapters required specific assumptions about the shape of the relevant population or treatment distributions. For example, when sample sizes are small, the validity of the two-sample t test and of the ANOVA F test rests on the assumption of normal population distributions. This chapter introduces some inferential procedures that do not depend on specific assumptions about the population distributions, such as normality. Such methods are described as nonparametric or distribution-free.

16.1 Distribution-Free Procedures for Inferences About a Difference Between Two Population or Treatment Means Using Independent Samples

One approach to making inferences about $\mu_1 - \mu_2$ when n_1 and n_2 are small is to assume that the two population or treatment response distributions are normal and then to use the two-sample t test or confidence interval presented in Section 11.1. In some situations, however, the normality assumption may not be reasonable. The validity of the procedures developed in this section does not depend on the normality of the population distributions. The procedures can be used to compare two populations or treatments when it is reasonable to assume that the population or treatment response distributions have the same shape and spread.

Basic Assumptions for Methods in This Section

The two population or treatment response distributions have the same shape and spread. The only possible difference between the distributions is that one may be shifted to one side or the other.

Distributions consistent with these assumptions are shown in Figure 16.1(a). The distributions shown in Figure 16.1(b) have different shapes and spreads and so the methods of this section would not be appropriate in this case. Inferences that involve comparing distributions with different shapes and spreads can be quite complicated; if you encounter that situation, good advice from a statistician is particularly important.

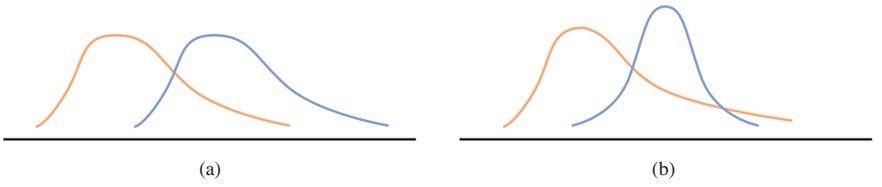


FIGURE 16.1
Two possible population distribution pairs: (a) same shape and spread, differing only in location; (b) very different shape and spread.

Procedures that do not require any overly specific assumptions about the population distributions are said to be **distribution-free or nonparametric**. The two-sample t test with small samples is not distribution-free because its appropriate use depends on the specific assumption of (at least approximate) normality.

Inferences about $\mu_1 - \mu_2$ are made using information from two independent random samples, one consisting of n_1 observations from the first population and the other consisting of n_2 observations from the second population. Suppose that the two population distributions are in fact identical (so that $\mu_1 = \mu_2$). In this case, each of the $n_1 + n_2$ observations is actually drawn from the same population distribution. The distribution-free procedure presented here is based on regarding the $n_1 + n_2$ observations as a single data set and assigning ranks to the ordered values. The assignment is easiest when there are no ties among the $n_1 + n_2$ values (each observation is different from every one of the other observations), so assume for the moment that

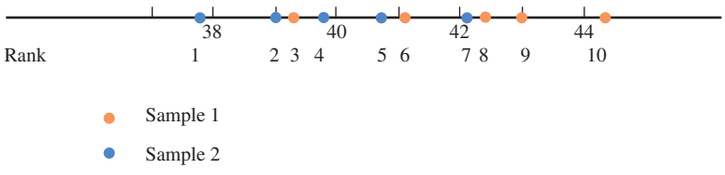
this is the case. Then the smallest among the $n_1 + n_2$ values receives rank 1, the second smallest rank 2, and so on, until finally the largest value is assigned rank $n_1 + n_2$. This procedure is illustrated in Example 16.1.

EXAMPLE 16.1 Comparing Fuel Efficiency

● An experiment to compare fuel efficiencies for two models of subcompact automobile was carried out by first randomly selecting $n_1 = 5$ cars of Model 1 and $n_2 = 5$ cars of Model 2. Each car was then driven from Phoenix to Los Angeles by a nonprofessional driver, after which the fuel efficiency (in miles per gallon) was determined. The resulting data, with observations in each sample ordered from smallest to largest, are given here:

Model 1	39.3	41.1	42.4	43.0	44.4
Model 2	37.8	39.0	39.8	40.7	42.1

The data and the associated ranks are shown in the following dotplot.



The ranks of the five observations in the first sample are 3, 6, 8, 9, and 10. If these five observations had all been larger than every value in the second sample, the corresponding ranks would have been 6, 7, 8, 9, and 10. On the other hand, if all five Sample 1 observations had been less than each value in the second sample, the ranks would have been 1, 2, 3, 4, and 5. The ranks of the five observations in the first sample might be any set of five numbers from among 1, 2, 3, … ,9, 10—there are actually 252 possibilities.

● Data set available online

Testing Hypotheses

Let's first consider testing

$$H_0\text{: } \mu_1 - \mu_2 = 0 \ (\mu_1 = \mu_2) \quad \text{versus} \quad H_a\text{: } \mu_1 - \mu_2 > 0 \ (\mu_1 > \mu_2)$$

If H_0 is true, all $n_1 + n_2$ observations in the two samples are actually drawn from identical population distributions. We would then expect that the observations in the first sample would be intermingled with those of the second sample when plotted along the number line. In this case, the ranks of the observations should also be intermingled. For example, with $n_1 = 5$ and $n_2 = 5$, the set of Sample 1 ranks 2, 3, 5, 8, 10 would be consistent with $\mu_1 = \mu_2$, as would the set 1, 4, 7, 8, 9. However, when $\mu_1 = \mu_2$, it would be quite unusual for all five values from Sample 1 to be larger than every value in Sample 2, resulting in the set 6, 7, 8, 9, 10 of Sample 1 ranks.

A convenient measure of the extent to which the ranks are intermingled is the sum of the Sample 1 ranks. These ranks in Example 16.1 were 3, 6, 8, 9, and 10, so

$$\text{rank sum} = 3 + 6 + 8 + 9 + 10 = 36$$

The largest possible rank sum when $n_1 = n_2 = 5$ is $6 + 7 + 8 + 9 + 10 = 40$. If μ_1 is much larger than μ_2, we would expect the rank sum to be near its largest possible value. This suggests that we reject H_0 for unusually large values of the rank sum.

Developing a test procedure requires information about the sampling distribution of the rank-sum statistic when H_0 is true. To illustrate this, consider again the case $n_1 = n_2 = 5$. There are 252 different sets of 5 from among the 10 ranks 1, 2, 3, … , 9, 10. The key point is that, when H_0 is true, any 1 of these 252 sets has the same chance of being the set of Sample 1 ranks as does any other set, because all 10 observations come from the same population distribution. The chance under H_0 that any particular set occurs is 1/252 (because the possibilities are equally likely).

Table 16.1 displays the 12 sets of Sample 1 ranks that yield the largest rank-sum values. Each of the other 240 possible rank sets has a rank-sum value less than 36. If we observe rank sum = 36, we can compute

$$P(\text{rank-sum} \geq 36 \text{ when } H_0 \text{ is true}) = 12/252 = .0476$$

TABLE 16.1 The 12 Rank Sets That Have the Largest Rank Sums When $n_1 = 5$ and $n_2 = 5$

Sample 1 Ranks	Rank Sum	Sample 1 Ranks	Rank Sum
6 7 8 9 10	40	5 6 7 9 10	37
5 7 8 9 10	39	2 7 8 9 10	36
4 7 8 9 10	38	3 6 8 9 10	36
5 6 8 9 10	38	4 5 8 9 10	36
3 7 8 9 10	37	4 6 7 9 10	36
4 6 8 9 10	37	5 6 7 8 10	36

That is, when H_0 is true, a rank sum at least as large as 36 would be observed only about 4.76% of the time. Thus, a test of H_0: $\mu_1 - \mu_2 = 0$ versus H_a: $\mu_1 - \mu_2 > 0$, based on $n_1 = 5$ and $n_2 = 5$ and a rank sum of 36, would have an associated P-value of .0476. It is this type of reasoning that allows us to reach a conclusion about whether or not to reject H_0.

The process of looking at different rank-sum sets to carry out a test can be quite tedious. Fortunately, information about both one- and two-tailed P-values associated with values of the rank-sum statistic has been tabulated. Chapter 16 Appendix Table 1 provides information on P-values for selected values of n_1 and n_2. For example, when n_1 and n_2 are both 5, Chapter 16 Appendix Table 1 tells us that, for an upper-tailed test with a rank-sum statistic value of 36, P-value $< .05$. This is consistent with the value of .0476 computed previously. Had the rank sum been 40, using the table, we would have concluded that the P-value was less than .01 (because 40 is greater than the tabled value of 39). The use of this appendix table is further illustrated in the examples that follow.

Summary of the Rank-Sum Test*

Null hypothesis: H_0: $\mu_1 - \mu_2 = 0$

Test statistic: rank sum = sum of ranks assigned to the observations in the first sample

(continued)

Alternative Hypothesis	Type of Test
H_a: $\mu_1 - \mu_2 > 0$	Upper-tailed
H_a: $\mu_1 - \mu_2 < 0$	Lower-tailed
H_a: $\mu_1 - \mu_2 \neq 0$	Two-tailed

Information about the *P*-value associated with this test is given in Chapter 16 Appendix Table 1.

Assumptions: 1. *The samples are independent random samples* OR *treatments are randomly assigned to individuals or objects* (or subjects randomly assigned to treatments).
2. The two *population or treatment response distributions have the same shape and spread.*

*This procedure is often called the *Wilcoxon rank-sum test* or the *Mann–Whitney test,* after the statisticians who developed it. Some sources use a slightly different (but equivalent) test statistic formula.

EXAMPLE 16.2 Parental Smoking and Infant Health

● The extent to which an infant's health is affected by parents' smoking is an important public health concern. The article "Measuring the Exposure of Infants to Tobacco Smoke" (*New England Journal of Medicine* [1984]: 1075–1078) reported on a study in which various measurements were taken both from a random sample of infants who had been exposed to household smoke and from a sample of unexposed infants. The following data consist of observations on urinary concentration of cotinine, a major metabolite of nicotine (the values constitute a subset of the original data and were read from a plot that appeared in the article):

Unexposed ($n_1 = 7$)	8	11	12	14	20	43	111	
Rank	1	2	3	4	5	7	11	
Exposed ($n_2 = 8$)	35	56	83	92	128	150	176	208
Rank	6	8	9	10	12	13	14	15

Do the data suggest that the mean cotinine level is higher for exposed than for unexposed infants? The investigators used the rank-sum test to analyze the data.

