Nonparametric Tests

9

Chapter Outline
9.1 Why Nonparametric Tests?
9.2 The Sign Test
9.3 The Wilcoxon Signed Rank Test
9.4 The Wilcoxon Rank Sum Test
9.5 The Kruskal-Wallis Test
9.6 The Friedman Test

In this chapter we present several statistics for testing whether or not probability distributions have the same medians. The use of these statistics does not require that the sample data follow any particular probability distribution, and, thus, there are no distributional parameters to be estimated. Because of these features, these tests are called distribution-free or nonparametric tests. We still assume that the data come from continuous distributions. We begin with justification of using distribution-free methods.

9.1 Why Nonparametric Tests?

The methods studied in the previous chapter were mostly concerned with data from a normal distribution. In many situations the data may consist of a number of ordered categories such as a subjective rating of the amount of pain relief (none, a little, a lot, total) a patient perceives after receiving a treatment. In other cases the data may simply be the presence or absence of a condition. In such cases the investigator may be unwilling to use a numerical scale but still wants to test a hypothesis related to the effect of a treatment or to the effects of two different treatments. The sign test discussed in this chapter can be used for situations with two outcomes. Other methods in this chapter are used with ordered data or with numerical data that do not follow the normal distribution.

The methods for testing the mean and proportion in the previous chapter are based on normality assumptions. If there is obvious nonnormality in the data, distribution-free methods can be used. In some cases we may suspect that the data do not follow the normal distribution, but we cannot determine the lack of normality for sure because the sample size is too small. Distribution-free methods can be used then and are also often used for small samples when the central limit theorem may not apply.

9.2 The Sign Test

The sign test is one of the oldest tests used in statistics. For example, in 1710, John Arbuthnot, a British physician and collaborator of Jonathan Swift, performed what was in effect a sign test on the sex ratio of births over an 82-year period (Stigler 1986).

As we saw in the last chapter, the sign test can be used to compare different interventions for matched pairs. Individuals were assigned to a pair based on age, sex, weight, and exercise level, and then one member within the pair was randomly assigned to diet 1 and the other member assigned to diet 2. The sign test was then used to determine which of the two diets was more likely to be associated with the greater weight loss for each pair. Another way of stating this null hypothesis is that each difference of weight losses has a median of zero. The sign test can also be used with a single population — for example, in the comparison of multiple measurements made on the same individual or one set of measurements compared with some hypothesized value, as shown in the next examples.

As we saw in the last chapter, the *p*-value of the sign test can be exactly determined from the binomial distribution, or we can approximate the *p*-value by using the normal approximation to the binomial when the sample size is large.

Example 9.1

One problem often encountered in research designs involving pre- and posttest measurements is the reversion or regression toward the mean effect (Samuels 1991). Briefly, persons scoring high on one test tend not to score as high on a subsequent test and low scorers on the first test tend to score higher on the next test — that is, the test scores tend to revert toward the mean score. Reversion toward the mean is important because of its possible effect on test results (Davis 1976; Nesselroade, Stigler, and Baltes 1980).

We consider the caloric intake for 33 boys selected from a larger study (McPherson et al. 1990). Table 9.1 shows the caloric intake for the boys for the first two of three randomly selected days during a two-week period. The more extreme — the seven highest and seven lowest — day 1 values are marked. We can examine whether or not there is reversion toward the mean. Based on the descriptive statistics shown in Table 9.1, it appears that there could be a reversion toward the mean effect

Table 9.1 Two days of caloric intake for 33 boys enrolled in two middle schools outside of Houston.[a]

ID	Day 1	Day 2	ID	Day 1	Day 2	ID	Day 1	Day 2
10	1,823	1,623	39	2,330	2,339	118	1,781[L]	1,844
11	2,007	1,748	40	2,436	2,189	120	2,748	2,104
13	1,053[L]	2,484	41	3,076[H]	2,431	127	2,348	2,122
14	4,322[H]	2,926	44	1,843	2,907	130	2,773[H]	3,236
16	1,753[L]	1,054	46	2,301	4,120	137	2,310	1,569
17	2,685	2,304	47	2,546	1,732	139	2,594	2,867
26	2,340	3,182	50	1,292[L]	810	141	1,898	1,236
27	3,532[H]	3,289	51	3,049[H]	2,573	145	2,400	2,554
30	2,842[H]	2,849	101	3,277[H]	2,185	148	2,011	1,566
32	2,074	3,312	105	2,039	1,905	149	1,645[L]	2,269
33	1,505[L]	1,925	107	2,000	1,797	150	1,723[L]	3,163

		Mean	
	Number	Day 1	Day 2
[L]Lowest values	7	1,536	1,936
[H]Highest values	7	3,267	2,784

[a]Selected from a larger study by McPherson et al. (1990)

here. The seven lowest values had a mean of 1,536 calories on day 1 compared with a mean of 1,936 calories on day 2 — an increase. The seven highest values had a mean of 3,267 calories on day 1 compared with a mean of 2,784 calories on day 2 — a decrease. However, we wish to go beyond a descriptive presentation of the sample in our consideration of the question. We wish to test a hypothesis about the population values.

