14 Nonparametric Statistics

CONTENTS

Where We've Been

- Presented methods for making inferences about means (Chapters 7–10) and for making inferences about the correlation between two quantitative variables (Chapter 11)
- These methods required that the data be normally distributed or that the sampling distributions of the relevant statistics be normally distributed.

Where We're Going

- Develop the need for inferential techniques that require fewer or less stringent assumptions than the methods of Chapters 7–10 and 11 (14.1)
- Introduce *nonparametric* tests that are based on ranks (i.e., on an ordering of the sample measurements according to their relative magnitudes) (14.2–14.7)
- Present a nonparametric test about the central tendency of a single population (14.2)
- Present a nonparametric test for comparing two populations with independent samples (14.3)
- Present a nonparametric test for comparing two populations with paired samples (14.4)
- Present a nonparametric test for comparing three or more populations using a designed experiment (14.5–14.6)
- Present a nonparametric test for rank correlation (14.7)

Statistics IN Action How Vulnerable Are New Hampshire Wells to Groundwater Contamination?

Methyl tert-butyl ether (commonly known as MTBE) is a volatile, flammable, colorless liquid manufactured by the chemical reaction of methanol and isobutylene. MTBE was first produced in the United States as a lead fuel additive (octane booster) in 1979 and then as an oxygenate in reformulated fuel in the 1990s. Unfortunately, MTBE was introduced into water-supply aquifers by leaking underground storage tanks at gasoline stations, thus contaminating the drinking water. Consequently, by late 2006 most (but not all) American gasoline retailers had ceased using MTBE as an oxygenate, and accordingly, U.S. production has declined. Despite the reduction in production, there is no federal standard for MTBE in public water supplies; therefore, the chemical remains a dangerous pollutant, especially in states like New Hampshire that mandate the use of reformulated gasoline.

A study published in *Environmental Science & Technology* (Jan. 2005) investigated the risk of exposure to MTBE through drinking water in New Hampshire. In particular, the study reported on the factors related to MTBE contamination in public and private New Hampshire wells. Data were collected on a sample of 223 wells. These data are saved in the **MTBE** file (part of which you analyzed in Exercise 2.19). One of the variables measured was MTBE level (micrograms per liter) in the well water. An MTBE value exceeding .2 microgram per liter on the measuring instrument is a detectable level of MTBE. Of the 223 wells, 70 had detectable levels of MTBE. (Although the other wells are below the detection limit of the measuring device, the MTBE values for these wells are recorded as .2 rather than 0.) The other variables in the data set are described in Table SIA14.1.

How contaminated are these New Hampshire wells? Is the level of MTBE contamination different for the two classes of wells? For the two types of aquifers? What environmental factors are related to the MTBE level of a groundwater well? These are just a few of the research questions addressed in the study.

The researchers applied several nonparametric methods to the data in order to answer the research questions. We demonstrate the use of this methodology in four *Statistics in Action Revisited* examples.

Statistics IN Action Revisited

- Testing the Median MTBE Level of Groundwater Wells (p. 14-7)
- Comparing the MTBE Levels of Different Types of Groundwater Wells (p. 14-14)
- Comparing the MTBE Levels of Different Types of Groundwater Wells (continued) (p. 14-30)
- Testing the Correlation of MTBE Level with Other Environmental Factors (p. 14-44)

Table SIA14.1 Variables Measured in the MTBE Contamination Study

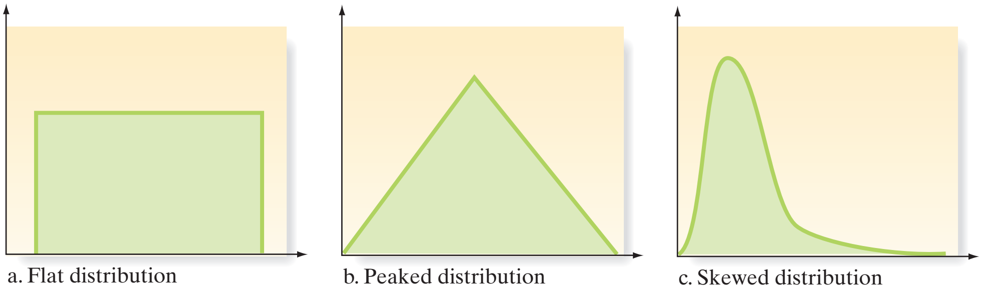
Variable Name	Type	Description	Units of Measurement, or Levels
CLASS	QL	Class of well	Public or Private
AQUIFER	QL	Type of aquifer	Bedrock or Unconsolidated
DETECTION	QL	MTBE detection status	Below limit or Detect
MTBE	QN	MTBE level	micrograms per liter
PH	QN	pH level	standard pH unit
DISSOXY	QN	Dissolved oxygen	milligrams per liter
DEPTH	QN	Well depth	meters
DISTANCE	QN	Distance to underground storage tank	meters
INDUSTRY	QN	Industries in proximity	Percent of industrial land within 500 meters of well

Data Set: MTBE

14.1 Introduction: Distribution-Free Tests

The confidence interval and testing procedures developed in Chapters 7–10 all involve making inferences about population parameters. Consequently, they are often referred to as **parametric statistical tests.** Many of these parametric methods (e.g., the small-sample t-test of Chapter 8 or the ANOVA F-test of Chapter 10) rely on the assumption that the data are sampled from a normally distributed population. When the data are normal, these tests are *most powerful*. That is, the use of such parametric tests maximizes power—the probability of the researcher correctly rejecting the null hypothesis.

Consider a population of data that is decidedly nonnormal. For example, the distribution might be flat, peaked, or strongly skewed to the right or left. (See Figure 14.1.) Applying the small-sample t-test to such a data set may lead to serious consequences. Since the normality assumption is clearly violated, the results of the t-test are unreliable. Specifically, (1) the probability of a Type I error (i.e., rejecting H_0 when it is true) may be larger than the value of α selected, and (2) the power of the test, $1 - \beta$, is not maximized.


a. Flat distribution

b. Peaked distribution

c. Skewed distribution

Figure 14.1
Some nonnormal distributions for which the t-statistic is invalid

A number of *nonparametric* techniques are available for analyzing data that do not follow a normal distribution. Nonparametric tests do not depend on the distribution of the sampled population; thus, they are called *distribution-free tests*. Also, nonparametric methods focus on the location of the probability distribution of the population, rather than on specific parameters of the population, such as the mean (hence the name "nonparametric").

Distribution-free tests are statistical tests that do not rely on any underlying assumptions about the probability distribution of the sampled population.

The branch of inferential statistics devoted to distribution-free tests is called **nonparametrics**.

Nonparametric tests are also appropriate when the data are nonnumerical in nature, but can be ranked.* For example, when taste-testing foods or in other types of consumer product evaluations, we can say that we like product A better than product B, and B better than C, but we cannot obtain exact quantitative values for the respective measurements. Nonparametric tests based on the ranks of measurements are called *rank tests*.

Nonparametric statistics (or tests) based on the ranks of measurements are called **rank statistics** (or **rank tests**).

In this chapter, we present several useful nonparametric methods. Keep in mind that these nonparametric tests are more powerful than their corresponding parametric counterparts in those situations where either the data are nonnormal or the data are ranked.

In Section 14.2, we develop a test for making inferences about the central tendency of a single population. In Sections 14.3 and 14.5, we present rank statistics for comparing two or more probability distributions using independent samples. In Sections 14.4 and 14.6, the matched-pairs and randomized block designs are used to make nonparametric comparisons of populations. Finally, in Section 14.7, we present a nonparametric measure of correlation between two variables.

Ethics IN Statistics

Consider a sampling problem where the assumptions required for the valid application of a parametric procedure (e.g., a t-test for a population mean) are clearly violated. Also, suppose the results of the parametric test lead you to a different inference about the target population than the corresponding nonparametric method. Intentional reporting of only the parametric test results is considered *unethical statistical practice*.

*Qualitative data that can be ranked in order of magnitude are called *ordinal* data.

14.2 Single-Population Inferences

In Chapter 8, we utilized the z- and t-statistics for testing hypotheses about a population mean. The z-statistic is appropriate for large random samples selected from "general" populations—that is, samples with few limitations on the probability distribution of the underlying population. The t-statistic was developed for small-sample tests in which the sample is selected at random from a *normal* distribution. The question is, How can we conduct a test of hypothesis when we have a small sample from a *nonnormal* distribution?

The **sign test** is a relatively simple nonparametric procedure for testing hypotheses about the central tendency of a nonnormal probability distribution. Note that we used the phrase *central tendency* rather than *population mean*. This is because the sign test, like many nonparametric procedures, provides inferences about the population *median* rather than the population mean μ. Denoting the population median by the Greek letter η, we know (Chapter 2) that η is the 50th percentile of the distribution (Figure 14.2) and, as such, is less affected by the skewness of the distribution and the presence of outliers (extreme observations). Since the nonparametric test must be suitable for all distributions, not just the normal, it is reasonable for nonparametric tests to focus on the more robust (less sensitive to extreme values) measure of central tendency: the median.

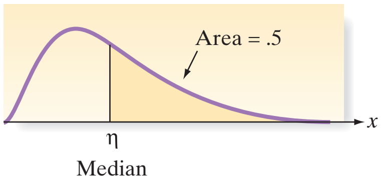


Figure 14.2
Location of the population median, η

For example, increasing numbers of both private and public agencies are requiring their employees to submit to tests for substance abuse. One laboratory that conducts such testing has developed a system with a normalized measurement scale in which values less than 1.00 indicate "normal" ranges and values equal to or greater than 1.00 are indicative of potential substance abuse. The lab reports a normal result as long as the median level for an individual is less than 1.00. Eight independent measurements of each individual's sample are made. One individual's results are shown in Table 14.1.

Table 14.1 Substance Abuse Test Results

.78	.51	3.79	.23	.77	.98	.96	.89

Data Set: SUBABUSE

If the objective is to determine whether the *population* median (i.e., the true median level if an infinitely large number of measurements were made on the same individual sample) is less than 1.00, we establish that as our alternative hypothesis and test

$$H_0: \eta = 1.00$$
$$H_a: \eta < 1.00$$

The one-tailed sign test is conducted by counting the number of sample measurements that "favor" the alternative hypothesis—in this case, the number that are less than 1.00. If the null hypothesis is true, we expect approximately half of the measurements to fall on each side of the hypothesized median, and if the alternative is true, we expect significantly more than half to favor the alternative—that is, to be less than 1.00. Thus,

Test statistic: S = Number of measurements less than 1.00, the null hypothesized median

If we wish to conduct the test at the $\alpha = .05$ level of significance, the rejection region can be expressed in terms of the observed significance level, or p-value, of the test:

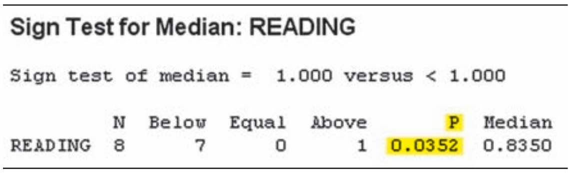
Rejection region: p-value $\leq .05$

In this example, $S = 7$ of the 8 measurements are less than 1.00. To determine the observed significance level associated with that outcome, we note that the number of measurements less than 1.00 is a binomial random variable (check the binomial characteristics presented in Chapter 4), and *if H_0 is true*, the binomial probability p that a measurement lies below (or above) the median 1.00 is equal to .5 (Figure 14.2). What is the probability that a result is *as contrary to or more contrary to* H_0 than the one observed? That is, what is the probability that 7 *or more* of 8 binomial measurements will result in Success (be less than 1.00) if the probability of Success is .5? Binomial Table II in Appendix A (with $n = 8$ and $p = .5$) indicates that

$$P(x \geq 7) = 1 - P(x \leq 6) = 1 - .965 = .035$$

Thus, the probability that at least 7 of 8 measurements would be less than 1.00 *if the true median were 1.00* is only .035. The p-value of the test is therefore .035.

This p-value can also be obtained from a statistical software package. The MINITAB printout of the analysis is shown in Figure 14.3, with the p-value highlighted. Since $p = .035$ is less than $\alpha = .05$, we conclude that this sample provides sufficient evidence to reject the null hypothesis. The implication of this rejection is that the laboratory can conclude at the $\alpha = .05$ level of significance that the true median level for the individual tested is less than 1.00. However, we note that one of the measurements, with a value of 3.79, greatly exceeds the others and deserves special attention. This large measurement is an outlier that would make the use of a t-test and its concomitant assumption of normality dubious. The only assumption necessary to ensure the validity of the sign test is that the probability distribution of measurements is continuous.



Sign Test for Median: READING

```
Sign test of median =  1.000 versus < 1.000

         N  Below  Equal  Above       P  Median
READING  8      7      0      1  0.0352  0.8350
```

Figure 14.3
MINITAB printout of sign test

The use of the sign test for testing hypotheses about population medians is summarized in the following box:

Sign Test for a Population Median η

One-Tailed Test

H_0: $\eta = \eta_0$

H_a: $\eta > \eta_0$ [or H_a: $\eta < \eta_0$]

Test statistic:

S = Number of sample measurements greater than η_0 [or S = number of measurements less than η_0]

Two-Tailed Test

H_0: $\eta = \eta_0$

H_a: $\eta \neq \eta_0$

Test statistic:

S = Larger of S_1 and S_2, where S_1 is the number of measurements less than η_0 and S_2 is the number of measurements greater than η_0

[*Note:* Eliminate observations from the analysis that are exactly equal to the hypothesized median, η_0.]

Observed significance level:

p-value $= P(x \geq S)$

Observed significance level:

p-value $= 2P(x \geq S)$

where x has a binomial distribution with parameters n and $p = .5$. (Use Table II, Appendix A.)

Rejection region: Reject H_0 if p-value $\leq \alpha$

Conditions Required for a Valid Application of the Sign Test

The sample is selected randomly from a continuous probability distribution.
[*Note:* No assumptions need to be made about the shape of the probability distribution.]

Recall that the normal probability distribution provides a good approximation of the binomial distribution when the sample size is large (i.e., when both $np \geq 15$ and

$nq \geq 15$). For tests about the median of a distribution, the null hypothesis implies that $p = .5$, and the normal distribution provides a good approximation if $n \geq 30$. (Note that for $n = 30$ and $p = .5, np = nq = 15$.) Thus, we can use the standard normal z-distribution to conduct the sign test for large samples. The large-sample sign test is summarized in the next box.

Large-Sample Sign Test for a Population Median η

One-Tailed Test	**Two-Tailed Test**
$H_0: \eta = \eta_0$	$H_0: \eta = \eta_0$
$H_a: \eta > \eta_0$ [or $H_a: \eta < \eta_0$]	$H_a: \eta \neq \eta_0$

$$\textit{Test statistic: } z = \frac{(S - .5) - .5n}{.5\sqrt{n}}$$

[*Note:* S is calculated as shown in the previous box. We subtract .5 from S as the "correction for continuity." The null-hypothesized mean value is $np = .5n$, and the standard deviation is

$$\sqrt{npq} = \sqrt{n(.5)(.5)} = .5\sqrt{n}$$

(See Chapter 5 for details on the normal approximation to the binomial distribution.)]

Rejection region: $z > z_\alpha$	*Rejection region*: $z > z_{\alpha/2}$

where tabulated z-values can be found in Table IV, Appendix A.

Example 14.1

Sign Test Application—Failure Times of MP3 Players

Problem A manufacturer of MP3 players has established that the median time to failure for its players is 5,250 hours of utilization. A sample of 40 MP3 players from a competitor is obtained, and the players are tested continuously until each fails. The 40 failure times range from 5 hours (a "defective" player) to 6,575 hours, and 24 of the 40 exceed 5,250 hours. Is there evidence that the median failure time of the competitor's product differs from 5,250 hours? Use $\alpha = .10$.

Solution The null and alternative hypotheses of interest are

$$H_0: \eta = 5{,}250 \text{ hours}$$
$$H_a: \eta \neq 5{,}250 \text{ hours}$$

Since $n \geq 30$, we use the standard normal z-statistic:

$$\textit{Test statistic: } z = \frac{(S - .5) - .5n}{.5\sqrt{n}}$$

Here, S is the maximum of S_1 (the number of measurements greater than 5,250) and S_2 (the number of measurements less than 5,250). Also,

$$\textit{Rejection region: } \quad z > 1.645, \text{ where } z_{\alpha/2} = z_{.05} = 1.645$$

Assumptions: The probability distribution of the failure times is continuous (time is a continuous variable), but nothing is assumed about its shape.

Since the number of measurements exceeding 5,250 is $S_2 = 24$, it follows that the number of measurements less than 5,250 is $S_1 = 16$. Consequently, $S = 24$, the greater of S_1 and S_2. The calculated z-statistic is therefore

$$z = \frac{(S - .5) - .5n}{.5\sqrt{n}} = \frac{23.5 - 20}{.5\sqrt{40}} = \frac{3.5}{3.162} = 1.11$$

The value of z is not in the rejection region, so we cannot reject the null hypothesis at the $\alpha = .10$ level of significance.

Look Back The manufacturer should not conclude, on the basis of this sample, that its competitor's MP3 players have a median failure time that differs from 5,250 hours. The manufacturer will not "accept H_0," however, since the probability of a Type II error is unknown.

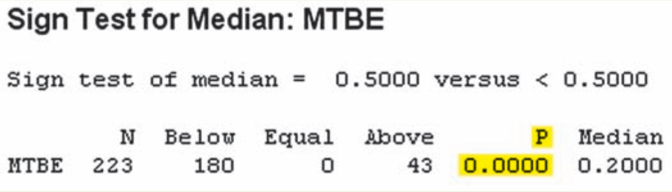
Now Work Exercise 14.5

The one-sample nonparametric sign test for a median provides an alternative to the *t*-test for small samples from nonnormal distributions. However, if the distribution is approximately normal, the *t*-test provides a more powerful test about the central tendency of the distribution.

Statistics IN Action Revisited — Testing the Median MTBE Level of Groundwater Wells

We return to the study of MTBE contamination of New Hampshire groundwater wells (p. 14-2). The Environmental Protection Agency (EPA) has not set a federal standard for MTBE in public water supplies; however, several states have developed their own standards. New Hampshire has a standard of 13 micrograms per liter; that is, no groundwater well should have an MTBE level that exceeds 13 micrograms per liter. Also, only half the wells in the state should have MTBE levels that exceed .5 microgram per liter. This implies that the median MTBE level should be less than .5. Do the data collected by the researchers provide evidence to indicate that the median level of MTBE in New Hampshire groundwater wells is less than .5 microgram per liter? To answer this question, we applied the sign test to the data saved in the MTBE file. The MINITAB printout is shown in Figure SIA14.1

We want to test H_0: $\eta = .5$ versus H_a: $\eta < .5$. According to the printout, 180 of the 223 sampled groundwater wells had MTBE levels below .5. Consequently, the test statistic value is $S = 180$. The one-tailed *p*-value for the test (highlighted on the printout) is .0000. Thus, the sign test is significant at $\alpha = .01$. Therefore, the data do provide sufficient evidence to indicate that the median MTBE level of New Hampshire groundwater wells is less than .5 microgram per liter.



Sign Test for Median: MTBE

```
Sign test of median =  0.5000 versus < 0.5000

         N  Below  Equal  Above       P  Median
MTBE   223    180      0     43  0.0000  0.2000
```

Figure SIA14.1
MINITAB sign test for MTBE data

Exercises 14.1–14.15

Understanding the Principles

14.1 Under what circumstances is the sign test preferred to the *t*-test for making inferences about the central tendency of a population?

14.2 What is the probability that a randomly selected observation exceeds the
a. Mean of a normal distribution?
b. Median of a normal distribution?
c. Mean of a nonnormal distribution?
d. Median of a nonnormal distribution?

Learning the Mechanics

14.3 Use Table II of Appendix A to calculate the following binomial probabilities:
a. $P(x \geq 6)$ when $n = 7$ and $p = .5$
b. $P(x \geq 5)$ when $n = 9$ and $p = .5$
c. $P(x \geq 8)$ when $n = 8$ and $p = .5$
d. $P(x \geq 10)$ when $n = 15$ and $p = .5$. Also, use the normal approximation to calculate this probability, and then compare the approximation with the exact value.
e. $P(x \geq 15)$ when $n = 25$ and $p = .5$. Also, use the normal approximation to calculate this probability, and then compare the approximation with the exact value.

14.4 Consider the following sample of 10 measurements, saved in the **LM14_4** file.

8.4	16.9	15.8	12.5	10.3	4.9	12.9	9.8	23.7	7.3

Use these data, the binomial tables (Table II, Appendix A), and $\alpha = .05$ to conduct each of the following sign tests:
a. H_0: $\eta = 9$ versus H_a: $\eta > 9$
b. H_0: $\eta = 9$ versus H_a: $\eta \neq 9$
c. H_0: $\eta = 20$ versus H_a: $\eta < 20$
d. H_0: $\eta = 20$ versus H_a: $\eta \neq 20$

e. Repeat each of the preceding tests, using the normal approximation to the binomial probabilities. Compare the results.

f. What assumptions are necessary to ensure the validity of each of the preceding tests?

14.5 NW Suppose you wish to conduct a test of the research hypothesis that the median of a population is greater than 80. You randomly sample 25 measurements from the population and determine that 16 of them exceed 80. Set up and conduct the appropriate test of hypothesis at the .10 level of significance. Be sure to specify all necessary assumptions.

Applying the Concepts—Basic

14.6 Caffeine in Starbucks coffee. Scientists at the University of Florida College of Medicine investigated the level of caffeine in 16-ounce cups of Starbucks coffee (*Journal of Analytical Toxicology*, Oct. 2003). In one phase of the experiment, cups of Starbucks Breakfast Blend (a mix of Latin American coffees) were purchased on six consecutive days from a single specialty coffee shop. The amount of caffeine in each of the six cups (measured in milligrams) is provided in the following table and saved in the **STARBUCKS** file.

564	498	259	303	300	307

a. Suppose the scientists are interested in determining whether the median amount of caffeine in Breakfast Blend coffee exceeds 300 milligrams. Set up the null and alternative hypotheses of interest.

b. How many of the cups in the sample have a caffeine content that exceeds 300 milligrams?

c. Assuming that $p = .5$, use the binomial table in Appendix A to find the probability that at least 4 of the 6 cups have caffeine amounts that exceed 300 milligrams.

d. On the basis of the probability you found in part **c,** what do you conclude about H_0 and H_a? (Use $\alpha = .05$.)

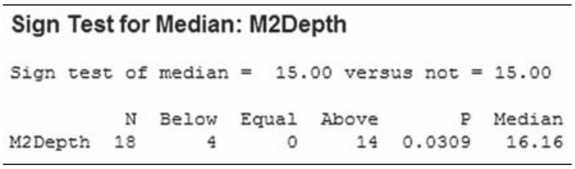
14.7 Cheek teeth of extinct primates. Refer to the *American Journal of Physical Anthropology* (Vol. 142, 2010) study of the characteristics of cheek teeth (e.g., molars) in an extinct primate species, Exercise 2.34 (p. 46). Recall that the researchers measured the dentary depth of molars (in millimeters) for 18 cheek teeth extracted from skulls. These depth measurements are reproduced in the next table (top, right) and saved in the **CHEEKTEETH** file. The researchers are interested in the median molar depth of all cheek teeth from this extinct primate species. In particular, they want to know if the population median differs from 15 mm.

a. Specify the null and alternative hypotheses of interest of the researchers.

b. Explain why the sign test is appropriate to apply in this case.

c. A MINITAB printout of the analysis is shown below. Locate the test statistic on the printout.

MINITAB output for Exercise 14.7



Sign Test for Median: M2Depth

```
Sign test of median =  15.00 versus not = 15.00

           N  Below  Equal  Above        P  Median
M2Depth   18      4      0     14   0.0309   16.16
```

18.12	19.48	19.36	15.94	15.83	19.70	15.76	17.00	16.20
13.96	16.55	15.70	17.83	13.25	16.12	18.13	14.02	14.04

Based on Boyer, D. M., Evans, A. R., and Jernvall, J. "Evidence of dietary differentiation among Late Paleocene–Early Eocene Plesiadapids (Mammalia, Primates)." *American Journal of Physical Anthropology*, Vol. 142, ©2010 (Table A3).

d. Find the p-value on the printout, and use it to draw a conclusion. Test using $\alpha = .05$.

14.8 Quality of white shrimp. In *The American Statistician* (May 2001), the nonparametric sign test was used to analyze data on the quality of white shrimp. One measure of shrimp quality is cohesiveness. Since freshly caught shrimp are usually stored on ice, there is concern that cohesiveness will deteriorate after storage. For a sample of 20 newly caught white shrimp, cohesiveness was measured both before and after storage on ice for two weeks. The difference in the cohesiveness measurements (before minus after) was obtained for each shrimp. If storage has no effect on cohesiveness, the population median of the differences will be 0. If cohesiveness deteriorates after storage, the population median of the differences will be positive.

a. Set up the null and alternative hypotheses to test whether cohesiveness will deteriorate after storage.

b. In the sample of 20 shrimp, there were 13 positive differences. Use this value to find the p-value of the test.

c. Make the appropriate conclusion (in the words of the problem) if $\alpha = .05$.

14.9 Emotional empathy in young adults. Refer to the *Journal of Moral Education* (June 2010) study of emotional empathy in young adults, Exercise 8.52 (p. 372). Recall that psychologists theorize that young female adults show more emotional empathy towards others than do males. To test the theory, each in a sample of 30 female college students responded to the following statement on emotional empathy: "I often have tender, concerned feelings for people less fortunate than me." Responses (i.e., empathy scores) ranged from 0 to 4, where 0 = "never" and 4 = "always." Suppose it is known that male college students have a median emotional empathy score of $\eta = 2.8$.

a. Specify the null and alternative hypothesis for testing whether female college students have a median emotional empathy scale score higher than 2.8.

b. Suppose that distribution of emotional empathy scores for the 30 female students is as shown in the table. Use this information to compute the test statistic.

Response (empathy score)	Number of Females
0	1
1	3
2	5
3	12
4	9

c. Find the observed significance level (p-value) of the test.

d. At $\alpha = .01$, what is the appropriate conclusion?

14.10 Crab spiders hiding on flowers. Refer to the *Behavioral Ecology* (Jan. 2005) field study on the natural camouflage of crab spiders, presented in Exercise 2.38 (p. 47). Ecologists collected a sample of 10 adult female crab spiders, each sitting on the yellow central part of a daisy, and measured the chromatic contrast between each spider and the flower. The contrast values for the 10 crab spiders are reproduced in the table and saved in the **SPIDER** file. (*Note*: The lower the contrast, the more difficult it is for predators to see the crab spider on the flower.) Recall that a contrast of 70 or greater allows bird predators to see the spider. Consider a test to determine whether the population median chromatic contrast of spiders on flowers is less than 70.

57	75	116	37	96	61	56	2	43	32

Based on Thery, M., et al. "Specific color sensitivities of prey and predator explain camouflage in different visual systems." *Behavioral Ecology*, Vol. 16, No. 1, Jan. 2005 (Table 1).

a. State the null and alternative hypotheses for the test of interest.
b. Calculate the value of the test statistic.
c. Find the p-value for the test.
d. At $\alpha = .10$, what is the appropriate conclusion? State your answer in the words of the problem.

Applying the Concepts—Intermediate

14.11 Lobster trap placement. Refer to the *Bulletin of Marine Science* (April 2010) observational study of lobster trap placement by teams fishing for the red spiny lobster in Baja California Sur, Mexico, Exercise 8.65 (p. 377). Trap spacing measurements (in meters) for a sample of seven teams of red spiny lobster fishermen are reproduced in the accompanying table (and saved in the **TRAPSPACE** file). In Exercise 8.65, you tested whether the average of the trap spacing measurements for the population of red spiny lobster fishermen fishing in Baja California Sur, Mexico, differs from 95 meters.

93	99	105	94	82	70	86

Based on Shester, G. G. "Explaining catch variation among Baja California lobster fishers through spatial analysis of trap-placement decisions." *Bulletin of Marine Science*, Vol. 86, No. 2, April 2010 (Table 1).

a. There is concern that the trap spacing data do not follow a normal distribution. If so, how will this impact the test you conducted in Exercise 8.65?
b. Propose an alternative nonparametric test to analyze the data.
c. Compute the value of the test statistic for the nonparametric test.
d. Find the p-value of the test.
e. Use the value of α you selected in Exercise 8.65 and give the appropriate conclusion.

14.12 Characteristics of a rockfall. Refer to the *Environmental Geology* (Vol. 58, 2009) simulation study of how far a block from a collapsing rockwall will bounce down a soil slope, Exercise 2.59 (p. 57). Recall that the variable of interest was *rebound length* (measured in meters) of the falling block. Based on the depth, location, and angle of block-soil impact marks left on the slope from an actual rockfall, the 13 rebound lengths shown in the table were estimated. (These data are saved in the **ROCKFALL** file.) Consider the following statement: "In all similar rockfalls, half of the rebound lengths will exceed 10 meters." Is this statement supported by the sample data? Test using $\alpha = .10$.

10.94	13.71	11.38	7.26	17.83	11.92	11.87	5.44	13.35
4.90	5.85	5.10	6.77					

Based on Paronuzzi, P. "Rockfall-induced block propagation on a soil slope, northern Italy." *Environmental Geology*, Vol. 58, 2009 (Table 2).

14.13 Freckling of superalloy ingots. Refer to the *Journal of Metallurgy* (Sept. 2004) study of freckling of superalloy ingots, presented in Exercise 2.187 (p. 101). Recall that freckles are defects that sometimes form during the solidification of the ingot. The freckle index for each of $n = 18$ superalloy ingots is shown in the next table and saved in the **FRECKLE** file. In the population of superalloy ingots, is there evidence to say that 50% of the ingots have a freckle index of 10 or higher? Test, using $\alpha = .01$.

30.1	22.0	14.6	16.4	12.0	2.4	22.2	10.0	15.1
12.6	6.8	4.1	2.5	1.4	33.4	16.8	8.1	3.2

Based on Yang, W. H., et al. "A freckle criterion for the solidification of superalloys with a tilted solidification front." *JOM: Journal of the Minerals, Metals and Materials Society,* Vol. 56, No. 9, Sept. 2004 (Table IV).

14.14 Study of guppy migration. To improve survival and reproductive success, many species of fish have an evolved migration history. In one migration study of guppy populations (*Zoological Science*, Vol. 6, 1989), adult female guppies were placed in the left compartment of an experimental aquarium tank divided in half by a glass plate. After the plate was removed, the numbers of fish passing through the slit from the left compartment to the right one, and vice versa, were monitored every minute for 30 minutes. If an equilibrium is reached (which is optimal for survival), the zoologists would expect about half the guppies to remain in the left compartment and half to remain in the right compartment. Consequently, if 80 guppies were placed in the aquarium, the median number of fish remaining in the left compartment should be 40. Data for a similar 30-minute experiment involving 80 guppies is shown in the table below and saved in the **GUPPY** file. (Each measurement represents the number of guppies in the left compartment at the end of a 1-minute interval.) Use the large-sample sign test to determine whether the median is less than 40. Test using $\alpha = .05$.