1. $\mu_1 - \mu_2$ is the difference between mean cotinine concentration for unexposed and exposed infants.
2. H_0: $\mu_1 - \mu_2 = 0$.
3. H_a: $\mu_1 - \mu_2 < 0$ (unexposed mean is less than exposed mean).
4. Significance level: $\alpha = .01$.
5. Test statistic: rank sum = sum of sample 1 ranks.
6. Assumptions: The authors of the article indicated that they believe the assumptions of the rank-sum test were reasonable.
7. Calculation: rank sum = 1 + 2 + 3 + 4 + 5 + 7 +11 = 33.
8. *P*-value: This is a lower-tailed test. With $n_1 = 7$ and $n_2 = 8$, Chapter 16 Appendix Table 1 tells us that *P*-value $< .05$ if the rank sum ≤ 41 and that *P*-value $< .01$ if the rank sum ≤ 36. Because the rank sum = 33, we conclude that *P*-value $< .01$.
9. Conclusion: Because the *P*-value is less than α (.01), we reject H_0 and conclude that there is convincing evidence that infants exposed to cigarette smoke have a higher mean cotinine level than unexposed infants.

● Data set available online

Many statistical computer packages can perform the rank-sum test and give exact *P*-values. Partial Minitab output for the data of Example 16.2 follows (Minitab uses the symbol W to denote the rank-sum statistic and the terms *ETA1* and *ETA2* in place of μ_1 and μ_2, but the test statistic value and the associated *P*-values are the same as for the test presented here):

```
Unexposed      N = 7      Median = 14.00
Exposed        N = 8      Median = 110.00
Point estimate for ETA1 – ETA2 is –79.00
95.7 Percent CI for ETA1 – ETA2 is (–156.00, –23.99)
W = 33.0
Test of ETA1 = ETA2 vs ETA1 < ETA2 is significant at 0.0046
```

The output indicates that $W = 33$ and that the test is significant at .0046. These two values, 33 and .0046, are the rank-sum statistic and the *P*-value, respectively. In Example 16.2, we used Chapter 16 Appendix Table 1 to determine that *P*-value $< .01$. This statement is consistent with the actual *P*-value given in the Minitab output.

The test procedure just described can be easily modified to handle a hypothesized value other than 0. Consider as an example testing H_0: $\mu_1 - \mu_2 = 5$. This hypothesis is equivalent to $(\mu_1 - 5) - \mu_2 = 0$. That is, if 5 is subtracted from each Population 1 value, then according to H_0, the distribution of the resulting values coincides with the Population 2 distribution. This suggests that, if the hypothesized value of 5 is first subtracted from each Sample 1 observation, the test can then be carried out as before.

To test H_0: $\mu_1 - \mu_2 =$ *hypothesized value*, subtract the hypothesized value from each observation in the first sample and then determine the ranks when these modified sample 1 values are combined with the n_2 observations from the second sample.

EXAMPLE 16.3 Parental Smoking and Infant Health Revisited

● Reconsider the cotinine concentration data introduced in Example 16.2. Suppose that a researcher wished to know whether mean concentration for exposed children exceeds that for unexposed children by more than 25. Recall that μ_1 denotes the mean concentration for unexposed children. The exposed mean exceeds the unexposed mean by 25 when $\mu_1 - \mu_2 = -25$ and by more than 25 when $\mu_1 - \mu_2 < -25$. The hypotheses of interest are therefore

$$H_0: \mu_1 - \mu_2 = -25 \quad \text{versus} \quad H_a: \mu_1 - \mu_2 < -25$$

These hypotheses can be tested by first subtracting -25 (or, equivalently, adding 25) to each Sample 1 observation.

Sample 1								
Unexposed	8	11	12	14	20	43	111	
Unexposed $-(-25)$	33	36	37	39	45	68	136	
Rank	1	3	4	5	6	8	12	
Sample 2								
Exposed	35	56	83	92	128	150	176	208
Rank	2	7	9	10	11	13	14	15

● Data set available online

1. $\mu_1 - \mu_2$ = difference in mean cotinine concentration for unexposed and exposed infants
2. H_0: $\mu_1 - \mu_2 = -25$
3. H_a: $\mu_1 - \mu_2 < -25$
4. Significance level: $\alpha = .01$
5. Test statistic: rank sum = sum of sample 1 ranks
6. Assumptions: Subtracting -25 does not change the shape or spread of a distribution, so if the assumptions were reasonable in Example 16.2, they are also reasonable here.
7. Calculation: rank sum = 1 + 3 + 4 + 5 + 6 + 8 + 12 = 39
8. *P*-value: This is a lower-tailed test. With $n_1 = 7$ and $n_2 = 8$, Chapter 16 Appendix Table 1 tells us that *P*-value $< .05$ if rank sum ≤ 41 and *P*-value $< .01$ if rank sum ≤ 36. Since rank sum = 39, we conclude that $.01 <$ *P*-value $< .05$.
9. Conclusion: Since *P*-value $> .01$, we fail to reject H_0. Sample evidence does not suggest that the mean concentration level for exposed infants is more than 25 higher than the mean for unexposed infants.

Frequently the $n_1 + n_2$ observations in the two samples are not all different from one another. When this occurs, the rank assigned to each observation in a tied group is the mean of the ranks that would be assigned if the values in the group all differed slightly from one another. Consider, for example, the 10 ordered values

5.6 6.0 6.0 6.3 6.8 7.1 7.1 7.1 7.9 8.2

If the two 6.0 values differed slightly from each other, they would be assigned ranks 2 and 3. Therefore, each one is assigned rank $(2 + 3)/2 = 2.5$. If the three 7.1 observations were all slightly different, they would receive ranks 6, 7, and 8, so each of the three is assigned rank $(6 + 7 + 8)/3 = 7$. The ranks for the above 10 observations are then

1 2.5 2.5 4 5 7 7 7 9 10

If the proportion of tied values is quite large, it is recommended that the rank-sum statistic be multiplied by a *correction factor*. Consult the references by Conover, Daniel, or Mosteller and Rourke for additional information.

Chapter 16 Appendix Table 1 contains information about *P*-values for the rank-sum test when $n_1 \leq 8$ and $n_2 \leq 8$. More extensive tables exist for other combinations of sample size, but with larger sample sizes, you may want to use a statistical software package to compute the value of the test statistic and the associated *P*-value. There is also a test procedure based on using a normal distribution to approximate the sampling distribution of the rank-sum statistic. This alternative procedure is often used when the two sample sizes are larger than 8.

A Confidence Interval for $\mu_1 - \mu_2$

A confidence interval based on the rank-sum statistic is not nearly as familiar to users of statistical methods as is the hypothesis-testing procedure. This is unfortunate, because the confidence interval is appropriate in the same situations as the rank-sum test. The assumption that the population distributions are normal, which is needed for the two-sample *t* interval with small samples, is not required.

The actual derivation of the rank-sum confidence interval is quite involved, and computing these intervals by hand can be tedious, so we rely on computer software.

The **rank-sum confidence interval for $\mu_1 - \mu_2$** is the interval consisting of all hypothesized values for which

H_0: $\mu_1 - \mu_2$ = hypothesized value

cannot be rejected when using a two-tailed test.

A 95% confidence interval consists of those hypothesized values for which the previous null hypothesis is not rejected by a test with significance level $\alpha = .05$. A 99% confidence interval is associated with a level $\alpha = .01$ test, and a 90% confidence interval is associated with a level $\alpha = .10$ test.

Thus, if H_0: $\mu_1 - \mu_2 = 100$ cannot be rejected at level .05, then 100 is included in the 95% confidence interval $\mu_1 - \mu_2$.

EXAMPLE 16.4 Strength of Bark Board

● The article "Some Mechanical Properties of Impregnated Bark Board" (*Forest Products Journal* [1977]: 31–38) reported the following observations on crushing strength for epoxy-impregnated bark board (Sample 1) and bark board impregnated with another polymer (Sample 2):

Sample 1	10,860	11,120	11,340	12,130	13,070	14,380
Sample 2	4,590	4,850	5,640	6,390	6,510	

A 95% confidence interval for $\mu_1 - \mu_2$ was requested using Minitab. The resulting output follows:

```
Sample 1    N = 6    Median = 11735
Sample 2    N = 5    Median = 5640
Point estimate for ETA1 – ETA2 is 6490
96.4 percent CI for ETA1 – ETA2 is (4730, 8480)
```

Remember that Minitab uses ETA1 − ETA2 in place of $\mu_1 - \mu_2$. Also note that it was not possible to construct an interval for an exact 95% confidence level. Minitab calculated a 96.4% confidence interval. The reported interval is (4730, 8480). Based on the sample information, we estimate that the difference in mean crushing strength for epoxy-impregnated bark board and board impregnated with a different polymer is between 4730 and 8480 psi.

EXERCISES 16.1 - 16.7

16.1 ● Urinary fluoride concentration (in parts per million) was measured for both a sample of livestock that had been grazing in an area previously exposed to fluoride pollution and a similar sample of livestock that had grazed in an unpolluted region. Do the accompanying data indicate strongly that the mean fluoride concentration for livestock grazing in the polluted region is larger than that for livestock grazing in the unpolluted region? Assume that the distributions of urinary fluoride concentration for both grazing areas have the same shape and spread, and use a level .05 rank-sum test.