If there is no reversion toward the mean effect here, of the boys with extreme day 1 values, the proportion of those whose day 2 values move in the direction of the mean should be equal to 0.50 (ignoring the possibility that the day 1 and day 2 values are the same). If there is reversion toward the mean, the proportion should be greater than 0.50. The null and alternative hypotheses are therefore

$$H_0: \pi = 0.50 \text{ versus } H_a: \pi > 0.50.$$

If there are few ties (a subject has same values for day 1 and day 2) in the data, convention is that these observation pairs are dropped from the data. For example, if one out of the 14 boys had the same day 1 and day 2 values, the sample size for the binomial would then be 13 instead of 14, reflecting the deletion of the tied pair. When there are many ties, indicating no difference in the day 1 and day 2 values, there is little reason to perform the test for the remaining untied pairs.

The population from which this sample is drawn consists of middle schools in a northern suburb of Houston. Although the population is limited, perhaps the results from this population can be generalized to boys in suburban middle schools throughout the United States, not just to those in one suburb of Houston. As was mentioned in Chapter 6, this generalization does not flow from statistical properties because we did not sample this larger population, but it is based on substantive considerations. If there are differences in dietary practices between the one Houston suburb and others, this generalization to the larger population is then questionable.

We conduct the test of hypothesis at the 0.05 level. The test statistic is the number of boys with an extreme day 1 value whose day 2 value moves toward the mean, which is found to be 10 from the data. The critical region for the test can be found from the binomial distribution. For larger sample sizes, the normal approximation to the binomial can be used. We could use Table B2 to find the probabilities for a binomial distribution with $n = 14$ and $\pi = 0.50$. We are interested only in the upper tail of the binomial distribution; therefore, we consider only values above the expected value of 7. Because we wish to perform the test at the 0.05 level, the rejection region consists of the values of 11 to 14. If 10 were included in the rejection region, the probability of Type I error would exceed the significance level of 0.05. Ten of the 14 boys with an extreme day 1 value had day 2 values that moved in the direction of the mean. Since 10 is not included in the rejection region, we fail to reject the null hypothesis in favor of the alternative at the 0.05 significance level. Although we failed to reject the null hypothesis, the *p*-value of this result is 0.0898.

What is the power of the test — that is, what is the probability of rejecting the null hypothesis when it should be rejected? As we saw in Chapter 8, to find a value for power, we must provide a specific alternative. Let us work with the alternative that π is 0.70. Then the power is easily found from Table B2 ($n = 14$, $\pi = 0.3$, for

$x = 3, 2, 1, 0$, which is equivalent to x = 11, 12, 13, 14 under $\pi = 0.7$). The power is 0.3552 (= 0.1943 + 0.1134 + 0.0407 + 0.0068), not a large value.

It is even easier to perform the sign test using a computer program (see **Program Note 9.1** on the website).

Example 9.2

Two major questions in interlaboratory testing programs are (1) whether or not the measuring instruments are properly calibrated and (2) whether or not the technicians are properly trained. The first question concerns the validity or bias issue, and the second question deals with the reliability or precision issue. In Chapter 2 we looked at an interlaboratory testing program of the CDC. They distributed a blood sample to over 100 randomly selected laboratories throughout the country and asked to measure the lead concentration. The test samples were created to contain the lead concentration of exactly 41 μg/dL (Hunter 1980). The average reported by all participating laboratories was 44 μg/dL with a large variability, ranging from 30 to 60. It appears that both the validity and reliability problems are present. The sign test can be used with the 100 measurements compared with the true value to examine the bias issue between laboratories. We consider the precision issue within one laboratory in the following example.

Suppose that the same CDC sample was mixed in other samples in 13 consecutive days and the following measurements are recorded. The laboratory director wants to examine the bias issue (calibration of instruments) quickly — that is, whether or not these measurements differ significantly from the true value of 41:

$$\underline{45} \quad \underline{43} \quad 40 \quad \underline{44} \quad \underline{49} \quad 36 \quad \underline{51} \quad \underline{46} \quad 35 \quad \underline{50} \quad 41 \quad 38 \quad \underline{47}.$$