32	21	24	30	28	32	35	30	26	30	28	28	33	26	34
34	29	43	36	38	34	34	41	47	36	38	42	34	42	33

Based on Terami, H., and Watanabe, M. "Excessive transitory migration of guppy populations III: Analysis of perception of swimming space and a mirror effect." *Zoological Science*, Vol. 6, 1989.

14.15 Minimizing tractor skidding distance. Refer to the *Journal of Forest Engineering* (July 1999) study of minimizing tractor skidding distances along a new road in a European forest, presented in Exercise 8.73 (p. 379). The skidding distances (in meters) were measured at 20 randomly selected road sites. The data are repeated in the accompanying table and saved in the **SKIDDING** file. In Exercise 8.73, you conducted a test of hypothesis for the population mean skidding distance. Now conduct a test to determine whether the population median skidding distance is more than 400 meters. Use $\alpha = .10$.

488	350	457	199	285	409	435	574	439	546
385	295	184	261	273	400	311	312	141	425

Based on Tujek, J., and Pacola, E. "Algorithms for skidding distance modeling on a raster Digital Terrain Model," *Journal of Forest Engineering*, Vol. 10, No. 1, July 1999 (Table 1).

14.3 Comparing Two Populations: Independent Samples

Suppose two independent random samples are to be used to compare two populations, but the t-test of Chapter 9 is inappropriate for making the comparison. We may be unwilling to make assumptions about the form of the underlying population probability distributions, or we may be unable to obtain exact values of the sample measurements. If the data can be ranked in order of magnitude in either of these cases, the **Wilcoxon rank sum test** (developed by Frank Wilcoxon) can be used to test the hypothesis that the probability distributions associated with the two populations are equivalent.

For example, consider an experimental psychologist who wants to compare reaction times for adult males under the influence of drug A with reaction times for those under the influence of drug B. Experience has shown that the populations of reaction-time measurements often possess probability distributions that are skewed to the right, as shown in Figure 14.4. Consequently, a t-test should not be used to compare the mean reaction times for the two drugs, because the normality assumption that is required for the t-test may not be valid.

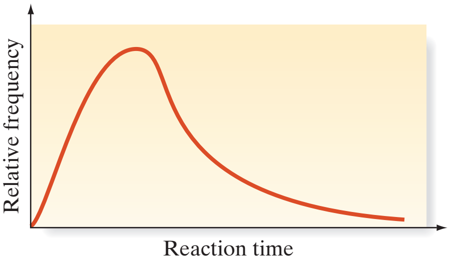


Figure 14.4
Typical probability distribution of reaction times

Suppose the psychologist randomly assigns seven subjects to each of two groups, one group to receive drug A and the other to receive drug B. The reaction time for each subject is measured at the completion of the experiment. These data (with the exception of the measurement for one subject in group A who was eliminated from the experiment for personal reasons) are shown in Table 14.2.

BIOGRAPHY

FRANK WILCOXON (1892–1965)

Wilcoxon Rank Tests

Frank Wilcoxon was born in Ireland, where his wealthy American parents were vacationing. He grew up in the family home in Catskill, New York, and then spent time working as an oil worker and tree surgeon in the back country of West Virginia. At age 25, Wilcoxon's parents sent him to Pennsylvania Military College, but he dropped out due to the death of his twin sister. Later, Wilcoxon earned degrees in chemistry from Rutgers (master's) and Cornell University (Ph.D.). After receiving his doctorate, Wilcoxon began work as a chemical researcher at the Boyce Thompson Institute for Plant Research. There, he began studying R. A. Fisher's (p. 477) newly issued *Statistical Methods for Research Workers*. In a now-famous 1945 paper, Wilcoxon presented the idea of replacing the actual sample data in Fisher's tests by their ranks and called the tests the rank sum test and signed-rank test. These tests proved to be inspirational to the further development of nonparametrics. After retiring from industry, Wilcoxon accepted a Distinguished Lectureship position at the newly created Department of Statistics at Florida State University. ■

Table 14.2 Reaction Times of Subjects under the Influence of Drug A or B

Drug A		Drug B	
Reaction Time (seconds)	Rank	Reaction Time (seconds)	Rank
1.96	4	2.11	6
2.24	7	2.43	9
1.71	2	2.07	5
2.41	8	2.71	11
1.62	1	2.50	10
1.93	3	2.84	12
		2.88	13

Data Set: DRUGS

The population of reaction times for either of the drugs—say, drug A—is that which could conceptually be obtained by giving drug A to all adult males. To compare the probability distributions for populations A and B, *we first rank the sample observations as though they were all drawn from the same population*. That is, we pool the measurements from both samples and then rank all the measurements from the smallest (a rank of 1) to the largest (a rank of 13). The results of this ranking process are also shown in Table 14.2.

If, on the one hand, the two populations were identical, we would expect the ranks to be *randomly mixed* between the two samples. If, on the other hand, one population tends to have longer reaction times than the other, we would expect the larger ranks to be mostly in one sample and the smaller ranks mostly in the other. Thus, the test statistic for the Wilcoxon test is based on the totals of the ranks for each of the two samples—that is, on the **rank sums.** When the sample sizes are equal, the greater the difference in the rank sums, the greater will be the weight of evidence indicating a difference between the probability distributions of the populations.

In the reaction-times example, we denote the rank sum for drug A by T_1 and that for drug B by T_2. Then

$$T_1 = 4 + 7 + 2 + 8 + 1 + 3 = 25$$
$$T_2 = 6 + 9 + 5 + 11 + 10 + 12 + 13 = 66$$

The sum of T_1 and T_2 will always equal $n(n + 1)/2$, where $n = n_1 + n_2$. So, for this example, $n_1 = 6, n_2 = 7$, and

$$T_1 + T_2 = \frac{13(13 + 1)}{2} = 91$$

Since $T_1 + T_2$ is fixed, a small value for T_1 implies a large value for T_2 (and vice versa) and a large difference between T_1 and T_2. Therefore, the smaller the value of one of the rank sums, the greater is the evidence indicating that the samples were selected from different populations.

The test statistic for this test is the rank sum for the smaller sample; or, in the case where $n_1 = n_2$, either rank sum can be used. Values that locate the rejection region for this rank sum are given in Table XII of Appendix A, a partial reproduction of which is shown in Table 14.3. The columns of the table represent n_1, the first sample size, and the rows represent n_2, the second sample size. *The T_L and T_U entries in the table are the boundaries of the lower and upper regions, respectively, for the rank sum associated with the sample that has fewer measurements*. If the sample sizes n_1 and n_2 are the same, either rank sum may be used as the test statistic. To illustrate, suppose $n_1 = 8$ and $n_2 = 10$. For a two-tailed test with $\alpha = .05$, we consult the table and find that the null hypothesis will be rejected if the rank sum of sample 1 (the sample with fewer measurements), T, is less than or equal to $T_L = 54$ *or* greater than or equal to $T_U = 98$. The Wilcoxon rank sum test is summarized in the next two boxes.

Table 14.3 Reproduction of Part of Table XII in Appendix A: Critical Values for the Wilcoxon Rank Sum Test

$\alpha = .025$ one-tailed; $\alpha = .05$ two-tailed

n_2 \ n_1	3		4		5		6		7		8		9		10	
	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U
3	5	16	6	18	6	21	7	23	7	26	8	28	8	31	9	33
4	6	18	11	25	12	28	12	32	13	35	14	38	15	41	16	44
5	6	21	12	28	18	37	19	41	20	45	21	49	22	53	24	56
6	7	23	12	32	19	41	26	52	28	56	29	61	31	65	32	70
7	7	26	13	35	20	45	28	56	37	68	39	73	41	78	43	83
8	8	28	14	38	21	49	29	61	39	73	49	87	51	93	54	98
9	8	31	15	41	22	53	31	65	41	78	51	93	63	108	66	114
10	9	33	16	44	24	56	32	70	43	83	54	98	66	114	79	131

Wilcoxon Rank Sum Test: Independent Samples*

Let D_1 and D_2 represent the probability distributions for populations 1 and 2, respectively.

One-Tailed Test

H_0: D_1 and D_2 are identical

H_a: D_1 is shifted to the right of D_2
[or H_a: D_1 is shifted to the left of D_2]

Test statistic:

T_1, if $n_1 < n_2$; T_2, if $n_2 < n_1$ (Either rank sum can be used if $n_1 = n_2$.)

Two-Tailed Test

H_0: D_1 and D_2 are identical

H_a: D_1 is shifted either to the left or to the right of D_2

Test statistic:

T_1, if $n_1 < n_2$; T_2, if $n_2 < n_1$ (Either rank sum can be used if $n_1 = n_2$.)
We will denote this rank sum as T.

(*continued*)

*Another statistic used to compare two populations on the basis of independent random samples is the *Mann–Whitney U-statistic*, a simple function of the rank sums. It can be shown that the Wilcoxon rank sum test and the Mann–Whitney U-test are equivalent.

Rejection region:	*Rejection region:*
T_1: $T_1 \geq T_U$ [or $T_1 \leq T_L$] T_2: $T_2 \leq T_L$ [or $T_2 \geq T_U$]	$T \leq T_L$ or $T \geq T_U$

where T_L and T_U are obtained from Table XII of Appendix A.

Ties: Assign tied measurements the average of the ranks they would receive if they were unequal, but occurred in successive order. For example, if the third-ranked and fourth-ranked measurements are tied, assign each a rank of $(3 + 4)/2 = 3.5$.

Conditions Required for a Valid Rank Sum Test

1. The two samples are random and independent.
2. The two probability distributions from which the samples are drawn are continuous.

Note that the assumptions necessary for the validity of the Wilcoxon rank sum test do not specify the shape or type of probability distribution. However, the distributions are assumed to be continuous so that the probability of tied measurements is 0 (see Chapter 5) and each measurement can be assigned a unique rank. In practice, however, rounding of continuous measurements will sometimes produce ties. As long as the number of ties is small relative to the sample sizes, the Wilcoxon test procedure will still have an approximate significance level of α. The test is not recommended to compare discrete distributions, for which many ties are expected.

Example 14.2

Applying The Rank Sum Test—Comparing Reaction Times of Two Drugs

Problem Do the data given in Table 14.2 provide sufficient evidence to indicate a shift in the probability distributions for drugs A and B—that is, that the probability distribution corresponding to drug A lies either to the right or to the left of the probability distribution corresponding to drug B? Test at the .05 level of significance.

Solution

H_0: The two populations of reaction times corresponding to drug A and drug B have the same probability distribution.

H_a: The probability distribution for drug A is shifted to the right or left of the probability distribution for drug B.*

Test statistic: Since drug A has fewer subjects than drug B, the test statistic is T_1, the rank sum of drug A's reaction times.

Rejection region: Since the test is two sided, we consult part a of Table XII for the rejection region corresponding to $\alpha = .05$. We will reject H_0 for $T_1 \leq T_L$ or $T_1 \geq T_U$. Thus, we will reject H_0 if $T_1 \leq 28$ or $T_1 \geq 56$.

Since T_1, the rank sum of drug A's reaction times in Table 14.2, is 25, it is in the rejection region. (See Figure 14.5.)† Therefore, there is sufficient evidence to reject H_0. This same conclusion can be reached with a statistical software package. The SAS printout of the analysis is shown in Figure 14.6. Both the test statistic ($T_1 = 25$) and one-tailed p-value ($p = .007$) are highlighted on the printout. The one-tailed p-value is less than $\alpha = .05$, leading us to reject H_0.

*The alternative hypotheses in this chapter will be stated in terms of a difference in the *location* of the distributions. However, since the shapes of the distributions may also differ under H_a, some of the figures (e.g., Figure 14.5) depicting the alternative hypothesis will show probability distributions with different shapes.

†Figure 14.5 depicts only one side of the two-sided alternative hypothesis. The other would show the distribution for drug A shifted to the right of the distribution for drug B.

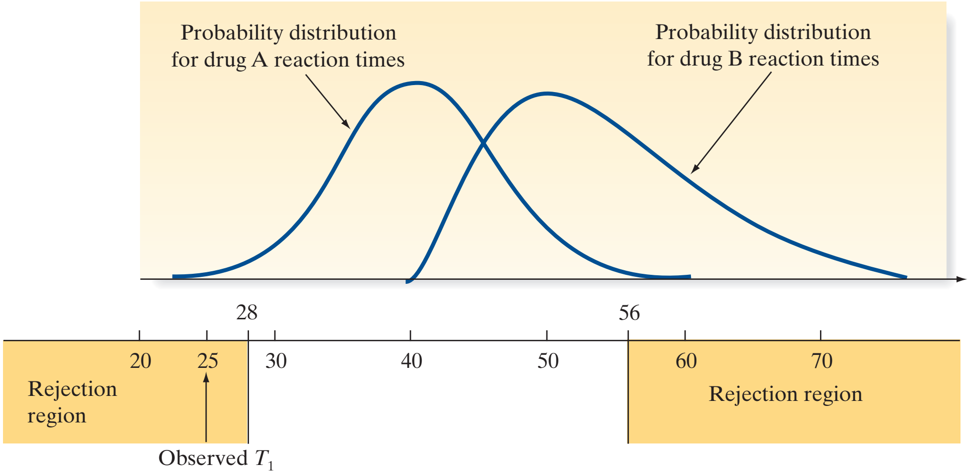
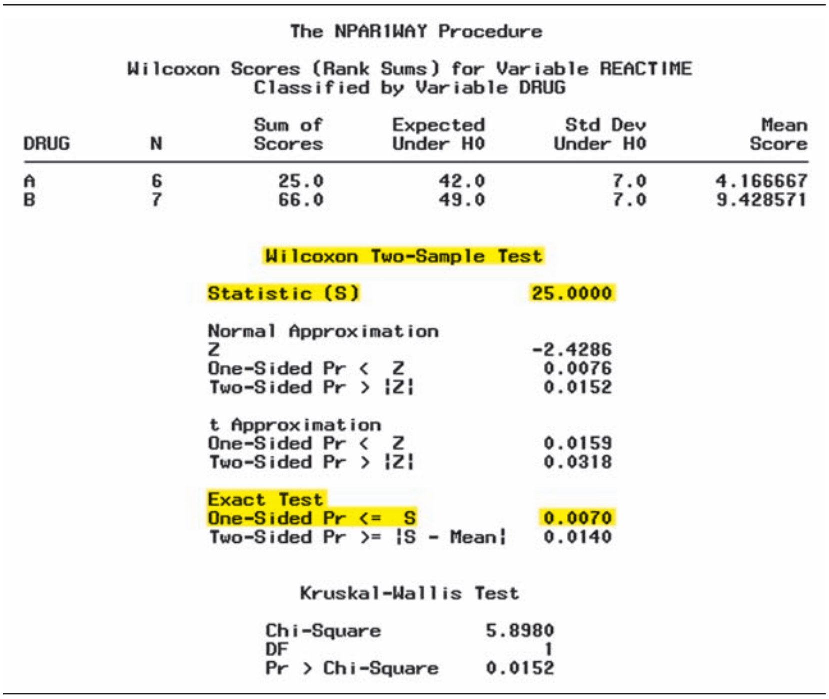


Figure 14.5
Alternative hypothesis and rejection region for Example 14.2.



```
                     The NPAR1WAY Procedure

           Wilcoxon Scores (Rank Sums) for Variable REACTIME
                     Classified by Variable DRUG

                        Sum of      Expected      Std Dev        Mean
DRUG          N         Scores      Under H0      Under H0       Score
_______________________________________________________________________
A             6          25.0         42.0          7.0        4.166667
B             7          66.0         49.0          7.0        9.428571


                     Wilcoxon Two-Sample Test

              Statistic (S)                25.0000

              Normal Approximation
              Z                            -2.4286
              One-Sided Pr <  Z             0.0076
              Two-Sided Pr > |Z|            0.0152

              t Approximation
              One-Sided Pr <  Z             0.0159
              Two-Sided Pr > |Z|            0.0318

              Exact Test
              One-Sided Pr <=  S            0.0070
              Two-Sided Pr >= |S - Mean|    0.0140


                       Kruskal-Wallis Test

                   Chi-Square          5.8980
                   DF                       1
                   Pr > Chi-Square     0.0152
```

Figure 14.6
SAS printout for Example 14.2

Look Back Our conclusion is that the probability distributions for drugs A and B are not identical. In fact, it appears that drug B tends to be associated with reaction times that are larger than those associated with drug A (because T_1 falls into the lower tail of the rejection region).

Now Work Exercise 14.20

Table XII in Appendix A gives values of T_L and T_U for values of n_1 and n_2 less than or equal to 10. When both sample sizes, n_1 and n_2, are 10 or larger, the sampling distribution of T_1 can be approximated by a normal distribution with mean

$$E(T_1) = \frac{n_1(n_1 + n_2 + 1)}{2}$$

and variance

$$\sigma^2_{T_1} = \frac{n_1 n_2(n_1 + n_2 + 1)}{12}$$

Therefore, for $n_1 \geq 10$ and $n_2 \geq 10$, we can conduct the Wilcoxon rank sum test using the familiar z-test of Chapters 8 and 9. The test is summarized in the following box:

The Wilcoxon Rank Sum Test for Large Samples ($n_1 \geq 10$ and $n_2 \geq 10$)

Let D_1 and D_2 represent the probability distributions for populations 1 and 2, respectively.

One-Tailed Test	**Two-Tailed Test**
H_0: D_1 and D_2 are identical	H_0: D_1 and D_2 are identical
H_a: D_1 is shifted to the right of D_2 (or H_a: D_1 is shifted to the left of D_2)	H_a: D_1 is shifted to the right or to the left of D_2

$$\textit{Test statistic: } z = \frac{T_1 - \dfrac{n_1(n_1 + n_2 + 1)}{2}}{\sqrt{\dfrac{n_1 n_2(n_1 + n_2 + 1)}{12}}}$$

Rejection region:	Rejection region:
$z > z_\alpha$ (or $z < -z_\alpha$)	$\lvert z \rvert > z_{\alpha/2}$

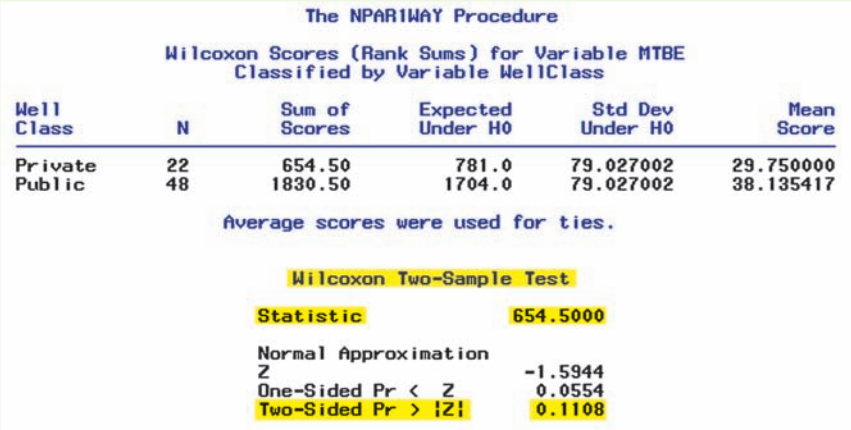
Statistics IN Action Revisited

Comparing the MTBE Levels of Different Types of Groundwater Wells

Refer to the study of MTBE contamination of New Hampshire groundwater wells (p. 14-2). One of the objectives of the study was to determine whether the level of MTBE contamination is different for private and public wells and for bedrock and unconsolidated aquifers. For this objective, the researchers focused on only the 70 sampled wells that had detectable levels of MTBE. They wanted to determine whether the distribution of MTBE levels in public wells is shifted above or below the distribution of MTBE levels in private wells and whether the distribution of MTBE levels in bedrock aquifers is shifted above or below the distribution of MTBE levels in unconsolidated aquifers.

To answer these questions, the researchers applied the Wilcoxon rank sum test for two independent samples. In the first analysis, public and private wells were compared; in the second analysis, bedrock and unconsolidated aquifers were compared. The SAS printouts for these analyses are shown in Figures SIA14.2 and SIA14.3, respectively. Both the test statistics and the two-tailed p-values are highlighted on the printouts.

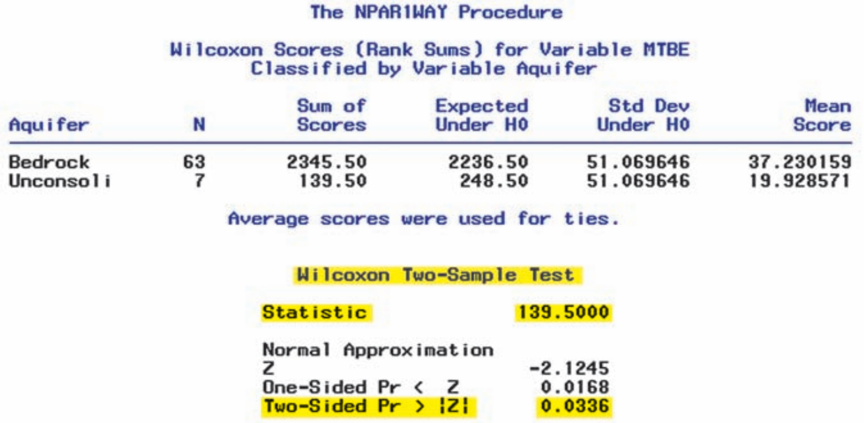
For the comparison of public and private wells in Figure SIA14.2, p-value = .1108. Thus, at $\alpha = .05$, there is insufficient evidence to conclude that the distribution of MTBE levels differs for public and private New Hampshire groundwater wells. Although public wells tend to have higher MTBE values than private wells (note the rank sums in Figure SIA14.2), the difference is not statistically significant.



```
                      The NPAR1WAY Procedure

          Wilcoxon Scores (Rank Sums) for Variable MTBE
                 Classified by Variable WellClass

Well                   Sum of      Expected       Std Dev          Mean
Class          N       Scores      Under H0      Under H0         Score
Private       22       654.50         781.0     79.027002     29.750000
Public        48      1830.50        1704.0     79.027002     38.135417

                 Average scores were used for ties.

                     Wilcoxon Two-Sample Test

               Statistic                654.5000

               Normal Approximation
               Z                         -1.5944
               One-Sided Pr <  Z          0.0554
               Two-Sided Pr > |Z|         0.1108
```

Figure SIA14.2
SAS rank sum test for comparing public and private wells

Statistics IN Action *(continued)*



```
                        The NPAR1WAY Procedure

             Wilcoxon Scores (Rank Sums) for Variable MTBE
                    Classified by Variable Aquifer

                      Sum of      Expected      Std Dev          Mean
Aquifer        N      Scores      Under H0      Under H0         Score

Bedrock       63     2345.50      2236.50     51.069646     37.230159
Unconsoli      7      139.50       248.50     51.069646     19.928571

                    Average scores were used for ties.

                      Wilcoxon Two-Sample Test

                   Statistic            139.5000

                   Normal Approximation
                   Z                     -2.1245
                   One-Sided Pr <  Z      0.0168
                   Two-Sided Pr > |Z|     0.0336
```

Figure SIA14.3
SAS rank sum test for comparing bedrock and unconsolidated aquifers

For the comparison of bedrock and unconsolidated aquifers in Figure SIA14.3, p-value $= .0336$. At $\alpha = .05$, there is sufficient evidence to conclude that the distribution of MTBE levels differs for bedrock and unconsolidated aquifers. Furthermore, the rank sums shown in Figure SIA14.3 indicate that bedrock aquifers have the higher MTBE levels.

[*Note:* Histograms of the MTBE levels for public wells, private wells, bedrock aquifers, and unconsolidated aquifers (not shown) reveal distributions that are highly skewed. Thus, application of the nonparametric rank sum test is appropriate.]

Exercises 14.16–14.33

Understanding the Principles

14.16 What is a rank sum?

14.17 *True or False*. If the rank sum for sample 1 is much larger than the rank sum for sample 2 when $n_1 = n_2$, then the distribution of population 1 is likely to be shifted to the right of the distribution of population 2.

4.18 What conditions are required for a valid application of the Wilcoxon rank sum test?

Learning the Mechanics

14.19 Specify the rejection region for the Wilcoxon rank sum test for independent samples in each of the following situations:

a. H_0: Two probability distributions, 1 and 2, are identical H_a: The probability distribution for population 1 is shifted to the right or left of the probability distribution for population 2

$n_1 = 7, n_2 = 8, \alpha = .10$

b. H_0: Two probability distributions, 1 and 2, are identical H_a: The probability distribution for population 1 is shifted to the right of the probability distribution for population 2

$n_1 = 6, n_2 = 6, \alpha = .05$

c. H_0: Two probability distributions, 1 and 2, are identical H_a: The probability distribution for population 1 is shifted to the left of the probability distribution for population 2

$n_1 = 7, n_2 = 10, \alpha = .025$

d. H_0: Two probability distributions, 1 and 2, are identical H_a: The probability distribution for population 1 is shifted to the right or left of the probability distribution for population 2

$n_1 = 20, n_2 = 20, \alpha = .05$

14.20 NW Suppose you want to compare two treatments, A and B. In particular, you wish to determine whether the distribution for population B is shifted to the right of the distribution for population A. You plan to use the Wilcoxon rank sum test.

a. Specify the null and alternative hypotheses you would test.

b. Suppose you obtained the following independent random samples of observations on experimental units subjected to the two treatments. These data are saved in the **LM14_20** file.

Sample A	37,	40,	33,	29,	42,	33,	35,	28,	34,
Sample B	65,	35,	47,	52					

Conduct a test of the hypotheses you specified in part **a.** Test, using $\alpha = .05$.

14.21 Suppose you wish to compare two treatments, A and B, on the basis of independent random samples of 15 observations selected from each of the two populations. If $T_1 = 173$, do the data provide sufficient evidence to indicate that distribution A is shifted to the left of distribution B? Test, using $\alpha = .05$.

14.22 Random samples of sizes $n_1 = 16$ and $n_2 = 12$ were drawn from populations 1 and 2, respectively. The measurements obtained are listed in the next table (p. 14-16) and saved in the **LM14_22** file.

Sample 1				Sample 2		
9.0	15.6	25.6	31.1	10.1	11.1	13.5
21.1	26.9	24.6	20.0	12.0	18.2	10.3
24.8	16.5	26.0	25.1	9.2	7.0	14.2
17.2	30.1	18.7	26.1	15.8	13.6	13.2

a. Conduct a hypothesis test to determine whether the probability distribution for population 2 is shifted to the left of the probability distribution for population 1. Use $\alpha = .05$.

b. What is the approximate p-value of the test of part **a**?

14.23 Independent random samples are selected from two populations. The data are shown in the following table and saved in the **LM14_23** file.

Sample 1		Sample 2		
15	16	5	9	5
10	13	12	8	10
12	8	9	4	

a. Use the Wilcoxon rank sum test to determine whether the data provide sufficient evidence to indicate a shift in the locations of the probability distributions of the sampled populations. Test, using $\alpha = .05$.

b. Do the data provide sufficient evidence to indicate that the probability distribution for population 1 is shifted to the right of the probability distribution for population 2? Use the Wilcoxon rank sum test with $\alpha = .05$.

Applying the Concepts—Basic

14.24 Short Message Service for cell phones. Short Message Service (SMS) is the formal name for the communication service that allows the interchange of short text messages between mobile telephone devices. About 75% of mobile phone subscribers worldwide send or receive SMS text messages. Consequently, SMS provides a opportunity for direct marketing. In *Management Dynamics* (2007), marketing researchers investigated the perceptions of college students towards SMS marketing. For one portion of the study, the researchers applied the Wilcoxon rank sum test to compare the distributions of the number of text messages sent and received during peak time for two groups of cell phone users: those on an annual contract and those with a pay-as-you-go option.

a. Specify the null hypothesis tested in the words of the problem.

b. Give the formula for the large-sample test statistic if there were 25 contract users and 40 pay-as-you-go users in the sample.

c. The Wilcoxon test results led the researchers to conclude "that contract users sent and received significantly more SMS's during peak time than pay-as-you-go users." Based on this information, draw a graph that is representative of the two SMS usage rate populations.

14.25 The X-Factor in golf performance. Many golf teaching professionals believe that a greater hip-to-shoulder differential angle during the early downswing—dubbed the "X-Factor"—leads to improved golf performance. The *Journal of Quantitative Analysis in Sports* (Vol. 5, 2009) published an article on the X-Factor and its relationship to golfing performance. The study involved 15 male golfers with a player handicap of 20 strokes or less. The golfers were divided into two groups: 8 golfers with a handicap of 10 strokes or less (low-handicapped group) and 7 golfers with a handicap between 12 and 20 strokes (high-handicapped group). The X-Factor, i.e., the hip-to-shoulder differential angle (in degrees), was measured for each golfer at the top of the backswing during his tee shot. The researchers hypothesized that low-handicapped golfers will tend to have higher X-factors than high-handicapped golfers. The researchers also discovered that the sample data were not normally distributed. Consequently, they applied a nonparametric test.

a. What nonparametric test is appropriate for analyzing these data?

b. Specify the null and alternative hypotheses of interest in the words of the problem.

c. Give the rejection region for this test, using $\alpha = .05$.

d. The researchers reported a p-value of .487. Use this result to draw a conclusion.

14.26 Bursting strength of bottles. Polyethylene terephthalate (PET) bottles are used for carbonated beverages. A critical property of PET bottles is their bursting strength (i.e., the pressure at which bottles filled with water burst when pressurized). In the *Journal of Data Science* (May 2003), researchers measured the bursting strength of PET bottles made from two different designs: an old design and a new design. The data (in pounds per square inch) for 10 bottles of each design are shown in the accompanying table and saved in the **PET** file. Suppose you want to compare the distributions of bursting strengths for the two designs.

Old Design	210	212	211	211	190	213	212	211	164	209
New Design	216	217	162	137	219	216	179	153	152	217

a. Rank all 20 observed pressures from smallest to largest, and assign ranks from 1 to 20.

b. Sum the ranks of the observations from the old design.

c. Sum the ranks of the observations from the new design.

d. Compute the Wilcoxon rank sum statistic.

e. Carry out a nonparametric test (at $\alpha = .05$) to compare the distribution of bursting strengths for the two designs.

14.27 Research on eating disorders. The "fear of negative evaluation" (FNE) scores for 11 female students known to suffer from the eating disorder bulimia and 14 female students with normal eating habits, first presented in Exercise 2.40 (p. 48), are reproduced in the next table (top of page 14-17). (Recall that the higher the score, the greater is the fear of a negative evaluation.) These data are saved in the **BULIMIA** file. Suppose you want to determine whether the distribution of the FNE scores for bulimic female students is shifted above the corresponding distribution for female students with normal eating habits.

a. Specify H_0 and H_a for the test.

b. Rank all 25 FNE scores in the data set from smallest to largest.

c. Sum the ranks of the 11 FNE scores for bulimic students.

d. Sum the ranks of the 14 FNE scores for students with normal eating habits.

e. Give the rejection region for a nonparametric test of the data if $\alpha = .10$.

f. Conduct the test and give the conclusion in the words of the problem.

