CHAPTER 14

Nonparametric Methods

CONTENTS

■ Example 14.1: Quality Control

A large company manufacturing rubber windshield wipers for use on automobiles was involved in a research project for improving the quality of their standard wiper. An engineer developed four types of chemical treatments that were thought to increase the lifetime of the wiper. An experiment was performed in which samples of 15 blades were treated with each of these chemical treatments and measured for the amount of wear (in mm) over a period of 2 h on a test machine. The results are shown in Table 14.1.

An analysis of variance was performed (see Chapter 6) to test for difference in average wear over the four treatments. The results are shown at the bottom of Table 14.1. The engineer, however, did not believe that the assumption of normality was valid (see Section 6.4). That is, she suspected that the error terms were probably distributed more like a uniform distribution. The histogram of the residuals given in Fig. 14.1 appears to justify the concern of the engineer.

Statistical Methods, DOI: 10.1016/B978-0-12-374970-3.00014-7

Table 14.1 Wear Data for Window Wipers for Four Teatments (in mm)

TREAT = 1	TREAT = 2	TREAT = 3	TREAT = 4
11.5	14.3	13.7	17.0
11.5	12.7	14.8	14.7
10.1	14.3	13.5	16.5
11.6	13.1	14.2	15.5
11.2	14.3	14.7	14.2
10.6	14.7	14.4	16.6
11.2	12.5	14.2	14.5
11.5	14.0	14.8	16.6
10.3	15.0	14.0	14.9
11.8	13.2	14.8	16.5
11.3	13.9	15.0	16.5
10.1	14.9	13.2	14.2
10.9	12.6	14.2	16.4
11.2	14.2	13.3	14.6
10.4	12.8	13.5	15.3

ANOVA for Wear Data

Source	df	SS	*F*	*Pr* > *F*
Treat	3	165.3	88.65	0.0001
Error	56	34.8		
Total	59	200.11		

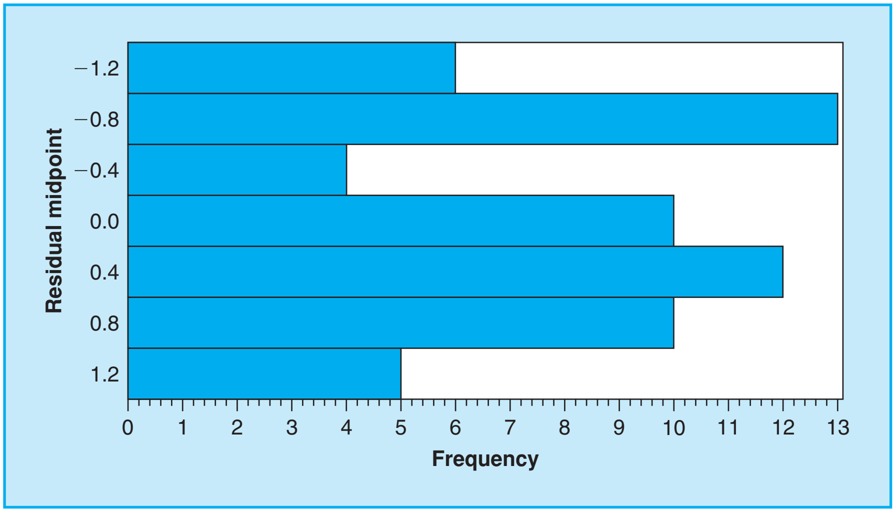


FIGURE 14.1

Histogram of Residuals.

To further check the assumption of normality, she performed a goodness of fit test and rejected the null hypothesis of normality (see Section 12.2). An approach for solving this problem is presented in the material covered in this chapter, and we will return to this example in Section 14.8. ■

14.1 INTRODUCTION

As Chapter 11 demonstrated, most of the statistical analysis procedures presented in Chapters 4 through 11 are based on the assumption that some form of linear model describes the behavior of a ratio or interval response variable. That is, the behavior of the response variable is approximated by a linear model and inferences are made on the parameters of that model. Because the primary focus is on the parameters, including those describing the distribution of the random error, statistical methods based on linear models are often referred to as "parametric" methods.

We have repeatedly noted that the correctness of an analysis depends to some degree on certain assumptions about the model. One major assumption is that the errors have a nearly normal distribution with a common variance, so that the normal, t, χ^2, and F distributions properly describe the distribution of the test statistics. This assures that the probabilities associated with the inferences are as stated by significance levels or p values. Fortunately, these methods are reasonably "robust" so that most estimates and test statistics give sufficiently valid results even if the assumptions on the model are not exactly satisfied (as they rarely are).

Obviously, there are situations in which the assumptions underlying an analysis are not satisfied and remedial methods such as transformations do not work, in which case there may be doubt as to whether the significance levels or confidence coefficients are really correct. Therefore, an alternative approach for which the correctness of the stated significance levels and confidence coefficients is not heavily dependent on rigorous distributional assumptions is needed. Such methods should not depend on the distribution of the random error nor necessarily make inferences on any particular parameter. Such procedures are indeed available and are generally called "nonparametric" or "distribution-free" methods. These procedures generally use simple, tractable techniques for obtaining exact error probabilities while not assuming any particular form of the model.

Many of the tests discussed in Chapter 12 may be considered nonparametric methods. For example, the contingency table analysis makes no assumptions about the underlying probability distribution. The p values for this test statistic were obtained by using a large sample χ^2 approximation. Obviously this type of methodology does not fit the "normal" theory of parametric statistics because the scale of measure used was at best nominal (categorical). As a matter of fact, one of the desirable characteristics of most nonparametric methods is that they do not require response variables to have an interval or ratio scale of measurement.

A brief introduction to the concept of nonparametric statistics was presented in Section 3.5 where we noted that a wide spectrum of methods is available when the assumptions on the model are not fulfilled. Examples of nonparametric tests were presented in Sections 4.5 and 5.6 where the tests were done on medians. In both of these examples, we noted that the distributions of the observations were skewed, therefore making the "parametric" t tests suspect. In this chapter, we present some additional examples of nonparametric methods.

Most classical nonparametric techniques are based on two fundamental principles. The first is the use of the ranks of the original data. The ranks (1 for the smallest, up to n for the largest) are not at all affected by the presence of outliers or skewed distributions. In fact, barring ties, the collection of ranks will be the same for any set of n data values! Hence, distributions of test statistics based on ranks do not depend on precise distributional assumptions such as normality. The second fundamental principle is that of a randomization test, as we illustrated in the discussion of Fisher's exact test (Section 12.4). This provides a way of assessing significance, again without making detailed assumptions regarding distributions. Each of these principles is discussed next.

14.1.1 Ranks

The methods presented in Chapter 12 were used to analyze response variables that are categorical in nature; that is, they are measured in a nominal scale. Of course, data of the higher order scales can be artificially converted to nominal scale, simply by grouping observations. That is, ordinal data and interval or ratio scale measurements can be "categorized" into nominal-looking data. Interval or ratio measurements can also be changed into ordinal scale measurements by simply ranking the observations.[1] A number of nonparametric statistical methods are, in fact, based on ranks. The methods presented in this chapter are mostly of this type. These methods work equally well on variables originally measured in the ordinal scale as well as on variables measured on ratio or interval scales and subsequently converted to ranks.

Ranks may actually be preferable to the actual data in many cases. For example, if numerical measurements assigned to the observations have no meaning by themselves, but only have meaning in a comparison with other observations, then the ranks convey all the available information. An example of this type of variable is the "scores" given to individual performers in athletic events such as gymnastics or diving. In such situations the measurements are essentially ordinal in scale from the beginning. Even when measurements are actually in the interval scale, the underlying probability distribution may be intractable. That is, we are not able to use the additional information in a statistical inference because we cannot evaluate the resulting sampling distribution. In this case, switching to ranks allows us to use the relatively simple distributions associated with ranks.

[1] This was illustrated in Chapter 1.