Polluted	21.3	18.7	23.0	17.1	16.8	20.9	19.7
Unpolluted	14.2	18.3	17.2	18.4	20.0		

16.2 ● A modification has been made to the process for producing a certain type of film. Because the modification involves extra cost, it will be incorporated only if sample data strongly indicate that the modification decreases

mean developing time by more than 1 second. Assuming that the developing-time distributions have the same shape and spread, use the rank-sum test at level .05 and the following data to test the appropriate hypotheses:

Original process	8.6	5.1	4.5	5.4	6.3	6.6	5.7	8.5
Modified process	5.5	4.0	3.8	6.0	5.8	4.9	7.0	5.7

16.3 ● The study reported in "Gait Patterns During Free Choice Ladder Ascents" (*Human Movement Science* [1983]: 187–195) was motivated by publicity concerning the increased accident rate for individuals climbing ladders. A number of different gait patterns were used by subjects climbing a portable straight ladder according to specified instructions. The following data consist of climbing times for seven subjects who used a lateral gait and six subjects who used a four-beat diagonal gait:

Lateral gait	0.86	1.31	1.64	1.51	1.53	1.39	1.09
Diagonal gait	1.27	1.82	1.66	0.85	1.45	1.24	

a. Use the rank-sum test to decide whether the data suggest a difference in the mean climbing times for the two gaits.

b. Interpret the 95% confidence interval for the difference between the mean climbing times given in the following Minitab output:

```
lateral      N = 7     Median = 1.3900
diagonal     N = 6     Median = 1.3600
Point estimate for ETA1 – ETA2 is 0.0400
96.2 percent C.I. for ETA1 – ETA2 is (–0.4300, 0.3697)
```

16.4 ● A blood lead level of 70 mg /ml has been commonly accepted as safe. However, researchers have noted that some neurophysiological symptoms of lead poisoning appear in people whose blood lead levels are below 70 mg/ml. The article "Subclinical Neuropathy at Safe Levels of Lead Exposure" (*Archives of Environmental Health* [1975]: 180–183) gave the following nerve-conduction velocities for a group of workers who were exposed to lead in the workplace but whose blood lead levels were below 70 mg/ml and for a group of controls who had no exposure to lead:

Exposed to lead	46	46.5	43	41	38	36	31
Control	54	50.5	46	45	44	42	41

Use a level .05 rank-sum test to determine whether there is a significant difference in mean conduction velocity between workers exposed to lead and those not exposed to lead.

16.5 ● The effectiveness of antidepressants in treating the eating disorder bulimia was examined in the article "Bulimia Treated with Imipramine: A Placebo-Controlled Double-Blind Study" (*American Journal of Psychology* [1983]: 554–558). A group of patients diagnosed with bulimia were randomly assigned to one of two treatment groups, one receiving imipramine and the other a placebo. One of the variables recorded was binge frequency. The authors chose to analyze the data using a rank-sum test because it makes no assumption of normality. They stated that "because of the wide range of some measures, such as frequency of binges, the rank sum is more appropriate and somewhat more conservative." Data on number of binges during one week that are consistent with the findings of the article are given in the following table:

Placebo	8	3	15	3	4	10	6	4
Imipramine	2	1	2	7	3	12	1	5

Do these data strongly suggest that imipramine is effective in reducing the mean number of binges per week? Use a level .05 rank-sum test.

16.6 ● In an experiment to compare the bond strength of two different adhesives, each adhesive was used in five bondings of two surfaces, and the force necessary to separate the surfaces was determined for each bonding. For Adhesive 1, the resulting values were 229, 286, 245, 299, and 259, whereas the Adhesive 2 observations were 213, 179, 163, 247, and 225. Let μ_1 and μ_2 denote the mean bond strengths of Adhesives 1 and 2, respectively. Interpret the 90% distribution-free confidence interval estimate of $\mu_1 - \mu_2$ given in the Minitab output shown here:

```
adhes. 1     N = 5     Median = 259.00
adhes. 2     N = 5     Median = 213.00
Point estimate for ETA1 – ETA2 is 61.00
90.5 Percent C.I. for ETA1 – ETA2 is (16.00, 95.98)
```

16.7 ● The article "A Study of Wood Stove Particulate Emissions" (*Journal of the Air Pollution Control Association* [1979]: 724–728) reported the following data on burn time (in hours) for samples of oak and pine:

Oak	1.72	0.67	1.55	1.56	1.42	1.23	1.77	0.48
Pine	0.98	1.40	1.33	1.52	0.73	1.20		

An estimate of the difference between mean burn time for oak and mean burn time for pine is desired. Interpret the interval given in the following Minitab output:

```
Oak  N = 8     Median = 1.4850
Pine N = 6     Median = 1.2650
Point estimate for ETA1 – ETA2 is 0.2100
95.5 Percent C.I. for ETA1 – ETA2 is (0.4998, 0.5699)
```

16.2 Distribution-Free Procedures for Inferences About a Difference Between Two Population or Treatment Means Using Paired Samples

In Section 11.2, the paired t test and paired t confidence interval were used to make inferences about μ_d, the population mean difference. These methods are appropriate when it is reasonable to assume that the difference population (from which the sample differences were randomly selected) is normal in shape. Since this may not always be the case, in this section we present an alternate test procedure, called the *signed-rank test*, and an associated confidence interval. These procedures are also based on the sample differences, but their validity requires only that the difference distribution be symmetric in shape. Symmetry is a weaker condition than normality (any normal distribution is symmetric, but there are many symmetric distributions that are not normal), so the signed-rank procedures are more widely valid than are the paired t procedures. Since the signed-rank procedures do not depend on specific distributional assumptions such as normality, they are distribution-free. A sufficient condition for the difference distribution to be symmetric is that the two populations (from which the first and second observations in each pair are drawn) are identical with respect to shape and spread.

As with the paired t test, we begin by forming differences. Next, the absolute values of the differences are assigned ranks (this amounts to ignoring any negative signs when ranking). We then associate a + or a − sign with each rank, depending on whether the corresponding difference is positive or negative. For example, with $n = 5$, the differences might be

$$-17 \quad 12 \quad 3 \quad 10 \quad -6$$

The ordered absolute differences are then

$$3 \quad 6 \quad 10 \quad 12 \quad 17$$

and the corresponding signed ranks are

$$1 \quad -2 \quad 3 \quad 4 \quad -5$$

negative because the corresponding difference is negative

If there are ties in the differences, the average of the appropriate ranks is assigned, as was the case with the rank-sum test in Section 16.1.

The signed-rank test statistic for testing H_0: $\mu_d = 0$ is the **signed-rank sum,** which is the sum of the signed ranks. A large positive sum suggests that $\mu_d > 0$, since, if this were the case, most differences would be positive and larger in magnitude than the few negative differences; most of the ranks, and especially the larger ones would then be positively signed. Similarly, a large negative sum would suggest $\mu_d < 0$. A signed-rank sum near zero would be compatible with H_0: $\mu_d = 0$.

EXAMPLE 16.5 Blood Pressure and Kidney Disease

● Treatment of terminal renal failure involves surgical removal of a kidney (a nephrectomy). The paper "Hypertension in Terminal Renal Failure, Observations Pre-and Post-Bilateral Nephrectomy" (*Journal of Chronic Diseases* [1973]: 471–501) gave the accompanying blood pressure readings for five terminal renal patients before and 2 months after surgery.

● Data set available online

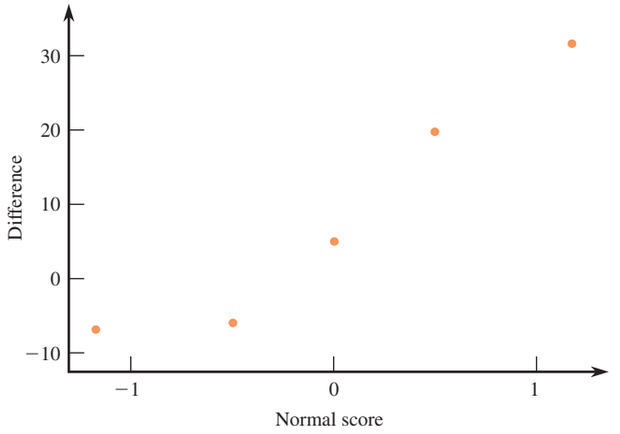
Patient	1	2	3	4	5
Before Surgery	107	102	95	106	112
After Surgery	87	97	101	113	80
Difference	20	5	−6	−7	32

We can determine whether the mean blood pressure before surgery exceeds the mean blood pressure 2 months after surgery by testing

$$H_0\colon\ \mu_1 - \mu_2 = 0 \quad \text{versus} \quad H_a\colon\ \mu_1 - \mu_2 > 0$$

where μ_1 denotes the mean diastolic blood pressure for patients in renal failure and μ_2 denotes the mean blood pressure for patients 2 months after surgery (equivalent hypotheses are $H_0\colon \mu_d = 0$ and $H_a\colon \mu_d > 0$ where μ_d is the mean difference in blood pressure).

A normal probability plot for this set of differences follows. Since the plot appears to be more S-shaped than linear, the assumption of a normal difference distribution is questionable. If it is reasonable to assume that the difference distribution is symmetric, a test based on the signed ranks can be used.



The absolute values of the differences and the corresponding ranks are as follows.

Absolute Difference	5	6	7	20	32
Rank	1	2	3	4	5

Associating the appropriate sign with each rank then yields signed ranks 1, −2, −3, 4, and 5, and a signed-rank sum of $1 - 2 - 3 + 4 + 5 = 5$.

The largest possible value for this sum would be 15, occurring only when all differences are positive. There are 32 possible ways to associate signs with ranks 1, 2, 3, 4, and 5, and 10 of them have rank sums of at least 5. When the null hypothesis $H_0\colon \mu_d = 0$ is true, each of the 32 possible assignments is equally likely to occur, and so

$$P(\text{signed-rank sum} \geq 5 \text{ when } H_0 \text{ is true}) = \frac{10}{32} = .3125$$

Therefore, the observed sum of 5 is compatible with H_0—it does not provide evidence that H_0 should be rejected.

Testing Hypotheses Using Signed Ranks

Suppose we are interested in testing H_0: $\mu_d = 0$. Given a set of n pairs of observations, ranking the absolute differences requires using ranks 1 to n. Since each rank could then be designated as either a plus or a minus, there are 2^n different possible sets of signed ranks. When the null hypothesis is true, each of the 2^n signed rankings has the same chance of occurring. Examining these different signed rankings and the associated sums gives information about how the signed-rank sum behaves when the null hypothesis is true. In particular, by looking at the distribution of the sum when H_0 is true, we can determine which values are unusual enough to suggest rejection of H_0.

For example, when $n = 5$ there are 2^5 different signed-rank sets. A few of these and the associated sums are

1	2	3	4	5	sum = 15
−1	2	−3	4	5	sum = 7
−1	−2	3	4	−5	sum = −1

By systematically listing all 32 possible signed rankings, the following information is obtained:

Signed-rank sum	15	13	11	9	7	5	3	1
Number of rankings yielding sum	1	1	1	2	2	3	3	3
Signed-rank sum	−1	−3	−5	−7	−9	−11	−13	−15
Number of rankings yielding sum	3	3	3	2	2	1	1	1

If we were to reject H_0: $\mu_1 - \mu_2 = 0$ in favor of H_a: $\mu_1 - \mu_2 \neq 0$ whenever we observed a signed-rank sum greater than or equal to 13 or less than or equal to −13, the probability of incorrect rejection would be $\frac{4}{32} = .125$ (since 4 of the possible signed rankings result in sums in the rejection region). Therefore, when $n = 5$, using the indicated rejection region gives a test with significance level .125.