Data for Exercise 14.27

Bulimic Students	21	13	10	20	25	19	16	21	24	13	14			
Normal Students	13	6	16	13	8	19	23	18	11	19	7	10	15	20

Based on Randles, R.H. "On neutral responses (zeros) in the sign test and ties in the Wilcoxon–Mann–Whitney test." *American Statistician,* Vol. 55, No. 2, May 2001 (Figure 3).

14.28 Children's recall of TV ads. Refer to the *Journal of Advertising* (Spring 2006) study of children's recall of television advertisements, presented in Exercise 9.15 (p. 424). Two groups of children were shown a 60-second commercial for Sunkist Fun Fruit Rock-n-Roll Shapes. One group (the A/V group) was shown both the audio and video portions of the ad; the other group (the video-only group) was shown only the video portion of the commercial. The number out of 10 specific items from the ad recalled correctly by each child is shown in the accompanying table. (These data are saved in the **FUNFRUIT** file.) Recall that the researchers theorized that children who receive an audiovisual presentation will have the same level of recall as those who receive only the visual aspects of the ad. Consider testing the researchers' theory, using the Wilcoxon rank sum test.

A/V group: 0 4 6 6 1 2 2 6 6 4 1 2 6 1 3 0 2 5 4 5
Video-only group: 6 3 6 2 2 4 7 6 1 3 6 2 3 1 3 2 5 2 4 6

Based on Maher, J. K., Hu, M. Y., and Kolbe, R. H. "Children's recall of television ad elements." *Journal of Advertising*, Vol. 35, No. 1, Spring 2006 (Table 1).

a. Set up the appropriate null and alternative hypotheses for the test.
b. Find the value of the test statistic.
c. Give the rejection region for $\alpha = .10$.
d. Make the appropriate inference. What can you say about the researchers' theory?

Applying the Concepts–Intermediate

14.29 Is honey a cough remedy? Refer to the *Archives of Pediatrics and Adolescent Medicine* (Dec. 2007) study of honey as a children's cough remedy, Exercise 2.32 (p. 45). Recall that 70 children who were ill with an upper respiratory tract infection were given either a dosage of dextromethorphan (DM)—an over-the-counter cough medicine—or a similar dose of honey. Parents then rated their children's cough symptoms and the improvement in total cough symptoms score was determined for each child. The data (improvement scores) are reproduced in the accompanying table and saved in the **HONEYCOUGH** file. The researchers concluded that "honey may be a preferable treatment for the cough and sleep difficulty associated with childhood upper respiratory tract infection." Use the nonparametric method presented in this section to analyze the data (use $\alpha = .05$). Do you agree with the researchers?

Honey Dosage: 12 11 15 11 10 13 10 4 15 16 9 14 10
6 10 8 11 12 12 8 12 9 11 15 10 15
9 13 8 12 10 8 9 5 12
DM Dosage: 4 6 9 4 7 7 7 9 12 10 11 6 3
4 9 12 7 6 8 12 12 4 12 13 7 10
13 9 4 4 10 15 9

Based on Paul, I. M., et al. "Effect of honey, dextromethorphan, and no treatment on nocturnal cough and sleep quality for coughing children and their parents." *Archives of Pediatrics and Adolescent Medicine*, Vol. 161, No. 12, Dec. 2007 (data simulated).

14.30 Does rudeness really matter in the workplace? Refer to the *Academy of Management Journal* (Oct. 2007) study on rudeness in the workplace, Exercise 9.26 (p. 427). Recall that 98 college students enrolled in a management course were randomly assigned to one of two experimental conditions: rudeness condition (where the students were berated by a facilitator for being irresponsible and unprofessional) and control group (no facilitator comments). Each student was asked to write down as many uses for a brick as possible in five minutes. The data, saved in the **RUDE** file, are reproduced below.

Control Group:
1 24 5 16 21 7 20 1 9 20 19 10 23 16 0 4 9 13
17 13 0 2 12 11 7 1 19 9 12 18 5 21 30 15 4 2
12 11 10 13 11 3 6 10 13 16 12 28 19 12 20 3 11
Rudeness Condition:
4 11 18 11 9 6 5 11 9 12 7 5 7 3 11 1 9 11
10 7 8 9 10 7 11 4 13 5 4 7 8 3 8 15 9 16
10 0 7 15 13 9 2 13 10

a. Show that although the data for the rudeness condition are approximately normally distributed, the control group data are skewed.
b. Conduct the appropriate nonparametric test (at $\alpha = .01$) to determine if the true median performance level for students in the rudeness condition is lower than the true median performance level for students in the control group.
c. Explain why the parametric two-sample test conducted in Exercise 9.26 is appropriate even though the data for both groups are not normally distributed. (Note that the nonparametric and parametric tests yield the same conclusions.)

14.31 Computer-mediated communication study. Computer-mediated communication (CMC) is a form of interaction that heavily involves technology (e.g., instant messaging, e-mail). A study was conducted to compare relational intimacy in people interacting via CMC with people meeting face-to-face (FTF) (*Journal of Computer-Mediated Communication*, Apr. 2004). Participants were 48 undergraduate students, of whom half were randomly assigned to the CMC group and half to the FTF group. Each group was given a task that required communication among its group members. Those in the CMC group communicated via the "chat" mode of instant-messaging software; those in the FTF group met in a conference room. The variable of interest, relational intimacy score, was measured (on a seven-point scale) for each participant after each of three different meetings. Scores for the first meeting are given in the accompanying table and saved in the **INTIMACY** file. The researchers hypothesized that the relational intimacy

CMC
group: 4 3 3 4 3 3 3 3 4 4 3 4 3 3 2 4 2 4 5 4 4 4 5 3
FTF
group: 5 4 4 4 3 3 3 4 3 3 3 3 4 4 4 4 4 3 3 3 4 4 2 4

Note: Data simulated from descriptive statistics provided in article.

scores for participants in the CMC group will tend to be lower than the relational intimacy scores for participants in the FTF group.

a. Which nonparametric procedure should be used to test the researchers' hypothesis?
b. Specify the null and alternative hypotheses of the test.
c. Give the rejection region for the test, using $\alpha = .10$.
d. Conduct the test and give the appropriate conclusion in the context of the problem.

14.32 Brood-parasitic birds. The term *brood-parasitic intruder* is used to describe a bird that searches for and lays eggs in a nest built by a bird of another species. For example, the brown-headed cowbird is known to be a brood parasite of the smaller willow flycatcher. Ornithologists theorize that those flycatchers which recognize, but do not vocally react to, cowbird calls are more apt to defend their nests and less likely to be found and parasitized. In a study published in *The Condor* (May 1995), each of 13 active flycatcher nests was categorized as parasitized (if at least one cowbird egg was present) or nonparasitized. Cowbird songs were taped and played back while the flycatcher pairs were sitting in the nest prior to incubation. The vocalization rate (number of calls per minute) of each flycatcher pair was recorded. The data for the two groups of flycatchers are given in the table and saved in the **COWBIRD** file. Do the data suggest (at $\alpha = .05$) that the vocalization rates of parasitized flycatchers are higher than those of nonparasitized flycatchers?

Parasitized	Not Parasitized
2.00	1.00
1.25	1.00
8.50	0
1.10	3.25
1.25	1.00
3.75	.25
5.50	

Based on Uyehara, J. C., and Narins, P. M. "Nest defense by Willow Flycatchers to brood-parasitic intruders." *The Condor*, Vol. 97, No. 2, May 1995, p. 364 (Figure 1).

14.33 Family involvement in homework. A study of the impact of the interactive Teachers Involve Parents in Schoolwork (TIPS) program was published in the *Journal of Educational Research* (July/Aug. 2003). A sample of 128 middle school students were assigned to complete TIPS homework assignments, while 98 students were assigned traditional, noninteractive homework assignments (ATIPS). At the end of the study, all students reported on the level of family involvement in their homework on a five-point scale (0 = Never, 1 = Rarely, 2 = Sometimes, 3 = Frequently, 4 = Always). The data for the science, math, and language arts homework are saved in the **HWSTUDY** file. (The first five and last five observations in the data set are reproduced in the accompanying table.)

a. Why might a nonparametric test be the most appropriate test to apply in order to compare the levels of family involvement in homework assignments of TIPS and ATIPS students?
b. Conduct a nonparametric analysis to compare the involvement in science homework assignments of TIPS and ATIPS students. Use $\alpha = .05$.
c. Repeat part **b** for mathematics homework assignments.
d. Repeat part **b** for language arts homework assignments.

Homework Condition	Science	Math	Language
ATIPS	1	0	0
ATIPS	0	1	1
ATIPS	0	1	0
ATIPS	1	2	0
ATIPS	1	1	2
⋮	⋮	⋮	⋮
TIPS	2	3	2
TIPS	1	4	2
TIPS	2	4	2
TIPS	4	0	3
TIPS	2	0	1

Source: Van Voorhis, F. L. "Interactive homework in middle school: Effects on family involvement and science achievement." *Journal of Educational Research*, 96(6), 2003, pp. 323–338. Reprinted with permission from Frances Van Voorhis.

14.4 Comparing Two Populations: Paired Difference Experiment

Nonparametric techniques may also be employed to compare two probability distributions when a paired difference design is used. For example, consumer preferences for two competing products are often compared by having each of a sample of consumers rate both products. Thus, the ratings have been paired on each consumer. Following is an example of this type of experiment.

For some paper products, softness is an important consideration in determining consumer acceptance. One method of determining softness is to have judges give a sample of the products a softness rating. Suppose each of 10 judges is given a sample of two products that a company wants to compare. Each judge rates the softness of each product on a scale from 1 to 20, with higher ratings implying a softer product. The results of the experiment are shown in Table 14.4.

Since this is a paired difference experiment, we analyze the differences between the measurements. (See Section 9.3.) However, a nonparametric approach developed by Wilcoxon requires that we calculate the ranks of the absolute values of the differences between the measurements (i.e., the ranks of the differences after removing any minus signs). *Note that tied absolute differences (e.g., the two differences of 4) are assigned the average of the ranks they would receive if they were unequal, but successive, measurements (e.g., 4.5, the average of the ranks 4 and 5).* After the absolute differences are ranked,

Table 14.4 **Softness Ratings of Paper**

Judge	Product A	Product B	Difference $(A - B)$	Absolute Value of Difference	Rank of $(A - B)$ Absolute Value
1	12	8	4	4	4.5
2	16	10	6	6	7
3	8	9	−1	1	1
4	10	8	2	2	2
5	19	12	7	7	8
6	14	17	−3	3	3
7	12	4	8	8	9
8	10	6	4	4	4.5
9	12	17	−5	5	6
10	16	4	12	12	10

$T_{\pm}$ = Sum of positive ranks = 45
T_{-} = Sum of negative ranks = 10

Data Set: SOFTPAPER

the sum of the ranks of the positive differences of the original measurements, T_+, and the sum of the ranks of the negative differences of the original measurements, T_-, are computed. (The ranks of the negative differences are highlighted in Table 14.4.)

We are now prepared to test the nonparametric hypotheses:

H_0: The probability distributions of the ratings for products A and B are identical.

H_a: The probability distributions of the ratings differ (in location) for the two products. (Note that this is a two-sided alternative and that it implies a two-tailed test.)

Test statistic: T = Smaller of the positive and negative rank sums T_+ and T_-

The smaller the value of T, the greater is the evidence indicating that the two probability distributions differ in location. The rejection region for T can be determined by consulting Table XIII in Appendix A, (part of which is shown in Table 14.5). This table gives a value T_0 for both one-tailed and two-tailed tests for each value of n, the number of matched pairs. For a two-tailed test with $\alpha = .05$, we will reject H_0 if $T \leq T_0$. You can see in Table 14.5 that the value of T_0 which locates the boundary of the rejection region

Table 14.5 **Reproduction of Part of Table XIII of Appendix A: Critical Values for the Wilcoxon Paired Difference Signed Rank Test**

One-Tailed	Two-Tailed	$n = 5$	$n = 6$	$n = 7$	$n = 8$	$n = 9$	$n = 10$
$\alpha = .05$	$\alpha = .10$	1	2	4	6	8	11
$\alpha = .025$	$\alpha = .05$		1	2	4	6	8
$\alpha = .01$	$\alpha = .02$			0	2	3	5
$\alpha = .005$	$\alpha = .01$				0	2	3
		$n = 11$	$n = 12$	$n = 13$	$n = 14$	$n = 15$	$n = 16$
$\alpha = .05$	$\alpha = .10$	14	17	21	26	30	36
$\alpha = .025$	$\alpha = .05$	11	14	17	21	25	30
$\alpha = .01$	$\alpha = .02$	7	10	13	16	20	24
$\alpha = .005$	$\alpha = .01$	5	7	10	13	16	19
		$n = 17$	$n = 18$	$n = 19$	$n = 20$	$n = 21$	$n = 22$
$\alpha = .05$	$\alpha = .10$	41	47	54	60	68	75
$\alpha = .025$	$\alpha = .05$	35	40	46	52	59	66
$\alpha = .01$	$\alpha = .02$	28	33	38	43	49	56
$\alpha = .005$	$\alpha = .01$	23	28	32	37	43	49
		$n = 23$	$n = 24$	$n = 25$	$n = 26$	$n = 27$	$n = 28$
$\alpha = .05$	$\alpha = .10$	83	92	101	110	120	130
$\alpha = .025$	$\alpha = .05$	73	81	90	98	107	117
$\alpha = .01$	$\alpha = .02$	62	69	77	85	93	102
$\alpha = .005$	$\alpha = .01$	55	61	68	76	84	92

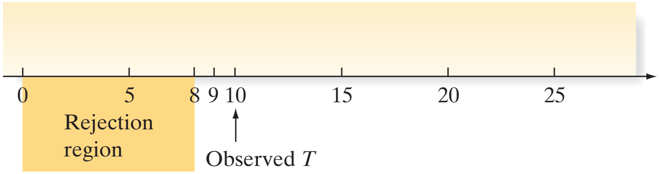


Figure 14.7
Rejection region for paired difference experiment

for the judges' ratings for $\alpha = .05$ and $n = 10$ pairs of observations is 8. Thus, the rejection region for the test (see Figure 14.7) is

$$\textit{Rejection region:} \quad T \le 8 \quad \text{for } \alpha = .05$$

Since the smaller rank sum for the paper data, $T_- = 10$, does not fall within the rejection region, the experiment has not provided sufficient evidence indicating that the two paper products differ with respect to their softness ratings at the $\alpha = .05$ level.

Note that if a significance level of $\alpha = .10$ had been used, the rejection region would have been $T \le 11$ and we would have rejected H_0. In other words, the samples do provide evidence that the probability distributions of the softness ratings differ at the $\alpha = .10$ significance level.

The **Wilcoxon signed rank test** is summarized in the next box. Note that the difference measurements are assumed to have a continuous probability distribution so that the absolute differences will have unique ranks. Although tied (absolute) differences can be assigned ranks by averaging, in order to ensure the validity of the test, the number of ties should be small relative to the number of observations.

Wilcoxon Signed Rank Test for a Paired Difference Experiment

Let D_1 and D_2 represent the probability distributions for populations 1 and 2, respectively.

One-Tailed Test	**Two-Tailed Test**
H_0: D_1 and D_2 are identical	H_0: D_1 and D_2 are identical
H_a: D_1 is shifted to the right of D_2 [or H_a: D_1 is shifted to the left of D_2]	H_a: D_1 is shifted either to the left or to the right of D_2

Calculate the difference within each of the n matched pairs of observations. Then rank the absolute value of the n differences from the smallest (rank 1) to the highest (rank n), and calculate the rank sum T_- of the negative differences and the rank sum T_+ of the positive differences. [*Note:* Differences equal to 0 are eliminated, and the number n of differences is reduced accordingly.]

Test statistic: T_-, the rank sum of the negative differences [or T_+, the rank sum of the positive differences]	*Test statistic:* T, the smaller of T_+ or T_-
Rejection region: $T_- \le T_0$ [or $T_+ \le T_0$]	*Rejection region:* $T \le T_0$

where T_0 is given in Table XIII in Appendix A.

Ties: Assign tied absolute differences the average of the ranks they would receive if they were unequal, but occurred in successive order. For example, if the third-ranked and fourth-ranked differences are tied, assign both a rank of $(3 + 4)/2 = 3.5$.

Conditions Required for a Valid Signed Rank Test

1. The sample of differences is randomly selected from the population of differences.
2. The probability distribution from which the sample of paired differences is drawn is continuous.

Example 14.3

Applying the Signed Rank Test—Comparing Two Crime Prevention Plans

Problem Suppose the police commissioner in a small community must choose between two plans for patrolling the town's streets. Plan A, the less expensive plan, uses voluntary citizen groups to patrol certain high-risk neighborhoods. In contrast, plan B would utilize police patrols. As an aid in reaching a decision, both plans are examined by 10 trained criminologists, each of whom is asked to rate the plans on a scale from 1 to 10. (High ratings imply a more effective crime prevention plan.) The city will adopt plan B (and hire extra police) only if the data provide sufficient evidence that criminologists tend to rate plan B more effective than plan A. The results of the survey are shown in Table 14.6. Do the data provide evidence at the $\alpha = .05$ level that the distribution of ratings for plan B lies above that for plan A? Use the Wilcoxon signed rank test to answer the question.

Table 14.6 Effectiveness Ratings by 10 Qualified Crime Prevention Experts

	Plan		Difference	
Crime Prevention Expert	A	B	(A − B)	Rank of Absolute Difference
1	7	9	−2	4.5
2	4	5	−1	2
3	8	8	0	(Eliminated)
4	9	8	1	2
5	3	6	−3	6
6	6	10	−4	7.5
7	8	9	−1	2
8	10	8	2	4.5
9	9	4	5	9
10	5	9	−4	7.5

Positive rank sum $= T_+ = 15.5$

Data Set: CRIMEPLAN

Solution The null and alternative hypotheses are as follows:

H_0: The two probability distributions of effectiveness ratings are identical

H_a: The effectiveness ratings of the more expensive plan (B) tend to exceed those of plan A

Observe that the alternative hypothesis is one sided (i.e., we only wish to detect a shift in the distribution of the B ratings to the right of the distribution of A ratings); therefore, it implies a one-tailed test of the null hypothesis. (See Figure 14.8.) If the alternative hypothesis is true, the B ratings will tend to be larger than the paired A ratings, more negative differences in pairs will occur, T_- will be large, and T_+ will be small. Because Table XIII is constructed to give lower-tail values of T_0, we will use T_+ as the test statistic and reject H_0 for $T_+ \leq T_0$.

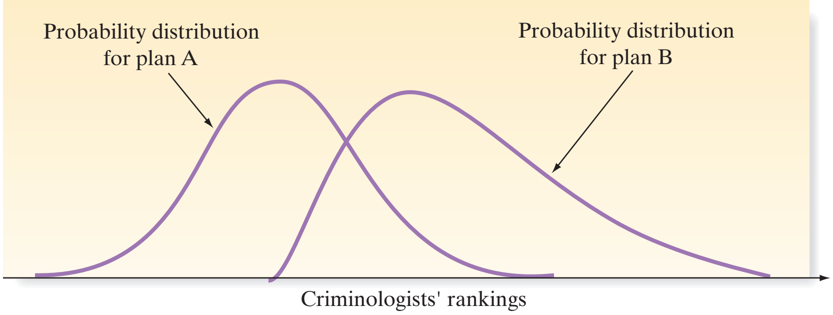


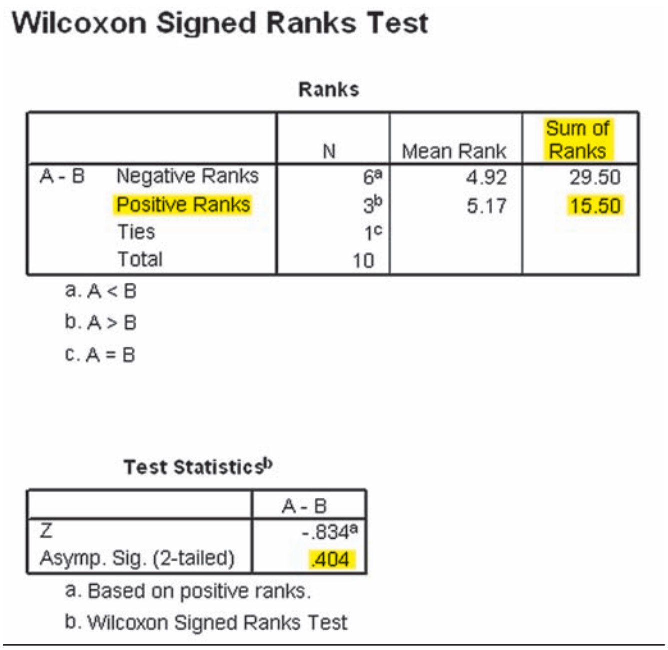
Figure 14.8
The alternative hypothesis for Example 14.3

Because a paired difference design was used (both plans were evaluated by the same criminologist), we first calculate the difference between the rating for each expert. The differences in ratings for the pairs (A − B) are shown in Table 14.6. Note that one of the differences equals 0. Consequently, we eliminate this pair from the ranking and reduce the number of pairs to $n = 9$. Looking in Table XIII, we have $T_0 = 8$ for a one-tailed test with $\alpha = .05$ and $n = 9$. Therefore, the test statistic and rejection region for the test are

Test statistic: T_+, the positive rank sum

Rejection region: $T_+ \leq 8$

Summing the ranks of the positive differences (highlighted) in Table 14.6, we find that $T_+ = 15.5$. Since this value exceeds the critical value, $T_0 = 8$, we conclude that the sample provides insufficient evidence at the $\alpha = .05$ level to support the alternative hypothesis. The commissioner *cannot* conclude that the plan utilizing police patrols tends to be rated higher than the plan using citizen volunteers. That is, on the basis of this study, extra police will not be hired.



Wilcoxon Signed Ranks Test

Ranks

		N	Mean Rank	Sum of Ranks
A - B	Negative Ranks	6[a]	4.92	29.50
	Positive Ranks	3[b]	5.17	15.50
	Ties	1[c]		
	Total	10		

a. A < B

b. A > B

c. A = B

Test Statistics[b]

	A - B
Z	-.834[a]
Asymp. Sig. (2-tailed)	.404

a. Based on positive ranks.

b. Wilcoxon Signed Ranks Test

Figure 14.9
SPSS printout for Example 14.3

Look Back An SPSS printout of the analysis, shown in Figure 14.9, confirms the preceding conclusion. Both the test statistic and two-tailed p-value are highlighted on the printout. Since the one-tailed p-value, $.404/2 = .202$, exceeds $\alpha = .05$, we fail to reject H_0.

Now Work Exercise 14.37

As is the case for the rank sum test for independent samples, the sampling distribution of the signed rank statistic can be approximated by a normal distribution when the number n of paired observations is large (say, $n \geq 25$). The large-sample z-test is summarized in the following box:

Wilcoxon Signed Rank Test for Large Samples ($n \geq 25$)

Let D_1 and D_2 represent the probability distributions for populations 1 and 2, respectively.

One-Tailed Test	**Two-Tailed Test**
H_0: D_1 and D_2 are identical H_a: D_1 is shifted to the right of D_2 [or H_a: D_1 is shifted to the left of D_2]	H_0: D_1 and D_2 are identical H_a: D_1 is shifted either to the left or to the right of D_2

$$\textit{Test statistic: } z = \frac{T_+ - [n(n+1)/4]}{\sqrt{[n(n+1)(2n+1)]/24}}$$

Rejection region:
$z > z_\alpha$ [or $z < -z_\alpha$]

Rejection region:
$|z| > z_{\alpha/2}$

Assumptions: The sample size n is greater than or equal to 25. Differences equal to 0 are eliminated and the number n of differences is reduced accordingly. Tied absolute differences receive ranks equal to the average of the ranks they would have received had they not been tied.

Exercises 14.34–14.52

Understanding the Principles

14.34 Explain the difference between the one- and two-tailed versions of the Wilcoxon signed rank test for the paired difference experiment.

14.35 In order to conduct the Wilcoxon signed rank test, why do we need to assume that the probability distribution of differences is continuous?

Learning the Mechanics

14.36 Specify the test statistic and the rejection region for the Wilcoxon signed rank test for the paired difference design in each of the following situations:

a. H_0: Two probability distributions, A and B, are identical
H_a: The probability distribution for population A is shifted to the right or left of the probability distribution for population B
$n = 20, \alpha = .10$

b. H_0: Two probability distributions, A and B, are identical
H_a: The probability distribution for population A is shifted to the right of the probability distribution for population B
$n = 39, \alpha = .05$

c. H_0: Two probability distributions, A and B, are identical
H_a: The probability distribution for population A is shifted to the left of the probability distribution for population B
$n = 7, \alpha = .005$

14.37 NW Suppose you want to test a hypothesis that two treatments, A and B, are equivalent against the alternative hypothesis that the responses for A tend to be larger than those for B. You plan to use a paired difference experiment and to analyze the resulting data with the Wilcoxon signed rank test.

a. Specify the null and alternative hypotheses you would test.

b. Suppose the paired difference experiment yielded the data in the accompanying table. (These data are saved in the **LM14_37** file.) Conduct the test, of part **a.** Test using $\alpha = .025$.

Pair	A	B	Pair	A	B
1	54	45	6	77	75
2	60	45	7	74	63
3	98	87	8	29	30
4	43	31	9	63	59
5	82	71	10	80	82

14.38 Suppose you wish to test a hypothesis that two treatments, A and B, are equivalent against the alternative that the responses for A tend to be larger than those for B.

a. If the number of pairs equals 25, give the rejection region for the large-sample Wilcoxon signed rank test for $\alpha = .05$.

b. Suppose that $T_+ = 273$. State your test conclusions.

c. Find the p-value for the test and interpret it.

14.39 A paired difference experiment with $n = 30$ pairs yielded $T_+ = 354$.

a. Specify the null and alternative hypotheses that should be used in conducting a hypothesis test to determine whether the probability distribution for population A is located to the right of that for population B.

b. Conduct the test of part **a,** using $\alpha = .05$.

c. What is the approximate p-value of the test of part **b**?

d. What assumptions are necessary to ensure the validity of the test you performed in part **b**?

14.40 A random sample of nine pairs of measurements is shown in the following table (saved in the **LM14_40** file).

Pair	Sample Data from Population 1	Sample Data from Population 2
1	8	7
2	10	1
3	6	4
4	10	10
5	7	4
6	8	3
7	4	6
8	9	2
9	8	4

a. Use the Wilcoxon signed rank test to determine whether the data provide sufficient evidence to indicate that the probability distribution for population 1 is shifted to the right of the probability distribution for population 2. Test, using $\alpha = .05$.

b. Use the Wilcoxon signed rank test to determine whether the data provide sufficient evidence to indicate that the probability distribution for population 1 is shifted either to the right or to the left of the probability distribution for population 2. Test, using $\alpha = .05$.

Applying the Concepts—Basic

14.41 Treating psoriasis with the "Doctorfish of Kangal." Refer to the *Evidence-Based Research in Complementary and Alternative Medicine* (Dec. 2006) study of treating psoriasis with ichthyotherapy, presented in Exercise 2.133 (p. 85). (Recall that the therapy is also known as the "Doctorfish of Kangal," since it uses fish from the hot pools of Kangal, Turkey, to feed on skin scales.) In the study, 67 patients diagnosed with psoriasis underwent three weeks of ichthyotherapy. The Psoriasis Area Severity Index (PASI) of each patient was measured both before and after treatment. (The lower the PASI score, the better is the skin condition.) Before- and after-treatment PASI scores were compared with the use of the Wilcoxon signed rank test.

a. Explain why the PASI scores should be analyzed with a test for paired differences.

b. Refer to the box plots shown in Exercise 2.133. Give a reason that the researchers opted to use a nonparametric test to compare the PASI scores.

c. The p-value for the Wilcoxon signed ranks test was reported as $p < .0001$. Interpret this result, and comment on the effectiveness of ichthyotherapy in treating psoriasis.

14.42 Computer-mediated communication study. Refer to the *Journal of Computer-Mediated Communication* (Apr. 2004) study comparing people who interact via computer-mediated communication (CMC) with those who meet face-to-face (FTF), presented in Exercise 14.31 (p. 14-17). Relational intimacy scores (measured on a seven-point scale) were obtained for each participant after each of three different meetings. The researchers hypothesized that relational intimacy scores for participants in the CMC group will tend to be higher at the third meeting than at the first meeting; however, they hypothesize that there are no differences in scores between the first and third meetings for the FTF group.

a. Explain why a nonparametric Wilcoxon signed ranks test is appropriate for analyzing the data.

b. For the CMC group comparison, give the null and alternative hypotheses of interest.

c. Give the rejection region (at $\alpha = .05$) for conducting the test mentioned in part **b.** Recall that there were 24 participants assigned to the CMC group.

d. For the FTF group comparison, give the null and alternative hypotheses of interest.

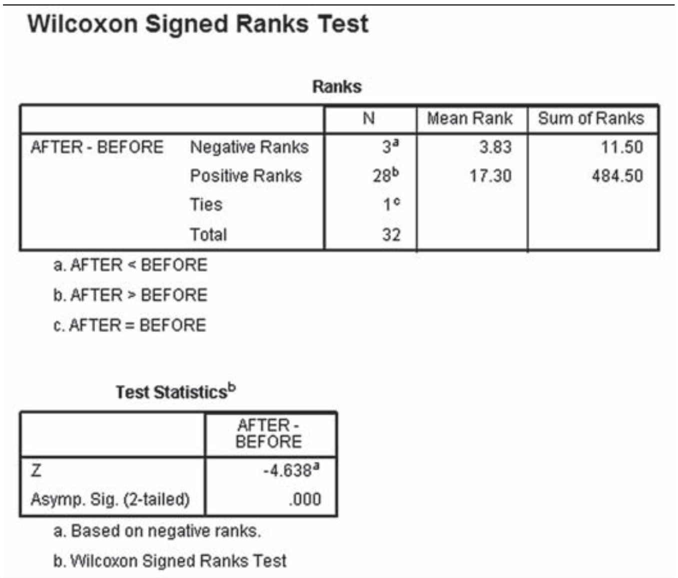
e. Give the rejection region (at $\alpha = .05$) for conducting the test mentioned in part **d.** Recall that there were 24 participants assigned to the FTF group.