To convert interval or ratio data into ranks, we must have a consistent procedure for ordering data. This ordering is called ranking and the ranking procedure normally used in statistics orders data from "smallest" to "largest" with a "1" being the smallest and an "n" being the largest (where n is the size of the data set being ranked). This ranking does not necessarily imply a numerical relationship, but may represent another ordinality such as "good," "better," and "best," or "sweet" to "sour," "preferred" to "disliked," or some other relative ranking. In other words, any ratio, interval, or ordinal variable can usually be converted to ranks.

As indicated in Section 1.3, a special problem in ranking occurs when there are "ties," that is, when a variable contains several identically recorded values. As a practical solution, ties are handled by assigning mean ranks to tied values. While the methodology of rank-based nonparametric statistics usually assumes no ties, a reasonably small number of ties have minimal effect on the usefulness of the resulting statistics.

14.1.2 Randomization Tests

To understand randomization tests, we must first recall how critical regions and p values are constructed for classic tests, such as the independent samples t test (Section 5.2). For the pooled t test, the test statistic is

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{s_p^2\,(1/n_1 + 1/n_2)}}.$$

Assume that we have a large basket filled with slips of paper, each with a number written on it. Assume further that these numbers come from a normal distribution. We draw a random sample of $n_1 + n_2$ observations from the basket, and randomly assign n_1 of them to group 1, and the others to group 2. For our data set, we calculate t and write down that value. Then we put those slips back, shake up the basket, and repeat the process a huge number of times. Afterward, we histogram the list of t values. The result should look very like the Student's t distribution with $n_1 + n_2 - 2$ degrees of freedom.

The process we have described mimics the situation when the null hypothesis for the independent samples t test is correct and the underlying assumptions are valid. Then the two samples are essentially being drawn from the same basket. Based on the empirical distribution shown in the histogram, we can judge whether a value of t from an actual data set is unusual or not. If our observed t is in the α most unusual region of the distribution, we would claim that our t is inconsistent with the assumption that the samples came from the same basket. Alternatively, we could calculate the proportion of values in our experiment that are as or more unusual than the observed $|t|$, which would give us an empirical estimate of the p value. We could do this for any choice of test statistic, but the t statistic is particularly good at detecting differences in population means.

Under the assumption of normality, the distribution of t is known mathematically. This saves us an extremely tedious process. But what if the parent distribution, that is, the distribution of the values on the slips of paper, is not normal? There are several options. If it is reasonable that the values follow some other parametric distribution, for example the Poisson, then we might computerize the process described earlier. This is called a Monte Carlo study.

If we do not know what the parent distribution is like, it is reasonable to use the data itself as a guess for that distribution. In a randomization test, we would fill a basket with the exact values seen in our combined data set. Then we would write down all the possible ways to split those values into group 1 (with n_1 values) and group 2 (with n_2 values), calculating and recording the values of the test statistic. Our samples would be constructed without replacement, so that we are enumerating all the ways to permute the data into two separate groups of the specified sizes. For this reason, randomization tests are also known as **permutation tests**. The number of possible splits of the data can be calculated using the formula for combinations given in Section 2.3.

The development of nonparametric statistics predates the advent of modern computers. Naturally, the focus was on test statistics that were quick to compute. If those statistics only depended on the ranks, then the enumeration could be done just once for any given n_1 and n_2, because all data sets of that size with no ties would have the same ranks. Hence, randomization tests came to be thought of as a natural partner with the use of ranks. In fact, however, randomization tests can be developed for many test statistics whether or not they are based on ranks.

In practice, the enumeration of all the possible splits of the data is too time-consuming. For small samples with no ties, the early nonparametric statisticians developed tables of the distributions. Some of these are discussed in this chapter. For larger samples, asymptotic approximations were developed. These are also presented for a few of the most common nonparametric tests. Most data will contain ties, and then the tables are no longer accurate, though they can be approximately correct if the number of ties is small. Fortunately, most statistical software will carry out the enumeration for you, calculating an exact p value.

When the data sets become moderately large, even modern desktop computers will find the enumeration too time-consuming. Instead, the software will draw a large random sample of permutations, counting the number of times in the sample that the calculated test statistic is as or more extreme than that observed in the data. This gives an estimate of the p value. This is called an **approximate randomization test**.

We will comment on the randomization principles behind a few of the tests presented in later sections, but space limitations prevent a full development. For more information on randomization tests and their mathematical cousin, the bootstrap, see Higgins (2004).

14.1.3 Comparing Parametric and Nonparametric Procedures

Before presenting specific rank-based nonparametric test procedures, we must understand how parametric and nonparametric tests compare.

- *Power:* Conversion of a set of interval or ratio values to ranks involves a loss of information. This loss of information usually results in a loss of power in a hypothesis test. As a result, most rank-based nonparametric tests will be less powerful when used on interval or ratio data, especially if the distributional assumptions of parametric tests are not severely violated. Of course, we cannot compare parametric and nonparametric tests if the observed variables are already nominal or ordinal in scale.
- *Interpretation:* By definition, nonparametric tests do not state hypotheses in terms of parameters; hence inferences for such tests must be stated in other terms. For some of these tests the inference is on the median, but for others the tests may specify location or difference in distribution. Fortunately, this is not usually a problem, because the research hypothesis can be adjusted if a nonparametric test is to be used.
- *Application:* As will be seen, comprehensive rank-based methodology analogous to that for the linear model is not available.[2] Instead nonparametric tests are usually designed for a specific experimental situation. In fact, nonparametric methods are most frequently applied to small samples and simple experimental situations, where violations of assumptions are likely to have more serious consequences.
- *Computing:* For modern computers, the process of ranking is quite fast. The computing difficulty lies in the enumeration of the possible rearrangements under the null hypothesis. There are specially developed packages for nonparametric analysis based on full enumeration in small samples and approximate randomization in large samples (e.g., Resampling Stats or StatXact). Most multipurpose statistical software will implement this for at least the most common nonparametric techniques.
- *Robustness:* While few assumptions are made for nonparametric methods, these methods are not uniformly insensitive to all types of violations of assumptions nor are they uniformly powerful against all alternative hypotheses.
- *Experimental design:* The use of nonparametric methods does not eliminate the need for careful planning and execution of the experiment or other data accumulation. Good design principles are important regardless of the method of analysis.

[2] It has been suggested that converting the response variable to ranks and then performing standard linear model analyses will produce acceptable results. Of course, such analyses do not produce the usual estimates of means and other parameters. This is briefly discussed in Section 14.8; a more comprehensive discussion is found in Conover and Iman (1981).

The various restrictions and disadvantages of nonparametric methods would appear to severely limit their usefulness. This is not the case, however, and they should be used (and usually are) when the nature of the population so warrants. Additionally, many of the nonparametric methods are extremely easy to perform, especially on small data sets. They are therefore an attractive alternative for "on the spot" analyses. In fact, one of the earlier books on nonparametric methods is called *Some Rapid Approximate Statistical Procedures* (Wilcoxon and Wilcox, 1964).

The following sections discuss some of the more widely used rank-based nonparametric hypothesis testing procedures. In the illustration of these procedures, we use some examples previously analyzed with parametric techniques. The purpose of this is to compare the two procedures. Of course, in actual practice, only one method should be applied to any one set of data.

We emphasize that the methods presented in this chapter represent only a few of the available nonparametric techniques. If none of these are suitable, additional nonparametric or other robust methods may be found in the literature, such as in Huber (1981), or in texts on the subject, such as Conover (1999).

14.2 ONE SAMPLE

In Section 4.5 we considered an alternative approach to analyzing some income data that had a single extreme observation. This approach was based on the fact that the median is not affected by extreme observations. Recall that in this example we converted the income values into either a "success" (if above the specified median) or a "failure" (if below), and the so-called sign test was based on the proportion of successes. In other words, the test was performed on a set of data that were converted from the ratio to the nominal scale.