For values of n larger than 5, finding the exact distribution of the signed-rank sum when H_0 is true is tedious and time-consuming, so tables have been developed. For selected sample sizes, Chapter 16 Appendix Table 2 gives critical values for levels of significance closest to the usual choices of .01, .05, and .10.

Summary of the Signed-Rank Test*

Null hypothesis: H_0: $\mu_d = 0$

Test statistic: signed-rank sum

Alternative Hypothesis	Rejection Region
H_0: $\mu_d > 0$	signed-rank sum ≥ critical value
H_0: $\mu_d < 0$	signed-rank sum ≤ −critical value
H_0: $\mu_d \neq 0$	signed-rank sum ≥ critical value or signed-rank sum ≤ −critical value

Selected critical values are given in Chapter 16 Appendix Table 2.

Assumptions: 1. The samples are paired.
2. The population difference distribution is symmetric.

*Alternative forms of the test statistic sometimes used are the sum of the positive ranks, the sum of the negative ranks, or the smaller of the sum of positive ranks and the sum of negative ranks. However, Chapter 16 appendix Table 2 should not be used to obtain critical values for these statistics.

EXAMPLE 16.6 Competitive Swimming

● Some swimming races are won by less than .001 second. As a result, a technique that might give a competitive swimmer even a slight edge is given careful consideration. To determine which of two racing starts, the hole entry or the flat entry, is faster, the authors of the paper "Analysis of the Flat vs. the Hole Entry" (*Swimming Technique* [Winter 1980]: 112–117) studied 10 college swimmers. A number of variables were measured for each type of start. The data for time to water entry appear here.

Swimmer	1	2	3	4	5	6	7	8	9	10
Flat Entry	1.13	1.11	1.18	1.26	1.16	1.41	1.43	1.25	1.33	1.36
Hole Entry	1.07	1.03	1.21	1.24	1.33	1.42	1.35	1.32	1.31	1.33
Difference	.06	.08	−.03	.02	−.17	−.01	.08	−.07	.02	.03

The authors of the paper used a signed-rank test with a .05 significance level to determine if there is a difference between the mean time to water entry for the two entry methods. Ordering the absolute differences results in the following assignment of signed ranks.

Difference	−.01	.02	.02	.03	−.03	.06	−.07	.08	.08	−.17
Signed Rank	−1	2.5	2.5	4.5	−4.5	6	−7	8.5	8.5	−10

1. Let μ_d denote the mean difference in time to water entry for flat and hole entry.
2. H_0: $\mu_d = 0$
3. H_a: $\mu_d \neq 0$
4. Test statistic: signed-rank sum
5. With $n = 10$ and $\alpha = .05$, Chapter 16 Appendix Table 2 gives 39 as the critical value for a two-tailed test. Therefore, H_0 will be rejected if either the signed-rank sum ≥ 39 or signed-rank sum ≤ -39.
6. Signed-rank sum $= -1 + 2.5 + \ldots + (-10) = 10$
7. Since 10 does not fall in the rejection region, we do not reject H_0. There is not sufficient evidence to indicate that the mean time to water entry differs for the two methods.

Example 16.7 illustrates how zero differences are handled when performing a signed-rank test. Since zero is considered to be neither positive nor negative, zero values are generally excluded from a signed-rank analysis, and the sample size is reduced accordingly.

EXAMPLE 16.7 Vitamin B_{12} Levels

● Two assay methods for measuring the level of vitamin B_{12} in red blood cells were compared in the paper "Noncobalimin Vitamin B_{12} Analogues in Human Red Cells, Liver and Brain" (*American Journal of Clinical Nutrition* [1983]: 774–777). Blood

● Data set available online

samples were taken from 15 healthy adults, and, for each blood sample, the B_{12} level was determined using both methods. The resulting data are given here.

Subject	1	2	3	4	5	6	7	8
Method 1	204	238	209	277	197	227	207	205
Method 2	204	238	198	253	180	209	217	204
Difference	0	0	11	24	17	18	−10	1

Subject	9	10	11	12	13	14	15
Method 1	131	282	76	194	120	92	114
Method 2	137	250	82	165	79	100	107
Difference	−6	32	−6	29	41	−8	7

We assume that the difference distribution is symmetric and proceed with a signed-rank test to determine whether there is a significant difference between the two methods for measuring B_{12} content. A significance level of .05 will be used.

Two of the observed differences are zero. Eliminating the two zeros reduces the sample size from 15 to 13. Ordering the nonzero absolute differences results in the following assignment of signed ranks.

Difference	1	−6	−6	7	−8	−10	11	17	18	24	29	32	41
Signed Rank	1	−2.5	−2.5	4	−5	−6	7	8	9	10	11	12	13

1. Let μ_d denote the mean difference in B_{12} determination for the two methods.
2. H_0: $\mu_d = 0$
3. H_a: $\mu_d \neq 0$
4. Test statistic: signed-rank sum
5. The form of H_0 indicates that a two-tailed test should be used. With $n = 13$ and $\alpha = .05$, Chapter 16 Appendix Table 2 gives a critical value of 57 (corresponding to an actual significance level of .048). Therefore, H_0 will be rejected if either the signed-rank sum ≥ 57 or signed-rank sum ≤ -57.
6. Signed-rank sum $= 1 + (-2.5) + (-2.5) + \ldots + 13 = 59$
7. Since 59 falls in the rejection regions, H_0 is rejected in favor of H_a. We conclude that there is a significant difference in measured B_{12} levels for the two assay methods.

The procedure for testing H_0: $\mu_d = 0$ just described can be easily adapted to test H_0: $\mu_d =$ hypothesized value, where the hypothesized value is something other than zero.

To test H_0: $\mu_d =$ hypothesized value, subtract the hypothesized value from each difference prior to assigning signed ranks.

EXAMPLE 16.8 Treating Dyskinesia

● Tardive dyskinesia is a syndrome that sometimes follows long-term use of antipsychotic drugs. Symptoms include abnormal involuntary movements. In an experiment to evaluate the effectiveness of the drug Deanol in reducing symptoms, Deanol and a placebo treatment were each administered for 4 weeks to 14 patients. A Total Severity Index (TSI) score was used to measure improvement (larger TSI scores indicate greater improvement). The accompanying data come from "Double-Blind Evaluation of Deanol in Tardive Dyskinesia" (*Journal of the American Medical Association* [1978]: 1997–1998). Let's use these data and a significance level of .01 to determine if the mean TSI score for people treated with Deanol exceeds the mean placebo TSI score by more than 1.

TSI Scores

Patient	1	2	3	4	5	6	7
Deanol	12.4	6.8	12.6	13.2	12.4	7.6	12.1
Placebo	9.2	10.2	12.2	12.7	12.1	9.0	12.4
Difference	3.2	−3.4	.4	.5	.3	−1.4	−.3

Patient	8	9	10	11	12	13	14
Deanol	5.9	12.0	1.1	11.5	13.0	5.1	9.6
Placebo	5.9	8.5	4.8	7.8	9.1	3.5	6.4
Difference	0.0	3.5	−3.7	3.7	3.9	1.6	3.2

1. Let μ_d denote the mean difference in TSI score for Deanol and the placebo treatment.
2. H_0: $\mu_d = 1$
3. H_a: $\mu_d > 1$
4. Test statistic: signed-rank sum
5. The form of H_0 indicates that an upper-tailed test should be used. With $n = 14$ and $\alpha = .01$, Chapter 16 Appendix Table 2 gives a critical value of 73. Therefore, H_0 will be rejected if the signed-rank sum equals or exceeds 73.
6. Subtracting 1 from each difference results in the following set of values.

   2.2 −4.4 −.6 −.5 −.7 −2.4 −1.3 −1 2.5 −4.7 2.7 2.9 .6 2.2

   Ordering these values and associating signed ranks yields:

Sign	−	−	+	−	−	−	+
Absolute Difference	.5	.6	.6	.7	1	1.3	2.2
Signed Rank	−1	−2.5	2.5	−4	−5	−6	7.5

Sign	+	−	+	+	+	−	−
Absolute Difference	2.2	2.4	2.5	2.7	2.9	4.4	4.7
Signed Rank	7.5	−9	10	11	12	−13	−14

● Data set available online

Then the signed-rank sum = −1 + (−2.5) + 2.5 + … + (−14) = −4.

7. Since $-4 < 73$, we fail to reject H_0. There is not sufficient evidence to indicate that the mean TSI score for the drug Deanol exceeds the mean TSI score for a placebo treatment by more than 1.

A Normal Approximation

Signed-rank critical values for sample sizes up to 20 are given in Chapter 16 Appendix Table 2. For larger sample sizes, the distribution of the signed-rank statistic when H_0 is true can be approximated by a normal distribution.

If $n > 20$, the distribution of the signed-rank sum when H_0 is true is well approximated by the normal distribution with mean 0 and standard deviation $\sqrt{n(n+1)(2n+1)/6}$. This implies that the standardized statistic

$$z = \frac{\text{signed-rank sum}}{\sqrt{n(n+1)(2n+1)/6}}$$

has approximately a standard normal distribution. This z statistic can be used as a test statistic and the associated *P*-value can be determined using the z table.

EXAMPLE 16.9 Chronic Airflow Obstruction

● The exercise capability of people suffering chronic airflow obstruction (CAO) is severely limited. In order to determine maximum exercise ventilation under two different experimental conditions, 21 patients suffering from CAO exercised to exhaustion under each condition. Ventilation was then measured. The accompanying data are from "Exercise Performance with Added Dead Space in Chronic Airflow Obstruction" (*Journal of Applied Physiology* [1984]: 1020–1023).