14.43 Healing potential of handling museum objects. Refer to the *Museum & Society* (Nov. 2009) study of the healing potential of handling museum objects, Exercise 9.39 (p. 436). Recall that the health status of each of 32 hospital patients was recorded both before and after handling a museum object (such as an archaeological artifact or brass etching). The simulated data (measured on a 100-point scale) are reproduced in the next table and saved in the **MUSEUM** file. The Wilcoxon signed rank test was applied to the data, with the results shown in the accompanying SPSS printout.

a. Use the information in the printout to find the large-sample Wilcoxon signed rank test statistic.

b. Does handling a museum object have a positive impact on a sick patient's well-being? Test using $\alpha = .01$.

Session	Before	After	Session	Before	After
1	52	59	17	65	65
2	42	54	18	52	63
3	46	55	19	39	50
4	42	51	20	59	69
5	43	42	21	49	61
6	30	43	22	59	66
7	63	79	23	57	61
8	56	59	24	56	58
9	46	53	25	47	55
10	55	57	26	61	62
11	43	49	27	65	61
12	73	83	28	36	53
13	63	72	29	50	61
14	40	49	30	40	52
15	50	49	31	65	70
16	50	64	32	59	72



Wilcoxon Signed Ranks Test

Ranks

		N	Mean Rank	Sum of Ranks
AFTER - BEFORE	Negative Ranks	3[a]	3.83	11.50
	Positive Ranks	28[b]	17.30	484.50
	Ties	1[c]		
	Total	32		

a. AFTER < BEFORE

b. AFTER > BEFORE

c. AFTER = BEFORE

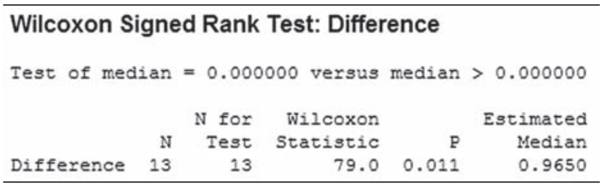
Test Statistics[b]

	AFTER - BEFORE
Z	-4.638[a]
Asymp. Sig. (2-tailed)	.000

a. Based on negative ranks.

b. Wilcoxon Signed Ranks Test

14.44 Impact of red light cameras on car crashes. Refer to the June 2007 Virginia Department of Transportation (VDOT) study of a newly adopted photo-red-light enforcement program, Exercise 9.47 (p. 438). Recall that the VDOT provided crash data both before and after installation of red light cameras at several intersections. The data (measured as the number of crashes caused by red light running per intersection per year) for 13 intersections in Fairfax County, Virginia, are reproduced in the next table (p. 14-25) and saved in the **REDLIGHT** file. The VDOT wants to determine if the photo-red enforcement program is effective in reducing red-light-running crash incidents at intersections. Use the nonparametric Wilcoxon singed rank test (and the accompanying MINITAB printout below) to analyze the data for the VDOT.



Wilcoxon Signed Rank Test: Difference

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

                 N for   Wilcoxon            Estimated
             N    Test  Statistic       P       Median
Difference  13      13       79.0   0.011       0.9650
```

Data for Exercise 14.44

Intersection	Before Camera	After Camera
1	3.60	1.36
2	0.27	0
3	0.29	0
4	4.55	1.79
5	2.60	2.04
6	2.29	3.14
7	2.40	2.72
8	0.73	0.24
9	3.15	1.57
10	3.21	0.43
11	0.88	0.28
12	1.35	1.09
13	7.35	4.92

Based on Virginia Transportation Research Council, "Research Report: The Impact of Red Light Cameras (Photo-Red Enforcement) on Crashes in Virginia," June 2007.

14.45 Reading comprehension strategies of elementary school children. An investigation of the reading comprehension strategies employed by good and average elementary school readers was the topic of research published in *The Reading Matrix* (April 2004). Both good and average readers were recruited on the basis of their scores on a midterm language test. Each group was evaluated on how often its members employed each of eight different reading strategies. The accompanying table (saved in the **READSTRAT** file) gives the proportion of times the reading group used each strategy (called the Factor Specificity Index, or FSI score). The researchers conducted a Wilcoxon signed rank test to compare the FSI score distributions of good and average readers.

	FSI Scores	
Strategy	Good Readers	Average Readers
Word meaning	.38	.32
Words in context	.29	.25
Literal comprehension	.42	.25
Draw inference from single string	.60	.26
Draw inference from multiple string	.45	.31
Interpretation of metaphor	.32	.14
Find salient or main idea	.21	.03
Form judgment	.73	.80

Based on Ahmed, S., and Asraf, R. M. "Making sense of text: Strategies used by good and average readers." *The Reading Matrix*, Vol. 4, No. 1, April 2004 (Table 2).

a. State H_0 and H_a for the desired test of hypothesis.

b. For each strategy, compute the difference between the FSI scores of good and average readers.

c. Rank the absolute values of the differences.

d. Calculate the value of the signed rank test statistic.

e. Find the rejection region for the test, using $\alpha = .05$.

f. Make the appropriate inference in the words of the problem.

14.46 NHTSA new car crash tests. Refer to the National Highway Traffic Safety Administration (NHTSA) new-car crash test data saved in the **CRASH** file. In Exercise 9.42 (p. 437), you compared the chest injury ratings of drivers and front-seat passengers by using the Student's *t*-procedure for matched pairs. Suppose you want to make the comparison for only those cars which have a driver's rating of five stars (the highest rating). The data for these 18 cars are listed in the accompanying table and saved in the **CRASH** file. Now consider analyzing the data by using the Wilcoxon signed rank test.

	Chest Injury Rating			Chest Injury Rating	
Car	Driver	Passenger	Car	Driver	Passenger
1	42	35	10	36	37
2	42	35	11	36	37
3	34	45	12	43	58
4	34	45	13	40	42
5	45	45	14	43	58
6	40	42	15	37	41
7	42	46	16	37	41
8	43	58	17	44	57
9	45	43	18	42	42

a. State the null and alternative hypotheses.

b. Use a statistical software package to find the signed rank test statistic.

c. Give the rejection region for the test, using $\alpha = .01$.

d. State the conclusion in practical terms. Report the *p*-value of the test.

Applying the Concepts—Intermediate

14.47 Ethical sensitivity of teachers towards racial intolerance. Refer to the *Journal of Moral Education* (March 2010) study of the effectiveness of a program to encourage teachers to embrace racial tolerance, Exercise 9.46 (p. 438). Recall that the level of racial tolerance was measured for each teacher before (pretest) and after (posttest) the teachers participated in an all-day workshop on cultural competence. The original sample included 238 high school teachers. The table below lists the pretest and posttest scores for a smaller sample of 10 high school teachers. These data are saved in the **TOLERANCE** file. The researchers conducted a paired-difference test to gauge the effectiveness of the program. Use the smaller sample to conduct the appropriate nonparametric test at $\alpha = .01$. What do you conclude?

Teacher	Pretest	Posttest
1	53	74
2	73	80
3	70	94
4	72	78
5	77	78
6	81	84
7	73	71
8	87	88
9	61	63
10	76	83

14.48 Sea turtles and beach nourishment. According to the National Oceanic and Atmospheric Administration's Office of Protected Species, sea turtle nesting rates have declined in all parts of the southeastern United States over the past 10 years. Environmentalists theorize that beach nourishment may improve the nesting rates of these turtles. (Beach nourishment involves replacing the sand on the beach in order to extend the high-water line seaward.) A study was undertaken to investigate the effect of beach nourishment on sea turtle nesting rates in Florida (Aubry Hershorin, unpublished doctoral dissertation, University of Florida, 2010). For one part of the study, eight beach zones were sampled in Jacksonville, Florida. Each beach zone was nourished by the Florida Fish and Wildlife Conservation Commission between 2000 and 2008. Nesting densitites (measured as nests per linear meter) were recorded both before and after nourishing at each of the eight beach zones. The data (saved in the **NESTDEN** file) are listed in the following table. Conduct a Wilcoxon signed rank test to compare the sea turtle nesting densities before and after beach nourishing. Use $\alpha = .05$.

Beach Zone	Before Nourishing	After Nourishing
401	0	0.003595
402	0.001456	0.007278
403	0	0.003297
404	0.002868	0.003824
405	0	0.002198
406	0	0.000898
407	0.000626	0
408	0	0

14.49 Concrete pavement response to temperature. Civil engineers at West Virginia University have developed a three-dimensional model to predict the response of jointed concrete pavement to temperature variations (*The International Journal of Pavement Engineering*, Sept. 2004). To validate the model, its predictions were compared with field measurements of key concrete stress variables taken at a newly constructed highway. One variable measured was slab top transverse strain (i.e., change in length per unit length per unit time) at a distance of 1 meter from the longitudinal joint. The 5-hour changes (8:20 P.M. to 1:20 A.M.) in slab top transverse strain for six days are listed in the accompanying table and saved in the **SLABSTRAIN** file. Analyze the data, using a nonparametric test. Is there a shift in the change in transverse strain distributions between field measurements and the model? Test, using $\alpha = .05$.

		Change in Transverse Strain	
Day	Change in Temperature (°C)	Field Measurement	3D Model
Oct. 24	−6.3	−58	−52
Dec. 3	13.2	69	59
Dec. 15	3.3	35	32
Feb. 2	−14.8	−32	−24
Mar. 25	1.7	−40	−39
May. 24	−.2	−83	−71

Based on Shoukry, S., William, G., and Riad, M. "Validation of 3DFE model of jointed concrete pavement response to temperature variations." *International Journal of Pavement Engineering*, Vol. 5, No. 3, Sept. 2004 (Table IV).

14.50 Neurological impairment of POWs. Eleven prisoners of war during the war in Croatia were evaluated for neurological impairment after their release from a Serbian detention camp (*Collegium Antropologicum*, June 1997). All 11 experienced blows to the head and neck and/or loss of consciousness during imprisonment. Neurological impairment was assessed by measuring the amplitude of the visual evoked potential (VEP) in both eyes at two points in time: 157 days and 379 days after their release. (The higher the VEP value, the greater the neurological impairment.) The data on the 11 POWs are shown in the accompanying table and saved in the **POWVEP** file. Determine whether the VEP measurements of POWs 379 days after their release tend to be greater than the VEP measurements of POWs 157 days after their release. Test, using $\alpha = .05$.

POW	157 Days after Release	379 Days after Release
1	2.46	3.73
2	4.11	5.46
3	3.93	7.04
4	4.51	4.73
5	4.96	4.71
6	4.42	6.19
7	1.02	1.42
8	4.30	8.70
9	7.56	7.37
10	7.07	8.46
11	8.00	7.16

Based on Vrca, A., et al. "The use of visual evoked potentials to follow-up prisoners of war after release from detention camps." *Collegium Antropologicum,* Vol. 21, No. 1, June 1997, p. 232. (Data simulated from information provided in Table 3.)

14.51 Treatment for tendon pain. Refer to the *British Journal of Sports Medicine* (Feb. 1, 2004) study of chronic Achilles tendon pain, presented Exercise 10.68 (p. 516). Recall that each in a sample of 25 patients with chronic Achilles tendinosis was treated with heavy-load eccentric calf muscle training. Tendon thickness (in millimeters) was measured both before and following the treatment of each patient. The experimental data are reproduced in the next table and saved in the **TENDON** file. Use a nonparametric test to determine whether the treatment for tendonitis tends to reduce the thickness of tendons. Test using $\alpha = .10$.

Patient	Before Thickness *(millimeters)*	After Thickness *(millimeters)*
1	11.0	11.5
2	4.0	6.4
3	6.3	6.1
4	12.0	10.0
5	18.2	14.7
6	9.2	7.3
7	7.5	6.1
8	7.1	6.4
9	7.2	5.7
10	6.7	6.5
11	14.2	13.2
12	7.3	7.5
13	9.7	7.4
14	9.5	7.2
15	5.6	6.3
16	8.7	6.0
17	6.7	7.3
18	10.2	7.0

Data for Exercise 14.51 (*continued*)

Patient	Before Thickness (*millimeters*)	After Thickness (*millimeters*)
19	6.6	5.3
20	11.2	9.0
21	8.6	6.6
22	6.1	6.3
23	10.3	7.2
24	7.0	7.2
25	12.0	8.0

Based on Ohberg, L., et al. "Eccentric training in patients with chronic Achilles tendinosis: Normalized tendon structure and decreased thickness at follow up." *British Journal of Sports Medicine*, Vol. 38, No. 1, Feb. 1, 2004 (Table 2).

Applying the Concepts—Advanced

14.52 Bowler's hot hand. Is the probability of a bowler rolling a strike higher after he has thrown four consecutive strikes? An investigation into the phenomenon of a "hot hand" in bowling was published in *The American Statistician* (Feb. 2004). Frame-by-frame results were collected on 43 professional bowlers from the 2002–2003 Professional Bowlers Association (PBA) season. For each bowler, the researchers calculated the proportion of strikes rolled after bowling four consecutive strikes and the proportion after bowling four consecutive nonstrikes. The data on 4 of the 43 bowlers, saved in the **HOTBOWLER** file, are shown in the following table.

	Proportion of Strikes	
Bowler	After Four Strikes	After Four Nonstrikes
Paul Fleming	.683	.432
Bryon Smith	.684	.400
Mike DeVaney	.632	.421
Dave D'Entremont	.610	.529

Source: Dorsey-Palmateer, R., and Smith, G. "Bowlers' hot hands." *American Statistician,* Vol. 58, No. 1, Feb. 2004 (Table 3). Reprinted with permission from *The American Statistician*. Copyright 2004 by the American Statistical Association. All rights reserved.

a. Do the data on the sample of four bowlers provide support for the "hot hand" theory in bowling? Explain.

b. When the data on all 43 bowlers are used, the p-value for the hypothesis test is approximately 0. Interpret this result.

14.5 Comparing Three or More Populations: Completely Randomized Design

In Chapter 10, we used an analysis of variance and the F-test to compare the means of k populations (treatments) on the basis of random sampling from populations that were normally distributed with a common variance σ^2. We now present a nonparametric technique for comparing the populations—the **Kruskal-Wallis H-test**–that requires no assumptions concerning the population probability distributions.

Suppose a health administrator wants to compare the unoccupied bed space for three hospitals located in the same city. She randomly selects 10 different days from the records of each hospital and lists the number of unoccupied beds for each day. (See Table 14.7.) Because the number of unoccupied beds per day may occasionally be quite large, it is conceivable that the population distributions of data may be skewed to the right and that this type of data may not satisfy the assumptions necessary for a parametric comparison of the population means. We therefore use a nonparametric analysis and base our comparison on the rank sums for the three sets of sample data. Just as with two independent samples (Section 14.3), the ranks are computed for each observation according to the relative magnitude of the measurements *when the data for all the samples are combined.* (See Table 14.7.) Ties are treated as they were for the Wilcoxon rank sum and signed rank tests, by assigning the average value of the ranks to each of the tied observations.

Table 14.7 Number of Available Beds

Hospital 1		Hospital 2		Hospital 3	
Beds	Rank	Beds	Rank	Beds	Rank
6	5	34	25	13	9.5
38	27	28	19	35	26
3	2	42	30	19	15
17	13	13	9.5	4	3
11	8	40	29	29	20
30	21	31	22	0	1
15	11	9	7	7	6
16	12	32	23	33	24
25	17	39	28	18	14
5	4	27	18	24	16
	$R_1 = 120$		$R_2 = 210.5$		$R_3 = 134.5$

Data Set: HOSPBEDS

We test

H_0: The probability distributions of the number of unoccupied beds are the same for all three hospitals

H_a: At least two of the three hospitals have probability distributions of the number of unoccupied beds that differ in location

If we denote the three sample rank sums by R_1, R_2, and R_3, then the test statistic is given by

$$H = \frac{12}{n(n+1)} \sum n_j(\overline{R}_j - \overline{R})^2$$

where n_j is the number of measurements in the jth sample and n is the total sample size $(n = n_1 + n_2 + \ \dots\ + n_k)$, $\overline{R}_j$ is the mean rank corresponding to sample j, and $\overline{R}$ is the mean of all the ranks [i.e., $\overline{R} = {}^1\!/_2(n+1)$]. The H-statistic measures the extent to which the k samples differ with respect to their relative ranks. Thus, $H = 0$ if all samples have the same mean rank and H becomes increasingly large as the distance between the sample mean ranks grows.

If the null hypothesis is true, the distribution of H in repeated sampling is approximately a χ^2 (chi-square) distribution. This approximation of the sampling distribution of H is adequate as long as one of the k sample sizes exceeds 5. (See the references for more detail.) The degrees of freedom corresponding to the approximate sampling distribution of H will always be $(k-1)$—one less than the number of probability distributions being compared. Because large values of H support the alternative hypothesis that the populations have different probability distributions, the rejection region for the test is located in the upper tail of the χ^2 distribution.

Example 14.4

Applying the Kruskal-Wallis Test to Compare Available Hospital Beds

Problem Consider the data in Table 14.7. Recall that a health administrator wants to compare the unoccupied bed space of the three hospitals. Apply the Kruskal-Wallis H-test to the data. What conclusion can you draw? Test using $\alpha = .05$.

Solution As stated previously, the administrator wants to test

H_0: The distributions of the number of unoccupied beds are the same for the three hospitals

H_a: At least two of the three hospitals have unoccupied bed distributions that differ in location

For the data in Table 14.7, we have $k = 3$ samples with $n_1 = n_2 = n_3 = 10$ and $n = 30$. The rank sums are $R_1 = 120$, $R_2 = 210.5$, and $R_3 = 134.5$; consequently, $\overline{R}_1 = 12.0$, $\overline{R}_2 = 21.05$, and $\overline{R}_3 = 13.45$. Also, the mean of all the ranks is $\overline{R} = (31)/2 = 15.5$. Substituting these values in the test statistic formula, we have

$$\textit{Test statistic}: H = \frac{12}{30(31)}[10(12.0 - 15.5)^2 + 10(21.05 - 15.5)^2 + 10(13.45 - 15.5)^2]$$

$$= \frac{12}{30(31)}[472.55] = 6.097$$

Now, when $k = 3$, the test statistic has a χ^2 distribution with $(k-1) = 2$ df. For $\alpha = .05$, we consult Table VII of Appendix B and find $\chi^2_{.05} = 5.99147$. Therefore,

Rejection region: $H > 5.99147$ (see Figure 14.10)

Conclusion: Because $H = 6.097$ exceeds the critical value of 5.99147, we reject the null hypothesis and conclude that at least one of the three hospitals has a distribution of unoccupied beds that is shifted above the distributions for the other hospitals. That is, at least one of the hospitals tends to have a larger number of unoccupied beds than the others.

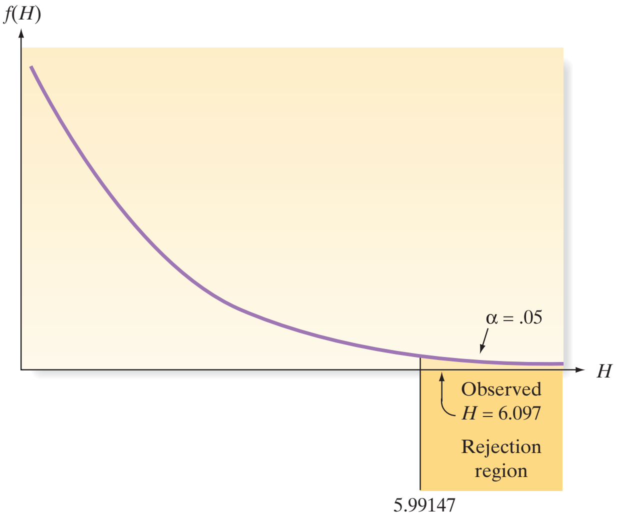
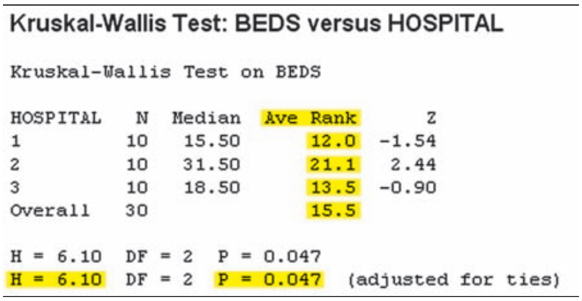


Figure 14.10
Rejection region for the comparison of three probability distributions

Look Back The same conclusion can be reached from a computer printout of the analysis. The mean ranks, test statistic, and p-value of the nonparametric test are highlighted on the MINITAB printout shown in Figure 14.11. Because $\alpha = .05$ exceeds $p\text{-value} = .047$, there is sufficient evidence to reject H_0.



```
Kruskal-Wallis Test: BEDS versus HOSPITAL

Kruskal-Wallis Test on BEDS

HOSPITAL    N  Median  Ave Rank      Z
1          10   15.50      12.0  -1.54
2          10   31.50      21.1   2.44
3          10   18.50      13.5  -0.90
Overall    30              15.5

H = 6.10  DF = 2  P = 0.047
H = 6.10  DF = 2  P = 0.047  (adjusted for ties)
```

Figure 14.11
MINITAB Kruskal-Wallis test comparing three hospitals

Now Work Exercise 14.56

The Kruskal-Wallis H-test for comparing more than two probability distributions is summarized in the next box. Note that we can use the Wilcoxon rank sum test of Section 14.3 to compare a pair of populations (selected a priori) if the Kruskal–Wallis H-test supports the alternative hypothesis that at least two of the probability distributions differ.*

Kruskal-Wallis *H*-Test for Comparing *k* Probability Distributions

H_0: The k probability distributions are identical

H_a: At least two of the k probability distributions differ in location

Test statistic:† $H = \dfrac{12}{n(n+1)} \sum n_j(\overline{R}_j - \overline{R})^2$

where

n_j = Number of measurements in sample j

R_j = Rank sum for sample j, where the rank of each measurement is computed according to its relative magnitude in the totality of data for the k samples

(*continued*)

*A method similar to the multiple-comparison procedure of Chapter 10 can be used to rank the treatment medians. This nonparametric multiple comparisons of medians will control the experimentwise error rate selected by the analyst. [See Daniel (1990) and Dunn (1964) for details.]

†An alternative but equivalent formula for the test statistic is $H = \dfrac{12}{n(n+1)} \sum \dfrac{\mathrm{R}_\mathrm{j}^2}{\mathrm{n}_\mathrm{j}} - 3(\mathrm{n} + 1)$.

$\overline{R}_j = R_j/n_j =$ Mean rank sum for the j^{th} sample

$\overline{R} =$ Mean of all ranks $= (n + 1)/2$

$n =$ Total sample size $= n_1 + n_2 + \cdots + n_k$

Rejection region: $H > \chi^2_\alpha$ with $(k - 1)$ degrees of freedom.

Ties: Assign tied measurements the average of the ranks they would receive if they were unequal, but occurred in successive order. For example, if the third-ranked and fourth-ranked measurements are tied, assign both a rank of $(3 + 4)/2 = 3.5$. The number of ties should be small relative to the total number of observations.

Conditions Required for the Valid Application of the Kruskal-Wallis Test

1. The k samples are random and independent.
2. There are five or more measurements in each sample.
3. The k probability distributions from which the samples are drawn are continuous.

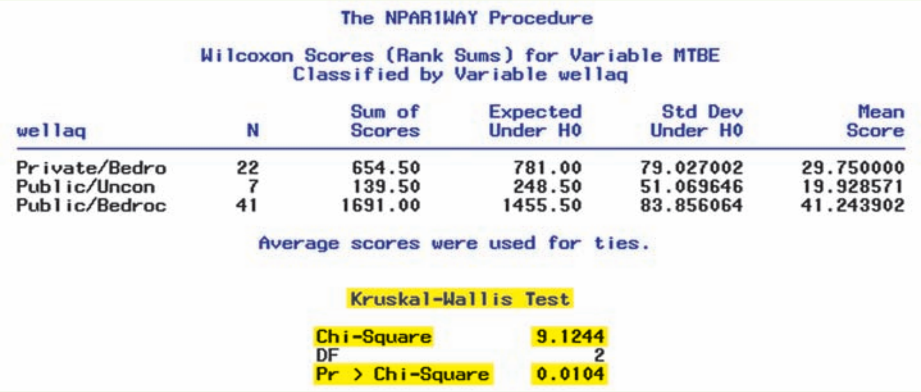
Statistics IN Action Revisited

Comparing the MTBE Levels of Different Types of Groundwater Wells (continued)

In the previous *Statistics in Action Revisited* (p. 14-14), we demonstrated the use of Wilcoxon rank sum tests to compare the MTBE distributions of public and private groundwater wells and of bedrock and unconsolidated aquifers. The environmental researchers also investigated how the combination of well class and aquifer affected the MTBE levels of the 70 wells in the **MTBE** file that had detectable levels of MTBE. Although there are four possible combinations of well class and aquifer, data were available for only three: Private/bedrock, Public/bedrock, and Public/unconsolidated.

The distributions of MTBE levels for these three groups of wells were compared with the use of the Kruskal-Wallis nonparametric test for independent samples. The SAS printout for the analysis is shown in Figure SIA14.4. The test statistic is $H = 9.12$ and the p-value is .0104 (highlighted). At $\alpha = .05$, there is sufficient evidence to indicate differences in the distributions of MTBE levels of the three class-aquifer types. (However, at $\alpha = .01$, no significant differences are found.) On the basis of the mean rank sum scores shown on the printout, it appears that public wells with bedrock aquifers have the highest levels of MTBE contamination.

Data Set: MTBE



```
                     The NPAR1WAY Procedure

          Wilcoxon Scores (Rank Sums) for Variable MTBE
                  Classified by Variable wellaq

                          Sum of      Expected       Std Dev          Mean
wellaq           N        Scores      Under H0      Under H0         Score

Private/Bedro   22        654.50        781.00     79.027002     29.750000
Public/Uncon     7        139.50        248.50     51.069646     19.928571
Public/Bedroc   41       1691.00       1455.50     83.856064     41.243902

                 Average scores were used for ties.


                       Kruskal-Wallis Test

                   Chi-Square          9.1244
                   DF                       2
                   Pr > Chi-Square     0.0104
```

Figure SIA14.4

SAS Kruskal-Wallis test for comparing MTBE levels of wells

Exercises 14.53–14.66

Understanding the Principles

14.53 Under what circumstances does the χ^2 distribution provide an appropriate characterization of the sampling distribution of the Kruskal-Wallis *H*-statistic?

14.54 Which of the following results would lead you to conclude that the treatments in a balanced completely randomized design have distributions that differ in location?

a. The rank sums for all treatments are about equal.

b. The rank sum for one treatment is much larger than the rank sum for all other treatments.

Learning the Mechanics

14.55 Suppose you want to use the Kruskal-Wallis *H*-test to compare the probability distributions of three populations. The following data (saved in the **LM14_55** file) represent independent random samples selected from the three populations:

I:	34	56	65	59	82	70	45
II:	24	18	27	41	34	42	33
III:	72	101	91	76	80	75	

a. What experimental design was used?

b. Specify the null and alternative hypotheses you would test.

c. Specify the rejection region you would use for your hypothesis test at $\alpha = .01$.

d. Conduct the test at $\alpha = .01$.

14.56 Data were collected from three populations—*A, B,* and *C,*—by means of a completely randomized design. The following describes the sample data:

$$n_A = n_B = n_C = 15$$

$$R_A = 235 \quad R_B = 439 \quad R_C = 361$$

a. Specify the null and alternative hypotheses that should be used in conducting a test of hypothesis to determine whether the probability distributions of populations A, B, and C differ in location.

b. Conduct the test of part **a.** Use $\alpha = .05$.

c. What is the approximate *p*-value of the test of part **b**?

Applying the Concepts—Basic

14.57 Dog behavior on walks. Researchers at the School of Veterinary Science, University of Liverpool (United Kingdom), conducted a field study to investigate the frequency and nature of interactions of pet dogs with other dogs (*Applied Animal Behaviour Science*, June 2010). The behavior of pet dogs being walked by their owners was observed at several popular dog-walking areas. When a pet dog encountered one or more other dogs on the walk, the length of the interaction was recorded (in seconds). The interaction episodes were classified into three groups according to the number of dogs encountered (1 dog, 2 dogs, or at least 3 dogs). The researchers compared the distributions of the interaction lengths for the three groups using the Kruskal-Wallis *H*-test.

a. Set up the null and alternative hypotheses for the test.

b. At $\alpha = .05$, what is the rejection region?

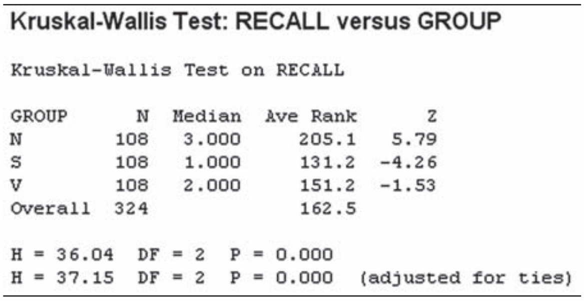
c. The test statistic was reported as $H = 1.1$, with an associated *p*-value of .60. What conclusion can you draw from these results?

14.58 Study of recall of TV commercials. Refer to the *Journal of Applied Psychology* (June 2002) study of the recall of the content of television commercials, presented in Exercise 10.33 (p. 495). In a designed experiment, 324 adults were randomly assigned to one of three viewer groups: (1) Watch a TV program with a violent content code (V) rating, (2) watch a show with a sex content code (S) rating, and, (3) watch a neutral TV program. The number of brand names recalled in the commercial messages was recorded for each participant, and the data are saved in the **TVADRECALL** file.

a. Give the null and alternative hypotheses for a Kruskal-Wallis test applied to the data.

b. The results of the nonparametric test are shown below in the MINITAB printout. Locate the test statistic and *p*-value on the printout.

c. Interpret the results of part **b,** using $\alpha = .01$. What can the researchers conclude about the three groups of TV ad viewers?



Kruskal-Wallis Test: RECALL versus GROUP

```
Kruskal-Wallis Test on RECALL

GROUP       N  Median  Ave Rank      Z
N         108   3.000     205.1   5.79
S         108   1.000     131.2  -4.26
V         108   2.000     151.2  -1.53
Overall   324             162.5