Of course the conversion of the variable from a ratio to a nominal scale with only two values implies a loss of information; hence, the resulting test is likely to have less power. However, converting a nominal variable to ranks preserves more of the information and thus a test based on ranks should provide more power. One such test is known as the Wilcoxon signed rank test.

The Wilcoxon signed rank test is used to test that a distribution is symmetric about some hypothesized value, which is equivalent to the test for location. We illustrate with a test of a hypothesized median, which is performed as follows:

1. Rank the magnitudes (absolute values) of the deviations of the observed values from the hypothesized median, adjusting for ties if they exist.
2. Assign to each rank the sign ($+$ or $-$) of the deviation (thus, the name "signed rank").
3. Compute the sum of positive ranks, $T(+)$, or negative ranks, $T(-)$, the choice depending on which is easier to calculate. The sum of $T(+)$ and $T(-)$ is $n(n+1)/2$, so either can be calculated from the other.
4. Choose the smaller of $T(+)$ and $T(-)$, and call this T.

5. Since the test statistic is the minimum of $T(+)$ and $T(-)$, the critical region consists of the left tail of the distribution, containing a probability of at most $\alpha/2$. For small samples ($n \leq 50$), critical values are found in Appendix Table A.9. If n is large, the sampling distribution of T is approximately normal with

$$\mu = n\,(n+1)/4, \quad \text{and}$$
$$\sigma^2 = n\,(n+1)(2n+1)/24,$$

which can be used to compute a z statistic for the hypothesis test.

■ Example 14.2

Example 4.7, particularly the data in Table 4.5, concerned a test for the mean family income of a neighborhood whose results were unduly influenced by an extreme outlier. A test for the median was used to overcome the influence of that observation. We now use that example to illustrate the Wilcoxon signed rank test. The hypothesis of interest is

H_0: the distribution on incomes is symmetric about 13.0,

with a two-tailed alternative,

H_1: the distribution is symmetric about some other value.

Table 14.2 Deviations from the Median and Signed Ranks

Obs	Diff	Signed Rank	Obs	Diff	Signed Rank
1	4.1	18	11	2.7	14
2	−0.3	−4.5	12	80.4	20
3	3.5	16	13	1.9	13
4	1.0	11	14	0.0	1
5	1.2	12	15	0.8	10
6	−0.7	−9	16	3.2	15
7	0.2	2.5	17	0.6	8
8	0.3	4.5	18	−0.2	−2.5
9	4.9	19	19	0.4	6
10	−0.5	−7	20	3.6	17

Solution

The deviations of the observed values from 13.0 (the specified H_0 value) are given in Table 14.2 in the column labeled “Diff,” followed by the signed ranks corresponding to the differences. Note that several ties are given average ranks, and that zero is arbitrarily given a positive sign. A quick inspection shows that there are fewer negative signed ranks so we first compute $T(-)$:

$$T(-) = 4.5 + 9 + 7 + 2.5 = 23.$$

The total sum of n ranks is $(n)(n + 1)/2$; hence, it follows that $T(+) + T(-) = (20)(21)/2 = 210$. Thus $T(+) = 210 - 23 = 187$. The test statistic is the smaller of the two, $T = T(-) = 23$. From Appendix Table A.9, using $n = 20$ and $\alpha = 0.01$, we see that the critical value is 37. We reject H_0 if the calculated value is less than 37; hence, we reject the hypothesis and conclude that the population is not symmetric about 13.0.

Alternately, we can use the large sample normal approximation. Under the null hypothesis, T is approximately normally distributed with

$$\mu = (20)(21)/4 = 105, \quad \text{and}$$
$$\sigma^2 = (20)(21)(41)/24 = 717.5;$$

hence $\sigma = 26.79$. These values are used to compute the test statistic

$$z = \frac{23 - 105}{26.79} = -3.06.$$

Using Appendix Table A.1, we find a (two-tailed) p value of approximately 0.002; hence, the null hypothesis is readily rejected. However, the sample is rather small; hence, the p value calculated from the large sample approximation should not be taken too literally.

The p value obtained for the sign test in Section 4.5 was 0.012. Thus, for $\alpha = 0.01$ the Wilcoxon signed rank test rejected the null hypothesis while the sign test did not.[3] ■

Some texts recommend discarding zero differences, such as the one arbitrarily assigned a positive value in Table 14.2. This discard is done before the ranking, and the test statistic computed using the remaining observations with the correspondingly smaller sample size. See Higgins (2004) for a discussion.

A popular application of the signed rank test is for comparing means from paired samples. In this application the differences between the pairs are computed as is done for the paired t test (Section 5.4). The hypothesis to be tested is that the distribution of differences is symmetric about 0.

[3] The problem as stated in Example 4.7 had a one-sided alternative while the procedure for the two-sided alternative is presented here since it is more general.

■ Example 14.3

To determine the effect of a special diet on activity in small children, 10 children were rated on a scale of 1 to 20 for degree of activity during lunch hour by a school psychologist. After 6 weeks on the special diet, the children were rated again. The results are give in Table 14.3. We test the hypothesis that the distribution of differences is symmetric about 0 against the alternative that it is not.

Table 14.3 Effect of Diet on Activity

Child	Before Rating	After Rating	|Difference|	Signed Rank
1	19	11	8	−10
2	14	15	1	+1
3	20	17	3	−3.5
4	6	12	6	+8
5	12	8	4	−5
6	4	9	5	+6.5
7	10	7	3	−3.5
8	13	6	7	−9
9	15	10	5	−6.5
10	9	11	2	+2

Solution

The sum of the positive ranks is $T(+) = 17.5$; hence $T(-) = 55 - 17.5 = 37.5$. Using $\alpha = 0.05$, the rejection region is for the smaller of $T(+)$ and $T(-)$ to be less than 8 (from Appendix Table A.9). Using $T(+)$ as our test statistic, we cannot reject the null hypothesis, so we conclude that there is insufficient evidence to conclude that the diet affected the level of activity. ■

The Randomization Approach for Example 14.3

Because this data contains ties and is a small sample, we might request an exact p value computed using a randomization test. How should the randomization be done? That the values are paired by child is an inherent feature of this data, and we must maintain it. When we randomize, the only possibility is that the before-and-after values within each child might switch places. This would cause the signs on the rank to switch, though it would not disturb the magnitude of the rank. Hence, we would need to list all the $2^{10} = 1024$ possible sets where the signed ranks in Table 14.3 are free to reverse their signs. For each of these hypothetical (or pseudo) data sets, we compute the pseudo-value of T. In 33.2% of the sets, the pseudo-T is at or below our observed value of 17.5. Hence, our p value is 0.3320, which agrees with the value from the SAS System's `Proc UNIVARIATE`.

CASE STUDY 14.1

Gumm *et al.* (2009) studied the preferences of female zebrafish for males with several possible fin characteristics. Each zebrafish can be *long fin*, *short fin*, or *wildtype*. Do females have a preference for a particular fin type? In each trial, a female zebrafish (the focal individual) was placed in the central part of an aquarium. At one end, behind a divider, was a male of one of the fin types. At the other end, behind a divider, was a male of a contrasting fin type. The males are referred to as the stimulus fish. The researchers recorded the amount of time each female spent in the vicinity of each stimulus fish, yielding two measurements for each trial.

We would prefer to use a paired *t* test to compare the preference of females for one type of fin versus the other. However, the authors state:

The data were not normally distributed after all attempts at transformations and thus nonparametric statistics were used for within treatment analysis. Total time spent with each stimulus was compared within treatments with a Wilcoxon-Signed Rank test.

The results of their analysis are summarized as follows:

Treatment	Wilcoxon Signed Rank Test
wildtype female: wildtype vs. long fin male	$n = 19, z = -1.81, p = 0.86$
wildtype female: short fin vs. wildtype male	$n = 20, z = -0.131, p = 0.90$
long fin female: wildtype vs. long fin male	$n = 20, z = -2.427, p = 0.02$
short fin female: short fin vs. wildtype male	$n = 20, z = -0.08, p = 0.45$

(Note the inconsistency in the p value for short fin females.) The authors conclude:

The preference for males with longer fins was observed only in females that also have long fins. This unique preference for longer fins by long fin females may suggest that the mutation controlling the expression of the long fin trait is also playing a role in controlling female association preferences.