If the measuring instrument is properly calibrated, one would expect half the measurements to be above the value of 41 and half to be below the value. If there is a bias problem, the proportion should be greater or less than 0.50. The null and alternative hypotheses are therefore

$$H_0: \pi = 0.50 \text{ versus } H_a: \pi \neq 0.50.$$

In this case, 48 values are above 41 (underscored values) and four are below (8 positives and 4 negatives). One value is exactly 41, and we cannot assign a sign. We drop this observation from the data following the usual practice and analyze twelve observations for which we can determine a positive or negative sign. Thus, the test statistic, the number of positive signs, has a value of 8. We conduct the sign test at the 0.50 level. The critical region for the test can be found from the binomial distribution. For larger sample sizes, the normal approximation to the binomial can be used. We use Table B2 to find the probabilities for a binomial distribution with $n = 12$ and $\pi = 0.50$. Since it is a two-tailed test, we need to look at both tails of the distribution. The p-value of this test is then twice of the probability that X is 8 or more — that is, 2(0.1209 + 0.0537 + 0.0161 + 0.0029 + 0.0002), which gives 0.39. As the p-value is greater than 0.05, we fail to reject the null hypothesis of no

difference in favor of the alternative hypothesis. There is no evidence that the early measurements are significantly different from the true value of 41.

We must realize how much information the sign test discards — a value of 42 is treated exactly the same as a value of 50 or 51. If the data are normal, the *t* test makes better use of the data and gets more out of the data. There are also other nonparametric tests that use more of the information in the data and those will be discussed in the following sections.

The sign test is easy to perform as the test statistic is simply a count of the occurrences of some event — for example, a move toward the mean or a positive difference. The test can also be used with nonnumerical data — for example, in situations in which the outcome is that the subject does or does not feel better. The simplicity of the test is attractive, but with numeric data, in ignoring the magnitude of the values, the sign test does not use all the information in the data. The other tests in this chapter use more of the available information in the data.

9.3 The Wilcoxon Signed Rank Test

Another much more recently developed test that can be used to examine whether or not there is reversion toward the mean in the data in Example 9.1 is the Wilcoxon Signed Rank (WSR) test. An American statistician, Frank Wilcoxon, who worked in the chemical industry, developed this test in 1945. Unlike the sign test which can be used with nonnumeric data, the WSR test requires that the differences in the paired data come from a continuous distribution.

To apply the WSR test to examine whether or not there is reversion toward the mean, we prepare the data as follows: The data for the 14 boys with an extreme day 1 value are shown in Table 9.2. In this table, the differences between day 1 and day 2 values are shown as either a change in the direction of the mean (+) or away from the mean (−). If the day 1 and day 2 values for a boy are the same, then we cannot assign a sign, and

Table 9.2 Days 1 and 2 caloric intakes for the 14 boys with the more extreme caloric intakes on day 1.

					Rank	
ID	Day 1	Day 2	Change (+) Toward the Mean	Change (−) Away from the Mean	+	−
13	1,053	2,484	1,431		13	
14	4,322	2,926	1,396		12	
16	1,753	1,054		699		10
27	3,532	3,289	243		3	
30	2,842	2,849		7		1
33	1,505	1,925	420		4	
41	3,076	2,431	645		9	
50	1,292	810		482		7
51	3,049	2,573	476		6	
101	1,277	2,185	1,092		11	
118	1,781	1,844	63		2	
130	2,773	3,236		463		5
149	1,645	2,269	624		8	
150	1,723	3,163	1,440		14	
				Sum of Ranks	82	23

such a pair would be excluded from the analysis. The absolute differences are ranked from smallest to largest, and the ranks are summed separately for those changes in the direction of mean and for those changes away from the mean. We use R_{WSR} to represent the signed rank sum statistic for the positive differences — in this case, those changes toward the mean.

We now consider the logic behind the testing of R_{WSR}. When there are n observations or pairs of data, the sum of the ranks is the sum of the integers from 1 to n and that sum is $n(n + 1)/2$. The average rank for an observation is therefore $(n + 1)/2$.

The null hypothesis is that the differences have a median of zero and the alternative hypothesis that the median is not equal to zero for a two-sided test or greater (or smaller) than zero for a one-sided test. If the null hypothesis is true, the distribution of the differences will be symmetric, and there should be $n/2$ positive differences and $n/2$ negative differences. Therefore, if the null hypothesis is true, the sum of the ranks for positive (or negative) differences, R_{WSR}, should be $(n/2)$ times the average rank: $(n/2)(n + 1)/2 = n(n + 1)/4$.

The test statistic is the sum of the ranks of positive (or negative) differences, R_{WSR}. For a small sample, Table B9 ($n < 30$) provides boundaries for the critical region for the sum of the ranks of the positive (or negative) differences. To give an idea how these boundaries were determined, let us consider five pairs of observations. The boundaries result from the enumeration of possible outcomes as shown in Table 9.3.

Table 9.3 Positive ranks for a sample of size 5 for 0, 1, and 2 positive ranks.