H = 36.04  DF = 2  P = 0.000
H = 37.15  DF = 2  P = 0.000  (adjusted for ties)
```

14.59 Effect of scopolamine on memory. Refer to the *Behavioral Neuroscience* (Feb. 2004) study of the drug scopolamine's effects on memory for word-pair association presented in Exercise 10.56 (p. 505). Recall that a completely randomized design with three groups was used: Group 1 subjects were injected with scopolamine, group 2 subjects were injected with a placebo, and group 3 subjects were not given any drug. The response variable was number of word pairs recalled. The data on all 28 subjects are reproduced in the following table and saved in the **SCOPOLAMINE** file.

Group 1 (Scopolamine):	5	8	8	6	6	6	6	8	6	4	5	6
Group 2 (Placebo):	8	10	12	10	9	7	9	10				
Group 3 (No drug):	8	9	11	12	11	10	12	12				

a. Rank the data for all 28 observations from smallest to largest.

b. Sum the ranks of the observations from group 1.

c. Sum the ranks of the observations from group 2.

d. Sum the ranks of the observations from group 3.

e. Use the rank sums from parts **b–d** to compute the Kruskal-Wallis H-statistic.

f. Carry out the Kruskal-Wallis nonparametric test (at $\alpha = .05$) to compare the distributions of number of word pairs recalled for the three groups.

g. Recall from Exercise 10.56 that the researchers theorized that group 1 subjects would tend to recall the fewest number of words. Use the Wilcoxon rank sum test to compare the word recall distributions of group 1 and group 2. (Use $\alpha = .05$.)

14.60 Commercial eggs produced from different housing systems. Refer to the *Food Chemistry* (Vol. 106, 2008) study of commercial eggs produced from different housing systems for chickens, Exercise 10.117 (p. 543). Recall that the four housing systems investigated were (1) cage, (2) barn, (3) free range, and (4) organic. Twenty-eight commercial grade A eggs were randomly selected from supermarkets—10 of which were produced in cages, 6 in barns, 6 with free range, and 6 organically. A number of quantitative characteristics were measured for each egg, including penetration strength (Newtons). The data (simulated from summary statistics provided in the journal article) are given in the accompanying table and saved in the **EGGS** file.

Cage:	36.9 39.2 40.2 33.0 39.0 36.6 37.5 38.1 37.8 34.9
Free:	31.5 39.7 37.8 33.5 39.9 40.6
Barn:	40.0 37.6 39.6 40.3 38.3 40.2
Organic:	34.5 36.8 32.6 38.5 40.2 33.2

a. Rank the observations in the data set from 1 to 28.

b. Sum the ranks of the data for each housing system.

c. Use the rank sums to find the Kruskal-Wallis test statistic.

d. Based on the result, part **c**, what do you infer about the strength distributions of the four housing systems?

Applying the Concepts—Intermediate

14.61 Relieving pain with hypnosis. Rehabilitation medicine researchers at the University of Washington investigated whether virtual reality hypnosis can relieve pain in trauma patients (*International Journal of Clinical and Experimental Hypnosis*, Vol. 58, 2010). Study participants were 20 patients treated at a major Level 1 trauma center. The patients were randomly assigned to one of three treatment groups: (1) VRH—virtual reality hypnosis with posthypnotic suggestions for pain reduction, (2) VRD—virtual reality distraction from pain without hypnotic suggestions for pain reduction, and (3) CONTROL—no virtual reality hypnosis, but standard care. Pain intensity was measured (on a 100-point scale) prior to treatment and one hour after treatment. The differences in pain intensity levels (before minus after) are listed in the next table (top of the next column) and saved in the **PAINHYP** file.

a. Conduct a nonparametric test to determine whether the distribution of differences in pain intensity levels differs for the three treatments. Test using $\alpha = .05$. What do you conclude?

b. Combine the patients in the VRD and CONTROL groups into a single treatment group (called nonposthypnotic suggestion). Compare the VRH treatment patients to the patients in this new group using the appropriate nonparametric test (using $\alpha = .05$). What do you conclude?

VRH	VRD	CONTROL
–20	–12	51
–56	63	21
–34	12	8
0	–7	0
16	29	4
0		
–14		
–7		
–44		
43		
–11		

14.62 Energy expenditure of laughter. Refer to the *International Journal of Obesity* (Jan. 2007) study of the physiological changes that accompany laughter, presented in Exercise 8.30 (p. 365). Recall that pairs of subjects watched film clips designed to evoke laughter. In addition to heart rate, the researchers measured the duration of laughter (seconds per minute) and the energy expenditure (kilojoules per minute) of each pair during the laughing period. The subject pairs were then divided into four groups (quartiles)—0–5, 6–10, 10–20, and more than 20 seconds per minute—on the basis of duration of laughter. The energy expenditure values for the 45 subject pairs in the study are shown in the next table and saved in the **LAUGHTER** file. (The data are simulated on the basis of reported summary statistics.) The researchers compared the energy expenditure distributions across the four laughter duration groups by means of the Kruskal-Wallis test.

0–5 sec/min	6–10 sec/min	10–20 sec/min	> 20 sec/min
0.10	0.43	0.11	0.88
0.94	0.46	0.62	0.52
0.44	1.01	0.11	0.70
0.07	1.11	1.71	2.50
0.10	0.13	0.60	1.12
0.08	0.02	0.08	0.70
0.06	0.36	0.24	1.20
0.05	0.18	0.58	0.56
0.01	0.40	0.19	0.36
0.13	0.09	1.09	0.22
0.04	1.29	2.09	0.50
	0.50		

Based on Buchowski, M. S., et al. "Energy expenditure of genuine laughter." *International Journal of Obesity*, Vol. 31, No. 1, January 2007 (Figure 4).

a. State H_0 and H_a for the desired test of hypothesis.

b. Find the rejection region for the test, using $\alpha = .10$.

c. Compute the value of the test statistic.

d. On the basis of the results from parts **b** and **c,** what is the appropriate conclusion?

e. Compare the 0–5 and > 20 quartile groups, using the Wilcoxon rank sum test. What do you infer about the relationship between energy expenditure and duration of laughter?

f. Demonstrate why the researchers employed a nonparametric test on the data.

14.63 Restoring self-control when intoxicated. Refer to the *Experimental and Clinical Psychopharmacology* (February 2005) study of self-control when intoxicated, presented in Exercise 10.34 (p. 495). After memorizing two lists of words (20 words on a green list and 20 words on a red list), students were randomly assigned to one of four different treatment groups. Students in Group A received two alcoholic drinks. Students in Group AC had caffeine powder dissolved in their alcoholic drinks. Group AR also received two alcoholic drinks, but received a monetary award for correct responses. Students in Group P (the placebo group) were told that they would receive alcohol, but instead received two drinks containing a carbonated beverage (with a few drops of alcohol on the surface to provide an alcoholic scent). After consuming their drinks and resting for 25 minutes, the students performed a word completion task. Their scores (simulated on the basis of summary information from the article) are reported in the accompanying table and saved in the **DRINKERS** file. (*Note*: A score represents the difference between the proportion of correct responses on the green list of words and the proportion of incorrect responses on the red list of words.) Compare the task score distributions of the four groups, using an appropriate nonparametric test at $\alpha = .05$. What can you infer about the four groups of students?

AR	AC	A	P
.51	.50	.16	.58
.58	.30	.10	.12
.52	.47	.20	.62
.47	.36	.29	.43
.61	.39	−.14	.26
.00	.22	.18	.50
.32	.20	−.35	.44
.53	.21	.31	.20
.50	.15	.16	.42
.46	.10	.04	.43
.34	.02	−.25	.40

Based on Grattan-Miscio, K. E., and Vogel-Sprott, M. "Alcohol, intentional control, and inappropriate behavior: Regulation by caffeine or an incentive." *Experimental and Clinical Psychopharmacology*, Vol. 13, No. 1, February 2005 (Table 1).

14.64 Estimating the age of glacial drifts. Refer to the *American Journal of Science* (Jan. 2005) study of the chemical makeup of buried tills (glacial drifts) in Wisconsin, presented in Exercise 10.37 (p. 496). Recall that till specimens were obtained from five different boreholes (labeled UMRB-1, UMRB-2, UMRB-3, SWRA, and SD), and the ratio of aluminum to beryllium was measured for each specimen. The data are reproduced in the accompanying table and saved in the **TILLRATIO** file. Conduct a nonparametric analysis of variance of the data, using $\alpha = .10$. Interpret the results.

UMRB-1:	3.75	4.05	3.81	3.23	3.13	3.30	3.21
UMRB-2:	3.32	4.09	3.90	5.06	3.85	3.88	
UMRB-3:	4.06	4.56	3.60	3.27	4.09	3.38	3.37
SWRA:	2.73	2.95	2.25				
SD:	2.73	2.55	3.06				

Based on *American Journal of Science*, Vol. 305, No. 1, Jan. 2005, p. 16 (Table 2).

14.65 The "name game." Refer to the *Journal of Experimental Psychology–Applied* (June 2000) study of different methods of learning names, presented in Exercise 10.36 (p. 496). Recall that three groups of students used different methods to learn the names of the other students in their group. Group 1 used the "simple name game," Group 2 used the "elaborate name game," and Group 3 used "pairwise introductions." The tables (saved in the **NAMEGAME** file) lists the percentage of names recalled (after one year) for each student respondent.

Simple Name Game													
24	43	38	65	35	15	44	44	18	27	0	38	50	31
7	46	33	31	0	29	0	0	52	0	29	42	39	26
51	0	42	20	37	51	0	30	43	30	99	39	35	19
24	34	3	60	0	29	40	40						

Elaborate Name Game													
39	71	9	86	26	45	0	38	5	53	29	0	62	0
1	35	10	6	33	48	9	26	83	33	12	5	0	0
25	36	39	1	37	2	13	26	7	35	3	8	55	50

Pairwise Intro													
5	21	22	3	32	29	32	0	4	41	0	27	5	9
66	54	1	15	0	26	1	30	2	13	0	2	17	14
5	29	0	45	35	7	11	4	9	23	4	0	8	2
18	0	5	21	14									

Based on Morris, P. E., and Fritz, C. O. "The name game: Using retrieval practice to improve the learning of names." *Journal of Experimental Psychology—Applied*, Vol. 6, No. 2, June 2000 (data simulated from Figure 1).

a. Consider an analysis-of-variance F-test to determine whether the mean percentages of names recalled differ for the three name-retrieval methods. Demonstrate that the ANOVA assumptions are likely to be violated.

b. Use a nonparametric test to compare the distributions of the percentages of names recalled for the three name-retrieval methods. Use $\alpha = .05$.

14.66 Is honey a cough remedy? Refer to the *Archives of Pediatrics and Adolescent Medicine* (Dec. 2007) study of honey as a children's cough remedy, Exercise 14.29 (p. 14-17). In addition to the two experimental groups of children with an upper respiratory tract infection—one which was given a dosage of dextromethorphan (DM) and the other a similar dose of honey—a third group of children received no dosage (control group). The cough symptoms improvement scores for the children are reproduced in the accompanying table and saved in the **HONEYCOUGH** file. Conduct a nonparametric test to compare the distributions of cough improvement scores for the three dosage groups. Use $\alpha = .01$.

Honey	12	11	15	11	10	13	10	4	15	16	9	14	10	6
Dosage:	10	8	11	12	12	8	12	9	11	15	10	15	9	13
	8	12	10	8	9	5	12							
DM	4	6	9	4	7	7	7	9	12	10	11	6	3	4
Dosage:	9	12	7	6	8	12	12	4	12	13	7	10	13	9
	4	4	10	15	9									
No Dosage	5	8	6	1	0	8	12	8	7	7	1	6	7	7
(Control):	12	7	9	7	9	5	11	9	5	6	8	8	6	7
	10	9	4	8	7	3	1	4	3					

Based on Paul, I. M., et al. "Effect of honey, dextromethorphan, and no treatment on nocturnal cough and sleep quality for coughing children and their parents." *Archives of Pediatrics and Adolescent Medicine*, Vol. 161, No. 12, Dec. 2007 (data simulated).

14.6 Comparing Three or More Populations: Randomized Block Design

In Section 10.4 we employed an analysis of variance to compare k population (treatment) means when the data were collected using a randomized block design. The *Friedman* F_r*-test* provides another method for testing to detect a shift in location of a set of k populations that have the same spread (or, scale).* Like other nonparametric tests, it requires no assumptions concerning the nature of the populations other than the capacity of individual observations to be ranked.

Consider the problem of comparing the reaction times of subjects under the influence of different drugs produced by a pharmaceutical firm. When the effect of a drug is short-lived (there is no carryover effect) and when the drug effect varies greatly from person to person, it may be useful to employ a *randomized block design*. Using the subjects as blocks, we would hope to eliminate the variability among subjects and thereby increase the amount of information in the experiment. Suppose that three drugs, A, B, and C, are to be compared using a randomized block design. Each of the three drugs is administered to the *same subject*, with suitable time lags between the three doses. The order in which the drugs are administered is randomly determined for each subject. Thus, one drug would be administered to a subject and its reaction time would be noted; then, after a sufficient length of time, the second drug administered; etc.

Suppose six subjects are chosen and that the reaction times for each drug are as shown in Table 14.8. To compare the three drugs, we rank the observations within each subject (block) and then compute the rank sums for each of the drugs (treatments). Tied observations within blocks are handled in the usual manner by assigning the average value of the ranks to each of the tied observations.

Table 14.8 Reaction Time for Three Drugs

Subject	Drug A	Rank	Drug B	Rank	Drug C	Rank
1	1.21	1	1.48	2	1.56	3
2	1.63	1	1.85	2	2.01	3
3	1.42	1	2.06	3	1.70	2
4	2.43	2	1.98	1	2.64	3
5	1.16	1	1.27	2	1.48	3
6	1.94	1	2.44	2	2.81	3
		$R_1 = 7$		$R_2 = 12$		$R_3 = 17$

Data Set: REACTION2

The null and alternative hypotheses are

H_0: Populations of reaction times are identically distributed for all 3 drugs

H_a: At least two of the drugs have probability distributions of reaction times that differ in location

The **Friedman F_r-statistic,** which is based on the rank sums for each treatments, measures the extent to which the k samples differ with respect to their relative ranks within the blocks. The formula for F_r is

$$F_r = \frac{12b}{k(k+1)}\sum(\overline{R}_j - \overline{R})^2$$

where b is the number of blocks, k is the number of treatments, $\overline{R}_j$ is the mean rank corresponding to treatment j, and $\overline{R}$ is the mean of all the ranks [(i.e., $\overline{R} = \frac{1}{2}(k+1)$]. You can see that the F_r-statistic is 0 if all treatments have the same mean rank and becomes increasingly large as the distance between the sample mean ranks grows.

As for the Kruskal-Wallis H-statistic, the Friedman F_r-statistic has approximately a χ^2 sampling distribution with $(k - 1)$ degrees of freedom. Empirical results show the

*The Friedman F_r-test was developed by the Nobel Prize–winning economist Milton Friedman.

approximation to be adequate if either b or k exceeds 5. The Friedman F_r-test for a randomized block design is summarized in the next box.

Friedman F_r-Test for a Randomized Block Design

H_0: The probability distributions for the k treatments are identical

H_a: At least two of the probability distributions differ in location*

$$\textit{Test statistic*}: F_r = \frac{12b}{k(k+1)}\sum(\overline{R}_j - \overline{R})^2$$

where

b = Number of blocks

k = Number of treatments

R_j = Rank sum of the jth treatment, where the rank of each measurement is *computed relative to its position within its own block*

Rejection region: $F_r > \chi^2_\alpha$ with $(k-1)$ degrees of freedom

Ties: Assign tied measurements within a block the average of the ranks they would receive if they were unequal but occurred in successive order. For example, if the third-ranked and fourth-ranked measurements are tied, assign each a rank of $(3+4)/2 = 3.5$. The number of ties should be small relative to the total number of observations.

Conditions Required for a Valid Friedman F_r-Test

1. The treatments are randomly assigned to experimental units within the blocks.
2. The measurements can be ranked within blocks.
3. The k probability distributions from which the samples within each block are drawn are continuous.

Example 14.5

Applying the Friedman Test to Compare Drug Reaction Times

Problem Consider the data in Table 14.8. Recall that a pharmaceutical firm wants to compare the reaction times of subjects under the influence of three different drugs that it produces. Apply the Friedman F_r-test to the data. What conclusion can you draw? Test using $\alpha = .05$.

Solution As stated previously, the firm wants to test

H_0: The population distributions of reaction times are identical for the three drugs

H_a: At least two of the three drugs have reaction time distributions that differ in location

For the data in Table 14.8, we have $k = 3$ treatments (drugs) and $b = 6$ blocks (subjects). The treatment rank sums are $R_1 = 7$, $R_2 = 12$, and $R_3 = 17$; consequently, $\overline{R}_1 = 7/6 = 1.167$,$\overline{R}_2 = 12/6 = 2.0$, and $\overline{R}_3 = 17/6 = 2.833$. Also, the mean of all the ranks is $\overline{R} = (3+1)/2 = 2.0$. Substituting these values in the test statistic formula, we have

$$\textit{Test statistic}: F_r = \frac{12(6)}{(3)(4)}[(1.167 - 2.0)^2 + (2.0 - 2.0)^2 + (2.833 - 2.0)^2]$$
$$= 6(1.388) = 8.33$$

*An alternative but equivalent formula for the test statistic is $F_r = \frac{12}{bk(k+1)}\sum R_j^2 - 3b(k+1)$.

Now, when $k = 3$, the test statistic has a χ^2 distribution with $(k - 1) = 2$ df. For $\alpha = .05$, we consult Table VII of Appendix B and find a $\chi^2_{.05} = 5.99147$. Therefore,

Rejection region: $H > 5.99147$ (see Figure 14.12)

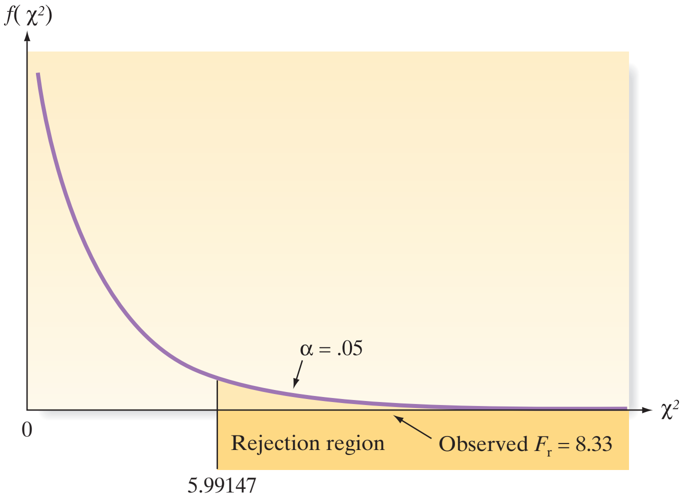
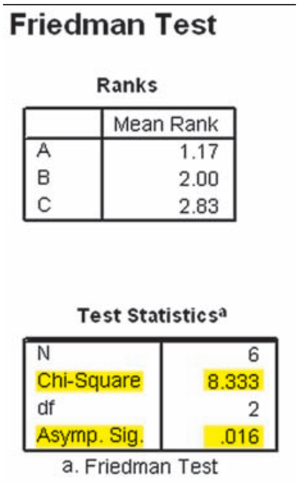


Figure 14.12
Rejection region for reaction time example



Friedman Test

Ranks

	Mean Rank
A	1.17
B	2.00
C	2.83

Test Statistics[a]

N	6
Chi-Square	8.333
df	2
Asymp. Sig.	.016

a. Friedman Test

Figure 14.13
SPSS Friedman test printout

Conclusion: Because $H = 8.33$ exceeds the critical value of 5.99, we reject the null hypothesis and conclude that at least two of the three drugs have distributions of reaction times that differ in location. That is, at least one of the drugs tends to yield reaction times that are faster than the others.

An SPSS printout of the nonparametric analysis, shown in Figure 14.13, confirms our inference. Both the test statistic and p-value are highlighted on the printout. Because p-value $= .016$ is less than our selected $\alpha = .05$, there is evidence to reject H_0.

Look Back Clearly, the assumptions for this test—that the measurements are ranked within blocks and that the number of blocks (subjects) is greater than 5—are satisfied. However, we must be sure that the treatments are randomly assigned to blocks. For the procedure to be valid, we assume that the three drugs are administered in a random order to each subject. If this were not true, the difference in the reaction times for the three drugs might be due to the order in which the drugs are given.

Now Work Exercise 14.70

Exercises 14.67–14.79

Understanding the Principles

14.67 Which of the following statements correctly describes how to rank the data in a randomized block design?

a. For each treatment, rank the data across the blocks from smallest to largest.

b. For each block, rank the data across the treatments from smallest to largest.

14.68 What conditions are required for a valid application of the Friedman F_r-test?

Learning the Mechanics

14.69 Data were collected under a randomized block design with four treatments ($A, B, C,$ and D) and $b = 6$. The following rank sums were obtained:

$$R_A = 11 \quad R_B = 21 \quad R_C = 21 \quad R_D = 7$$

a. How many blocks were used in the experimental design?

b. Specify the null and alternative hypotheses that should be used in conducting a hypothesis test to determine whether the probability distributions for at least two of the treatments differ in location.

c. Conduct the test of part **b.** Use $\alpha = .10$.

d. What is the approximate p-value of the test of part **c**?

14.70 NW Suppose you have used a randomized block design to help you compare the effectiveness of three different treatments: $A, B,$ and C. You obtained the data given in the next table (saved in the **LM14_70** file) and plan to conduct a Friedman F_r-test.

Data for Exercise 14.70

	Treatment		
Block	A	B	C
1	9	11	18
2	13	13	13
3	11	12	12
4	10	15	16
5	9	8	10
6	14	12	16
7	10	12	15

a. Specify the null and alternative hypotheses you will test.
b. Specify the rejection region for the test. Use $\alpha = .10$.
c. Conduct the test and interpret the results.

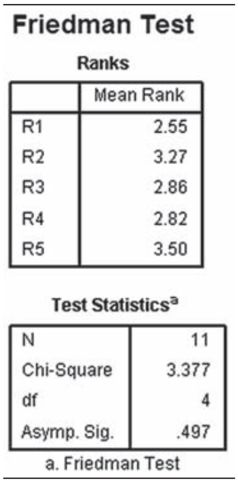
14.71 An experiment was conducted under a randomized block design with four treatments and six blocks. The ranks of the measurements within each block are shown in the accompanying table (saved in the **LM14_71** file). Use the Friedman F_r-test for a randomized block design to determine whether the data provide sufficient evidence to indicate that at least two of the treatment probability distributions differ in location. Test, using $\alpha = .05$.

	Block					
Treatment	1	2	3	4	5	6
1	3	3	2	3	2	3
2	1	1	1	2	1	1
3	4	4	3	4	4	4
4	2	2	4	1	3	2

Applying the Concepts—Basic

14.72 A new method of evaluating health care research reports. *The Open Dentistry Journal* (Vol. 4, 2010) published a study on a revised tool for assessing research reports in health care. (See Exercise 10.70, p. 517.) Recall that the assessment tool was validated on five systematic reviews (named R1, R2, R3, R4, and R5) on rheumatoid arthritis. For each review, scores on the 11 items in the assessment tool (all measured on a 4-point scale) were obtained. The data, saved in the **RAMSTAR** file, are repeated in the table below.

a. One goal of the study was to compare the distributions of scores of the five reviews. Set up the null and alternative hypothesis for this test.
b. Explain why the data should be analyzed using a nonparametric randomized block ANOVA.
c. For item #1, rank the scores for the five systematic reviews (R1, R2, R3, R4, and R5).
d. Repeat part **c** for each of the remaining items..
e. Sum the ranks for the R1 scores.
f. Repeat part **e** for the R2, R3, R4, and R5 scores.
g. Use the rank sums to calculate the Friedman F_r test statistic.
h. Find the rejection region for the test using $\alpha = .10$.
i. Formulate the appropriate conclusion for the test.
j. An SPSS printout of the analysis is shown below. Locate the p-value on the printout. Does this result confirm your conclusion in part **i**?



```
Friedman Test

   Ranks
|      | Mean Rank |
|------|-----------|
| R1   |      2.55 |
| R2   |      3.27 |
| R3   |      2.86 |
| R4   |      2.82 |
| R5   |      3.50 |

   Test Statistics^a
| N           |    11 |
| Chi-Square  | 3.377 |
| df          |     4 |
| Asymp. Sig. |  .497 |
   a. Friedman Test
```

14.73 Stress in cows prior to slaughter. Refer to the *Applied Animal Behaviour Science* (June 2010) study of stress in cows prior to slaughter, Exercise 10.71 (p. 517). In the experiment, recall that the heart rate (beats per minute) of a cow was measured at four different preslaughter phases—(1) first phase of visual contact with pen mates, (2) initial isolation from pen mates for prepping, (3) restoration of visual contact with pen mates, and (4) first contact with human prior to slaughter. Thus, a randomized block design was employed. The simulated data for eight cows are reproduced in the table (p. 14-38) and saved in the **COWSTRESS** file. Consider applying the nonparametric Friedman test to determine whether the heart rate distributions differ for cows in the four preslaughter phases. A MINITAB printout of the analysis follows the data.

a. Locate the rank sums on the printout.
b. Use the rank sums to calculate the F_r test statistic. Does the result agree with the value shown on the MINITAB printout?
c. Locate the p-value of the test on the printout.
d. Provide the appropriate conclusion in the words of the problem if $\alpha = .05$.

Data for Exercise 14.72

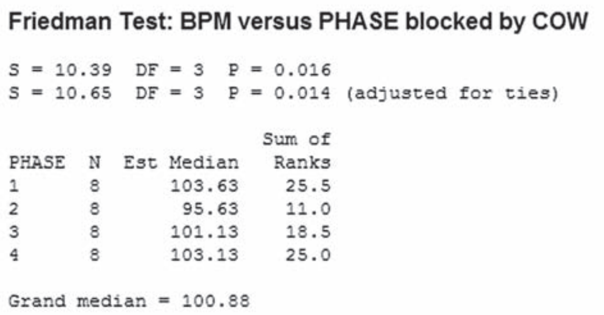
Review	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11
R1	4.0	1.0	4.0	2.0	3.5	3.5	3.5	3.5	1.0	1.0	1.0
R2	3.5	2.5	4.0	4.0	3.5	4.0	3.5	2.5	3.5	1.5	1.0
R3	4.0	4.0	3.5	4.0	1.5	2.5	3.5	3.5	2.5	1.5	1.0
R4	3.5	2.0	4.0	4.0	2.0	4.0	3.5	3.0	3.5	1.0	1.0
R5	3.5	4.0	4.0	3.0	2.5	4.0	4.0	4.0	2.5	1.0	2.5

Based on Kung, J., et al. "From systematic reviews to clinical recommendations to clinical-based health care: Validation of revised assessment of multiple systematic reviews (R-AMSTAR) for grading of clinical relevance." *The Open Dentistry Journal,* Vol. 4, 2010 (Table 2).

Data for Exercise 14.73

	Phase			
Cow	1	2	3	4
1	124	124	109	107
2	100	98	98	99
3	103	98	100	106
4	94	91	98	95
5	122	109	114	115
6	103	92	100	106
7	98	80	99	103
8	120	84	107	110

MINITAB Output for Exercise 14.73



Friedman Test: BPM versus PHASE blocked by COW

```
S = 10.39  DF = 3  P = 0.016
S = 10.65  DF = 3  P = 0.014 (adjusted for ties)

                        Sum of
PHASE  N  Est Median   Ranks
1      8      103.63    25.5
2      8       95.63    11.0
3      8      101.13    18.5
4      8      103.13    25.0