14.3 TWO INDEPENDENT SAMPLES

The Mann–Whitney test (also called the Wilcoxon two-sample test) is a rank-based nonparametric test for comparing the location of two populations using independent samples. Note that this test does not specify an inference to any particular parameter of location. Using independent samples of n_1 and n_2, respectively, the test is conducted as follows:

1. Rank all $(n_1 + n_2)$ observations as if they came from one sample, adjusting for ties.
2. Compute T, the sum of ranks for the smaller sample.
3. Compute $T' = (n_1 + n_2)(n_1 + n_2 + 1)/2 - T$, the sum of ranks for the larger sample. This is necessary to assure a two-tailed test.
4. For small samples $(n_1 + n_2 \leq 30)$, compare the smaller of T and T' with the rejection region consisting of values less than or equal to the critical values given in Appendix Table A.10. If either T or T' falls in the rejection region, we reject the null hypothesis. Note that even though this is a two-tailed test, we only use the lower quantiles of the tabled distribution.
5. For large samples, the statistic T or T' (whichever is smaller) has an approximately normal distribution with

$$\mu = n_1(n_1 + n_2 + 1)/2 \quad \text{and}$$
$$\sigma^2 = n_1 n_2(n_1 + n_2 + 1)/12.$$

The sample size n_1 should be taken to correspond to whichever value, T or T', has been selected as the test statistic.

These parameter values are used to compute a test statistic having a standard normal distribution. We then reject the null hypothesis if the value of the test statistic is smaller than $-z_{\alpha/2}$. Modifications are available when there are a large number of ties (for example, Conover, 1999).

The procedure for a one-sided alternative hypothesis depends on the direction of the hypothesis. For example, if the alternative hypothesis is that the location of population 1 has a smaller value than that of population 2 (a one-sided hypothesis), then we would sum the ranks from sample 1 and use that sum as the test statistic. We would reject the null hypothesis of equal distributions if this sum is less than the $\alpha/2$ quantile of the table. If the one-sided alternative hypothesis is the other direction, we would use the sum of ranks from sample 2 with the same rejection criteria.

■ Example 14.4

Because the taste of food is impossible to quantify, results of tasting experiments are often given in ordinal form, usually expressed as ranks or scores. In this experiment two types of hamburger substitutes were tested for quality of taste. Five sample hamburgers of type A and five of type B were scored from best (1) to worst (10). Although these responses may appear to be ratio variables (and are often analyzed using this definition), they are more appropriately classified as being in the ordinal scale. The results of the taste test are given in Table 14.4. The hypotheses of interest are

H_0: the types of hamburgers have the same quality of taste, and

H_1: they have different quality of taste.

Table 14.4 Hamburger Taste Test

Type of Burger	Score
A	1
A	2
A	3
B	4
A	5
A	6
B	7
B	8
B	9
B	10

Solution

Because the responses are ordinal, we use the Mann–Whitney test. Using these data we compute

$$T = 1 + 2 + 3 + 5 + 6 = 17 \quad \text{and}$$
$$T' = 10(11)/2 - 17 = 38.$$

Choosing $\alpha = 0.05$ and using Appendix Table A.10, we reject H_0 if the smaller of T or T' is less than or equal to 17. The computed value of the test statistic is 17; hence we reject the null hypothesis at $\alpha = 0.05$, and conclude that the two types differ in quality of taste. If we had to choose one or the other, we would choose burger type A based on the fact that it has the smaller rank sum. ■

Randomization Approach to Example 14.4

Since this data set does not contain any ties, Appendix Table A.10 is accurate. If we wished a p value, we could enumerate all the $10!/(5!\,5!) = 252$ ways the ranks 1 through 10 could be split into two groups of five each. Listing the corresponding pseudo-value of T would show that there were 3.17% of them at or less than 17. Hence, the exact p value is 0.0317, which agrees with the value from SAS System's `PROC NPAR1WAY`. Using the normal asymptotic approximation gives $z = 2.193$, with a p value of 0.028, which is surprisingly close given the small sample size.

14.4 MORE THAN TWO SAMPLES

The extension to more than two independent samples provides a nonparametric analog for the one-way analysis of variance, which can be used with a completely randomized design experiment or a t sample observational study. That is, we test the null hypothesis that t independent samples come from t populations with identical distributions against the alternative that they do not, with the primary differences being in location. A test for this hypothesis is provided by a rank-based nonparametric test called the Kruskal–Wallis t sample test. The procedure for this test follows the same general pattern as that for two samples. The Kruskal–Wallis test is conducted in the following manner:

1. Rank all observations. Denote the ijth rank by R_{ij}.
2. Sum the ranks for each sample (treatment), denote these totals by T_i.
3. The test statistic is

$$H = \frac{1}{S^2}\left[\sum \frac{T_i^2}{n_i} - \frac{n(n+1)^2}{4}\right],$$

where

$$S^2 = \frac{1}{n-1}\left[\sum R_{ij}^2 - \frac{n(n+1)^2}{4}\right],$$

and where the R_{ij} are the actual ranks,[4] and n_i are the sizes of the ith sample, and $n = \sum n_i$. If no ties are present in the ranks, then the test statistic takes on the simpler form

$$H = \frac{12}{n(n+1)} \sum \frac{T_i^2}{n_i} - 3(n+1).$$

For a select group of small sample sizes, there exist specialized tables of rejection regions for H. For example, some exact tables are given in Iman *et al.* (1975). Usually, however, approximate values based on the χ^2 distribution with $(t-1)$ degrees of freedom are used. This test is similar to the Mann–Whitney in that it uses only one tail of the distribution of the test statistic. Therefore, we would reject H_0 if the value of H exceeded the α level of the χ^2 distribution with $(t-1)$ degrees of freedom. If this hypothesis is rejected, we would naturally like to be able to determine where the differences are. Since no parameters such as means are estimated in this procedure, we cannot construct contrasts or use differences in means to isolate those populations that differ. Therefore, we will use a pairwise comparison method based on the average ranks. This is done in the following manner.

We infer at the α level of significance that the locations of the response variable for factor levels i and j differ if

$$\left|\frac{T_i}{n_i} - \frac{T_j}{n_j}\right| > t_{\alpha/2}\sqrt{S^2\left(\frac{n-1-H}{n-t}\right)\left(\frac{1}{n_i}+\frac{1}{n_j}\right)},$$

where $t_{\alpha/2}$ is the $\alpha/2$ critical value from the t distribution with $(n-t)$ degrees of freedom.

This procedure does not attempt to control for the experiment-wise error rate. However, if we proceed with these comparisons only if the overall test is significant, then we have protected our experiment-wise error rate in somewhat the same manner as Fisher's LSD in the one-way ANOVA (Section 6.5). More sophisticated approaches similar to Tukey's HSD can also be implemented (see Higgins, 2004).

■ Example 14.5

A psychologist is trying to determine whether there is a difference in three methods of training six-year-old children to learn a foreign language. A random selection of 10 six-year-old children with similar backgrounds is assigned to each of three different methods. Method 1 uses the traditional teaching format. Method 2 uses repeated listening to tapes of the language along with classroom instruction. Method 3 uses videotapes exclusively. At the end of a 6-week period, the children

[4] If there are no ties, $\sum R_{ij}^2$ is more easily computed by $[n\,(n+1)(2n+1)]/6$. This is also a rather good approximation if there are few ties.