Number of Positive Ranks	Possible Ranks	Sum of Positive Ranks	Sum of Negative Ranks
0		0	15
1	1	1	14
	2	2	13
	3	3	12
	4	4	11
	5	5	10
2	1, 2	3	12
	1, 3	4	11
	1, 4	5	10
	1, 5	6	9
	2, 3	5	10
	2, 4	6	9
	2, 5	7	8
	3, 4	7	8
	3, 5	8	7
	4, 5	9	6

In Table 9.3, there is no need to show the sum of ranks for 3, 4, and 5 positive ranks because their values are already shown under the sum of the negative rank column. For example, when there are 0 positive ranks, there are 5 negative ranks with a sum of 15. But the sum of 5 positive ranks must also be 15. When there is 1 positive rank, there are 4 negative ranks with the indicated sums. But these are also the sum for the possibilities with 4 positive ranks. The same reasoning applies for 2 and 3 positive ranks.

Based on Table 9.3, we can form Table 9.4, which shows all the possible values of the sum and their relative frequency of occurrence. Using Table 9.4, we see that the smallest rejection region for a two-sided test is 0 or 15, and this gives the probability of

Table 9.4 All possible sums and their relative frequency.

Sum	Frequency	Relative Frequency
0 or 15	1	0.031
1 or 14	1	0.031
2 or 13	1	0.031
3 or 12	2	0.063
4 or 11	2	0.063
5 or 10	3	0.094
6 or 9	3	0.094
7 or 8	3	0.094

a Type I error of 0.062. Thus, in Table B9, there is no rejection region shown for a sample size of 5 and a significance level of 0.05. If the test of interest were a one-sided test, then it would be possible to have a Type I error probability less than 0.05.

Example 9.3

Let us return to the data prepared for the 14 pairs in Table 9.2. We shall perform the test at the 0.05 significance level, the same level used in the sign test. Since this is a one-sided test, we read the boundary above $\alpha \leq 0.05$ under one-sided comparisons shown at the bottom of the table, which is equivalent to $\alpha \leq 0.10$ under two-sided comparisons. Using the row $n = 14$, the critical values are (25, 80). Since our test statistic is 82, greater than 80, we reject the null hypothesis of no regression toward the mean in favor of the alternative that there is regression toward the mean.

This result is inconsistent with the result of the sign test in Example 9.1 and reflects the greater power of the WSR test. This greater power is due to the use of more of the information in the data by the WSR test compared to the sign test. The WSR test incorporates the fact that the average rank for the four changes away from the mean is 5.75 (= [1 + 5 + 7 + 10]/4), less than the average rank of 7.50. This lower average rank of these four changes, along with the fact that there were only four changes away from the mean, caused the WSR test to be significant. The sign test used only the number of changes toward the mean, not the ranks of these changes, and was not significant. Although the sign test failed to reject the null hypothesis, its p-value of 0.0898 was not that different from 0.05.

In applying the WSR test, two types of ties can occur in the data. One type is that some observed values are the same as the hypothesized value or some paired observations are the same — that is, the differences are zero. If this type of tie occurs in an observational unit or pair, that unit or pair is deleted from the data set, and the sample size is reduced by one for every unit or pair deleted. Again, this procedure is appropriate when there are only a few ties in the data. If there are many ties of this type, there is little reason to perform the test.

The other type of tie occurs when two or more differences have exactly the same nonzero value. This has an impact on the ranking of the differences. In this case, convention is that the differences are assigned the same rank. For example, if two differences were tied as the smallest value, each would receive the rank of 1.5, the average

of ranks 1 and 2. If three differences were tied as the smallest value, each would receive the rank of 2, the average of ranks 1, 2, and 3. If there are few ties in the differences, the rank sum can still be used as the test statistic; however, the results of the test are now approximate. If there are many ties, an adjustment for the ties must be made (Hollander and Wolfe 1973), or one of the methods in the next chapter should be used.

Example 9.4

Let us apply the WSR test to the data in Example 9.2. The 13 measurements, the deviations from the true value of 41, and ranks of absolute differences are as follows:

Measures:	45	43	40	44	49	36	51	46	35	50	41	38	47
Differences:	+4	+2	−1	+3	+8	−5	+10	+5	−6	+9	0	−3	+6
Ranks:	5	2	1	3.5	10	6.5	12	6.5	8.5	11	—	3.5	8.5

Note that the average ranking procedure is used for the same values of absolute differences and the rank is not assigned to tenth observation.

Again the investigator wishes to test whether the repeated measurements are significantly different from the value of 41 at the 0.05 significance level. We delete one observation that has no rank. The test statistic (the sum of ranks for positive differences) is 58.5. Table B9 provides boundaries of the critical region. For $n = 12$ and $\alpha \leq 0.05$ under the two-sided comparison the boundaries are (13, 65). Since the test statistic is less than 65, we fail to reject the null hypothesis in favor of the alternative hypothesis at the 0.05 significance level. This conclusion is consistent with the result of the sign test in Example 9.2.