Grand median = 100.88
```

14.74 Conditions impeding farm production. A review of farmer involvement in agricultural research was presented in the *Journal of Agricultural, Biological, and Environmental Statistics* (Mar. 2001). In one study, each of six farmers ranked the level of farm production constraint imposed by five conditions: drought, pest damage, weed interference, farming costs, and labor shortage. The rankings, ranging from 1 (least severe) to 5 (most severe), and rank sums for the five conditions are listed in the table below and saved in the **FARM6** file.

a. Use the rank sums shown in the table to compute the Friedman F_r-statistic.

b. At $\alpha = .05$, find the rejection region for a test to compare the farmer opinion distributions for the five conditions.

c. Draw the proper conclusion in the words of the problem.

14.75 Impact study of distractions while driving. The consequences of performing verbal and spatial-imagery tasks while driving were studied and the results published in the *Journal of Experimental Psychology—Applied* (Mar. 2000). Twelve drivers were recruited to drive on a highway in Madrid, Spain. During the drive, each subject was asked to perform three different tasks: a verbal task (repeating words that begin with a certain letter), a spatial-imagery task (imagining letters rotated a certain way), and no mental task. Since each driver performed all three tasks, the design is a randomized block with 12 blocks (drivers) and 3 treatments (tasks). Using a computerized head-free eye-tracking system, the researchers kept track of the eye fixations of each driver on three different objects—the interior mirror, the side mirror, and the speedometer—and determined the proportion of eye fixations on the object. The researchers used the Friedman nonparametric test to compare the distributions of the eye fixation proportions for the three tasks.

a. Using $\alpha = .01$, find the rejection region for the Friedman test.

b. For the response variable Proportion of eye fixations on the interior mirror, the researchers determined the Friedman test statistic to be $F_r = 19.16$. Give the appropriate conclusion.

c. For the response variable Proportion of eye fixations on the side mirror, the researchers determined the Friedman test statistic to be $F_r = 7.80$. Give the appropriate conclusion.

d. For the response variable Proportion of eye fixations on the speedometer, the researchers determined the Friedman test statistic to be $F_r = 20.67$. Give the appropriate conclusion.

Applying the Concepts—Intermediate

14.76 "Topsy-turvy" seasons in college football. Refer to the *Chance* (Summer 2009) investigation into "topsy-turvy" college football seasons, Exercise 10.67 (p. 516). Recall that statisticians created a formula for determining a weekly "topsy-turvy" (TT) index, designed to measure the degree to which the top 25 ranked teams changed from the previous week. The greater the TT index, the greater the changes in the ranked teams. The statisticians calculated the TT index each week of the 15-week college football season for 6 recent seasons. In order to determine whether any of the 15 weeks in a season tend to be more or less topsy-turvy than others, the statisticians conducted an analysis of variance on the data using a randomized block design, where

Data for Exercise 14.74

		Condition				
		Drought	Pest Damage	Weed Interference	Farming Costs	Labor Shortage
Farmer	**1**	5	4	3	2	1
	2	5	3	4	1	2
	3	3	5	4	2	1
	4	5	4	1	2	3
	5	4	5	3	2	1
	6	5	4	3	2	1
	Rank sum	27	25	18	11	9

From Riley, J., and Fielding, W. J. "An illustrated review of some farmer participatory research techniques." *Journal of Agricultural, Biological, and Environmental Statistics*, Vol. 6, No. 1, Mar. 2001 (Table 1). Reprinted with permission from the International Biometric Society.

the 15 weeks were considered the treatments and the 6 seasons were the blocks.

a. Suppose the data (TT index values) are not normally distributed. How would this impact the ANOVA conducted in Exercise 10.67? Explain.

b. Give the null and alternative hypotheses for the Friedman test applied to the data.

c. Find the rejection region for the Friedman test using $\alpha = .01$.

d. Explain how you would calculate the Friedman test statistic for this data set.

e. Give a p-value of the test that would lead you to conclude that no one week is any more "topsy-turvy" than any other week.

14.77 Effect of massage on boxers. Refer to the *British Journal of Sports Medicine* (Apr. 2000) experiment to investigate the effect of massage on boxing performance, presented in Exercise 10.72 (p. 518) and saved in the **BOXING** file. Recall that the punching power (in newtons) of each of eight amateur boxers was measured after each of four rounds: (M1), round 1 following a pre-bout sports massage; (R1), round 1 following a prebout period of rest; (M5), round 5 following a sports massage between rounds; and (R5), round 5 following a period of rest between rounds. The data are reproduced in the table in the previous column. Use the appropriate nonparametric test to compare the punching power means of the four interventions. Compare the results with those of Exercise 10.72.

		Intervention			
		M1	R1	M5	R5
Boxer	**1**	1243	1244	1291	1262
	2	1147	1053	1169	1177
	3	1247	1375	1309	1321
	4	1274	1235	1290	1285
	5	1177	1139	1233	1238
	6	1336	1313	1366	1362
	7	1238	1279	1275	1261
	8	1261	1152	1289	1266

Based on Hemmings, B., Smith, M., Graydon, J., and Dyson, R. "Effects of massage on physiological restoration, perceived recovery, and repeated sports performance." *British Journal of Sports Medicine*, Vol. 34, No. 2, Apr. 2000 (adapted from Table 3).

14.78 Plants and stress reduction. Refer to the Kansas State study designed to investigate the effects of plants on human stress levels, Exercise 10.73 (p. 518). Recall that finger temperatures for each of ten students in a dimly lit room were recorded under three experimental conditions: presence of a live plant, presence of a plant photo, and absence of a plant (either live or photo). For example, one student's finger measured 95.6° in the "Live Plant" condition, 92.6° in the "Plant Photo" condition, and 96.6° in the "No Plant" condition. The data for all ten students are saved in the **PLANTS** file. Analyze the data using a nonparametric procedure. Do students' finger temperatures depend on the experimental condition?

Based on data from Elizabeth Schreiber, Department of Statistics, Kansas State University, Manhattan, Kansas.

14.79 Absentee rates at a jeans plant. Refer to Exercise 10.74 (p. 518) and the *New Technology, Work, and Employment* (July 2001) study of daily worker absentee rates at a jeans plant. Nine weeks were randomly selected and the absentee rate (percentage of workers absent) determined for each day (Monday through Friday) of the workweek. For example, the absentee rates for the five days of the first week selected are: 5.3, .6, 1.9, 1.3, and 1.6, respectively. The data for all nine weeks are saved in the **JEANS** file. Use statistical software to conduct a nonparametric analysis of the data to compare the distributions of absentee rates for the five days of the week.

Based on Jean J. Boggis, "The eradication of leisure." *New Technology, Work, and Employment,* Vol. 16, No. 2, July 2001 (Table 3).

14.7 Rank Correlation

Suppose 10 new paintings are shown to two art critics and each critic ranks the paintings from 1 (best) to 10 (worst). We want to determine whether the critics' ranks are related. Does a correspondence exist between their ratings? If a painting is ranked high by critic 1, is it likely to be ranked high by critic 2? Or do high rankings by one critic correspond to low rankings by the other? That is, are the rankings of the critics *correlated*?

If the rankings are as shown in the "Perfect Agreement" columns of Table 14.9, we immediately notice that the critics agree on the rank of every painting. High ranks correspond to high ranks and low ranks to low ranks. This is an example of a *perfect positive correlation* between the ranks. In contrast, if the rankings appear as shown in the "Perfect Disagreement" columns of Table 14.9, then high ranks for one critic correspond to low ranks for the other. This is an example of *perfect negative correlation*.

In practice, you will rarely see perfect positive or perfect negative correlation between the ranks. In fact, it is quite possible for the critics' ranks to appear as shown in Table 14.10. Note that these rankings indicate some agreement between the critics, but not perfect agreement, thus pointing up a need for a measure of rank correlation.

Table 14.9 Rankings of 10 Paintings by Two Critics

	Perfect Agreement		Perfect Disagreement	
Painting	Critic 1	Critic 2	Critic 1	Critic 2
1	4	4	9	2
2	1	1	3	8
3	7	7	5	6
4	5	5	1	10
5	2	2	2	9
6	6	6	10	1
7	8	8	6	5
8	3	3	4	7
9	10	10	8	3
10	9	9	7	4

Table 14.10 Rankings of Paintings: Less-than-Perfect Agreement

	Critic		Difference between Rank 1 and Rank 2	
Painting	1	2	d	d^2
1	4	5	−1	1
2	1	2	−1	1
3	9	10	−1	1
4	5	6	−1	1
5	2	1	1	1
6	10	9	1	1
7	7	7	0	0
8	3	3	0	0
9	6	4	2	4
10	8	8	0	0
				$\Sigma d^2 = 10$

Spearman's rank correlation coefficient, r_s, provides a measure of correlation between ranks. The formula for this measure of correlation is given in the next box. We also give a formula that is identical to r_s when there are no ties in rankings; this formula provides a good approximation to r_s when the number of ties is small relative to the number of pairs.

Note that if the ranks for the two critics are identical, as in the second and third columns of Table 14.9, the differences between the ranks will all be 0. Thus,

$$r_s = 1 - \frac{6\sum d^2}{n(n^2 - 1)} = 1 - \frac{6(0)}{10(99)} = 1$$

That is, *perfect positive correlation* between the pairs of ranks is characterized by a Spearman correlation coefficient of $r_s = 1$. When the ranks indicate perfect disagreement, as in the fourth and fifth columns of Table 14.9, $\Sigma d_i^2 = 330$ and

$$r_s = 1 - \frac{6(330)}{10(99)} = -1.$$

Thus, *perfect negative correlation* is indicated by $r_s = -1$.

BIOGRAPHY CHARLES E. SPEARMAN (1863–1945)

Spearman's Correlation

London-born Charles Spearman was educated at Leamington College before joining the British Army. After 20 years as a highly decorated officer, Spearman retired from the army and moved to Germany to begin his study of experimental psychology at the University of Leipzig. At the age of 41, he earned his Ph.D. and ultimately became one of the most influential figures in the field of psychology. Spearman was the originator of the classical theory of mental tests and developed the "two-factor" theory of intelligence. These theories were used to develop and support the "Plus-Elevens" tests in England: exams administered to British 11-year-olds that predict whether they should attend a university or a technical school. Spearman was greatly influenced by the works of Francis Galton (p. 552); consequently, he developed a strong statistical background. While conducting his research on intelligence, he proposed the rank-order correlation coefficient—now called "Spearman's correlation coefficient." During his career, Spearman spent time at various universities, including University College (London), Columbia University, Catholic University, and the University of Cairo (Egypt). ■

Spearman's Rank Correlation Coefficient

$$r_s = \frac{\text{SS}_{uv}}{\sqrt{\text{SS}_{uu}\text{SS}_{vv}}}$$

where

$$\text{SS}_{uv} = \sum(u_i - \bar{u})(v_i - \bar{v}) = \sum u_i v_i - \frac{\left(\sum u_i\right)\left(\sum v_i\right)}{n}$$

$$\text{SS}_{uu} = \sum(u_i - \bar{u})^2 = \sum u_i^2 - \frac{\left(\sum u_i\right)^2}{n}$$

$$\text{SS}_{vv} = \sum(v_i - \bar{v})^2 = \sum v_i^2 - \frac{\left(\sum v_i\right)^2}{n}$$

u_i = Rank of the ith observation in sample 1
v_i = Rank of the ith observation in sample 2
n = Number of pairs of observations (number of observations in each sample)

Shortcut Formula for r_s*

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

where
$d_i = u_i - v_i$ (difference in the ranks of the ith observations for samples 1 and 2)
n = number of pairs of observations (number of observations in each sample)

For the data of Table 14.10,

$$r_s = 1 - \frac{6\sum d^2}{n(n^2 - 1)} = 1 - \frac{6(10)}{10(99)} = 1 - \frac{6}{99} = .94$$

The fact that r_s is *close* to 1 indicates that the critics tend to agree, but the agreement is not perfect.

*The shortcut formula is not exact when there are tied measurements, but it is a good approximation when the total number of ties is not large relative to n.

The value of r_s always falls between -1 and $+1$, with $+1$ indicating perfect positive correlation and -1 indicating a perfect negative correlation. The closer r_s falls to $+1$ or -1, the greater the correlation between the ranks. Conversely, the nearer r_s is to 0, the less is the correlation.

Note that the concept of correlation implies that two responses are obtained for each experimental unit. In the art critics example, each painting received two ranks (one from each critic) and the objective of the study was to determine the degree of positive correlation between the two rankings. Rank correlation methods can be used to measure the correlation between any pair of variables. If two variables are measured on each of n experimental units, we rank the measurements associated with each variable separately. Ties receive the average of the ranks of the tied observations. Then we calculate the value of r_s for the two rankings. This value measures the rank correlation between the two variables. We illustrate the procedure in Example 14.6.

Example 14.6

Spearman's Rank Correlation—Smoking versus Babies' Weights

Problem A study is conducted to investigate the relationship between cigarette smoking during pregnancy and the weights of newborn infants. The 15 women smokers who make up the sample kept accurate records of the number of cigarettes smoked during their pregnancies, and the weights of their children were recorded at birth. The data are given in Table 14.11.

Table 14.11 Data and Calculations for Example 14.6

Woman	Cigarettes per Day	Rank	Baby's Weight (pounds)	Rank	d	d^2
1	12	1	7.7	5	−4	16
2	15	2	8.1	9	−7	49
3	35	13	6.9	4	9	81
4	21	7	8.2	10	−3	9
5	20	5.5	8.6	13.5	−8	64
6	17	3	8.3	11.5	−8.5	72.25
7	19	4	9.4	15	−11	121
8	46	15	7.8	6	9	81
9	20	5.5	8.3	11.5	−6	36
10	25	8.5	5.2	1	7.5	56.25
11	39	14	6.4	3	11	121
12	25	8.5	7.9	7	1.5	2.25
13	30	12	8.0	8	4	16
14	27	10	6.1	2	8	64
15	29	11	8.6	13.5	−2.5	6.25
					Total = 795	

Data Set: NEWBORN

a. Calculate and interpret Spearman's rank correlation coefficient for the data.

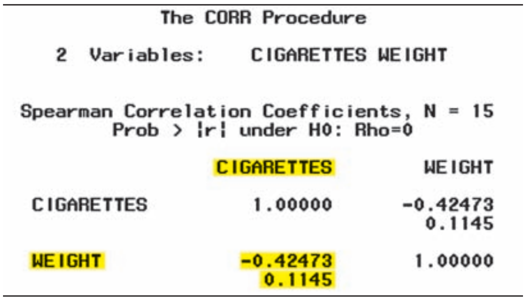
b. Use a nonparametric test to determine whether level of cigarette smoking and weights of newborns are negatively correlated for all smoking mothers. Use $\alpha = .05$.

Solution

a. We first rank the number of cigarettes smoked per day, assigning a 1 to the smallest number (12) and a 15 to the largest (46). Note that the two ties receive the averages of their respective ranks. Similarly, we assign ranks to the 15 babies' weights. Since the number of ties is relatively small, we will use the shortcut formula to calculate r_s. The differences d between the ranks of the babies' weights and the ranks of the number of cigarettes smoked per day are shown in Table 14.11. The squares of the differences, d^2, are also given. Thus,

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6(795)}{15(15^2-1)} = 1 - 1.42 = -.42$$

The value of r_s can also be obtained by computer. A SAS printout of the analysis is shown in Figure 14.14. The value of r_s, highlighted on the printout, agrees (except for rounding) with our hand-calculated value, $-.42$. The negative correlation coefficient indicates that in this sample an increase in the number of cigarettes smoked per day is *associated with* (but is not necessarily the *cause of*) a decrease in the weight of the newborn infant.



```
                  The CORR Procedure

   2  Variables:      CIGARETTES WEIGHT

 Spearman Correlation Coefficients, N = 15
         Prob > |r| under H0: Rho=0

                       CIGARETTES        WEIGHT

 CIGARETTES               1.00000      -0.42473
                                         0.1145

 WEIGHT                  -0.42473       1.00000
                           0.1145
```

Figure 14.14
SAS Spearman correlation printout for Example 14.6

b. If we define ρ as the **population rank correlation coefficient** [i.e., the rank correlation coefficient that could be calculated from all (x, y) values in the population], we can determine whether level of cigarette smoking and weights of newborns are negatively correlated by conducting the following test:

H_0: $\rho = 0$ (no population correlation between ranks)

H_a: $\rho < 0$ (negative population correlation between ranks)

Test statistic: r_s (the *sample* Spearman rank correlation coefficient)

To determine a rejection region, we consult Table XIV in Appendix A, which is partially reproduced in Table 14.12. Note that the left-hand column gives values of n, the number of pairs of observations. The entries in the table are values for an upper-tail rejection region, since only positive values are given. Thus, for $n = 15$ and $\alpha = .05$, the value .441 is the boundary of the upper-tailed rejection region, so $P(r_s > .441) = .05$ if H_0: $\rho = 0$ is true. Similarly, for negative values of r_s, we have $P(r_s < -.441) = .05$ if $\rho = 0$. That is, we expect to see $r_s < -.441$ only 5% of the time if there is really no relationship between the ranks of the variables.

Table 14.12 Reproduction of Part of Table XIV in Appendix A: Critical Values of Spearman's Rank Correlation Coefficient

n	$\alpha = .05$	$\alpha = .025$	$\alpha = .01$	$\alpha = .005$
5	.900	–	–	–
6	.829	.886	.943	–
7	.714	.786	.893	–
8	.643	.738	.833	.881
9	.600	.683	.783	.833
10	.564	.648	.745	.794
11	.523	.623	.736	.818
12	.497	.591	.703	.780
13	.475	.566	.673	.745
14	.457	.545	.646	.716
15	.441	.525	.623	.689
16	.425	.507	.601	.666
17	.412	.490	.582	.645
18	.399	.476	.564	.625
19	.388	.462	.549	.608
20	.377	.450	.534	.591

The lower-tailed rejection region is therefore

$$\textit{Rejection region } (\alpha = .05)\text{:} \quad r_s < -.441$$

Since the calculated $r_s = -.42$ is not less than $-.441$, we cannot reject H_0 at the $\alpha = .05$ level of significance. That is, this sample of 15 smoking mothers provides insufficient evidence to conclude that a negative correlation exists between the number of cigarettes smoked and the weight of newborns for the populations of measurements corresponding to all smoking mothers. This does not, of course, mean that no relationship exists. A study using a larger sample of smokers and taking other factors into account (father's weight, sex of newborn child, etc.) would be more likely to reveal whether smoking and the weight of a newborn child are related.

Look Back The two-tailed p-value of the test (.1145) is highlighted on the SAS printout, shown in Figure 14.14. Since the lower-tailed p-value, $.1145/2 = .05725$, exceeds $\alpha = .05$, our conclusion is the same: Do not reject H_0.

Now Work Exercise 14.85

A summary of Spearman's nonparametric test for correlation is given in the following box:

Spearman's Nonparametric Test for Rank Correlation

One-Tailed Test	**Two-Tailed Test**
H_0: $\rho = 0$ H_a: $\rho > 0$ $[$or H_a: $\rho < 0]$	H_0: $\rho = 0$ H_a: $\rho \neq 0$

Test statistic: $\mathrm{r_s}$, the sample rank correlation (see the formulas for calculating $\mathrm{r_s}$)

Rejection region: $r_s > r_{s,\alpha}$ $[$or $r_s < -r_{s,\alpha}$ when H_a: $\rho < 0]$	*Rejection region*: $\lvert r_s \rvert > r_{s,\alpha/2}$
where $r_{s,\alpha}$ is the value from Table XIV corresponding to the upper-tail area α and n pairs of observations	where $r_{s,\alpha/2}$ is the value from Table XIV corresponding to the upper-tail area $\alpha/2$ and n pairs of observations

Ties: Assign tied measurements the average of the ranks they would receive if they were unequal, but occurred in successive order. For example, if the third-ranked and fourth-ranked measurements are tied, assign each a rank of $(3 + 4)/2 = 3.5$. The number of ties should be small relative to the total number of observations.

Conditions Required for a Valid Spearman's Test

1. The sample of experimental units on which the two variables are measured is randomly selected.
2. The probability distributions of the two variables are continuous.

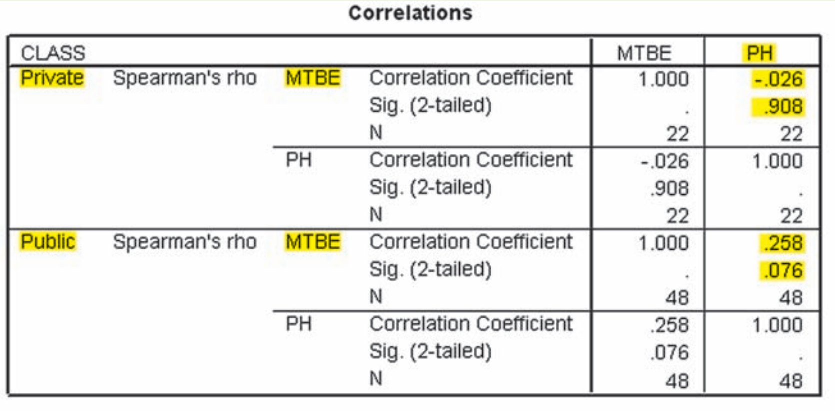
Statistics IN Action Revisited

Testing the Correlation of MTBE Level with Other Environmental Factors

Refer again to the *Environmental Science & Technology* (Jan. 2005) investigation of the MTBE contamination of drinking water in New Hampshire (p. 14-2). The environmental researchers also wanted an estimate of the correlation between the MTBE level of a groundwater well and each of the other environmental variables listed in Table SIA14.1. Since the MTBE level is not normally distributed, they employed Spearman's rank correlation method. Also, because earlier analyses indicated that public and private wells have different MTBE distributions, the rank correlations were computed separately for each well class. SPSS printouts for this analysis are shown in Figures SIA14.5a–e. The values of r_s (and associated p-values) are highlighted on the printouts. Our interpretations follow:

MTBE vs. pH level (Figure SIA14.5a). For private wells, $r_s = -.026$ (p-value $= .908$). Thus, there is a low negative association between MTBE level and pH level for private wells—an association that is not significantly different from 0 (at $\alpha = .10$). For public wells, $r_s = .258$ (p-value $= .076$). Consequently, there is a low positive association (significant

Statistics IN Action
(continued)



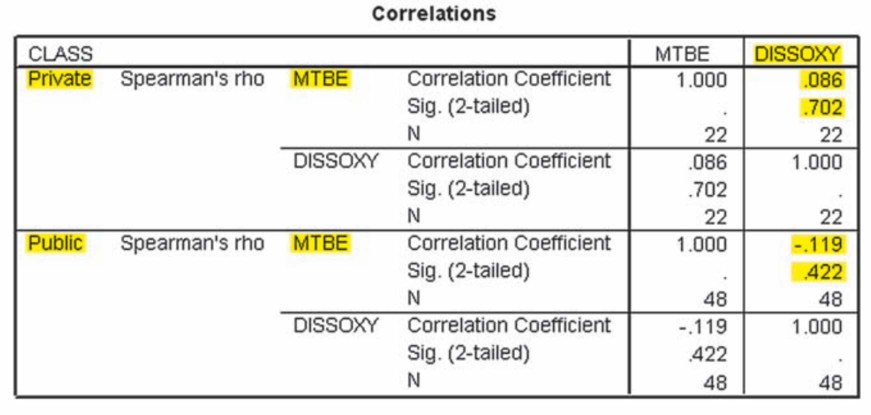
Correlations

CLASS				MTBE	PH
Private	Spearman's rho	MTBE	Correlation Coefficient	1.000	-.026
			Sig. (2-tailed)	.	.908
			N	22	22
		PH	Correlation Coefficient	-.026	1.000
			Sig. (2-tailed)	.908	.
			N	22	22
Public	Spearman's rho	MTBE	Correlation Coefficient	1.000	.258
			Sig. (2-tailed)	.	.076
			N	48	48
		PH	Correlation Coefficient	.258	1.000
			Sig. (2-tailed)	.076	.
			N	48	48

Figure SIA14.5a
SPSS Spearman rank correlation test: MTBE and pH level

difference from 0 at $\alpha = .10$) for public wells between MTBE level and pH level.

MTBE vs. Dissolved oxygen (*Figure SIA14.5b*). For private wells, $r_s = .086$ (p-value $= .702$). For public wells, $r_s = -.119$ (p-value $= .422$). Thus, there is a low positive association between MTBE level and dissolved oxygen for private wells, but a low negative association between MTBE level and dissolved oxygen for public wells. However, neither rank correlation is significantly different from 0 (at $\alpha = .10$).

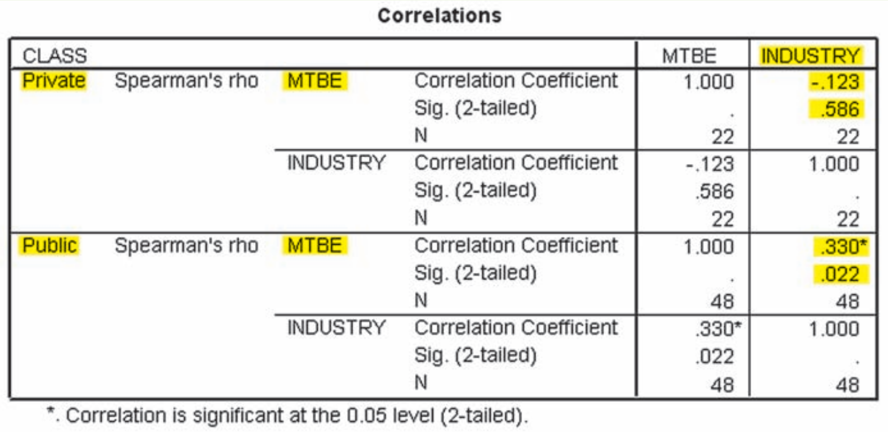


Correlations

CLASS				MTBE	DISSOXY
Private	Spearman's rho	MTBE	Correlation Coefficient	1.000	.086
			Sig. (2-tailed)	.	.702
			N	22	22
		DISSOXY	Correlation Coefficient	.086	1.000
			Sig. (2-tailed)	.702	.
			N	22	22
Public	Spearman's rho	MTBE	Correlation Coefficient	1.000	-.119
			Sig. (2-tailed)	.	.422
			N	48	48
		DISSOXY	Correlation Coefficient	-.119	1.000
			Sig. (2-tailed)	.422	.
			N	48	48

Figure SIA14.5b
SPSS Spearman rank correlation test: MTBE and dissolved oxygen

MTBE vs. Industry percentage (*Figure SIA14.5c*). For private wells, $r_s = -.123$ (p-value $= .586$). This low negative association between MTBE level and industry percentage for private wells is not significantly different from 0 (at $\alpha = .10$). For public wells, $r_s = .330$ (p-value $= .022$). Consequently, there is a low positive association (significantly different from 0 at $\alpha = .10$) for public wells between MTBE level and industry percentage.



Correlations

CLASS				MTBE	INDUSTRY
Private	Spearman's rho	MTBE	Correlation Coefficient	1.000	-.123
			Sig. (2-tailed)	.	.586
			N	22	22
		INDUSTRY	Correlation Coefficient	-.123	1.000
			Sig. (2-tailed)	.586	.
			N	22	22
Public	Spearman's rho	MTBE	Correlation Coefficient	1.000	.330*
			Sig. (2-tailed)	.	.022
			N	48	48
		INDUSTRY	Correlation Coefficient	.330*	1.000
			Sig. (2-tailed)	.022	.
			N	48	48

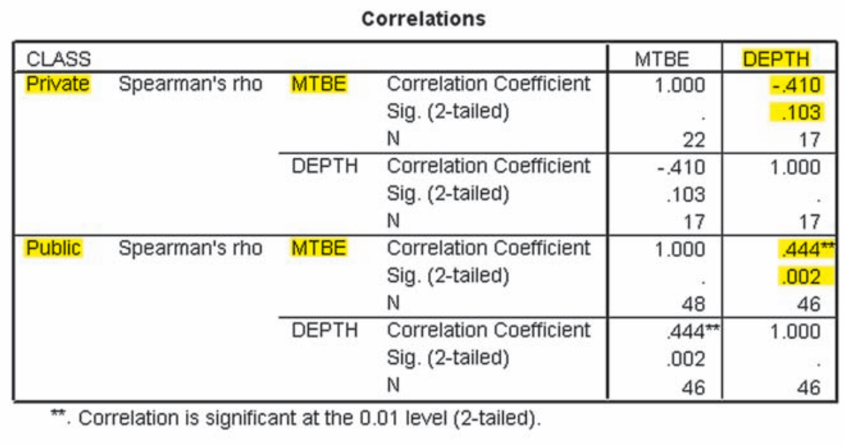
*. Correlation is significant at the 0.05 level (2-tailed).

Figure SIA14.5c
SPSS Spearman rank correlation test: MTBE and industry percentage

(*continued*)

Statistics IN Action
(continued)

MTBE vs. Depth of well (*Figure SIA14.5d*). For private wells, $r_s = -.410$ (p-value $= .103$). This low negative association between MTBE level and depth for private wells is not significantly different from 0 (at $\alpha = .10$). For public wells, $r_s = .444$ (p-value $= .002$). Consequently, there is a low positive association (significantly different from 0 at $\alpha = .10$) for public wells between MTBE level and depth.



Correlations

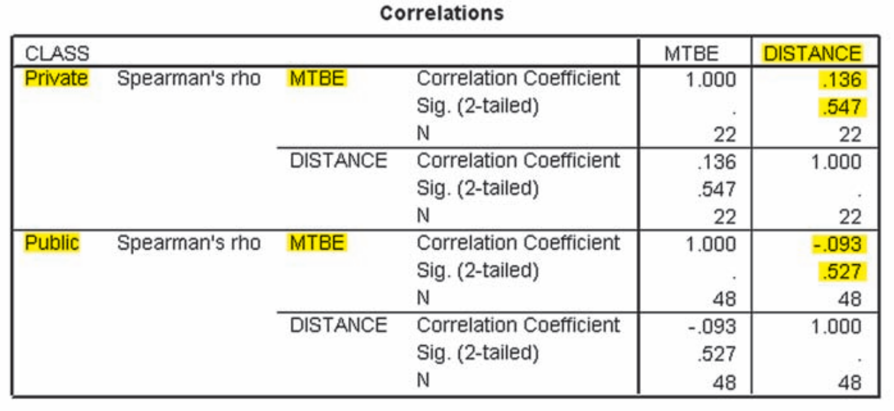
CLASS				MTBE	DEPTH
Private	Spearman's rho	MTBE	Correlation Coefficient	1.000	-.410
			Sig. (2-tailed)	.	.103
			N	22	17
		DEPTH	Correlation Coefficient	-.410	1.000
			Sig. (2-tailed)	.103	.
			N	17	17
Public	Spearman's rho	MTBE	Correlation Coefficient	1.000	.444**
			Sig. (2-tailed)	.	.002
			N	48	46
		DEPTH	Correlation Coefficient	.444**	1.000
			Sig. (2-tailed)	.002	.
			N	46	46

**. Correlation is significant at the 0.01 level (2-tailed).

Figure SIA14.5d
SPSS Spearman rank correlation test: MTBE and depth

MTBE vs. Distance from underground tank (*Figure SIA14.5e*). For private wells, $r_s = .136$ (p-value $= .547$). For public wells, $r_s = -.093$ (p-value $= .527$). Thus, there is a low positive association between MTBE level and distance for private wells, but a low negative association between MTBE level and distance for public wells. However, neither rank correlation is significantly different from 0 (at $\alpha = .10$).

In sum, the only significant rank correlations were for public wells, where the researchers discovered low positive associations of MTBE level with pH level, industry percentage, and depth of the well.



Correlations

CLASS				MTBE	DISTANCE
Private	Spearman's rho	MTBE	Correlation Coefficient	1.000	.136
			Sig. (2-tailed)	.	.547
			N	22	22
		DISTANCE	Correlation Coefficient	.136	1.000
			Sig. (2-tailed)	.547	.
			N	22	22
Public	Spearman's rho	MTBE	Correlation Coefficient	1.000	-.093
			Sig. (2-tailed)	.	.527
			N	48	48
		DISTANCE	Correlation Coefficient	-.093	1.000
			Sig. (2-tailed)	.527	.
			N	48	48

Figure SIA14.5e
SPSS Spearman rank correlation test: MTBE and distance

Exercises 14.80–14.97

Understanding the Principles

14.80 What is the value of r_s when there is perfect negative rank correlation between two variables? Perfect positive rank correlation?

14.81 What conditions are required for a valid Spearman's test?

Learning the Mechanics

14.82 Use Table XIV of Appendix A to find each of the following probabilities:

a. $P(r_s > .508)$ when $n = 22$
b. $P(r_s > .448)$ when $n = 28$
c. $P(r_s \leq .648)$ when $n = 10$
d. $P(r_s < -.738$ or $r_s > .738)$ when $n = 8$

14.83 Specify the rejection region for Spearman's nonparametric test for rank correlation in each of the following situations:

a. H_0: $\rho = 0$, H_a: $\rho \neq 0$, $n = 10$, $\alpha = .05$
b. H_0: $\rho = 0$, H_a: $\rho > 0$, $n = 20$, $\alpha = .025$
c. H_0: $\rho = 0$, H_a: $\rho < 0$, $n = 30$, $\alpha = .01$

14.84 Compute Spearman's rank correlation coefficient for each of the following pairs of sample observations:

a.

x	33	61	20	19	40
y	26	36	65	25	35

b.

x	89	102	120	137	41
y	81	94	75	52	136

c.

x	2	15	4	10
y	11	2	15	21

d.

x	5	20	15	10	3
y	80	83	91	82	87

14.85 NW The following sample data, saved in the **LM14_85** file, were collected on variables x and y:

x	0	3	0	−4	3	0	4
y	0	2	2	0	3	1	2

a. Specify the null and alternative hypotheses that should be used in conducting a hypothesis test to determine whether the variables x and y are correlated.
b. Conduct the test of part **a,** using $\alpha = .05$.
c. What is the approximate p-value of the test of part **b**?
d. What assumptions are necessary to ensure the validity of the test of part **b**?

Applying the Concepts—Basic

14.86 Mongolian desert ants. Refer to the *Journal of Biogeography* (Dec. 2003) study of ants in Mongolia, presented in Exercise 11.22 (p. 562). Data on annual rainfall, maximum daily temperature, and number of ant species recorded at each of 11 study sites are reproduced in the table below and saved in the **GOBIANTS** file.

Site	Region	Annual Rainfall (mm)	Max. Daily Temp. (°C)	Number of Ant Species
1	Dry Steppe	196	5.7	3
2	Dry Steppe	196	5.7	3
3	Dry Steppe	179	7.0	52
4	Dry Steppe	197	8.0	7
5	Dry Steppe	149	8.5	5
6	Gobi Desert	112	10.7	49
7	Gobi Desert	125	11.4	5
8	Gobi Desert	99	10.9	4
9	Gobi Desert	125	11.4	4
10	Gobi Desert	84	11.4	5
11	Gobi Desert	115	11.4	4

Based on Pfeiffer, M., et al. "Community organization and species richness of ants in Mongolia along an ecological gradient from steppe to Gobi desert." *Journal of Biogeography,* Vol. 30, No. 12, Dec. 2003 (Tables 1 and 2).

a. Consider the data for the five sites in the Dry Steppe region only. Rank the five annual rainfall amounts. Then rank the five maximum daily temperature values.
b. Use the ranks from part **a** to find and interpret the rank correlation between annual rainfall (y) and maximum daily temperature (x).
c. Repeat parts **a** and **b** for the six sites in the Gobi Desert region.
d. Now consider the rank correlation between the number of ant species (y) and annual rainfall (x). Using all the data, compute and interpret Spearman's rank correlation statistic.

14.87 Extending the life of an aluminum smelter pot. Refer to the *American Ceramic Society Bulletin* (Feb. 2005) study of the lifetime of an aluminum smelter pot, presented in Exercise 11.24 (p. 563). Since the life of a smelter pot depends on the porosity of the brick lining, the researchers measured the apparent porosity and the mean pore diameter of each of six bricks. The data, saved in the **SMELTPOT** file, are reproduced in the following table:

Brick	Apparent Porosity (%)	Mean Pore Diameter (micrometers)
A	18.8	12.0
B	18.3	9.7
C	16.3	7.3
D	6.9	5.3
E	17.1	10.9
F	20.4	16.8

Based on Bonadia, P., et al. "Aluminosilicate refractories for aluminum cell linings." *American Ceramic Society Bulletin*, Vol. 84, No. 2, Feb. 2005 (Table II).

a. Rank the apparent porosity values for the six bricks. Then rank the six pore diameter values.
b. Use the ranks from part **a** to find the rank correlation between apparent porosity (y) and mean pore diameter (x). Interpret the result.
c. Conduct a test for positive rank correlation. Use $\alpha = .01$.

14.88 Lobster fishing study. Refer to the *Bulletin of Marine Science* (April 2010) study of teams of fishermen fishing for the red spiny lobster in Baja California Sur, Mexico, Exercise 11.55 (p. 576). Recall that two variables measured for each of 8 teams from the Punta Abreojos (PA) fishing cooperative were total catch of lobsters (in kilograms) during the season and average percentage of traps allocated per day to exploring areas of unknown catch (called *search frequency*). These data, saved in the **TRAPSPACE** file, are reproduced in the table.

Total Catch	Search Frequency
2,785	35
6,535	21
6,695	26
4,891	29
4,937	23
5,727	17
7,019	21
5,735	20

From Shester, G. G. "Explaining catch variation among Baja California lobster fishers through spatial analysis of trap-placement decisions." *Bulletin of Marine Science,* Vol. 86, No. 2, April 2010 (Table 1). Reprinted with permission from the University of Miami, *Bulletin of Marine Science*.

a. Rank the total catch values from 1 to 8.
b. Rank the search frequency values from 1 to 8.
c. Use the ranks, parts **a** and **b,** to compute Spearman's rank correlation coefficient.
d. Based on the result, part **c,** is there sufficient evidence to indicate that total catch is negatively rank correlated with search frequency? Test using $\alpha = .05$.

14.89 Effect of massage on boxers. Refer to the *British Journal of Sports Medicine* (Apr. 2000) study of the effect of massaging boxers between rounds, presented in Exercise 11.60

(p. 577). Two variables measured on the boxers were blood lactate level (y) and the boxer's perceived recovery (x). The data for 16 five-round boxing performances are reproduced in the table and saved in the **BOXING2** file.

Blood Lactate Level	Perceived Recovery
3.8	7
4.2	7
4.8	11
4.1	12
5.0	12
5.3	12
4.2	13
2.4	17
3.7	17
5.3	17
5.8	18
6.0	18
5.9	21
6.3	21
5.5	20
6.5	24

Based on Hemmings, B., Smith, M., Graydon, J., and Dyson, R. "Effects of massage on physiological restoration, perceived recovery, and repeated sports performance." *British Journal of Sports Medicine,* Vol. 34, No. 2, Apr. 2000 (data adapted from Figure 3).

a. Rank the values of the 16 blood lactate levels.
b. Rank the values of the 16 perceived recovery values.
c. Use the ranks from parts **a** and **b** to compute Spearman's rank correlation coefficient. Give a practical interpretation of the result.
d. Find the rejection region for a test to determine whether y and x are rank correlated. Use $\alpha = .10$.
e. What is the conclusion of the test you conducted in part **d**? State your answer in the words of the problem.

14.90 Assessment of biometric recognition methods. Biometric technologies have been developed to detect or verify an individual's identity. These methods are based on physiological characteristics (called *biometric signatures*), such as facial features, the iris of the eye, fingerprints, the voice, the shape of the hand, and the gait. In *Chance* (Winter 2004), four biometric recognition algorithms were compared. All four were applied to 1,196 biometric signatures, and "match" scores were obtained. The Spearman correlation between match scores for each possible pair of algorithms was determined. The rank correlation matrix is as follows:

Method	I	II	III	IV
I	1	.189	.592	.340
II		1	.205	.324
III			1	.314
IV				1

a. Locate the largest rank correlation and interpret its value.
b. Locate the smallest rank correlation and interpret its value.

14.91 Childhood obesity study. Refer to the *Journal of Education and Human Development* (Vol. 3, 2009) study of the eating patterns of families of overweight preschool children, Exercise 12.62 (p. 652). The body mass index for each in a sample of 10 overweight children and their parents were determined. These data, saved in the **BMI** file, are reproduced in the next table.

Child	Parent
17.10	24.62
17.15	24.70
17.20	25.70
17.24	25.80
17.25	26.20
17.30	26.30
17.32	26.60
17.40	26.80
17.60	27.20
17.80	27.35

Based on Seal, N., and Seal, J. "Eating patterns of the rural families of overweight preschool children: A pilot study." *Journal of Education and Human Development*, Vol. 3, No. 1, 2009 (Figure 1).

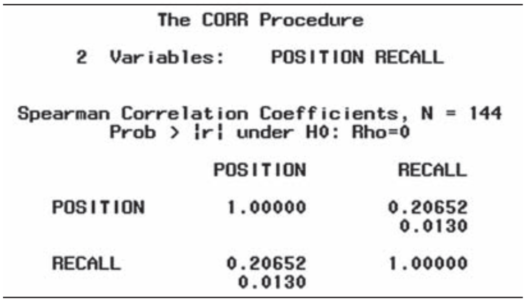
a. Demonstrate that the rank correlation between the BMI values in the sample is $r_s = 1$.
b. Interpret the result, part **a.**
c. Why should a researcher avoid concluding that there is a perfect linear relationship between BMI values in overweight children and their parents?

Applying the Concepts—Intermediate

14.92 The "name game" Refer to the *Journal of Experimental Psychology—Applied* (June 2000) study in which the "name game" was used to help groups of students learn the names of other students in the group, presented in Exercise 11.30 (p. 565). Recall that one goal of the study was to investigate the relationship between proportion y of names recalled by a student and position (order x) of the student during the game. The data for 144 students in the first eight positions are saved in the **NAMEGAME2** file. (The first five and last five observations in the data set are listed below.) A SAS printout follows.

Position	Recall
2	0.04
2	0.37
2	1.00
2	0.99
2	0.79
⋮	⋮
9	0.72
9	0.88
9	0.46
9	0.54
9	0.99

Based on Morris, P. E., and Fritz, C. O. "The name game: Using retrieval practice to improve the learning of names." *Journal of Experimental Psychology—Applied,* Vol. 6, No. 2, June 2000 (data simulated from Figure 2).