Table 14.5 Data and Ranks for Example 14.5

TEACHING METHOD					
1		2		3	
y	Rank	y	Rank	y	Rank
78	12.5	70	2.5	60	1
80	14	72	5.5	70	2.5
83	16	73	7	71	4
86	17	74	8.5	72	5.5
87	18	75	10	74	8.5
88	19	78	12.5	76	11
90	20	82	15		
		95	21		
$n_1 = 7$		$n_2 = 8$		$n_3 = 6$	
$T_1 = 116.5$		$T_2 = 82.0$		$T_3 = 32.5$	

were given identical, standardized exams. The exams were scored, with high scores indicating a better grasp of the language. Because of attrition, method 1 had 7 students finishing, method 2 had 8, and method 3 only 6. It is, however, important to note that we must assume that attrition was unrelated to performance. The data and associated ranks are given in Table 14.5.

Solution

Although the test scores may be considered ratio variables, concerns about the form of the distribution suggest the use of the Kruskal–Wallis nonparametric method. Since there are few ties, we will use the simpler form of the test statistic, resulting in

$$\begin{aligned} H &= \left[\frac{12}{(21)(22)}\right]\left(\frac{116.5^2}{7} + \frac{82.0^2}{8} + \frac{32.5^2}{6}\right) - 3(22) \\ &= 10.76. \end{aligned}$$

From Appendix Table A.3, we see that $\chi^2(2)$ for $\alpha = 0.05$ is 5.99; hence we reject the null hypothesis of equal location and conclude that there is a difference in the distributions of test scores for the different teaching methods.

To determine where the differences lie, we perform the multiple comparison procedure based on the average ranks discussed in the preceding. Using the ranks in Table 14.5 we obtain $\sum R_{ij}^2 = 3309$, so that[5]

$$S^2 = (1/20)[3309 - 21(22)^2/4] = 38.4.$$

[5] Using the shortcut formula for $\sum R_{ij}^2$ gives 3311.

The mean ranks are

Method 1: $116.5/7 = 16.64$,
Method 2: $82.0/8 = 10.25$, and
Method 3: $32.5/6 = 5.42$.

From Appendix Table A.2, the appropriate t value for a 5% significance level is 2.101. We will compare the difference between method 1 and method 2 with

$$(2.101)\sqrt{38.4\left(\frac{20-10.76}{18}\right)\left(\frac{1}{8}+\frac{1}{7}\right)} = 4.83.$$

The mean rank difference between methods 1 and 2 has a value of 6.39, which exceeds this quantity; hence we conclude the distributions of test scores for methods 1 and 2 may be declared different. Similarly, for comparing methods 1 and 3 the mean difference of 11.22 exceeds the required value of 5.18; hence we conclude that the distributions of scores differ. Finally, the mean difference between methods 2 and 3 is 4.83, which is less than the required difference of 5.03; hence there is insufficient evidence to declare different distributions between methods 2 and 3. The psychologist can conclude that the results of using method 1 differ from those of both the other methods, but that the effect of the other two may not. ■

Randomization Approach to Example 14.5

Since this data contains ties and is of modest size, we might prefer to calculate a p value using an exact enumeration of all the possibilities. There are $21!/(7!\,8!\,6!) = 349{,}188{,}840$ ways to rearrange the observed ranks into three groups of 7, 8, and 6 observations. Since the list is so long, we could adopt the alternate strategy of an approximate randomization test. We would use a random number generator to produce 10,000 random rearrangements of the ranks, tabulating the resulting pseudo-values for H. The estimated p value would be the proportion of times that values in the sample meet or exceed the observed value of 10.76. The SAS System's `PROC NPAR1WAY` will execute this, finding a p value in the vicinity of 0.0015. The precise value of the approximate p value will depend on the random selection of the rearrangements.

We have noted that the Kruskal–Wallis test is primarily designed to detect differences in "location" among the populations. In fact, theoretically, the Kruskal–Wallis test requires that the underlying distribution of each of the populations be identical in shape, differing only by their location. Fortunately, the test is rather insensitive to moderate differences in the shape of the underlying distributions, and this assumption can be relaxed in all but the most extreme applications. However, it is not useful for detecting differences in variability among populations having similar locations.

There are many nonparametric tests available for the analysis of t independent samples designed for a wide variety of alternative hypotheses. For example, there are tests to detect differences in scale (or shape) of the distributions, tests to detect differences in the skewness (symmetry) of the distributions, and tests to detect differences in the kurtosis (convexity) of the distributions. There are also so-called omnibus tests that detect any differences in the distributions, no matter what that difference may be. A good discussion of many of these tests can be found in Boos (1986).

14.5 RANDOMIZED BLOCK DESIGN

Data from a randomized block design may be analyzed by a nonparametric rank-based method known as the Friedman test. The Friedman test for the equality of treatment locations in a randomized block design is implemented as follows:

1. Rank treatment responses within each block, adjusting in the usual manner for ties. These ranks will go from 1 to t, the number of treatments, in each block. These are denoted R_{ij}.
2. Obtain the sum of ranks for each treatment. This means that we add one rank value from each block, for a total of b (the number of blocks) ranks. Call this sum R_i for the ith treatment.
3. The test statistic is

$$T^* = (b-1)\frac{\left[B - \frac{bt(t+1)^2}{4}\right]}{A-B},$$

where $A = \sum\sum R_{ij}^2$, which, if there are no ties, simplifies to

$$A = bt(t+1)(2t+1)/6$$

and $B = \frac{1}{b}\sum R_i^2$.

The test statistic, T^*, is compared to the F distribution with $[t-1, (b-1)(t-1)]$ degrees of freedom.

Some references give the Friedman test statistic as

$$T_1 = \frac{12}{bt(t+1)}\sum R_i^2 - 3b(t+1),$$

where t and b represent the number of treatments and blocks, respectively. This test statistic is compared with the χ^2 distribution with $(t-1)$ degrees of freedom. However, the T^* test statistic using the F distribution has been shown to be superior to the χ^2 approximation (Iman and Davenport 1980), and we therefore recommend the use of that statistic.

Pairwise comparisons can be performed using the R_i in the following manner. For a significance level of α, we can declare that the distributions of treatments i and j differ in location if

$$|R_i - R_j| > t_{\alpha/2}\sqrt{\frac{2b(A-B)}{(b-1)(t-1)}},$$

where $t_{\alpha/2}$ has $(b-1)(t-1)$ degrees of freedom.

■ Example 14.6

Responses given in terms of proportions will follow a scaling of the binomial distribution, which can be quite nonnormal and also exhibit heterogeneous variances. This experiment is concerned with the effectiveness of five weed killers. The experiment was conducted in a randomized block design with five treatments and three blocks, which corresponded to plots in the test area. The response is the percentage of weeds killed. The hypothesis that the killers (treatments) have equal effects on weeds is tested against an alternative that there are some differences. The data are given in Table 14.6, along with the ranks in parentheses.

Table 14.6 Percentage of Weeds Killed

	BLOCKS			
Treatment	**1**	**2**	**3**	R_i
1	16 (4.5)	51 (5)	11 (4.5)	14.0
2	1 (1)	29 (4)	2 (2)	7.0
3	16 (4.5)	24 (3)	11 (4.5)	12.0
4	4 (2.5)	11 (2)	5 (3)	7.5
5	4 (2.5)	1 (1)	1 (1)	4.5

Note: Ranks are in parentheses.

Solution

The Friedman test is appropriate for this example. Using the ranks from Table 14.6 we obtain the values

$$\mathrm{A} = 163.5 \quad \text{and} \quad \mathrm{B} = 155.17.$$

The test statistic is

$$T^* = 2\left[\frac{155.17 - \frac{(3)(5)(6)^2}{4}}{163.5 - 155.17}\right] = 4.84.$$

Table 14.7 Differences among Treatments

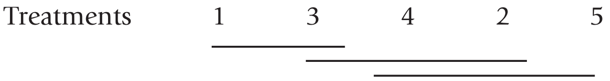
Treatments	Differences	Significant or Not
1 vs 5	$14 - 4.5 = 9.5$	yes
3 vs 5	$12 - 4.5 = 7.5$	yes
4 vs 5	$7.5 - 4.5 = 3$	no
2 vs 5	$7 - 4.5 = 2.5$	no
1 vs 2	$14 - 7 = 7$	yes
3 vs 2	$12 - 7 = 5$	no
4 vs 2	$7.5 - 7 = 0.5$	no
1 vs 4	$14 - 7.5 = 6.5$	yes
3 vs 4	$12 - 7.5 = 4.5$	no
1 vs 3	$14 - 12 = 2$	no

The null hypothesis is rejected if the test statistic is in the rejection region of the F distribution with 4 and 8 degrees of freedom. Using the 0.05 level of significance in Appendix Table A.4 we find the critical value of 3.84. Therefore, we reject the null hypothesis and conclude there is a difference among the killers tested.