Patient	1	2	3	4	5	6	7	8	9	10	11
Condition 1	62	57	56	55	50.5	50	47.2	43.5	40	40	41
Condition 2	52	46	51	52.4	55	51	43	40	34.2	34	33
Difference	10	11	5	2.6	−4.5	−1	4.2	3.5	5.8	6	8

Patient	12	13	14	15	16	17	18	19	20	21
Condition 1	33	31	28	27.1	27.5	27	25	19.2	17.5	12
Condition 2	32	38	26	28	28	18	21	18	16	15
Difference	1	−7	2	−.9	−.5	9	4	1.2	1.5	−3

Do these data suggest that the mean ventilation is different for the two experimental conditions? Let's analyze the data using a level .05 signed-rank test.

1. Let μ_d denote the mean difference in ventilation between experimental conditions 1 and 2.
2. H_0: $\mu_d = 0$
3. H_a: $\mu_d \neq 0$

● Data set available online

4. $\alpha = .05$
5. Test statistic:

$$z = \frac{\text{signed-rank sum}}{\sqrt{n(n+1)(2n+1)/6}}$$

6. Assumptions: The difference distribution is symmetric.
7. Ordering the absolute differences and assigning signed ranks yields

$$\begin{array}{cccccccccccc} -1 & -2 & -3.5 & 3.5 & 5 & 6 & 7 & 8 & -9 & 10 & 11 \\ 12 & -13 & 14 & 15 & 16 & -17 & 18 & 19 & 20 & 21 & \end{array}$$

The signed-rank sum is $-1 + (-2) + \ldots + 21 = 140$, and the denominator of z is

$$\sqrt{\frac{n(n+1)(2n+1)}{6}} = \sqrt{\frac{(21)(22)(43)}{6}} = 57.54$$

so

$$z = \frac{140}{57.54} = 2.43$$

8. Using Appendix Table 2, P-value $= 2P(z > 2.43) = 2(.0075) = .015$
9. Since P-value $\leq \alpha$ we reject H_0 in favor of H_a. The sample data do suggest that the mean ventilation rate differs for the two experimental conditions.

Comparing the Paired t and Signed-Rank Tests

In order for the paired t test to be an appropriate method of analysis, it must be assumed that the underlying difference distribution is normal. Proper use of the signed-rank test requires only that the difference distribution be symmetric. Since a normal distribution is symmetric, when the distribution of differences is normal, either the paired t or the signed-rank test could be used. In this case, however, for a fixed significance level and sample size, the paired t test gives a slightly smaller type II error probability and a slightly higher power. Therefore, when the assumption of a normal difference distribution is met, the paired t test would be the preferred method for testing hypotheses about $\mu_1 - \mu_2$ using paired data. However, when the difference distribution is symmetric but not necessarily normal, the signed-rank test is a better choice.

A Distribution-Free Confidence Interval for μ_d

The distribution-free confidence interval for $\mu_1 - \mu_2$ discussed in Section 16.1 consisted of all hypothesized values for which H_0: $\mu_1 - \mu_2 =$ hypothesized value could not be rejected by the rank-sum test. Similarly, the signed rank-sum confidence interval consists of those values for which H_0: $\mu_d =$ hypothesized value cannot be rejected by the signed-rank test. Unfortunately, in order to see the relation between the test procedure and the confidence interval formula clearly, the test statistics must first be expressed in a different form, one that involves taking averages of all pairs of sample differences. The details are tedious, so we ask you to accept that the procedure described below is correct (or consult a good reference on distribution-free procedures).

A signed-rank confidence interval for μ_d is based on all possible pairwise averages of sample differences (including the average of each difference with itself). The confidence interval has the form

(d^{th} smallest average, d^{th} largest average)

The value of d is obtained from Chapter 16 Appendix Table 3 and depends on the specified confidence level and sample size.

EXAMPLE 16.10 Growth Hormone Levels

● Elevated levels of growth hormone are characteristic of diabetic control. The paper "Importance of Raised Growth Hormone Levels in Medicating the Metabolic Derangements of Diabetes" (*New England Journal of Medicine* [March 29, 1981]: 810–815) reported the results of a comparison of growth hormone levels (mg/mL) for a conventional treatment and an insulin pump treatment for diabetes. Five diabetic patients participated in the study, with each patient receiving both treatments over a period of time. The resulting data are given. It would be useful to estimate the difference between mean growth hormone levels for the two treatments.

Patient	1	2	3	4	5
Conventional	10	16	17	20	10
Pump	9	7	8	8	6
Difference	1	9	9	12	4

To compute the required pairwise averages, it is convenient to arrange the differences along the top and left of a rectangular table. Then the averages of the corresponding pairs of differences can be calculated and entered at the intersection of each row and column on or above the diagonal of the table.

		Difference				
		1	4	9	9	12
Difference	1	1	2.5	5	5	6.5
	4	—	4	6.5	6.5	8
	9	—	—	9	9	10.5
	9	—	—	—	9	10.5
	12	—	—	—	—	12

Arranging the pairwise averages in order yields

1 2.5 4 5 5 6.5 6.5 6.5 8 9 9 9 10.5 10.5 12

With a sample size of 5 and a 90% confidence level, Chapter 16 Appendix Table 3 gives $d = 2$ (corresponding to an actual confidence level of 87.5%). The confidence interval for $\mu_d = \mu_1 - \mu_2$ is then determined by selecting the 2nd smallest and the 2nd largest of the pairwise averages. For this example , the 90% confidence interval is (2.5, 10.5).

● Data set available online

As you can see, the calculations required to obtain the pairwise averages can be tedious, especially for larger sample sizes. Fortunately, many of the standard computer packages calculate both the signed-rank sum and the signed-rank confidence interval. An approximate 90% signed-rank confidence interval from Minitab is as follows:

Wilcoxon Signed Rank CI

N	Estimated Median	Achieved Confidence	Confidence Interval
5	6.50	89.4	(2.50, 10.50)

The Signed-Rank Test for Single-Sample Problems

Although we have introduced the signed-rank test in a two-sample context, it can also be used to test H_0: μ = hypothesized value, where μ is the mean value of a single population. In this setting, rather than forming the differences and then associating signed ranks, a single sample is used and the hypothesized value from H_0 is subtracted from each observed sample value. Signed ranks are then associated with the resulting values. The rest of the test procedure (test statistics and rejection region) remains the same.

EXERCISES 16.8 - 16.18

16.8 ● The effect of a restricted diet in the treatment of autistic children was examined in the paper "Gluten, Milk Proteins, and Autism: Dietary Intervention Effects on Behavior and Peptide Secretion" (*Journal of Applied Nutrition* [1991]: 1–8). Ten children with autistic syndrome participated in the study. Peptide secretion was measured before diet restrictions and again after a period of restricted diet. The resulting data follow. Do these data suggest that the restricted diet was successful in reducing mean peptide secretion? Use the signed-rank test.

Subject	Before	After	Subject	Before	After
1	25	10	6	50	19
2	22	9	7	15	8
3	84	29	8	41	19
4	84	7	9	19	14
5	60	2	10	27	11

16.9 ● Peak force (N) on the hand was measured just prior to impact and just after impact on a backhand drive for six advanced tennis players. The resulting data, from the paper "Forces on the Hand in the Tennis One-Handed Backhand" (*International Journal of Sport Biomechanics* [1991]: 282–292), are given in the accompanying table. Use the signed-rank test to determine if the mean postimpact force is greater than the mean preimpact force by more than 6.

Player	Preimpact	Postimpact
1	26.7	38.2
2	44.3	47.2
3	53.9	61.0
4	26.4	34.3
5	47.6	64.9
6	43.1	44.2

16.10 ● In an experiment to study the way in which different anesthetics affected plasma epinephrine concentration, 10 dogs were selected and concentration was measured while they were under the influence of the anesthetics isoflurane and halothane ("Sympathoadrenal and Hemodynamic Effects of Isoflurane, Halothane, and Cyclopropane in Dogs" *Anesthesiology* [1974]: 465–470). The resulting data are as follows.

Dog	1	2	3	4	5	6	7	8	9	10
Isoflurane	.51	1.00	.39	.29	.36	.32	.69	.17	.33	.28
Halothane	.30	.39	.63	.38	.21	.88	.39	.51	.32	.42

Use a level .05 signed-rank test to see whether the mean epinephrine concentration differs for the two anesthetics. What assumption must be made about the epinephrine concentration distributions?

Bold exercises answered in back ● Data set available online ✦ Video Solution available

16.11 ● The accompanying data refer to the concentration of the radioactive isotope strontium-90 in samples of nonfat and 2% fat milk from five dairies. Do the data strongly support the hypothesis that mean strontium-90 concentration is higher for 2% fat milk than for nonfat milk? Use a level .05 signed-rank test.

Dairy	1	2	3	4	5
Nonfat	6.4	5.8	6.5	7.7	6.1
2% fat	7.1	9.9	11.2	10.5	8.8

16.12 ● Both a gravimetric and a spectrophotometric method are under consideration for determining phosphate content of a particular material. Six samples of the material are obtained, each is split in half, and a determination is made on each half using one of the two methods, resulting in the following data. Use an approximate 95% distribution-free confidence interval to estimate the mean difference for the two techniques. Interpret the interval.

Sample	1	2	3	4	5	6
Gravimetric	54.7	58.5	66.8	46.1	52.3	74.3
Spectrophotometric	55.0	55.7	62.9	45.5	51.1	75.4

16.13 ● The paper "Growth Hormone Treatment for Short Stature" (*New England Journal of Medicine* [October 27, 1983]: 1016–1022) gives the accompanying data for height velocity before growth hormone therapy and during growth hormone therapy for 14 children with hypopituitarism.

Child	1	2	3	4	5	6	7
Before	5.3	3.8	5.6	2.0	3.5	1.7	2.6
During	8.0	11.4	7.6	6.9	7.0	9.4	7.9

Child	8	9	10	11	12	13	14
Before	2.1	3.0	5.5	5.4	2.1	3.0	2.4
During	7.4	7.4	7.5	11.8	6.4	8.8	5.0

a. Use a level .05 signed-rank test to decide if growth hormone therapy is successful in increasing the mean height velocity.

b. What assumption about the height velocity distributions must be made in order for the analysis in Part (a) to be valid?

16.14 ● The paper "Analysis of the Flat vs. the Hole Entry" cited in Example 16.6 also gave the data below on time from water entry to first stroke (below) and initial velocity. The authors of the paper used signed-rank tests to analyze the data.

a. Use a level .01 test to decide whether there is a significant difference in mean time from entry to first stroke for the two entry methods.

b. Do the data suggest a difference in mean initial velocity for the two entry methods? Use a level .05 signed-rank test.