```
                    The CORR Procedure

          2  Variables:     POSITION RECALL


Spearman Correlation Coefficients, N = 144
        Prob > |r| under H0: Rho=0

                      POSITION         RECALL

   POSITION            1.00000        0.20652
                                       0.0130

   RECALL              0.20652        1.00000
                        0.0130
```

a. To properly apply the parametric test for correlation on the basis of the Pearson coefficient of correlation, *r* (Section 11.6), both the *x* and *y* variables must be normally distributed. Demonstrate that this assumption is violated for these data. What are the consequences of the violation?

b. Find Spearman's rank correlation coefficient on the accompanying SAS printout and interpret its value.

c. Find the observed significance level for testing for zero rank correlation on the SAS printout, and interpret its value.

d. At $\alpha = .05$, is there sufficient evidence of rank correlation between proportion *y* of names recalled by a student and position (order *x*) of the student during the game?

14.93 Study of child bipolar disorders. Psychiatric researchers at the University of Pittsburgh Medical Center have developed a new test for measuring manic symptoms in pediatric bipolar patients (*Journal of Child and Adolescent Psychopharmacology*, Dec. 2003). The new test is called the Kiddie Schedule for Affective Disorders and Schizophrenia-Mania Rating Scale (KSADS-MRS). The new test was compared with the standard test, the Clinical Global Impressions—Bipolar Scale (CGI-BP). Both tests were administered to a sample of 18 pediatric patients before and after they were treated for manic symptoms. The changes in the test scores are recorded in the accompanying table and saved in the **MANIA** file.

Patient	Change in KSADS-MRS (%)	Improvement in CGI-BP
1	80	6
2	65	5
3	20	4
4	−15	4
5	−50	4
6	20	3
7	−30	3
8	−70	3
9	−10	2
10	−25	2
11	−35	2
12	−65	2
13	−65	2
14	−70	2
15	−80	2
16	−90	2
17	−95	2
18	−90	1

Based on Axelson, D. et al. "A preliminary study of the Kiddie Schedule for Affective Disorders and Schizophrenia for School-Age Children Mania Rating Scale for children and adolescents." *Journal of Child and Adolescent Psychopharmacology,* Vol. 13, No. 4, Dec. 2003 (Figure 2).

a. The researchers used Spearman's statistic to measure the correlation between the changes in the two test scores. Compute the value of r_s.

b. Is there sufficient evidence (at $\alpha = .05$) of positive rank correlation between the two test score changes in the population of all pediatric patients with manic symptoms?

14.94 Do nice guys finish first or last? Refer to the *Nature* (March 20, 2008) study of whether the saying "nice guys finish last" applies to the competitive corporate world, Exercise 11.18 (p. 561). Recall that college students repeatedly played a version of the game "prisoner's dilemma," where competitors choose cooperation, defection, or costly punishment. At the conclusion of the games, the researchers recorded the average payoff and the number of times punishment was used for each player. The data in the table, saved in the **PUNISH** file, are representative of the data obtained in the study. The researchers concluded that "punishers tend to have lower payoffs." Do you agree? Use Spearman's rank correlation statistic to support your conclusion.

Punish	Payoff
0	0.50
1	0.20
2	0.30
3	0.25
4	0.00
5	0.30
6	0.10
8	–0.20
10	0.15
12	–0.30
14	–0.10
16	–0.20
17	–0.25

14.95 FCAT scores and poverty. Refer to the *Journal of Educational and Behavioral Statistics* (Spring 2004) analysis of the link between Florida Comprehensive Assessment Test (FCAT) scores and sociodemographic factors, presented in Exercise 11.26 (p. 564). Data on average math and reading FCAT scores of third graders, as well as the percentage of students below the poverty level, for a sample of 22 Florida elementary schools are saved in the **FCAT** file.

a. Compute and interpret Spearman's rank correlation between FCAT math score (*y*) and percentage (*x*) of students below the poverty level.

b. Compute and interpret Spearman's rank correlation between FCAT reading score (*y*) and percentage (*x*) of students below the poverty level.

c. Determine whether the value of r_s in part **a** would lead you to conclude that FCAT math score and percent below poverty level are negatively rank correlated in the population of all Florida elementary schools. Use $\alpha = .01$ to make your decision.

d. Determine whether the value of r_s in part **b** would lead you to conclude that FCAT reading score and per cent below poverty level are negatively rank correlated in the population of all Florida elementary schools. Use $\alpha = .01$ to make your decision.

14.96 Pain empathy and brain activity. Refer to the *Science* (Feb. 20, 2004) study on the relationship between brain activity and pain-related empathy in persons who watch others in pain, presented in Exercise 11.62 (p. 577). Recall that 16 female partners watched while painful stimulation was applied to the finger of their respective male partners. The two variables of interest were y = female's pain-related brain activity (measured on a scale ranging from −2 to 2) and x = female's score on the Empathic Concern Scale (0 to 25 points). The data, saved in the

BRAINPAIN file, are reproduced in the accompanying table. Use Spearman's rank correlation test to answer the research question, "Do people scoring higher in empathy show higher pain-related brain activity?"

Couple	Brain Activity (*y*)	Empathic Concern (x)
1	.05	12
2	−.03	13
3	.12	14
4	.20	16
5	.35	16
6	0	17
7	.26	17
8	.50	18
9	.20	18
10	.21	18
11	.45	19
12	.30	20
13	.20	21
14	.22	22
15	.76	23
16	.35	24

Based on Singer, T., et al. "Empathy for pain involves the affective but not sensory components of pain." *Science,* Vol. 303, Feb. 20, 2004. (Adapted from Figure 4.)

14.97 Public perceptions of health risks. Refer to the *Journal of Experimental Psychology: Learning, Memory, and Cognition* (July 2005) study of the ability of people to judge the risk of an infectious disease, presented in Exercise 12.73 (p. 655). Recall that the researchers asked German college students to estimate the number of people infected with a certain disease in a typical year. The median estimates, as well as the actual incidence for each in a sample of 24 infections, are reproduced in the accompanying table and saved in the **INFECTION** file.

a. Use graphs to demonstrate that the variables Actual incidence and Estimated incidence are not normally distributed.

b. Recall that the researchers used regression to model the relationship between Actual incidence and Estimated incidence. How does the result you found in part **a** affect this analysis?

c. Find Spearman's correlation coefficient for the two variables. Interpret this value.

d. Refer to part **c.** At $\alpha = .01$, is there a positive association between Actual incidence and Estimated incidence?

Infection	Actual Incidence	Estimated Incidence
Polio	0.25	300
Diphtheria	1	1000
Trachoma	1.75	691
Rabbit Fever	2	200
Cholera	3	17.5
Leprosy	5	0.8
Tetanus	9	1000
Hemorrhagic Fever	10	150
Trichinosis	22	326.5
Undulant Fever	23	146.5
Well's Disease	39	370
Gas Gangrene	98	400
Parrot Fever	119	225
Typhoid	152	200
Q Fever	179	200
Malaria	936	400
Syphilis	1514	1500
Dysentery	1627	1000
Gonorrhea	2926	6000
Meningitis	4019	5000
Tuberculosis	12619	1500
Hepatitis	14889	1000
Gastroenteritis	203864	37000
Botulism	15	37500

Based on Hertwig, R., Pachur, T., and Kurzenhauser, S. "Judgments of risk frequencies: Tests of possible cognitive mechanisms." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol. 31, No. 4, July 2005 (Table 1).

CHAPTER NOTES

Key Terms

Distribution-free tests 14-3
Friedman F_r-statistic 14-34
Kruskal-Wallis *H*-test 14-27
Nonparametrics 14-3
Parametric statistical tests 14-2
Population rank correlation coefficient 14-43
Rank statistics (or rank tests) 14-3
Rank sum 14-10
Sign test 14-4
Spearman's rank correlation coefficient 14-40
Wilcoxon rank sum test 14-10
Wilcoxon signed rank test 14-20

Key Symbols

η	Population median
S	Test statistic for sign test
T_1	Sum of ranks of observations in sample 1
T_2	Sum of ranks of observations in sample 2
T_L	Critical lower Wilcoxon rank sum value
T_U	Critical upper Wilcoxon rank sum value
T_+	Sum of ranks of positive differences of paired observations
T_-	Sum of ranks of negative differences of paired observations
T_0	Critical value of Wilcoxon signed ranks test
R_j	Rank sum of observations in sample *j*
H	Test statistic for Kruskal-Wallis test
F_r	Test statistic for Friedman test
r_s	Spearman's rank correlation coefficient
p	Population correlation coefficient

Key Ideas

Distribution-free Tests

Do not rely on assumptions about the probability distribution of the sampled population; are based on **rank statistics**

Nonparametrics

One-sample test for the population median: **sign test**
Test for *two independent samples:* **Wilcoxon rank sum test**
Test for *matched pairs:* **Wilcoxon signed rank test**
Test for a *completely randomized design:* **Kruskal-Wallis test**

Test for a *randomized block design:* **Friedman test**
Test for *rank correlation:* **Spearman's test**

Key Formulas

Sign test, large-sample test statistic:

$$z = \frac{(S - .5) - .5n}{.5\sqrt{n}}$$

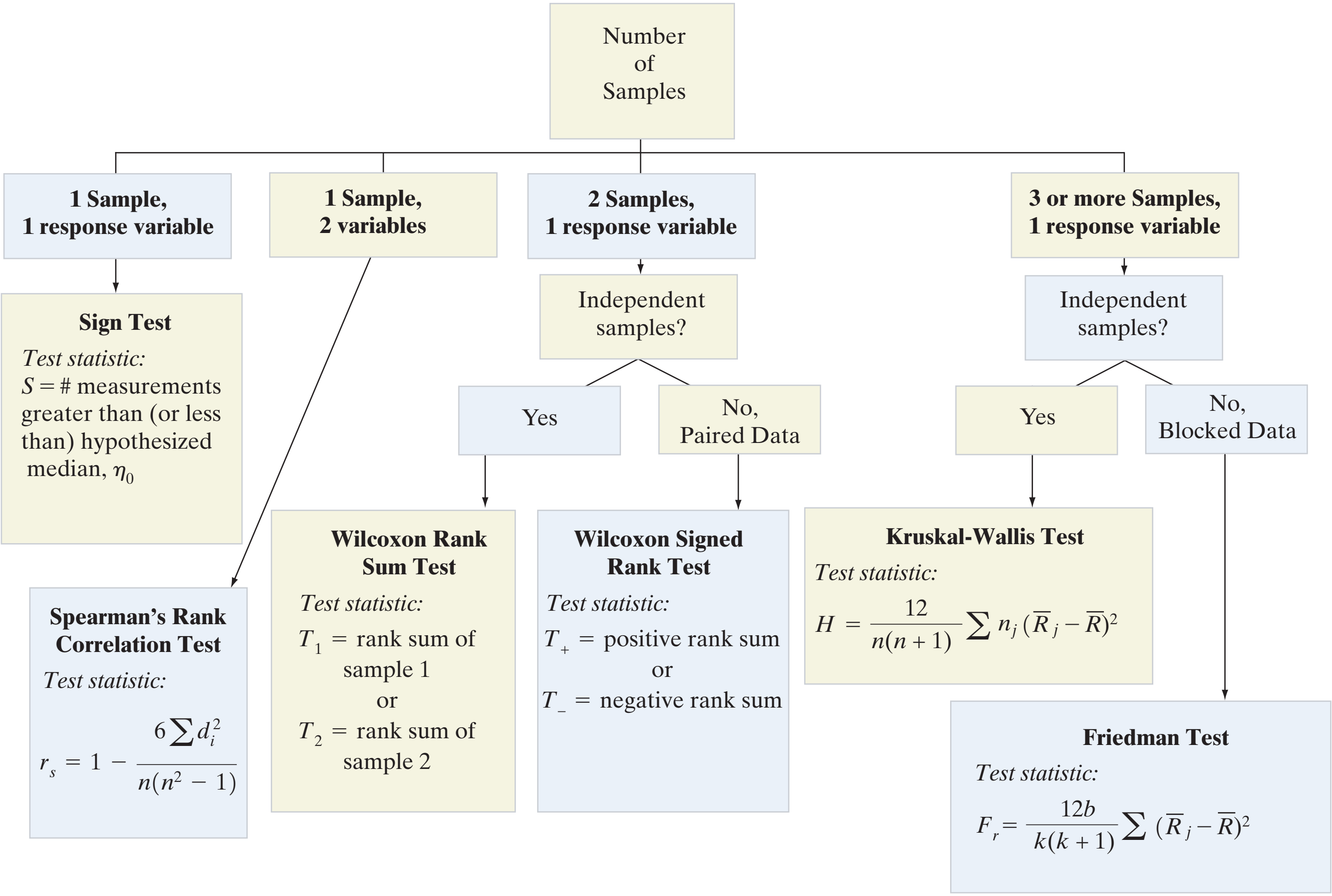
Wilcoxon rank sum test, large-sample test statistic:

$$z = \frac{T_1 - \dfrac{n_1(n_1 + n_2 + 1)}{2}}{\sqrt{\dfrac{n_1 n_2(n_1 + n_2 + 1)}{12}}}$$

Wilcoxon signed rank test, large-sample test statistic:

$$z = \frac{T_+ - \dfrac{n(n + 1)}{4}}{\sqrt{\dfrac{n(n + 1)(2n + 1)}{24}}}$$

Guide to Selecting a Nonparametric Method


Number of Samples
1 Sample, 1 response variable
1 Sample, 2 variables
2 Samples, 1 response variable
3 or more Samples, 1 response variable
Sign Test
Test statistic:
S = # measurements greater than (or less than) hypothesized median, η_0
Independent samples?
Yes
No, Paired Data
Independent samples?
Yes
No, Blocked Data
Spearman's Rank Correlation Test
Test statistic:
$r_s = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$
Wilcoxon Rank Sum Test
Test statistic:
T_1 = rank sum of sample 1
or
T_2 = rank sum of sample 2
Wilcoxon Signed Rank Test
Test statistic:
T_+ = positive rank sum
or
T_- = negative rank sum
Kruskal-Wallis Test
Test statistic:
$H = \frac{12}{n(n+1)} \sum n_j(\bar{R}_j - \bar{R})^2$
Friedman Test
Test statistic:
$F_r = \frac{12b}{k(k+1)} \sum (\bar{R}_j - \bar{R})^2$

Supplementary Exercises 14.98–14.124

Understanding the Principles

14.98 How does a nonparametric test differ from the parametric t- and F- tests of Chapters 8–10?

14.99 For each of the following, give the appropriate nonparametric test to apply:

a. Comparing two populations with independent samples
b. Making an inference about a population median
c. Comparing three or more populations with independent samples
d. Making an inference about rank correlation
e. Comparing two populations with matched pairs
f. Comparing three or more populations with a block design

Learning the Mechanics

14.100 The data for three independent random samples are shown in the table (top of page 14-52) and saved in the **LM14_100** file. It is known that the sampled populations are not normally distributed. Use an appropriate test to determine whether the data provide sufficient evidence to indicate that at least two of the populations differ in location. Use $\alpha = .05$.

Data for Exercise 14.100

Sample 1		Sample 2		Sample 3	
18	15	12	34	87	50
32	63	33	18	53	64
43		10		65	77

14.101 A random sample of nine pairs of observations is recorded on two variables x and y. The data are shown in the following table and saved in the **LM14_101** file.

Pair	x	y
1	19	12
2	27	19
3	15	7
4	35	25
5	13	11
6	29	10
7	16	16
8	22	10
9	16	18

a. Do the data provide sufficient evidence to indicate that ρ, the rank correlation between x and y, differs from 0? Test, using $\alpha = .05$.

b. Do the data provide sufficient evidence to indicate that the probability distribution for x is shifted to the right of that for y? Test, using $\alpha = .05$.

14.102 Two independent random samples produced the measurements listed in the next table. Do the data (saved in the **LM14_102** file) provide sufficient evidence to conclude that there is a difference between the locations of the probability distributions for the sampled populations? Test, using $\alpha = .05$.

Sample 1		Sample 2	
1.2	1.0	1.5	1.9
1.9	1.8	1.3	2.7
.7	1.1	2.9	3.5
2.5			

14.103 An experiment was conducted using a randomized block design with five treatments and four blocks. The data are shown in the accompanying table and saved in the **LM14_103** file. Do the data provide sufficient evidence to conclude that at least two of the treatment probability distributions differ in location? Test, using $\alpha = .05$.

	Block			
Treatment	1	2	3	4
1	75	77	70	80
2	65	69	63	69
3	74	78	69	80
4	80	80	75	86
5	69	72	63	77

Applying the Concepts—Basic

14.104 Reading Japanese books. Refer to the *Reading in a Foreign Language* (Apr. 2004) experiment to improve the Japanese reading comprehension levels of University of Hawaii students, presented in Exercise 9.17 (p. 424). Recall that 14 students participated in a 10-week extensive reading program in a second-semester Japanese course. The number of books read by each student and the student's course grade are repeated in the accompanying table and saved in the **JAPANESE** file. Consider a comparison of the distributions of number of books read by students who earn an "A" grade and those who earn a "B" or "C" grade.

Number of Books	Course Grade
53	A
42	A
40	A
40	B
39	A
34	A
34	A
30	A
28	B
24	A
22	C
21	B
20	B
16	B

Source: Hitosugi, C. I., and Day, R. R. "Extensive reading in Japanese." *Reading in a Foreign Language,* Vol. 16, No. 1, Apr. 2004 (Table 4). Reprinted with permission from the National Foreign Language Resource Center, University of Hawaii.

a. Rank all 14 observations from smallest to largest, and assign ranks from 1 to 14.

b. Sum the ranks of the observations for students with an "A" grade.

c. Sum the ranks of the observations for students with either a "B" or "C" grade.

d. Compute the Wilcoxon rank sum statistic.

e. Carry out a nonparametric test (at $\alpha = .10$) to compare the distribution of the number of books read by the two populations of students.

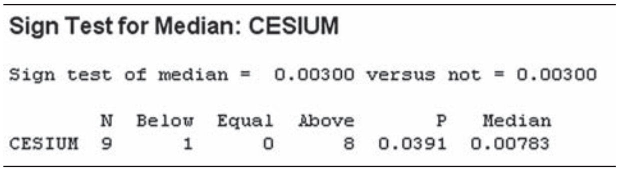
14.105 Radioactive lichen. Refer to the Lichen Radionuclide Baseline Research project to monitor the level of radioactivity in lichen, Exercise 8.71 (p. 379). Recall that University of Alaska researchers collected 9 lichen specimens and measured the amount of the radioactive element cesium-137 (in microcuries per milliliter) in each specimen. (The natural logarithms of the data values, saved in the **LICHEN** file, are listed in the next table.) In Exercise 8.71, you used the t-statistic to test whether the mean cesium amount in lichen differs from $\mu = .003$ microcurie per milliliter. Use the MINITAB printout (top of page 14-53) to conduct an alternative nonparametric test at $\alpha = .10$. Does the result agree with that of the t-test from Exercise 8.71? [*Note:* The values in the table were converted back to microcuries per milliliter to perform the analysis.]

Data for Exercise 14.105

Location			
Bethel	−5.50	−5.00	
Eagle Summit	−4.15	−4.85	
Moose Pass	−6.05		
Turnagain Pass	−5.00		
Wickersham Dome	−4.10	−4.50	−4.60

Based on Lichen Radionuclide Baseline Research Project, 2003, p. 25. *Orion*, University of Alaska–Fairbanks.

MINITAB output for Exercise 14.105



```
Sign Test for Median: CESIUM

Sign test of median =  0.00300 versus not = 0.00300

         N  Below  Equal  Above       P   Median
CESIUM   9      1      0      8  0.0391  0.00783
```

14.106 Social reinforcement of exercise. Two University of Georgia researchers studied the effect of social reinforcement on the duration of exercise in adolescents with moderate mental retardation (*Clinical Kinesiology*, Spring 1995). Eleven adolescents with IQs ranging from 32 to 61 were divided into two groups. All participated in a six-week exercise program. Group A (4 subjects) received verbal and social reinforcement during the program, while Group B (7 subjects) received verbal and social reinforcement and kept a self-record of individual performances. The researchers theorized that Group B subjects would exercise for longer periods than Group A. Upon completion of the exercise program, all 11 subjects participated in a run/walk "race" in which the goal was to complete as many laps as possible during a 15-minute period. The number of laps completed (to the nearest quarter lap) was used as a measure of the duration of exercise.

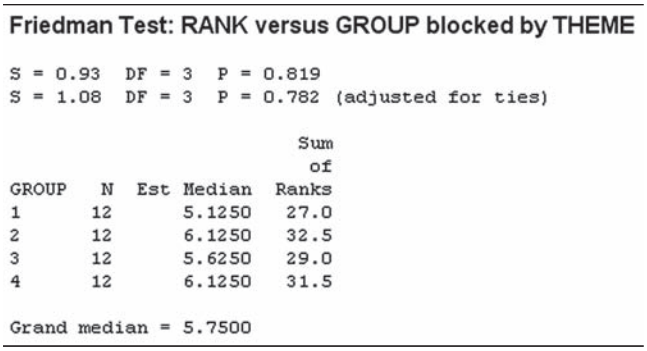
a. Specify the null and alternative hypotheses for a nonparametric analysis of the data.

b. The Kruskal-Wallis H-test was applied to the data. The researchers reported the test statistic as $H = 5.1429$ and the observed significance level of the test as p-value $= .0233$. Interpret these results.

c. Are the assumptions for the test carried out in part **b** satisfied? If not, propose an alternative nonparametric method for the analysis.

14.107 Thematic atlas topics. The regional atlas is an important educational resource that is updated on a periodic basis. One of the most critical aspects of a new atlas design is its thematic content. In a survey of atlas users (*Journal of Geography*, May/June 1995), a large sample of high school teachers in British Columbia ranked 12 thematic atlas topics for usefulness. The consensus rankings of the teachers (based on the percentage of teachers who responded that they "would definitely use" the topic) are saved in the **ATLAS** file. These teacher rankings were compared with the rankings a group of university geography alumni made three years earlier. Compare the distributions of theme rankings for the two groups with an appropriate nonparametric test. Use $\alpha = .05$. Interpret the results practically.

14.108 Thematic atlas topics. Refer to the *Journal of Geography*'s published rankings of regional atlas theme topics, presented in Exercise 14.107. In addition to high school teachers and university geography alumni, university geography students and representatives of the general public ranked the 12 thematic topics. The rankings of all four groups are saved in the **ATLAS2** file. A MINITAB analysis comparing the atlas theme-ranking distributions of the four groups is provided below.