To identify the nature of the differences we perform a multiple comparison test. We compare the pairwise differences among the R_i with

$$(2.306)\sqrt{\frac{(2)(3)(8.33)}{(2)(4)}} = 5.76.$$

The differences and conclusions of the multiple comparisons among the R_i are given in Table 14.7, where it is seen that treatment 1 differs from treatments 2, 4, and 5, and that treatment 3 differs from treatment 5. No other differences are significant. Using the traditional schematic (see discussion of post hoc comparisons in Section 6.5), the results can be presented as



14.6 RANK CORRELATION

The concept of correlation as a measure of association between two variables was presented in Section 7.6 where correlation was estimated by the Pearson product moment correlation coefficient. The value of this statistic is greatly influenced by extreme observations, and the test for significance is sensitive to deviations from normality. A correlation coefficient based on the ranked, rather than the originally observed, values would not be as severely affected by extreme or influential observations. One such rank-based correlation coefficient is obtained by simply using the formula given for the correlation coefficient in Section 7.6 on the ranks rather than the individual values of the observations. This rank-based correlation coefficient is known as Spearman's coefficient of rank correlation, which can, of course, also be used with ordinal variables. For reasonably large samples, the test statistic for determining the existence of significant correlation is the same as that for linear correlation given in Chapter 7,

$$F = (n-2)\,r^2/(1-r^2),$$

where r^2 is the square of the rank-based correlation coefficient.

Because the data consist of ranks, a shortcut formula exists for computing the Spearman rank correlation. This shortcut is useful for small data sets that have few ties. First, separately rank the observations in each variable (from 1 to n). Then for each observation compute the difference between the ranks of the two variables, ignoring the sign. Denote these differences as d_i. The correlation coefficient is then computed:

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

■ Example 14.7

The data from Exercise 2 of Chapter 1 described the abundance of waterfowl at different lakes. It was noted that the distributions of both waterfowl abundance and lake size were dominated by one very large lake. We want to determine the correlation between the water area (WATER) and the number of waterfowl (FOWL). The magnitude of the Pearson correlation is easily seen to be dominated by the values of the variables for the one large pond (observation 31) and may therefore not reflect the true magnitude of the relationship between these two variables.

Solution

The Spearman correlation may be a better measure of association for these variables. Table 14.8 gives the ranks of the two variables, labeled RWATER and RFOWL, and the absolute values of differences in the ranks, DIFF.

The correlation coefficient computed directly from the ranks is 0.490. Using the F statistic, we are able to test this correlation for significance. The p value for this test is 0.006, so we conclude that the correlation is in fact significant. The shortcut formula using the differences among ranks results in a correlation coefficient of 0.4996. The difference is due to a small number of ties in the data. Of course, for this large data set the special formula represents no savings in computational effort.

The Pearson correlation coefficient computed from the observed values results in a value of 0.885. The fact that this value is much larger than the Spearman correlation is the result of the highly skewed nature of the distributions of the variables in this data set. ■

Table 14.8 Waterfowl Data for Spearman Rank Correlation

OBS	RWATER	RFOWL	DIFF	OBS	RWATER	RFOWL	DIFF
1	20.5	8.5	12.0	27	6.5	8.5	2.0
2	6.5	24.0	17.5	28	28.0	50.0	22.0
3	20.5	42.5	22.0	29	33.0	19.0	14.0
4	46.0	36.0	10.0	30	50.0	8.5	41.5
5	20.5	8.5	12.0	31	52.0	52.0	0.0
6	51.0	37.0	14.0	32	20.5	8.5	12.0
7	15.5	31.0	15.5	33	12.0	29.0	17.0
8	15.5	8.5	7.0	34	28.0	31.0	3.0
9	33.0	28.0	5.0	35	6.5	8.5	2.0
10	28.0	33.0	5.0	36	6.5	8.5	2.0
11	20.5	8.5	12.0	37	15.5	42.5	27.0
12	47.5	48.0	0.5	38	6.5	19.0	12.5
13	6.5	25.5	19.0	39	24.5	8.5	16.0
14	39.0	49.0	10.0	40	41.0	46.0	5.0
15	45.0	22.0	23.0	41	33.0	41.0	8.0
16	24.5	35.0	10.5	42	39.0	44.0	5.0
17	12.0	21.0	9.0	43	33.0	8.5	24.5
18	47.5	40.0	7.5	44	6.5	25.5	19.0
19	33.0	8.5	24.5	45	39.0	51.0	12.0
20	28.0	38.0	10.0	46	42.5	39.0	3.5
21	12.0	27.0	15.0	47	44.0	47.0	3.0
22	15.5	34.0	18.5	48	1.5	8.5	7.0
23	6.5	17.0	10.5	49	1.5	8.5	7.0
24	49.0	19.0	30.0	50	42.5	45.0	2.5
25	36.0	31.0	5.0	51	37.0	8.5	28.5
26	28.0	23.0	5.0	52	20.5	8.5	12.0

14.7 THE BOOTSTRAP

Randomization tests tap the power of modern computers to provide a method for calculating p values in hypothesis tests. They are particularly adaptable to null hypotheses of "no relationship" between a dependent and independent variable. They do not, however, give easy access to confidence intervals for parameters. This is not surprising, as the structure of nonparametric tests in general is to avoid specific parameterizations of problems.

When we do have a natural parameter for which we need an interval estimate, we need another approach. Of course, if the usual distributional assumptions (normality) are reasonable, then the most powerful techniques are the classical ones presented in Chapters 4 through 11. When normality is not appropriate, it may be possible to implement a technique called the bootstrap. This method was originally intended to estimate standard errors when the parent distribution (from which the data came) is unknown or intractable. We will present only a very simple example. For more information, see Higgins (2004) or Efron (1982). This technique requires specialized software or advanced programming skills.

To motivate the bootstrap, we should examine the reasoning behind classical estimates of standard errors. Just as for randomization tests, the bootstrap attempts to mimic the classical process. Recall that the mean squared error is an estimate of the average size of the squared discrepancy between the estimate and the true value, where this average is over the population of possible samples of a particular size.

Assume we have a basket full of slips of paper, each with a value written on it. The values follow a very skewed distribution, and we wish to estimate the population median. As a point estimate, we draw a sample of n independent observations and calculate the sample median. To understand the reliability of this estimate, we need to calculate a standard error, which is the square root of the mean squared error.

Ideally, we could construct an experiment in which we repeat the process of drawing a sample of n independent observations and calculating the sample median a huge number of times. We could then calculate the squared error between each individual median and the true median for the entire basket, averaging these to obtain the mean squared error.

Of course, we cannot carry out the ideal experiment, because we do not have access to the true population that generated our data set. However, our sample is our best information on what the true basket looks like (if we are unwilling to assume normality or some other distribution). Hence, we will construct an artificial basket containing the values in our data set. For this artificial basket, we do know the population median. Using a computer, we mimic the process of selecting samples of n independent observations from this artificial basket. These are called pseudo-samples, and from each we calculate a pseudo-median. The pseudo-errors are the discrepancies between the pseudo-medians and the median of our artificial basket, which by design is the median observed in the actual data set. By calculating the average squared pseudo-errors, we have an estimate of the mean squared error of a median calculated from a sample of n when the parent distribution is similar to that observed in our sample.