For a large sample, the normal approximation is used. If there are at least 16 pairs of observations used in the calculations, R_{WSR} will approximately follow a normal distribution. As we just saw, the expected value of R_{WSR}, under the assumption that the null hypothesis is true, is $n(n+1)/4$, and its variance can be shown to be $n(n+1)(2n+1)/24$. Therefore, the statistic

$$\frac{|R_{WSR} - [n(n+1)/4]| - 0.5}{\sqrt{n(n+1)(2n+1)/24}}$$

approximately follows the standard normal distribution. The two vertical lines in the numerator indicate the absolute value of the difference — that is, regardless of the sign of the difference, it is now a positive value. The 0.5 term is the continuity correction term, required because the signed rank sum statistic is not a continuous variable.

Let us calculate the normal approximation to the pairs in Example 9.3. The expected value of R_{WSR} is 52.5 (= [14][15]/4), and the standard error is 15.93 ($= \sqrt{(14)(15)(29)/24}$). Therefore, the statistic's value is

$$\frac{|82 - 52.5| - 0.5}{15.93} = 1.82.$$

What is the probability that Z is greater than 1.82? This probability is found from Table B4 to be 0.0344. This agrees very closely with the exact *p*-value of 0.0338. The exact *p*-value is based on 554 of the 16,384 possible signed rank sums having a value of 82 or greater, applying the same logic illustrated in Tables 2 and 3 to the case $n = 14$. Thus, even though n is less than 16, the normal approximation worked quite well in this case. The WSR test can be performed by the computer (see **Program Note 9.1** on the website).

The sign and Wilcoxon Signed Rank tests are both used most frequently in the comparison of paired data, although they can be used with a single population to test that the median has a specified value. In the use of these tests with pre- and postintervention measurement designs, care must be taken to ensure that there are no extraneous factors that could have an impact during the study. Otherwise, the possibility of the confounding of the extraneous factor with the intervention variable is raised. In addition, the research designer must consider whether or not reversion to the mean is a possibility. If extraneous factors or reversion to the mean cannot be ruled out, the research design should be augmented to include a control group to help account for the effect of these possibilities.

9.4 The Wilcoxon Rank Sum Test

Another test developed by Wilcoxon is the Wilcoxon Rank Sum (WRS) test. This test is used to determine whether or not the probability that a randomly selected observation from one population being greater than a randomly selected observation from another population is equal to 0.5. This test is sometimes referred to as the Mann-Whitney test after Mann and Whitney, who later independently developed a similar test procedure for unequal sample sizes. The WRS test also requires that the data come from independent continuous distributions.

This test is appropriate for the following data situation. A nutritionist wishes to compare the proportion of calories from fat for boys in grades 5 and 6 and grades 7 and 8 that are shown in Table 9.5. Preparation of data involves (1) the ranking of all the observed values in the two groups from smallest to largest and (2) summing the ranks separately in each group. The ranks of these values are also shown in the table. We have rounded the proportions to three decimal places, and as a result there is one tie in the data. The tied values were the 16th and 17th smallest observations and hence were assigned the rank of 16.5, the average of 16 and 17. We could have used the fourth decimal place to break the tie, but we chose not to because we wanted to demonstrate how to calculate the ranks when there was a tie. The test statistic, R_{WRS}, is the sum of the ranks for the smaller sample ($n_1 = 14$) — in this case, for the 14 fifth- and sixth-grade boys. The total sample size n is 33 in this example.

If there were no differences in the magnitudes of the proportion of calories from fat variables in the two groups, the rank sum for the smaller sample would be the product of n_1 and the average rank of the n observations in the two groups — that is, $n_1(n + 1)/2$. For this example, the expected value of R_{WRS} under the null hypothesis of no difference would be 238. If the calculated R_{WRS} in Table 9.5 deviate greatly from 238 suggest that the null hypothesis of no difference in magnitudes should be rejected in favor of the alternative hypothesis that one group has larger values than the other. This test can be

Table 9.5 Proportion of calories from fat for boys in grades 5–6 and 7–8.

Grades 5–6		Grades 7–8	
Proportion from Fat	**Rank**	**Proportion from Fat**	**Rank**
0.365	21	0.311	13
0.437	30	0.278	6
0.248	4	0.282	8
0.424	26	0.421	25
0.403	23	0.426	28
0.337[a]	16.5	0.345	18
0.295	11	0.281	7
0.319	14	0.578	33
0.285	9	0.383	22
0.465	32	0.299	12
0.255	5	0.150	2
0.125	1	0.336	15
0.427	29	0.425	27
0.225	3	0.354	19
		0.337[b]	16.5
		0.289	10
		0.438	31
		0.411	24
		0.357	20
Sum of Ranks	224.5		336.5

[a]To four decimals, the value is 0.3373.
[b]To four decimals, the value is 0.3370.

done based on critical values shown in Table B10. In Table B10, the value 2α refers to the two-sided significance level and N_1 and N_2, respectively, refer to the number of observations in the smaller and larger groups. For a one-sided test at $\alpha = 0.05$, the page with $2\alpha = 0.10$ is used.