```
Friedman Test: RANK versus GROUP blocked by THEME

S = 0.93  DF = 3  P = 0.819
S = 1.08  DF = 3  P = 0.782 (adjusted for ties)

                              Sum
                               of
GROUP   N  Est Median       Ranks
1      12      5.1250        27.0
2      12      6.1250        32.5
3      12      5.6250        29.0
4      12      6.1250        31.5

Grand median = 5.7500
```

Based on C. P. Keller, et al. "Planning the next generation of regional atlases: Input from educators." *Journal of Geography*, Vol 94, No. 3, May/June 1995, p. 413 (Table 1).

a. Locate the rank sums on the printout.

b. Use the rank sums to find the Friedman F_r-statistic.

c. Locate the test statistic and the associated p-value on the printout.

d. Conduct the test and state the conclusion in the words of the problem.

14.109 Feeding habits of fish. Refer to the *Brain and Behavior Evolution* (Apr. 2000) study of the feeding behavior of blackbream fish, presented in Exercise 2.150 (p. 89). Recall that the zoologists recorded the number of aggressive strikes of two blackbream feeding at the bottom of an aquarium in the 10-minute period following the addition of food. The following table lists the weekly number of strikes and age of the fish (in days). These data are saved in the **BLACKBREAM** file.

Week	Number of Strikes	Age of Fish (days)
1	85	120
2	63	136
3	34	150
4	39	155
5	58	162
6	35	169
7	57	178
8	12	184
9	15	190

Based on Shand, J., et al. "Variability in the location of the retinal ganglion cell area centralis is correlated with ontogenetic changes in feeding behavior in the Blackbream, *Acanthopagrus 'butcher.'" Brain and Behavior,* Vol. 55, No. 4, Apr. 2000 (Figure H).

a. Find Spearman's correlation coefficient relating number of strikes (y) to age of fish (x).

b. Conduct a nonparametric test to determine whether number of strikes (y) and age (x) are negatively correlated. Test using $\alpha = .01$.

14.110 Organizational use of the Internet. Researchers from the United Kingdom and Germany attempted to develop a theoretically grounded measure of organizational Internet use (OIU) and published their results in *Internet Research* (Vol. 15, 2005). Using data collected from a sample of 77 Web sites, they investigated the link between OIU level (measured on a seven-point scale) and several observation-based indicators. Spearman's rank correlation coefficient (and associated *p*-values) for several indicators are shown in the next table.

Indicator	Correlation with OIU Level	
	r_s	*p*-value
Navigability	.179	.148
Transactions	.334	.023
Locatability	.590	.000
Information richness	−.115	.252
Number of files	.114	.255

Based on Brock, J. K., and Zhou, Y. "Organizational use of the internet." *Internet Research,* Vol. 15, No. 1, 2005 (Table IV).

a. Interpret each of the values of r_s given in the table.

b. Interpret each of the *p*-values given in the table. (Use $\alpha = .10$ to conduct each test.)

Applying the Concepts—Intermediate

14.111 Agent Orange and Vietnam Vets. Agent Orange, the code name for a herbicide developed for the U.S. armed forces in the 1960s, was found to be extremely contaminated with TCDD, or dioxin. During the Vietnam War, an estimated 19 million gallons of Agent Orange was used to destroy the dense plant and tree cover of the Asian jungle. As a result of this exposure, many Vietnam veterans have dangerously high levels of TCDD in their blood and adipose (fatty) tissue. A study published in *Chemosphere* (Vol. 20, 1990) reported on the TCDD levels of 20 Massachusetts Vietnam vets who were possibly exposed to Agent Orange. The TCDD amounts (measured in parts per trillion) in both plasma and fat tissue of the 20 vets are listed in the table in the previous column. The data are saved in the **TCDD** file.

Vet	Fat	Plasma
1	4.9	2.5
2	6.9	3.5
3	10.0	6.8
4	4.4	4.7
5	4.6	4.6
6	1.1	1.8
7	2.3	2.5
8	5.9	3.1
9	7.0	3.1
10	5.5	3.0
11	7.0	6.9
12	1.4	1.6
13	11.0	20.0
14	2.5	4.1
15	4.4	2.1
16	4.2	1.8
17	41.0	36.0
18	2.9	3.3
19	7.7	7.2
20	2.5	2.0

Based on Schecter, A., et al. "Partitioning of 2,3,7,8-chlorinated dibenzo-*p*-dioxins and dibenzofurans between adipose tissue and plasma lipid of 20 Massachusetts Vietnam veterans." *Chemosphere,* Vol. 20, Nos. 7–9, 1990, pp. 954–955 (Tables I and II).

a. Medical researchers consider a TCDD level of 3 parts per trillion (ppt) to be dangerously high. Do the data provide evidence (at $\alpha = .05$) to indicate that the median level of TCDD in the fat tissue of Vietnam vets exceeds 3 ppt?

b. Repeat part **a** for plasma.

c. Medical researchers also are interested in comparing the TCDD levels in fat tissue and plasma for Vietnam veterans. Specifically, they want to determine whether the distribution of TCDD levels in fat is shifted above or below the distribution of TCDD levels in plasma. Conduct this analysis (at $\alpha = .05$) and make the appropriate inference.

d. Find the rank correlation between the TCDD level in fat tissue and the TCDD level in plasma. Is there sufficient evidence (at $\alpha = .05$) of a positive association between the two TCDD measures?

14.112 Visual acuity of children. In a comparison of the visual acuity of deaf and hearing children, eye movement rates are taken on 10 deaf and 10 hearing children. The data are shown in the accompanying table and saved in the **EYEMOVE** file. A clinical psychologist believes that deaf children have greater visual acuity than hearing children. (The larger a child's eye movement rate, the more visual acuity the child possesses.)

Deaf Children		Hearing Children	
2.75	1.95	1.15	1.23
3.14	2.17	1.65	2.03
3.23	2.45	1.43	1.64
2.30	1.83	1.83	1.96
2.64	2.23	1.75	1.37

a. Use a nonparametric procedure to test the psychologist's claim at $\alpha = .05$.

b. Conduct the test by using the large-sample approximation for the nonparametric test. Compare the results with those found in part **a.**

14.113 Patent infringement case. Refer to the *Chance* (Fall 2002) study of a patent infringement case brought against Intel Corp., presented in Exercise 9.22 (p. 426). Recall that the case rested on whether a patent witness's signature was written on top of key text in a patent notebook or under the key text. Using an X-ray beam, zinc measurements were taken at several spots on the notebook page. The zinc measurements for three notebook locations—on a text line, on a witness line, and on the intersection of the witness and text lines—are reproduced in the following table and saved in the **PATENT** file.

Text line:	.335	.374	.440			
Witness line:	.210	.262	.188	.329	.439	.397
Intersection:	.393	.353	.285	.295	.319	

a. Why might the Student's *t*-procedure you applied in Exercise 9.22 be inappropriate for analyzing these data?

b. Use a nonparametric test (at $\alpha = .05$) to compare the distribution of zinc measurements for the text line with the distribution for the intersection.

c. Use a nonparametric test (at $\alpha = .05$) to compare the distribution of zinc measurements for the witness line with the distribution for the intersection.

d. Use a nonparametric test (at $\alpha = .05$) to compare the zinc measurements for all three notebook locations.

e. From the results you obtained in parts **b–d,** what can you infer about the mean zinc measurements at the three notebook locations?

14.114 Hematology tests on workers. The accompanying table (saved in the **LYMPHO** file) lists the lymphocyte count results from hematology tests administered to a sample of 50 West Indian or African workers. Test (at $\alpha = .05$) the hypothesis that the median lymphocyte count of all West Indian or African workers exceeds 20.

14	28	11
15	17	25
19	14	30
23	8	32
17	25	17
20	37	22
21	20	20
16	15	20
27	9	20
34	16	26
26	18	40
28	17	22
24	23	61
26	43	12
23	17	20
9	23	35
18	31	

Based on Royston, J. P. "Some techniques for assessing multivariate normality based on the Shapiro-Wilk W." *Applied Statistics*, Vol. 32, No. 2, pp. 121–133.

14.115 Preventing metal corrosion. Corrosion of different metals is a problem in many mechanical devices. Three sealers used to help retard the corrosion of metals were tested to see whether there were any differences among them. Samples of 10 different metal compositions were treated with each of the three sealers, and the amount of corrosion was measured after exposure to the same environmental conditions for one month. The data are given in the table and saved in the **CORRODE** file. Is there any evidence of a difference in the probability distributions of the amounts of corrosion among the three types of sealer? Use $\alpha = .05$.

		Sealer	
Metal	1	2	3
1	4.6	4.2	4.9
2	7.2	6.4	7.0
3	3.4	3.5	3.4
4	6.2	5.3	5.9
5	8.4	6.8	7.8
6	5.6	4.8	5.7
7	3.7	3.7	4.1
8	6.1	6.2	6.4
9	4.9	4.1	4.2
10	5.2	5.0	5.1

14.116 Aggressiveness of twins. Twelve sets of identical twins are given psychological tests to determine whether the firstborn of the twins tends to be more aggressive than the secondborn. The test scores are shown in the accompanying table, where the higher score indicates greater aggressiveness. Do the data (saved in the **AGGTWINS** file) provide sufficient evidence (at $\alpha = .05$) to indicate that the firstborn of a pair of twins is more aggressive than the other?

Set	Firstborn	Secondborn
1	86	88
2	71	77
3	77	76
4	68	64
5	91	96
6	72	72
7	77	65
8	91	90
9	70	65
10	71	80
11	88	81
12	87	72

14.117 Word association study. Three lists of words, representing three levels of abstractness, are randomly assigned to 21 experimental subjects so that 7 subjects receive each list. The subjects are asked to respond to each word on their list with as many associated words as possible within a given period. A subject's score is the total number of word associations, summing over all words in the list. Scores for each list are given in the accompanying table and saved in the **WORDLIST** file. Do the data provide sufficient evidence to indicate a difference (shift in location) between at least two of the probability distributions of the numbers of word associations that subjects can name for the three lists? Use $\alpha = .05$.

List 1	List 2	List 3
48	41	18
43	36	42
39	29	28
57	40	38
21	35	15
47	45	33
58	32	31

14.118 Eye pupil size and deception. An experiment was designed to study whether eye pupil size is related to a person's attempt at deception. Eight students were asked to respond verbally to a series of questions. Before the questioning began, the size of one of each student's pupils was noted and the students were instructed to answer some of the questions dishonestly. (The number of questions answered dishonestly was left to individual choice.) During questioning, the percentage increase in pupil size was recorded. Each student was then given a deception score based on the proportion of questions answered dishonestly. (High scores indicate a large number of deceptive responses.) The results are shown in the table (p. 14-56) and saved in the **DECEPEYE** file. Can you conclude that the percentage increase in eye pupil size is positively correlated with deception score? Use $\alpha = .05$.

Data for Exercise 14.118

Student	Deception Score	Percentage Increase in Pupil Size
1	87	10
2	63	6
3	95	11
4	50	7
5	43	0
6	89	15
7	33	4
8	55	5

14.119 Media coverage of the 9–11 attacks and public opinion. The terrorist attacks of September 11, 2001, and related events (e.g., the war in Iraq) have, and continue to receive, much media coverage. How has this coverage influenced the American public's concern about terrorism? This was the topic of research conducted by journalism professors at the University of Missouri (*International Journal of Public Opinion*, Winter 2004). Using random-digit dialing, they conducted a telephone survey of 235 Americans. Each person was asked to rate, on a scale of 1 to 5, his or her level of concern about each of eight topics: a long war, future terrorist attacks, the effect on the economy, the Israel-Palestine conflict, biological threats, air travel safety, war protests, and Afghan civilian deaths. The eight scores were summed to obtain a "public agenda" score. The respondents were also asked how many days per week they read the newspaper, watch the local television news, and watch national television news. The responses to these three questions were also summed to obtain a "media agenda" score. The researchers hypothesized that the public agenda score would be positively related to the media agenda score.

a. Spearman's rank correlation between the two scores was computed to be $r_s = .643$. Give a practical interpretation of this value.

b. The researchers removed the "length of war" question from the data and recomputed the "public agenda" score. Spearman's rank correlation between the public agenda and media agenda scores was then calculated as $r_s = .714$. Interpret this result.

c. Refer to part **b.** Conduct Spearman's test for positive rank correlation at $\alpha = .01$.

14.120 Fluoride in drinking water. Many water treatment facilities supplement the natural fluoride concentration with hydrofluosilicic acid in order to reach a target concentration of fluoride in drinking water. Certain levels are thought to enhance dental health, but very high concentrations can be dangerous. Suppose that one such treatment plant targets .75 milligram per liter (mg/L) for its water. The plant tests 25 samples each day to determine whether the median level differs from the target.

a. Set up the null and alternative hypotheses.

b. Set up the test statistic and rejection region, using $\alpha = .10$.

c. Explain the implication of a Type I error in the context of this application. A Type II error.

d. Suppose that one day's samples result in 18 values that exceed .75 mg/L. Conduct the test and state the appropriate conclusion in the context of this application.

e. When it was suggested to the plant's supervisor that a *t*-test should be used to conduct the daily test, she replied that the probability distribution of the fluoride concentrations was "heavily skewed to the right." Show graphically what she meant by this, and explain why this is a reason to prefer the sign test to the *t*-test.

14.121 Forums for tax litigation. In disagreements between the Internal Revenue Service (IRS) and taxpayers that end up in litigation, taxpayers are permitted by law to choose the court forum. Three trial courts are available: (1) U.S. Tax Court, (2) Federal District Court, and (3) U.S. Claims Court. Each court possesses different requirements and restrictions that make the choice an important one for the taxpayer. A study of taxpayers' choice of forum in litigating tax issues was published in the *Journal of Applied Business Research* (Fall 1996). In a random sample of 161 litigated tax disputes, the researchers measured the taxpayers' choice of forum (Tax, District, or Claims Court) and tax deficiency (i.e., the disputed amount, in dollars). One of the objectives of the study was to determine those factors taxpayers consider important in their choice of forum. If tax deficiency (called DEF by the researchers) is an important factor, then the mean DEF values for the three tax courts should be significantly different.

a. The researchers applied a nonparametric test rather than a parametric test to compare the DEF distributions of the three tax litigation forums. Give a plausible reason for their choice.

b. What nonparametric test is appropriate for this analysis? Explain.

c. The accompanying table summarizes the data analyzed by the researchers. Use the information in the table to compute the appropriate test statistic.

d. The observed significance level of the test was reported as p-value $= .0037$. Interpret this result fully.

Court Selected by Taxpayer	Sample Size	Sample Mean DEF	Rank Sum of DEF Values
Tax	67	$80,357	5,335
District	57	74,213	3,937
Claims	37	184,648	3,769

Based on Billing, B. A., Green, B. P., and Volz, W. H. "Selection of forum for litigated tax issues." *Journal of Applied Business Research*, Vol. 12, No. 4, Fall 1996, p. 38 (Table 2).

14.122 Ranking wines. Two expert wine tasters were asked to rank six brands of wine. Their rankings are shown in the following table and saved in the **WINETASTE** file. Do the data indicate a positive correlation in the rankings of the two experts? Test, using $\alpha = .10$.

Brand	Expert 1	Expert 2
A	6	5
B	5	6
C	1	2
D	3	1
E	2	4
F	4	3

14.123 Al Qaeda attacks on the United States. Refer to the *Studies in Conflict & Terrorism* (Vol. 29, 2006) analysis of recent incidents involving suicide terrorist attacks, presented in Exercise 2.173 (p. 98). The data in the accompanying table (saved in the **ALQAEDA** file) are the number of individual suicide bombings attacks for each in a sample of 21 recent incidents involving an attack against the United States by the Al Qaeda terrorist group. A counterterrorism expert claims that more than half of all Al Qaeda attacks against the United States involve two or fewer suicide bombings. Is there evidence to support this claim? Test at $\alpha = .05$.

1	1	2	1	2	4	1	1	1	1	2
3	4	5	1	1	1	2	2	2	1	

Source: Moghadam, A. "Suicide terrorism, occupation, and the globalization of martyrdom: A critique of *Dying to Win*," *Studies in Conflict & Terrorism*, Vol. 29, No. 8, 2006 (Table 3).

Critical Thinking Challenge

14.124 Self-managed work teams and family life. Refer to the *Quality Management Journal* (Summer 1995) study of self-managed work teams (SMWTs), presented in Exercise 9.139 (p. 467). Recall that the researchers investigated the connection between SMWT work characteristics and workers' perceptions of positive spillover into family life. (One group of workers reported positive spillover of work skills to family life, while another group did not report positive work spillover.) The data collected on 114 AT&T employees, saved in the **SPILLOVER** file, are described in the accompanying table. In Exercise 9.139, you compared the two groups of workers on each characteristic, using the parametric methods of Chapter 9. Reanalyze the data, this time using nonparametrics. Are the job-related characteristics most highly associated with positive work spillover the same as those identified in Exercise 9.139? Comment on the validity of the parametric and nonparametric results.

Characteristic	Variable
Information Flow	Use of creative ideas (seven-point scale)
Information Flow	Utilization of information (seven-point scale)
Decision Making	Participation in decisions regarding personnel matters (seven-point scale)
Job	Good use of skills (seven-point scale)
Job	Task identity (seven-point scale)
Demographic	Age (years)
Demographic	Education (years)
Demographic	Gender (male or female)

Activity Comparing Supermarket Prices (*continued*)

In Chapters 10 and 14, we discussed two methods of analyzing a randomized block design. When the populations have normal probability distributions and their variances are equal, we can employ the analysis of variance described in Chapter 10. Otherwise, we can use the Friedman F_r-test.

In the Activity of Chapter 10, we asked you to conduct a randomized block design to compare supermarket prices and to use an analysis of variance to interpret the data. Now use the Friedman F_r-test to compare the supermarket prices.

How do the results of the two analyses compare? Explain the similarity (or lack of similarity) between the two results.

References

Conover, W. J. *Practical Nonparametric Statistics*, 2nd ed. New York: Wiley, 1980.

Daniel, W. W. *Applied Nonparametric Statistics*, 2nd ed. Boston: PWS-Kent, 1990.

Dunn, O. J. "Multiple comparisons using rank sums." *Technometrics*, Vol. 6, 1964.

Friedman, M. "The use of ranks to avoid the assumption of normality implicit in the analysis of variance." *Journal of the American Statistical Association*, Vol. 32, 1937.

Gibbons, J. D. *Nonparametric Statistical Inference*, 4th ed. Boca Raton, FL: CRC Press, 2003.

Hollander, M., and Wolfe, D. A. *Nonparametric Statistical Methods*. 2nd ed. New York: Wiley, 1999.

Kruskal, W. H., and Wallis, W. A. "Use of ranks in one-criterion variance analysis." *Journal of the American Statistical Association*, Vol. 47, 1952.

Lehmann, E. L. *Nonparametrics: Statistical Methods Based on Ranks*. (revised). New York: Springer, 2006.

Marascuilo, L. A., and McSweeney, M. *Nonparametric and Distribution-Free Methods for the Social Sciences*. Monterey, CA: Brooks/Cole, 1977.

Wilcoxon, F., and Wilcox, R. A. "Some rapid approximate statistical procedures." The American Cyanamid Co., 1964.

Using Technology

MINITAB: Nonparametric Tests

Sign Test

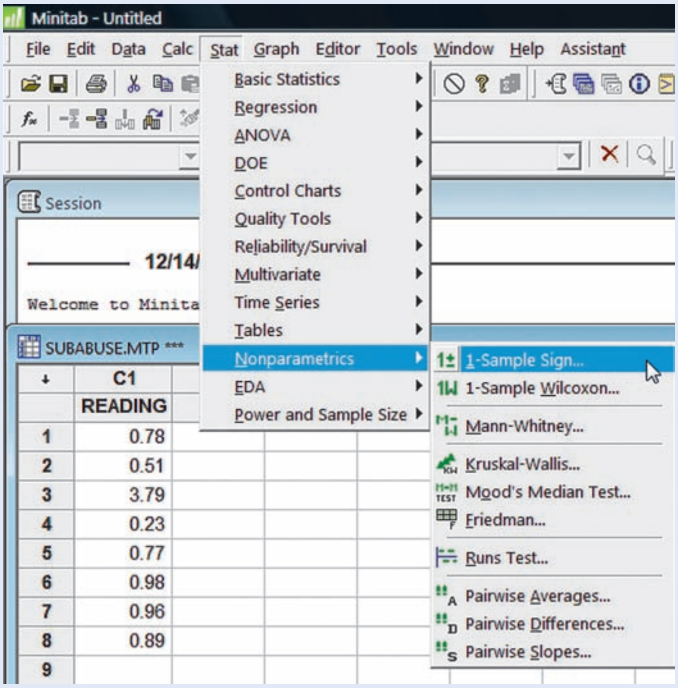
Step 1 Access the MINITAB worksheet file with the sample data. It should contain a single quantitative variable.

Step 2 Click on the "Stat" button on the MINITAB menu bar, then click on "Nonparametrics" and "1-Sample Sign," as shown in Figure 14.M.1.

Step 3 On the resulting dialog box (see Figure 14.M.2), enter the quantitative variable to be analyzed in the "Variables" box.

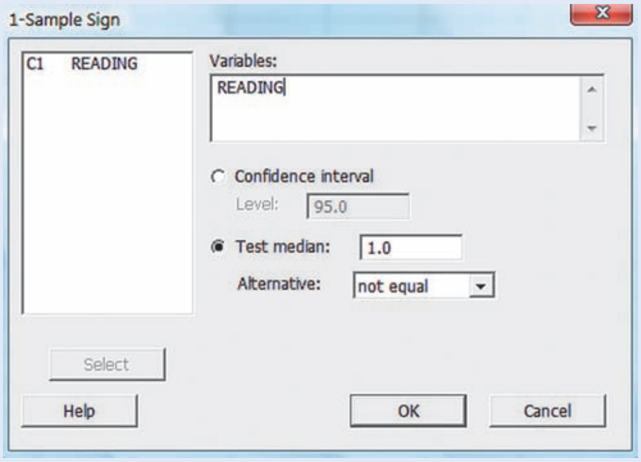
Step 4 Select the "Test median" option and specify the hypothesized value of the median and the form of the alternative hypothesis ("not equal", "less than", or "greater than").

Step 5 Click "OK" to generate the MINITAB printout.



↓	C1
	READING
1	0.78
2	0.51
3	3.79
4	0.23
5	0.77
6	0.98
7	0.96
8	0.89
9	

Figure 14.M.1
MINITAB nonparametric menu options



Variables: READING

○ Confidence interval
Level: 95.0

● Test median: 1.0
Alternative: not equal

Figure 14.M.2
MINITAB 1-sample sign dialog box

Rank Sum Test

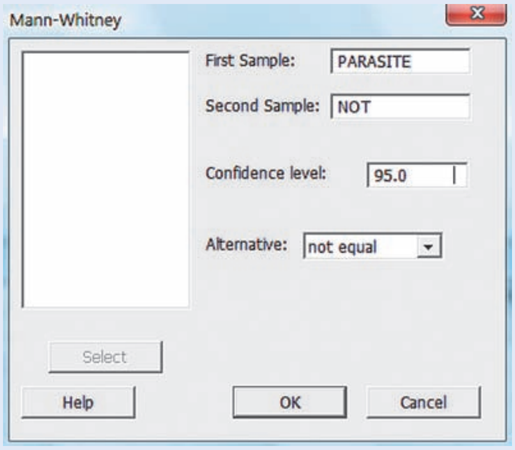
Step 1 Access the MINITAB worksheet file with the sample data. It should contain two quantitative variables, one for each of the two samples being compared.

Step 2 Click on the "Stat" button on the MINITAB menu bar, then click on "Nonparametrics" and "Mann-Whitney" (see Figure 14.M.1).

Step 3 On the resulting dialog box (see Figure 14.M.3), specify the variable for the first sample in the "First Sample" box and the variable for the second sample in the "Second Sample" box.

Step 4 Specify the form of the alternative hypothesis ("not equal," "less than," or "greater than").

Step 5 Click "OK" to generate the MINITAB printout.



First Sample: PARASITE
Second Sample: NOT
Confidence level: 95.0
Alternative: not equal

Figure 14.M.3
MINITAB Mann-Whitney (rank sum) test dialog box

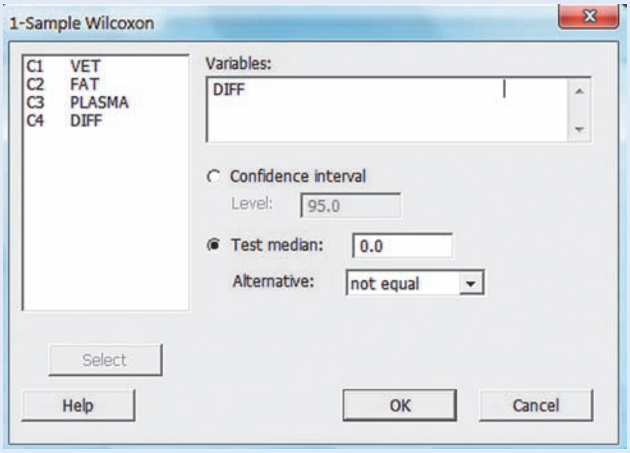
Signed Rank Test

Step 1 Access the MINITAB worksheet file with the matched-pairs data. It should contain two quantitative variables, one for each of the two groups being compared.

Step 2 Compute the difference between these two variables and save it in a column on the worksheet. (Use the "Calc" button on the MINITAB menu bar.)

Step 3 Click on the "Stat" button on the MINITAB menu bar, then click on "Nonparametrics" and "1-Sample Wilcoxon" (see Figure 14.M.1).

Step 4 On the resulting dialog box (see Figure 14.M.4), enter the variable representing the paired differences in the "Variables" box.



Variables: DIFF

○ Confidence interval
Level: 95.0

● Test median: 0.0
Alternative: not equal

Figure 14.M.4
MINITAB 1-sample Wilcoxon (signed rank) test dialog box

Step 5 Select the "Test median" option and specify the hypothesized value of the median as "0." Select the form of the alternative hypothesis ("not equal," "less than," or "greater than").

Step 6 Click "OK" to generate the MINITAB printout.

Kruskal-Wallis Test

Step 1 Access the MINITAB worksheet file that contains the completely randomized design data. It should contain one quantitative variable (the response, or dependent, variable) and one factor variable with at least two levels.

Step 2 Click on the "Stat" button on the MINITAB menu bar, then click on "Nonparametrics" and "Kruskal-Wallis" (see Figure 14.M.1).

Step 3 On the resulting dialog box (see Figure 14.M.5), specify the response variable in the "Response" box and the factor variable in the "Factor" box.

Step 4 Click "OK" to generate the MINITAB printout.

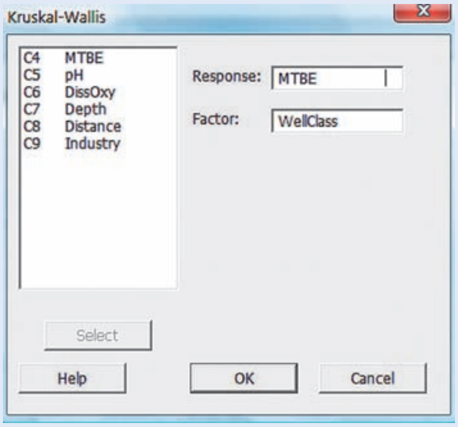

Response: MTBE
Factor: WellClass

Figure 14.M.5
MINITAB Kruskal-Wallis test dialog box

Friedman Test

Step 1 Access the MINITAB spreadsheet file that contains the randomized block design data. It should contain one quantitative variable (the response, or dependent, variable) and one factor variable and one blocking variable.

Step 2 Click on the "Stat" button on the MINITAB menu bar, then click on "Nonparametrics" and "Friedman" (see Figure 14.M.1).

Step 3 On the resulting dialog box (see Figure 14.M.6), specify the response, treatment, and blocking variables in the appropriate boxes.

Step 4 Click "OK" to generate the MINITAB printout.

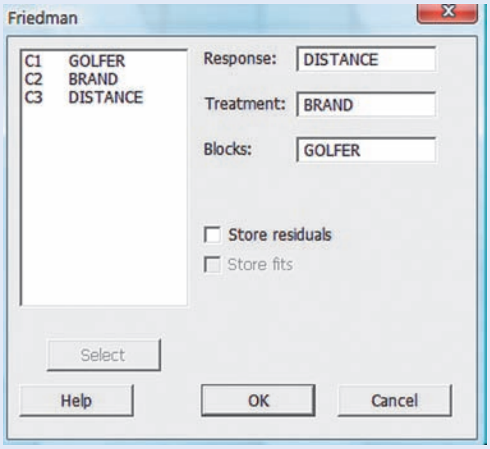

Response: DISTANCE
Treatment: BRAND
Blocks: GOLFER
☐ Store residuals
☐ Store fits

Figure 14.M.6
MINITAB Friedman test dialog box

Rank Correlation

Step 1 To obtain Spearman's rank correlation coefficient in MINITAB, you must first rank the values of the two quantitative variables of interest. Click the "Calc" button on the MINITAB menu bar and create two additional columns, one for the ranks of the *x*-variable and one for the ranks of the *y*-variable. (Use the "Rank" function on the MINITAB calculator as shown in Figure 14.M.7.)

Step 2 Click on the "Stat" button on the main menu bar, then click on "Basic Statistics" and "Correlation."

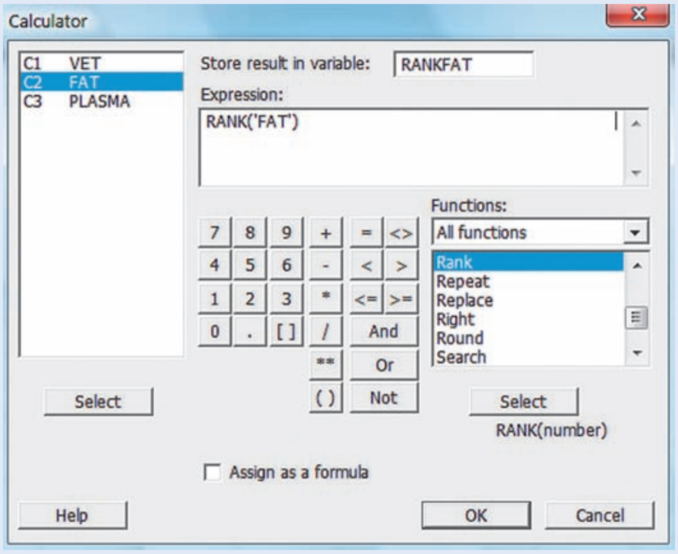

Store result in variable: RANKFAT
Expression: RANK('FAT')
Functions: All functions
☐ Assign as a formula

Figure 14.M.7
MINITAB calculator menu screen

Step 3 On the resulting dialog box (see Figure 14.M.8), enter the ranked variables in the "Variables" box and unselect the "Display *p*-values" option.

Step 4 Click "OK" to obtain the MINITAB printout. (You will need to look up the critical value of Spearman's rank correlation to conduct the test.)

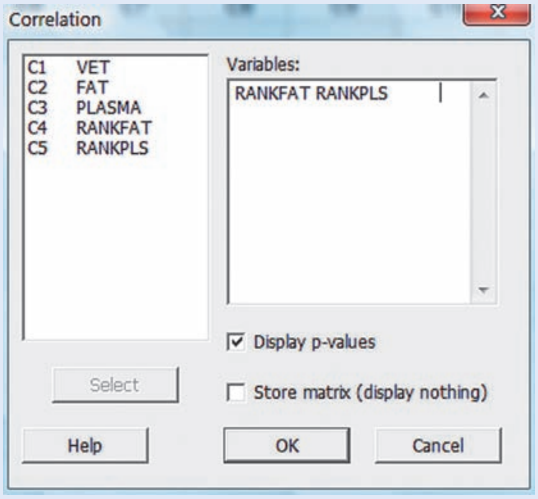

Variables: RANKFAT RANKPLS
☑ Display p-values
☐ Store matrix (display nothing)

Figure 14.M.8
MINITAB correlation dialog box