DATA FOR EXERCISE 16.14

Time from Entry to First Stroke

Swimmer	1	2	3	4	5	6	7	8	9	10
Hole	1.18	1.10	1.31	1.12	1.12	1.23	1.27	1.08	1.26	1.27
Flat	1.06	1.23	1.20	1.19	1.29	1.09	1.09	1.33	1.27	1.38

Initial Velocity

Swimmer	1	2	3	4	5	6	7	8	9	10
Hole	24.0	22.5	21.6	21.4	20.9	20.8	22.4	22.9	23.3	20.7
Flat	25.1	22.4	24.0	22.4	23.9	21.7	23.8	22.9	25.0	19.5

16.15 ● The paper "Effects of a Rice-rich versus Potato-rich Diet on Glucose, Lipoprotein, and Cholesterol Metabolism in Noninsulin-Dependent Diabetics" (*American Journal of Clinical Nutrition* [1984]: 598–606) gave the data below on cholesterol synthesis rate for eight diabetic subjects. Subjects were fed a standardized diet with potato or rice as the major carbohydrate source. Participants received both diets for specified periods of time, with cholesterol synthesis rate (mmol/day) measured at the end of each dietary period. The analysis presented in this paper used the signed-rank test. Use a test with significance level .05 to determine whether the mean cholesterol synthesis rate differs significantly for the two sources of carbohydrates.

16.16 ● The data below on pre- and postoperative lung capacities for 22 patients who underwent surgery as treatment for tuberculosis kyphosis of the spine appeared in the paper "Tuberculosis Kyphosis, Correction with Spinal Osteotomy, Halo-Pelvic Distractor, and Anterior and Posterior Fusion" (*Journal of Bone Joint Surgery* [1974]: 1419–1434). Do the data suggest that surgery increases the mean lung capacity? Use a level .05 large-sample signed-rank test.

16.17 Using the data of Exercise 16.13, estimate the mean difference in height velocity before and during growth hormone therapy with a 90% distribution-free confidence interval.

16.18 ● The signed-rank test can be adapted for use in testing H_0: μ = hypothesized value, where μ is the mean of a single population (see the last part of this section). Suppose that the time required to process a request at a bank's automated teller machine is recorded for each of 10 randomly selected transactions, resulting in the following times (in minutes); 1.4, 2.1, 1.9, 1.7, 2.4, 2.9, 1.8, 1.9, 2.6, 2.2. Use the one-sample version of the signed-rank test and a .05 significance level to decide if the data indicate that the mean processing time exceeds 2 minutes.

DATA FOR EXERCISE 16.15

Subject	1	2	3	4	5	6	7	8
Potato	1.88	2.60	1.38	4.41	1.87	2.89	3.96	2.31
Rice	1.70	3.84	1.13	4.97	.86	1.93	3.36	2.15

DATA FOR EXERCISE 16.16

Patient	1	2	3	4	5	6	7	8	9	10	11
Preoperative	1540	1160	1870	1980	1520	3155	1485	1150	1740	3260	4950
Postoperative	1620	1500	2220	2080	2160	3040	2030	1370	2370	4060	5070

Patient	12	13	14	15	16	17	18	19	20	21	22
Preoperative	1440	1770	2850	2860	1530	3770	2260	3370	2570	2810	2990
Postoperative	1680	1750	3730	3430	1570	3750	2840	3500	2640	3260	3100

16.3 Distribution-Free ANOVA

The validity of the F tests presented in Chapter 15 is based on the assumption that observations are selected from normal distributions, all of which have the same variance σ^2. When this is the case, the type I error probability is controlled at the desired level of significance a by using an appropriate F test. Additionally, the test has good ability to detect departures from the null hypothesis—its type II error probabilities are smaller than those for any other test.

There are two potential difficulties in using an F test when the basic assumptions are violated. One is that the actual level of significance may be different from what the investigator desires. This is because the test statistic will no longer have an F distribution, so P-values based on the F distribution may not be correct. A second problem is that the test may have rather large type II error probabilities, so that substantial departures from H_0 are likely to go undetected.

Studies have shown that when population or treatment distributions are only mildly nonnormal, neither of these difficulties is serious enough to warrant abandoning the F test. Statisticians say that the test is *robust* to small departures from normality. However, distributions that are either very skewed or have much heavier tails than the normal distribution do adversely affect the performance of the F test. Here we present test procedures that are valid (have a known type I error probability) when underlying population or treatment distributions are nonnormal, as long as they have the same shape and spread. These procedures are distribution-free because they are valid for a very wide class of distributions rather than just for a particular type of distribution, such as the normal. As was the case with the distribution-free rank-sum test discussed in Section 16.1, the distribution-free ANOVA procedures are based on ranks of the observations.

The Kruskal–Wallis Test for a Completely Randomized Design

As before, k denotes the number of populations or treatments being compared, and $\mu_1, \mu_2, \ldots, \mu_k$ represent the population or treatment means. The hypotheses to be tested are still

H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ versus H_a: at least two among the k means are different

Basic Assumption

The k population or treatment distributions all have the same shape and spread.

The distribution-free test to be described here is called the **Kruskal–Wallis (KW) test** after the two statisticians who developed it. Suppose that k independent random samples are available, one from each population or treatment. Again, let $n_1, n_2, \ldots, n_k$ denote the sample sizes, with $N = n_1 + n_2 + \ldots + n_k$. When H_0 is true, observations in all samples are selected from the same population or treatment-response distribution. Observations in the different samples should then be quite comparable in magnitude. However, when some μ's are different, some samples will consist mostly of relatively small values, whereas others will contain a preponderance of large values.

We assign rank 1 to the smallest observation among all N in the k samples, rank 2 to the next smallest, and so on (for the moment let's assume that there are no tied observations). The average of all ranks assigned is

$$\frac{1 + 2 + 3 + \ldots + N}{N} = \frac{N + 1}{2}$$

If all μ's are equal, the average of the ranks for each of the k samples should be reasonably close to $\frac{N + 1}{2}$ (since the observations will typically be intermingled, their ranks will be also). On the other hand, large differences between some of the μ's will usually result in some samples having average ranks much below $\frac{N + 1}{2}$ (those samples that contain mostly small observations), whereas others will have average ranks considerably exceeding $\frac{N + 1}{2}$. The KW statistic measures the discrepancy between the average rank in each of the k samples and the overall average $\frac{N + 1}{2}$.

DEFINITION

Let $\bar{r}_1$ denote the average of the ranks for observations in the first sample, $\bar{r}_2$ denote the average rank for observations in the second sample, and let $\bar{r}_3, \ldots, \bar{r}_k$ denote the analogous rank averages for samples $3, \ldots, k$. Then the KW statistic is

$$KW = \frac{12}{N(N + 1)}\left[n_1\left(\bar{r}_1 - \frac{N + 1}{2}\right)^2 + n_2\left(\bar{r}_2 - \frac{N + 1}{2}\right)^2 + \cdots + n_k\left(\bar{r}_k - \frac{N + 1}{2}\right)^2\right]$$

EXAMPLE 16.11 Starting Salaries

● To gain information on salaries for its graduates, suppose that a business school selected a random sample of students from each of the following four disciplines: (1) finance, (2) accounting, (3) marketing, and (4) business administration. Starting salary data (in thousands of dollars) are given in the accompanying table. Within each sample, values are listed in increasing order, and the corresponding rank among all $N = 22$ reported salaries appears below each observation.

1.	Finance salary:	59.4	59.8	60.3	62.3	63.9	
	Rank:	10	12	14	19	22	
2.	Accounting salary:	58.7	58.9	59.5	60.1	61.8	62.9
	Rank:	6	8	11	13	18	20
3.	Marketing salary:	56.7	57.6	58.2	60.4	61.4	63.4
	Rank:	1	4	5	15	17	21
4.	Business Administration salary:	56.9	57.3	58.8	59.1	61.0	
	Rank:	2	3	7	9	16	

The average rank in the first sample is

$$\bar{r}_1 = \frac{10 + 12 + 14 + 19 + 22}{5} = 15.4$$

● Data set available online

and the other rank averages are $\bar{r}_2 = 12.7$, $\bar{r}_3 = 10.5$, and $\bar{r}_4 = 7.4$. The average of all ranks assigned is

$$\frac{N+1}{2} = \frac{23}{2} = 11.5$$

so

$$KW = \frac{12}{(22)(23)}[5(15.4 - 11.5)^2 + 6(12.7 - 11.5)^2 + 6(10.5 - 11.5)^2 + 5(7.4 - 11.5)^2]$$
$$= \frac{12}{(22)(23)}(174.74) = 4.14$$

H_0 will be rejected when the value of KW is sufficiently large. To specify a critical value that controls the type I error probability, it is necessary to know how KW behaves when H_0 is true.

There are only a finite number of ways to assign the N ranks, and these all have the same chance of occurring when H_0 is true. Suppose all possibilities are enumerated, KW is computed for each one, and the 5% with the largest KW values are separated out. Rejecting H_0 when the observed allocation of ranks to samples falls within this 5% set then results in a test with significance level .05. The difficulty with this procedure is that unless N is small, the number of possibilities is quite large, and so enumeration is really out of the question. Fortunately, as long as no n_i is too small, there is an approximate result that saves the day. The approximation is based on a type of probability distribution called a **chi-square distribution**. As with t distributions, there is a different chi-square distribution for each different number of df. Unlike a t curve, a chi-square distribution is not symmetric, but instead looks rather like an F curve. Chapter 16 Appendix Table 4 gives upper-tail critical values for various chi-square distributions.

The Kruskal–Wallis Test

When H_0 is true and either

1. $k = 3$ and each n_i is at least 6
   or
2. $k \geq 4$ and each n_i is at least 5

the statistic KW has approximately a chi-squared distribution based on $k - 1$ df. A test with (approximate) level of significance α results from using KW as the test statistic and rejecting H_0 if $KW >$ chi-square critical value. The chi-square critical value is obtained from the $k - 1$ df row of Chapter 16 Appendix Table 4 in the column headed by the desired α.

EXAMPLE 16.12 Starting Salaries Revisited

Let's use the KW test at level .05 to analyze the salary data introduced in Example 16.11.