■ Example 14.2: Revisited

Interest is focused on the median income, and we have discussed testing the null hypothesis that the true median is 13. Suppose that we had no preconceived notions regarding the median, and simply wanted an interval estimate. The sample median was 13.90, but how far off might that be from the true population median?

Solution

The SAS System macro `jackboot.sas` was used to construct 1000 pseudo-samples of size 20 drawn at random (with replacement) from the observed data set. The standard deviation of the pseudo-medians was 0.66. A rough confidence interval for the median in the population would be

$$13.90 \pm 2 \times 0.66 = (12.58,\ 15.22).$$

■

The implementation of the bootstrap is far advanced beyond the simple idea presented here. There are a number of ways to use the bootstrap to estimate the possible bias in an estimator, and to refine the confidence intervals beyond the rough interval we have discussed.

The bootstrap is a powerful method for estimating standard errors in regression situations, especially for small to moderate samples where the distributions of the residuals appear nonnormal. An introduction to the bootstrap for regression is given in Higgins (2004).

14.8 CHAPTER SUMMARY

Solution to Example 14.1

The distribution of the residuals from the ANOVA model for Example 14.1 did not have the assumed normal probability distribution. This leads us to suspect the results of the F test, particularly the p value. This problem, however, does fit the criteria for the use of a Kruskal–Wallis test. The data, the ranks, and the result of using `PROC NONPAR1WAY` in SAS are given in Table 14.9. Note that the printout gives the Kruskal–Wallis test statistic along with the p value calculated from the χ^2 approximation. In this example, the p value is quite small so we reject the null hypothesis of equal treatment distributions.

Note that the output also gives the sums and means of the ranks (called scores). The sums are the $\sum R_i$ in the formula for the test statistic. Also provided are the expected sum and the standard deviations if the null hypothesis is true. These are identical because the sample sizes are equal (each is 15), and the null hypothesis is that of equality. That is, we expect all four of the treatments to have equal sums of ranks if the populations are identical.

The mean scores given in Table 14.9 can be used to make pairwise comparisons (Section 14.4). The least significant difference between average ranks for $\alpha = 0.05$ is 6.69. From Table 14.9 we can see that treatment 1 is significantly smaller than the other three, and that treatment 4 is significantly larger than the other three.

Table 14.9 Windshield Wipers

```
                      N P A R 1 W A Y   P R O C E D U R E

                 Wilcoxon Scores (Rank Sums) for Variable WEAR

                            Classified by Variable TRT

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

 1    15      120.000000    457.500000    58.5231375     8.0000000
 2    15      452.500000    457.500000    58.5231375    30.1666667
 3    15      516.000000    457.500000    58.5231375    34.4000000
 4    15      741.500000    457.500000    58.5231375    49.4333333

                        Average Scores were used for Ties

 Kruskal-Wallis Test (Chi - Square Approximation)
 CHISQ = 43.360    DF = 3    Pr > CHISQ = 0.0001
```

Treatments 2 and 3 are not significantly different. Since we wanted to minimize the amount of wear, chemical treatment number 1 seems to be the best.

It is interesting to note that these results are quite similar to those obtained by the analysis of variance. This is because, unlike highly skewed or fat-tailed distributions, the uniform distribution of the random error does not pose a very serious violation of asumptions.

Nonparametric methods provide alternative statistical methodology when assumptions necessary for the use of linear model-based methods fail as well as provide procedures for making inferences when the scale of mesurement is ordinal or nominal. Generally, nonparametric methods use functions of observations, such as ranks, and therefore make no assumptions about underlying probability distributions. Previous chapters have presented various nonparametric procedures (for example, Chapter 12) usually used for handling nominal scale data. This chapter discusses rank-based nonparametric methods for one, two, and more than two independent samples, paired samples, randomized block designs, and correlation.

Many nonparametric techniques give answers similar to classical analyses on the ranks of the data rather than the raw data. For example, applying a standard one-way ANOVA to the ranks rather than the raw data will yield something close to the Kruskal-Wallis test. However, the precise details of the test differ somewhat, because we know the population variance for the ranks. For the raw data, this population variance is unknown and has to be estimated using the MSE.

One of the most powerful ideas presented here is the use of a randomization test to assign a p value. We have applied this technique to calculate p values for the standard nonparametric test statistics. However, this is a very general technique, and gives you the option for creating test statistics that you feel are particularly appropriate. For example, rather than using the Wilcoxon rank sum statistic to compare the locations of two groups, you could use the difference in the sample medians as your test statistic. A randomization test would allow you to calculate a *p* value for the null hypothesis that the distributions are the same, with special sensitivity to the possibility that they differ with respect to their medians.

The bootstrap is also a powerful tool that allows you some creativity in your estimation. Although the original intent of the bootstrap was to develop confidence intervals, it can also be used to calculate *p* values. Note that bootstraps are very much oriented toward identifying parameters, whereas randomization tests are meant to test for a relationship of an unspecified (nonparametric) nature.

The bootstrap and randomization tests are not all-purpose techniques that supplant the classical inferences. Bear in mind that if the assumptions of a classical analysis (such as least squares regression) are appropriate, then the traditional techniques will not only be more powerful but substantially simpler to employ. Further, the bootstrap has a number of special tricks that have to be understood before it can be applied in any but the simplest situation.

For quantitative dependent variables, we recommend that the first choice for any analysis be one of the standard parametric techniques. If the structure of the variables or an analysis of the residuals reveals problems with assumptions such as normality, then we should try to find a set of transformations that make the assumptions more reasonable. Nonparametric techniques are a useful fallback if no transformation can be found.

14.9 CHAPTER EXERCISES

Concept Questions

1. Describe a randomization procedure for assigning a *p* value to a Spearman correlation coefficient. What hypothesis is being tested? Would the same procedure work for a Pearson correlation coefficient? *Hint*: your randomization procedure should mimic the process at work if the null hypothesis is correct.
2. Two different statisticians are evaluating the data given in Table 14.5. The first statistician uses a Kruskal-Wallis test applied to the values given in the table. The second statistician uses a Kruskal-Wallis test applied to the logarithm of the values. Both statisticians get exactly the same results. Why?

For problems 3 through 6, identify an appropriate nonparametric and parametric technique. If the results were significant, how would the conclusions differ?

3. Participants swallow an oral dose of calcium containing an isotopic tracer. None of the participants wants to have needles stuck in them more than once. So you recruit 40 participants, and randomly select eight to have samples drawn at 15 minutes, another eight to have samples drawn at 30 minutes, and so on at 45, 60, and 90 minutes. Does the typical amount of tracer in the bloodstream increase with time?

4. You survey consumer confidence (on an ordinal scale from 1 = low to 5 = high). The participants in your survey are also asked about their income and are classified into one of four income groups (1 = low, 2 = medium, 3 = high, 4 = out-of-sight). You believe that there will be differences in typical confidence by income group, but do not necessarily think it will be a trend.

5. An engineer has two meters for reading electrical resistance, and wishes to know whether they differ systematically in their readings. For 12 different circuits, the engineer records the reading using both meter A and meter B.

6. Four different income tax decreases have been proposed by Congress. You want to know whether these plans will differ in their impact on people's taxes. You randomly select 20 households and review each one's 2007 tax records, then work out their savings under each of the four plans. (For each of the 20 households, you will have four savings calculations.) You want to compare the plans to see how they tend to differ.

Exercises

1. In 11 test runs a brand of harvesting machine operated for 10.1, 12.2, 12.4, 12.4, 9.4, 11.2, 14.8, 12.6, 10.1, 9.2, and 11.0 h on a tank of gasoline.
   (a) Use the Wilcoxon signed rank test to determine whether the machine lives up to the manufacturer's claim of an average of 12.5 h on a tank of gasoline. (Use $\alpha = 0.05$.)
   (b) For the sake of comparison, use the one-sample t test and compare results. Comment on which method is more appropriate.