The critical regions, shown in Table B10, are determined in a similar manner to that for the Wilcoxon Signed Rank statistic. All possible arrangements of size n_1 of n ranks are listed, and the sum of n_1 ranks in each arrangement is found. The p-value of the R_{WRS} is then determined. For a two-sided test, if R_{WRS} is less than the expected sum, the p-value is twice the proportion of the rank sums that are less than or equal to the test statistic. If R_{WRS} is greater than the expected sum, the p-value is twice the proportion of the rank sums that are greater than or equal to R_{WRS}. For a lower tail, one-sided test, the p-value is the proportion of the rank sums that are less than or equal to R_{WRS}. For an upper tail, one-sided test, the p-value is the proportion of the rank sums that are greater than or equal to R_{WRS}.

As an example of determining the rejection region, consider a situation with four observations in each of two samples. The possible ranks are 1 through 8. Table 9.6 shows all possible arrangements of size 4 of these ranks, and Table 9.7 shows the relative frequency of the rank sums.

For a two-sided test that is to be performed at the 0.05 significance level, the rejection region consists of rank sums of 10 and 26. The probability of these two values is 0.0286, which is less than the 0.05 level. Including 11 and 25 in the rejection region increases the probability of the rejection region to 0.0571, which is greater than the 0.05 value. For a lower tail one-sided test to be performed at the 0.05 level, the rejection region is

Table 9.6 Listing of sets of size 4 from the ranks 1 to 8.

Set	Sum of Ranks	Set	Sum of Ranks
1,2,3,4	10	2,3,4,5	14
1,2,3,5	11	2,3,4,6	15
1,2,3,6	12	2,3,4,7	16
1,2,3,7	13	2,3,4,8	17
1,2,3,8	14	2,3,5,6	16
1,2,4,5	12	2,3,5,7	17
1,2,4,6	13	2,3,5,8	18
1,2,4,7	14	2,3,6,7	18
1,2,4,8	15	2,3,6,8	19
1,2,5,6	14	2,3,7,8	20
1,2,5,7	15	2,4,5,6	17
1,2,5,8	16	2,4,5,7	18
1,2,6,7	16	2,4,5,8	19
1,2,6,8	17	2,4,6,7	19
1,2,7,8	18	2,4,6,8	20
1,3,4,5	13	2,4,7,8	21
1,3,4,6	14	2,5,6,7	20
1,3,4,7	15	2,5,6,8	21
1,3,4,8	16	2,5,7,8	22
1,3,5,6	15	2,6,7,8	23
1,3,5,7	16	3,4,5,6	18
1,3,5,8	17	3,4,5,7	19
1,3,6,7	17	3,4,5,8	20
1,3,6,8	18	3,4,6,7	20
1,3,7,8	19	3,4,6,8	21
1,4,5,6	16	3,4,7,8	22
1,4,5,7	17	3,5,6,7	21
1,4,5,8	18	3,5,6,8	22
1,4,6,7	18	3,5,7,8	23
1,4,6,8	19	3,6,7,8	24
1,4,7,8	20	4,5,6,7	22
1,5,6,7	19	4,5,6,8	23
1,5,6,8	20	4,5,7,8	24
1,5,7,8	21	4,6,7,8	25
1,6,7,8	22	5,6,7,8	26

Table 9.7 Frequency and relative frequency of the rank sums for two samples of four observations each.

Rank Sum	Frequency	Relative Frequency
10 or 26	1	0.0143
11 or 25	1	0.0143
12 or 24	2	0.0286
13 or 23	3	0.0429
14 or 22	5	0.0714
15 or 21	5	0.0714
16 or 20	7	0.1000
17 or 19	7	0.1000
18	8	0.1143

10 and 11. It is not possible to perform the test at the 0.01 level because the probability of each rank sum in the Table 9.7 is greater than 0.01. The rejection region we have found here agrees with that shown in Table B10 ($2\alpha = 0.05$, $N_1 = 4$, $N_2 = 4$), the critical region for the WRS test at the 0.05 significance level.