1. Let μ_1, μ_2, μ_3, and μ_4 denote the mean starting salaries for all graduates in each of the four disciplines respectively.
2. H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
3. H_a: At least two of the four μ's are different
4. Test statistic:

$$KW = \frac{12}{N(N+1)}\left[n_1\left(\bar{r}_1 - \frac{N+1}{2}\right)^2 + n_2\left(\bar{r}_2 - \frac{N+1}{2}\right)^2 + \ldots + n_4\left(\bar{r}_4 - \frac{N+1}{2}\right)^2\right]$$

5. Rejection region: The number of df for the chi-squared approximation is $k - 1 = 3$. For $\alpha = .05$, Chapter 16 Appendix Table 4 gives 7.82 as the critical value. H_0 will be rejected if $KW > 7.82$.
6. We previously computed KW as 4.14.
7. The computed KW value 4.14 does not exceed the critical value 7.82, so H_0 should not be rejected. The data do not provide enough evidence to conclude that the mean starting salaries for the four disciplines are different.

When there are tied values in the data set, ranks are determined as they were for the rank-sum test—by assigning each tied observation in a group the average of the ranks they would receive if they all differed slightly from one another.

Rejection of H_0 by the KW test can be followed by the use of an appropriate multiple comparison procedure. Also, the most widely used statistical computer packages will perform a KW test.

The KW test does not require normality, but it does require equal population or treatment-response distribution variances (all distributions must have the same spread). If you encounter a data set for which variances appear to be quite different, you should consult a statistician for advice.

Friedman's Test for a Randomized Block Design

The validity of the randomized block F test rested on the assumption that the observations in the experiment were drawn from normal distributions with the same variance. The test described here, called Friedman's test, does not require normality.

Basic Assumption

Observations in the experiment are assumed to have been selected from distributions having exactly the same shape and spread, but the mean value may depend separately both on the treatment applied and on the block.

The hypotheses are

H_0: the mean value does not depend on which treatment is applied

versus

H_a: the mean value does depend on which treatment is applied

The rationale for Friedman's test is quite straightforward. The observations in each block are first ranked separately from 1 to k (since every treatment appears once, there

are k observations in any block). Then the rank averages $\bar{r}_1, \bar{r}_2, \ldots, \bar{r}_k$ for treatments 1, 2, … , k, respectively, are computed. When H_0 is false, some treatments will tend to receive small ranks in most blocks, whereas other treatments will tend to receive mostly large ranks. In this case the $\bar{r}$'s will tend to be rather different. On the other hand, when H_0 is true, all the $\bar{r}$'s will tend to be close to the same value $(k + 1)/2$, the average of the ranks 1, 2, … , k. The test statistic measures the discrepancy between the $\bar{r}$'s and $(k + 1)/2$. A large discrepancy suggests that H_0 is false.

Friedman's Test

After ranking observations separately from 1 to k within each of the l blocks, let $\bar{r}_1, \bar{r}_2, \ldots, \bar{r}_k$ denote the resulting rank averages for the k treatments. The test statistic is

$$F_r = \frac{12l}{k(k+1)}\left[\left(\bar{r}_1 - \frac{k+1}{2}\right)^2 + \left(\bar{r}_2 - \frac{k+1}{2}\right)^2 + \ldots + \left(\bar{r}_k - \frac{k+1}{2}\right)^2\right]$$

As long as l is not too small, when H_0 is true F_r has approximately a chi-squared distribution based on $k - 1$ df. The rejection region for a test that has approximate level of significance α is then $F_r >$ chi-square critical value.

EXAMPLE 16.13 High-Pressure Sales

● High-pressure sales tactics of door-to-door salespeople can be quite offensive. Many people succumb to such tactics, sign a purchase agreement, and later regret their actions. In the mid 1970s, the Federal Trade Commission implemented regulations clarifying and extending rights of purchasers to cancel such agreements. The accompanying data are a subset of the data given in the paper "Evaluating the FTC Cooling-Off Rule" (*Journal of Consumer Affairs* [1977]: 101–106). Individual observations are cancellation rates for each of nine salespeople (the blocks) during each of 4 years. Let's use Friedman's test at level .05 to see if mean cancellation rate depends on the year.

Salesperson

Cancellation Rate	1	2	3	4	5	6	7	8	9
1973	2.8	5.9	3.3	4.4	1.7	3.8	6.6	3.1	0.0
1974	3.6	1.7	5.1	2.2	2.1	4.1	4.7	2.7	1.3
1975	1.4	.9	1.1	3.2	.8	1.5	2.8	1.4	.5
1976	2.0	2.2	.9	1.1	.5	1.2	1.4	3.5	1.2

Salesperson

Rank	1	2	3	4	5	6	7	8	9	$\bar{r}_i$
1973	3	4	3	4	3	3	4	3	1	3.11
1974	4	2	4	2	4	4	3	2	4	3.22
1975	1	1	2	3	2	2	2	1	2	1.78
1976	2	3	1	1	1	1	1	4	3	1.89

● Data set available online

H_0: mean cancellation rate is the same for all four years
H_a: mean cancellation rates differ for at least two of the years
Test statistic: F_r
Rejection region: With $\alpha = .05$ and $k - 1 = 3$, chi-square critical value = 7.82. H_0 will be rejected at level of significance .05 if $F_r > 7.82$.

Computations: Using $\frac{k+1}{2} = 2.5$,

$$F_r = \frac{(12)(9)}{(4)(5)}[(3.11 - 2.5)^2 + (3.22 - 2.5)^2 + (1.78 - 2.5)^2 + (1.89 - 2.5)^2] = 9.62$$

Conclusion: Since $9.62 > 7.82$, H_0 is rejected in favor of H_a. Mean cancellation rate is not the same for all four years.

EXERCISES 16.19 - 16.25

16.19 The paper "The Effect of Social Class on Brand and Price Consciousness for Supermarket Products" (*Journal of Retailing* [1978]: 33–42) used the Kruskal–Wallis test to determine if social class (lower, middle, and upper) influenced the importance (scored on a scale of 1 to 7) attached to a brand name when purchasing paper towels. The reported value of the KW statistic was .17. Use a .05 significance level to test the null hypothesis of no difference in the mean importance score for the three social classes.

16.20 ● Protoporphyrin levels were determined for three groups of people—a control group of normal workers, a group of alcoholics with sideroblasts in their bone marrow, and a group of alcoholics without sideroblasts. The given data appeared in the paper "Erythrocyte Coproporphyrin and Protoporphyrin in Ethanol-Induced Sideroblastic Erythroporiesis" (*Blood* [1974]: 291–295). Do the data (see page 16-28) suggest that normal workers and alcoholics with and without sideroblasts differ with respect to mean protoporphyrin level? Use the KW test with a .05 significance level.

16.21 ● The given data on phosphorus concentration in topsoil for four different soil treatments appeared in the article "Fertilisers for Lotus and Clover Establishments on a Sequence of Acid Soils on the East Otago Uplands" (*New Zealand Journal of Experimental Agriculture* [1984]: 119–129). Use the KW test and a .01 significance level to test the null hypothesis of no difference in true mean phosphorus concentration for the four soil treatments.

Treatment	Concentration (mg/g)				
I	8.1	5.9	7.0	8.0	9.0
II	11.5	10.9	12.1	10.3	11.9
III	15.3	17.4	16.4	15.8	16.0
IV	23.0	33.0	28.4	24.6	27.7

16.22 ● The paper "Physiological Effects During Hypnotically Requested Emotions" (*Psychosomatic Medicine* [1963]: 334–343) reported data (see page 16-28) on skin potential (mV) when the emotions of fear, happiness, depression, and calmness were requested from each of eight subjects. Do the data suggest that the mean skin potential differs for the emotions tested? Use a significance level of .05.

16.23 ● In a test to determine if soil pretreated with small amounts of Basic-H makes the soil more permeable to water, soil samples were divided into blocks and each block received all four treatments under study. The treatments were (1) water with .001% Basic-H on untreated soil; (2) water without Basic-H on untreated soil; (3) water with Basic-H on soil pretreated with Basic-H; and (4) water without Basic-H, on soil pretreated with Basic-H. Using a significance level of .01, determine if mean permeability differs for the four treatments. (Data on page 16-28)

16.24 ● The following data on amount of food consumed (g) by eight rats after 0, 24, and 72 hours of food deprivation appeared in the paper "The Relation Between Differences in Level of Food Deprivation and Dominance in Food Getting in the Rat" (*Psychological Science* [1972]: 297–298). Do the data indicate a difference in the mean food consumption for the three experimental conditions? Use $\alpha = .01$.

	Rat							
Hours	1	2	3	4	5	6	7	8
0	3.5	3.7	1.6	2.5	2.8	2.0	5.9	2.5
24	5.9	8.1	8.1	8.6	8.1	5.9	9.5	7.9
72	13.9	12.6	8.1	6.8	14.3	4.2	14.5	7.9

16.25 ● The article "Effect of Storage Temperature on the Viability and Fertility of Bovine Sperm Diluted and Stored in Caprogen" (*New Zealand Journal of Agricultural Research* [1984]: 173–177) examined the effect of temperature on sperm survival. Survival data for various storage times are given below. Use Friedman's test with a .05 significance level to determine if survival is related to storage temperature. Regard time as the blocking factor.