2. Twelve adult males were put on a liquid diet in a weight-reducing plan. Weights were recorded before and after the diet. The data are shown in Table 14.10. Use the Wilcoxon signed rank test to ascertain whether the plan was successful. Do you think the use of this test is appropriate for this set of data? Comment.

Table 14.10 Data for Exercise 2

	SUBJECT											
	1	2	3	4	5	6	7	8	9	10	11	12
Before	186	171	177	168	191	172	177	191	170	171	188	187
After	188	177	176	169	196	172	165	190	165	180	181	172

Table 14.11 Data for Exercise 3

PROFESSOR	
A	B
74	75
78	80
68	87
72	81
76	72
69	73
71	80
74	76
77	68
71	78

3. The test scores shown in Table 14.11 were recorded by two different professors for two sections of the same course. Using the Mann–Whitney test and $\alpha = 0.05$, determine whether the locations of the two distributions are equal. Why might the median be a better measure of location than the mean for these data?

4. Inspection of the data for Exercise 11 in Chapter 5 suggests that the data may not be normally distributed. Redo the problem using the Mann–Whitney test. Compare the results with those obtained by the pooled t test.

5. Eight human molar teeth were sliced in half. For each tooth, one randomly chosen half was treated with a compound designed to slow loss of minerals; the other half served as a control. All tooth halves were then exposed to a demineralizing solution. The response is percent of mineral content remaining in the tooth enamel. The data are given in Table 14.12.

Table 14.12 Data for Exercise 5

	Mineral Content							
Control	66.1	79.3	55.3	68.8	57.8	71.8	81.3	54.0
Treated	59.1	58.9	55.0	65.9	54.1	69.0	60.2	55.5

(a) Perform the Wilcoxon signed rank test to determine whether the treatment maintained a higher mineral content in the enamel.
(b) Compute the paired t statistic and compare the results. Comment on the differences in the results.

Table 14.13 Data for Exercise 6

METHOD		
1	2	3
94	82	89
87	85	68
90	79	72
74	84	76
86	61	69
97	72	
	80	

6. Three teaching methods were tested on a group of 18 students with homogeneous backgrounds in statistics and comparable aptitudes. Each student was randomly assigned to a method and at the end of a 6-week program was given a standardized exam. Because of classroom space, the students were not equally allocated to each method. The results are shown in Table 14.13.
(a) Test for a difference in distributions of test scores for the different teaching methods using the Kruskal–Wallis test.
(b) If there are differences, explain the differences using a multiple comparison test.

7. Hail damage to cotton, in pounds per planted acre, was recorded for four counties for three years. The data are shown in Table 14.14. Using years as blocks use the Friedman test to determine whether there was a difference in hail damage among the four counties. If a difference exists, determine the nature of this difference with a multiple comparison test. Also discuss why this test was recommended.

8. To be as fair as possible, most county fairs employ more than one judge for each type of event. For example, a pie-tasting competition may have two judges testing each entered pie and ranking it according to preference. The Spearman rank correlation coefficient may be used to determine the consistency between the

Table 14.14 Data for Exercise 7

County	YEAR		
	1	2	3
P	49	141	82
B	13	64	8
C	175	30	7
R	179	9	7

Table 14.15 Ranking of Pies by Judges

Pie	Judge A	Judge B
1	4	5
2	7	6
3	5	4
4	8	9
5	10	8
6	1	1
7	2	3
8	9	10
9	3	2
10	6	7

judges (the interjudge reliability). In one such competition there were 10 pies to be judged. The results are given in Table 14.15.

(a) Calculate the Spearman correlation coefficient between the two judges' rankings.

(b) Test the correlation for significance at the 0.05 level.

9. An agriculture experiment was conducted to compare four varieties of sweet potatoes. The experiment was conducted in a completely randomized design with varieties as the treatment. The response variable was yield in tons per acre. The data are given in Table 14.16. Test for a difference in distributions of yields using the Kruskal–Wallis test. (Use $\alpha = 0.01$.)

Table 14.16 Yield of Sweet Potatoes

Variety A	Variety B	Variety C	Variety D
8.3	9.1	10.1	7.8
9.4	9.0	10.0	8.2
9.1	8.1	9.6	8.1
9.1	8.2	9.3	7.9
9.0	8.8	9.8	7.7
8.9	8.4	9.5	8.0
8.9	8.3	9.4	8.1

10. In a study of student behavior, a school psychologist randomly sampled four students from each of five classes. He then gave each student one of four different tasks to perform and recorded the time, in seconds, necessary to complete the assigned task. The data from the study are listed in Table 14.17. Using classes as

Table 14.17 Time to Perform Assigned Task

	TASK			
Class	**1**	**2**	**3**	**4**
1	43.2	45.8	45.4	44.7
2	48.3	48.7	46.9	48.8
3	56.6	56.1	55.3	54.6
4	72.0	74.1	89.5	82.7
5	88.0	88.6	91.5	88.2

blocks use the Friedman test to determine whether there is a difference in tasks. Use a level of significance of 0.10. Explain your results.

11. Table 14.18 shows the total number of birds of all species observed by bird-watchers for routes in three different cities observed at Christmas for each of the 25 years from 1965 through 1989.

Table 14.18 Bird Counts for Twenty-Five Years

	ROUTE				ROUTE		
Year	**A**	**B**	**C**	**Year**	**A**	**B**	**C**
65	138	815	259	78	201	1146	674
66	331	1143	202	79	267	661	494
67	177	607	102	80	357	729	454
68	446	571	214	81	599	845	270
69	279	631	211	82	563	1166	238
70	317	495	330	83	481	1854	98
71	279	1210	516	84	1576	835	268
72	443	987	178	85	1170	968	449
73	1391	956	833	86	1217	907	562
74	567	859	265	87	377	604	380
75	477	1179	348	88	431	1304	392
76	294	772	236	89	459	559	425
77	292	1224	570				

An inspection of the data indicates that the counts are not normally distributed. Since the responses are frequencies, a possible alternative is to use the square root transformation, but another alternative is to use a nonparametric method. Perform the analysis using the Friedman test. Compare results with those obtained in Exercise 10.10. Which method appears to provide the most useful results?

12. The ratings by respondents on the visual impact of wind farms (Table 12.25 for Exercise 12.16) are on an ordinal scale that makes rankings possible. Use a nonparametric test to compare the ratings from residents of Gigha to those of Kintyre. How does the interpretation of these results compare to the interpretation of the analysis in Exercise 12.16?

13. Table 5.1 summarizes stock prices (12/31/2007 and 12/31/2008) for stocks in the Consumer Staples and Financial categories of the S&P 500. Suppose you are unwilling to accept the assumption that the data come from normal distributions.
    (a) Within each category separately, is there evidence of a change in the location of the prices? What assumptions are required by the analysis?
    (b) Is there evidence that the typical change in the Consumer Staples category differs from that in the Financial category? What assumptions are required by the analysis?

14. Compare the variability in the test scores for the three teaching methods given in Table 14.5. To do this, implement a nonparametric version of Levene's test by first calculating the absolute differences of each value from its group *median*. Compare the typical magnitudes of the absolute differences using a nonparametric test from this chapter. What do you conclude?

Projects

1. **Lake Data Set**. The data set described in Appendix C.1 contains information on the county and summer nitrogen levels for each lake. Do the nitrogen levels appear to differ by county? This might happen either for geologic reasons, differing land use, or environmental restrictions. Attempt to answer this question first by using a classical parametric analysis. Evaluate the data to see if the assumptions of the analysis are satisfied. Then attempt to answer the question using a nonparametric technique. Do the analyses reach the same conclusion? What are the pros and cons of each analysis?

2. **Education Data**. The data set described in Appendix C.2 contains state average scores on the eighth grade NAEP test for mathematics, together with some economic information for each state. Assess the relationship between eighth grade math scores and median incomes using both a parametric and nonparametric measure. What are the pros and cons of each analysis? What is it about the data that might lead you to prefer the nonparametric measure?

This page intentionally left blank