Example 9.5

Now we return to the data regarding the proportion of calories coming from fat shown in Table 9.5. Let us perform the test of the null hypothesis of no difference in the magnitudes of the variable in the two independent populations at the 0.01 significance level. The alternative hypothesis is that there is a difference in the magnitudes. Since this is a two-sided test, extremely large or small values of the test statistic will cause us to reject the null hypothesis. The test statistic is the rank sum of the smaller sample, which is 224.5. Since the test is being performed at the 0.01 significance level, we use Table B10 ($2\alpha = 0.01$) with sample sizes of 14 and 19. The critical values are 168 and 308. If R_{WRS} is less than or equal to 168 or greater than or equal to 308, we reject the null hypothesis in favor of the alternative hypothesis. Since R_{WRS} is 224.5, a value not in the rejection region, we fail to reject the null hypothesis. Based on this test, there is no evidence that fifth- and sixth-grade boys differ from seventh- and eighth-grade boys in terms of the proportion of calories coming from fat.

Computer programs can be used to perform the Mann-Whitney test (see **Program Note 9.2** on the website).

Once we exceed the sample sizes shown in Table B10, or for both n_1 and n_2 greater than or equal to 8, we can use a normal distribution as an approximation for the distribution of the R_{WRS} statistic. As we just saw, the expected value of R_{WRS} is expressed in terms of the sample sizes. Let n_1 be the sample size of the smaller sample, n_2 be the sample size of the other sample, and n be their sum. The mean and variance of R_{WRS}, assuming that the null hypothesis is true, are $n_1(n+1)/2$ and $n_1n_2(n+1)/12$, respectively. Therefore, the statistic

$$\frac{|R_{WRS} - n_1(n+1)/2| - 0.5}{\sqrt{n_1 n_2 (n+1)/12}}$$

approximately follows the standard normal distribution. The 0.5 term is the continuity correction term, required since the rank sum statistic is not a continuous variable.

Let us calculate the normal approximation for the data in Example 9.5. The expected value of R_{WRS} is 238 (= 14[34]/2). The standard error is 27.453 ($=\sqrt{14*19*34/12}$). Therefore, the statistic's value is

$$\frac{|224.5 - 238| - 0.5}{27.453} = 0.4735.$$

Since this is a two-sided test, the p-value is twice the probability that a standard normal variable is greater than 0.4735. Using linear interpolation in Table B4, we find that

$$\Pr\{Z > 0.4735\} = 0.3179$$

and hence the p-value is twice that, or 0.6358.

If there are many ties between the data in the two groups, an adjustment for the ties should be made (Hollander and Wolfe 1973) or a procedure in the next chapter should be used in the analysis of the data.

9.5 The Kruskal-Wallis Test

The Wilcoxon Rank Sum test is limited to the consideration of two populations. In this section, a method for the comparison of the locations (medians) from two or more populations is presented. This method, the Kruskal-Wallis (KW) test, a generalization of the Wilcoxon test, is named after the two prominent American statisticians who developed it in 1952. The KW test also requires that the data come from continuous probability distributions. The hypothesis being tested by the KW statistic is that all the medians are equal to one another, and the alternative hypothesis is that the medians are not all equal.

We first introduce a data situation appropriate for this test. A study examined the effect of weight loss without salt restriction on blood pressure in overweight hypertensive patients (Reisin et al. 1978). Patients in the study all weighed at least 10 percent above their ideal weight, and all were hypertensive. The patients either were not taking any medication or were on medication that had not reduced their blood pressure below 140 mmHg systolic or 90 mmHg diastolic. Three groups of patients were formed. Group I consisted of patients who were not taking any antihypertensive medication and who were placed on a weight reduction program; Group II patients were also placed on a weight reduction program in addition to continuing their antihypertensive medication; and Group III patients simply continued with their antihypertensive medication. Patients already receiving medication were randomly assigned to Groups II or III. Patients were followed initially for two months, and the baseline value was the blood pressure reading at the end of the two-month period. Patients were then followed for four additional months. Changes in weight and blood pressure between Month 2 and Month 6 were measured.

Table 9.8 contains simulated values that are consistent with those reported in the study by Reisin et al. (1978). Besides using simulated values, the only data shown are from the female patients. We wish to determine whether or not there are differences in the median reductions in diastolic blood pressure in the populations of females from which these samples were drawn. To prepare the data for the test we rank all the simulated values in three groups from the smallest to the largest value (1 through 39 in this

Table 9.8 Simulated reductions (mmHg) in diastolic blood pressure for females from month two to month six of follow-up in each of the three treatment groups with ranks of simulated values and sums of ranks.

Only Weight Reduction ($n_1 = 8$)				**Medication and Weight Reduction ($n_2 = 15$)**				**Only Medication ($n_3 = 16$)**			
Simulated Values											
38	10	10	28	19	36	16	36	12	16	0	–12
6	8	33	8	38	28	36	22	14	16	–10	4
				42	24	40	34	–20	–6	18	16
				6	16	30		–14	6	–16	6
Ranks of Simulated Values											
36.5	15.5	15.5	28.5	25	34	21	34	17	21	7	4
10.5	13.5	31	13.5	36.5	28.5	34	26	18	21	5	8
				39	27	38	32	1	6	24	21
				10.5	21	30		3	10.5	2	10.5
Sums of Ranks											
$R_1 = 164.5$				$R_2 = 436.5$				$R_3 = 179$			

case) and sum the ranks separately in each group. Observations with the same value receive the same average rank as above. Table 9.8 also shows the ranks of the values and sums of ranks in each group.