Storage Temperature °C	Storage Time (hours)				
	6	24	48	120	168
15.6	61.9	59.6	57.0	58.8	53.7
21.1	62.5	60.0	57.4	59.3	54.9
26.7	60.7	55.5	54.5	53.3	45.3
32.2	59.9	48.6	42.6	36.6	24.8

DATA FOR EXERCISE 16.20

Group	Protoporphyrin Level (mg)														
Normal	22	27	47	30	38	78	28	58	72	56	30	39	53	50	36
Alcoholics with Sideroblasts	78	172	286	82	453	513	174	915	84	153	780				
Alcoholics without Sideroblasts	37	28	38	45	47	29	34	20	68	12	37	8	76	148	11

DATA FOR EXERCISE 16.22

	Subject							
Emotion	1	2	3	4	5	6	7	8
Fear	23.1	57.6	10.5	23.6	11.9	54.6	21.0	20.3
Happiness	22.7	53.2	9.7	19.6	13.8	47.1	13.6	23.6
Depression	22.5	53.7	10.8	21.1	13.7	39.2	13.7	16.3
Calmness	22.6	53.1	8.3	21.6	13.3	37.0	14.8	14.8

DATA FOR EXERCISE 16.23

	Treatment					Treatment			
Block	1	2	3	4	Block	1	2	3	4
1	37.1	33.2	58.9	56.7	6	25.3	19.3	48.8	37.1
2	31.8	25.3	54.2	49.6	7	23.7	17.3	47.8	37.5
3	28.0	20.2	49.2	46.4	8	24.4	17.0	40.2	39.6
4	25.9	20.3	47.9	40.9	9	21.7	16.7	44.0	35.1
5	25.5	18.3	38.2	39.4	10	26.2	18.3	46.4	36.5

Summary of Key Concepts and Formulas

TERM OR FORMULA	COMMENT
Rank sum = sum of sample 1 ranks	The distribution-free test statistic for testing H_0: $\mu_1 - \mu_2$ = hypothesized value using independent samples when it is reasonable to assume that the two populations have the same shape and spread.
Signed-rank sum = sum of signed ranks	The distribution-free test statistic for testing H_0: $\mu_d = 0$ using paired samples when it is reasonable to assume that the difference distribution is symmetric.
$z = \dfrac{\text{signed-rank sum}}{\sqrt{n(n+1)(2n+1)/6}}$	A large-sample test statistic for testing H_0: $\mu_d = 0$ using paired samples. This statistic has approximately a standard normal distribution when the sample size is greater than 20.
(d^{th} smallest average, d^{th} largest average)	A signed-rank confidence interval for μ_d based on all possible pairwise averages of sample differences.
$KW = \dfrac{12}{N(N+1)}\left[n_1\left(\bar{r}_1 - \dfrac{N+1}{2}\right)^2 + n_2\left(\bar{r}_2 - \dfrac{N+1}{2}\right)^2 + \cdots + n_k\left(\bar{r}_k - \dfrac{N+1}{2}\right)^2\right]$	Test statistic for testing H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ using data from a completely randomized design when it is reasonable to assume that the k population distributions all have the same shape and spread.
$F_r = \dfrac{12l}{k(k+1)}\left[\left(\bar{r}_1 - \dfrac{k+1}{2}\right)^2 + \left(\bar{r}_2 - \dfrac{k+1}{2}\right)^2 + \cdots + \left(\bar{r}_k - \dfrac{k+1}{2}\right)^2\right]$	Test statistic for testing the null hypothesis of no treatment effect using data from a randomized block design when it is reasonable to assume that the treatment distributions all have the same shape and spread.

Chapter 16 Appendix: Tables

TABLE 1 *P*-Value Information for the Rank-Sum Test

		Upper-tailed test		Lower-tailed test		Two-tailed test	
n_1	n_2	*P*-value $<$ .05 if rank sum is greater than or equal to	*P*-value $<$ .01 if rank sum is greater than or equal to	*P*-value $<$ .05 if rank sum is less than or equal to	*P*-value $<$ .01 if rank sum is less than or equal to	*P*-value $<$ .05 if rank sum is not between*	*P*-value $<$ .01 if rank sum is not between
3	3	15	—	6	—	—	—
3	4	17	—	7	—	18,6	—
3	5	20	21	7	6	21,6	—
3	6	22	24	8	6	23,7	—
3	7	24	27	9	6	26,7	27,6
3	8	27	29	9	7	28,8	30,6
4	3	21	—	11	—	—	—
4	4	24	26	12	10	25,11	—
4	5	27	30	13	10	29,11	30,10
4	6	30	33	14	11	32,12	34,10
4	7	33	36	15	12	35,13	37,11
4	8	36	40	16	12	38,14	41,11
5	3	29	30	16	15	30,15	—
5	4	32	35	18	15	34,16	35,15
5	5	36	39	19	16	37,18	39,16
5	6	40	43	20	17	41,19	44,16
5	7	43	47	22	18	45,20	48,17
5	8	47	51	23	19	49,21	52,18
6	3	37	39	23	21	38,22	—
6	4	41	44	25	22	43,23	45,21
6	5	46	49	26	23	47,25	50,22
6	6	50	54	28	24	52,26	55,23
6	7	54	58	30	26	56,28	60,24
6	8	58	63	32	27	61,29	65,25
7	3	46	49	31	28	48,29	48,28
7	4	51	54	33	30	53,31	55,29
7	5	56	60	35	31	58,33	61,30
7	6	61	65	37	33	63,35	67,31
7	7	66	71	39	34	68,37	72,33
7	8	71	76	41	36	73,39	78,34

(continued)

TABLE 1 (*continued*)

		Upper-tailed test		Lower-tailed test		Two-tailed test	
n_1	n_2	*P*-value $<$ .05 if rank sum is greater than or equal to	*P*-value $<$ .01 if rank sum is greater than or equal to	*P*-value $<$ .05 if rank sum is less than or equal to	*P*-value $<$ .01 if rank sum is less than or equal to	*P*-value $<$ .05 if rank sum is not between*	*P*-value $<$ .01 if rank sum is not between
8	3	57	59	39	37	58,38	60,36
8	4	62	66	42	38	64,40	67,37
8	5	68	72	44	40	70,42	73,39
8	6	73	78	47	42	76,44	80,40
8	7	79	84	49	44	81,47	86,42
8	8	84	90	52	46	87,49	92,44

*Including endpoints. For example, when $n_1 = 3$ and $n_2 = 4$, *P*-value $\geq .05$ if $6 \leq$ rank sum ≤ 18.

TABLE 2 Critical Values for the Signed-Rank Test

n	Significance Level for One-Tailed Test	Significance Level for Two-Tailed Test	Critical Value
5	.031	.062	15
	.062	.124	13
	.094	.188	11
6	.016	.032	21
	.031	.062	19
	.047	.094	17
	.109	.218	13
7	.008	.016	28
	.023	.046	24
	.055	.110	20
	.109	.218	16
8	.012	.024	32
	.027	.054	28
	.055	.110	24
	.098	.196	20
9	.010	.020	39
	.027	.054	33
	.049	.098	29
	.102	.204	23
10	.010	.020	45
	.024	.048	39
	.053	.106	33
	.097	.194	27
11	.009	.018	52
	.027	.054	44
	.051	.102	38
	.103	.206	30

(continued)

TABLE 2 (*continued*)

n	Significance Level for One-Tailed Test	Significance Level for Two-Tailed Test	Critical Value
12	.010	.020	58
	.026	.052	50
	.046	.092	44
	.102	.204	34
13	.011	.022	65
	.024	.048	57
	.047	.094	49
	.095	.190	39
14	.010	.020	73
	.025	.050	63
	.052	.104	53
	.097	.194	43
15	.011	.022	80
	.024	.048	70
	.047	.094	60
	.104	.208	46
16	.011	.022	88
	.025	.050	76
	.052	.104	64
	.096	.192	52
17	.010	.020	97
	.025	.050	83
	.049	.098	71
	.103	.206	55
18	.010	.020	105
	.025	.050	91
	.049	.098	77
	.098	.196	61
19	.010	.020	114
	.025	.050	98
	.052	.104	82
	.098	.196	66
20	.010	.020	124
	.024	.048	106
	.049	.098	90
	.101	.202	70

TABLE 3 Values of *d* for the Signed-Rank Confidence Interval

n	Confidence Level	*d*	*n*	Confidence Level	*d*
5	93.8	1	14	99.1	13
	87.5	2		95.1	22
6	96.9	1		89.6	27
	90.6	3	15	99.0	17
7	98.4	1		95.2	26
	96.9	2		90.5	31
	89.1	5	16	99.1	20
8	99.2	1		94.9	31
	94.5	5		89.5	37
	89.1	7	17	99.1	25
9	99.2	2		94.9	36
	94.5	7		90.2	42
	90.2	9	18	99.0	29
10	99.0	4		95.2	41
	95.1	9		90.1	48
	89.5	12	19	99.1	33
11	99.0	6		95.1	47
	94.6	12		90.4	54
	89.8	15	20	99.1	38
12	99.1	8		95.2	53
	94.8	15		90.3	61
	90.8	18			
13	99.0	11			
	95.2	18			
	90.6	22			

TABLE 4 Chi-Square Distribution Critical Values

	Significance Level			
df	.10	.05	.01	.001
1	2.71	3.84	6.64	10.83
2	4.61	5.99	9.21	13.82
3	6.25	7.82	11.34	16.27
4	7.78	9.49	13.28	18.47
5	9.24	11.07	15.09	20.52
6	10.64	12.59	16.81	22.46
7	12.02	14.07	18.48	24.32
8	13.36	15.51	20.09	26.12
9	14.68	16.92	21.67	27.88
10	15.99	18.31	23.21	29.59
11	17.28	19.68	24.72	31.26
12	18.55	21.03	26.22	32.91
13	19.81	22.36	27.69	34.53
14	21.06	23.68	29.14	36.12
15	22.31	25.00	30.58	37.70
16	23.54	26.30	32.00	39.25
17	24.77	27.59	33.41	40.79
18	25.99	28.87	34.81	42.31
19	27.20	30.14	36.19	43.82
20	28.41	31.41	37.57	45.31

Answers to Selected Odd-Numbered Exercises

16.1 rank sum = 53, P-value > 0.05, fail to reject H_0

16.3 **a.** rank sum = 48, P-value > 0.05, fail to reject H_0

16.5 rank sum = 83, P-value > 0.05, fail to reject H_0

16.7 The confidence interval indicates that the mean burning time of oak may be as much as 0.5699 hours longer than pine; but also that the mean burning time of oak may be as much as 0.4998 hours shorter than pine.

16.9 signed-rank sum = 7, 7 does not exceed the critical value of 13, fail to reject H_0

16.11 signed-rank sum = -15, -15 is less than the critical value of -15, reject H_0

16.13 **a.** signed-rank sum = 105, 105 exceeds the critical value of 53, reject H_0 **b.** Must assume that population distribution of differences in height velocities is symmetric.

16.15 signed-rank sum = 12, 12 does not exceed the critical value of 28, fail to reject H_0

16.17 (3.65, 5.55)

16.19 $KW = 0.17$, 0.17 does not exceed the critical value of 5.99, fail to reject H_0

16.21 $KW = 17.86$, 0.17 exceeds the critical value of 11.34, reject H_0

16.23 $F_r = 28.92$, 28.92 exceeds the critical value of 11.34, reject H_0

16.25 $F_r = 15$, 15 exceeds the critical value of 7.82, reject H_0