It is possible, although not feasible for any reasonable sample sizes, to explore the rationale underlying this test by examining the sums of the ranks as we had done in the Wilcoxon tests. Since it is not feasible to determine the distribution of the rank sums, Kruskal and Wallis suggested that H, a statistic defined in terms of n_i and R_i, the sample size and rank sum for the ith group, be used as the test statistic. The definition of H is

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$$

where n is the sum of the group sample sizes and k is the number of groups. This statistic follows the chi-square distribution with $(k-1)$ degrees of freedom when the null hypothesis is true. The statistic H follows the chi-square distribution because H can be shown to be proportional to the sample variance of the rank sums which follows a chi-square distribution. Thus, we reject the null hypothesis when the observed value of H is greater than $\chi^2_{k\text{-}1,1\text{-}\alpha}$, and we fail to reject the null hypothesis otherwise.

Example 9.6

Let us conduct the KW test for the weight loss data in Table 9.8. To calculate the test statistic H we need to use the rank sums for the three groups shown, and we must also choose the significance level for the test. Let us perform the test at the 0.10 significance level.

We already have the information required to calculate H. The observed value of H is

$$H = \frac{12}{39(39+1)}\left(\frac{164.5^2}{8} + \frac{436.5^2}{15} + \frac{179^2}{16}\right) - 3(39+1) = 19.133.$$

If the null hypothesis, equality of medians, is true, H follows the chi-square distribution with 2 degrees of freedom. Since 19.133 is greater than 4.61 ($=\chi^2_{2,\,0.90}$), we reject the null hypothesis in favor of the alternative. From Table B7, we see that the p-value of H is less than 0.005. There appears to be a difference in the effects of the different interventions on diastolic blood pressure. Weight reduction can play an important role in blood pressure reduction for overweight patients.

Computer programs can be used to perform the Kruskal-Wallis test (see **Program Note 9.3** on the website).

9.6 The Friedman Test

While the Kruskal-Wallis test is designed to compare k independent groups, the Friedman test is for comparing k dependent groups. The groups are no longer independent when matched samples are assigned to k comparison groups. Referring to the experimental designs discussed in Chapter 6, the Kruskal-Wallis test is suitable for a completely randomized design, and the Friedman test is for a randomized block design. A

distribution-free test for the randomized block design was given by Friedman (1937), and this test is a generalization of the sign test to more than two groups.

The Friedman test starts with ranking of observed values within blocks. The test statistic T suggested by Friedman is defined in terms of the sum of ranks for the ith comparison groups, R_i; the number of blocks, b; and the number of comparison groups, k, as follows:

$$T = \frac{12}{bk(k+1)} \sum_{i=1}^{k} R_i^2 - 3b(k+1).$$

It is similar to H statistic for the Kruskal-Wallis test, but the ranking procedure is different. The ranking in the Friedman test is done separately within blocks recognizing the randomized block design, whereas the Kruskal-Wallis test is based on a single overall ranking reflecting the completely randomized design. The T statistic follows the chi-square distribution with $(k - 1)$ degrees of freedom when the null hypothesis is true. As in the case of H, we reject the null hypothesis when the observed value of T is greater than $\chi^2_{k-1,1-\alpha}$, and otherwise we fail to reject the null hypothesis.

Example 9.7

Effectiveness of insecticides is evaluated based on a randomized block design (Steel and Tome 1960). Four blocks of fields were used for this study. The numbers of living adult plum curculios emerging from separate caged areas of soil treated by five different insecticides and a control (check) were counted. The data in this example are count data ranging from 0 to 217, and the assumptions of normality and equality of variance may be in doubt. Therefore, we decided to use the Friedman to test the null hypothesis at the 0.05 significance level. Table 9.9 shows the data and the ranks within each block.

Based on the data in Table 9.9, the test statistic is calculated as follows:

$$T = \frac{12}{4(6)(6+1)}(12^2 + 6.5^2 + 5.5^2 + 20^2 + 16^2 + 24^2) - 3(4)(6+1) = 19.5.$$

Since the test statistic of 19.5 is greater than the critical value of 11.07 (= $\chi^2_{5,0.95}$), we reject the null hypothesis in favor of the alternative hypothesis that the treatment groups are different.

Table 9.9 The number of living adult plum curculios emerging from caged areas treated by different insecticides (the number in parentheses are ranks within each block).

	Insecticides					
Block	**Lindane**	**Dieldrin**	**Aldrin**	**EPN**	**Chlordane**	**Check**
1	14	7	6	95	37	212
	(3)	(2)	(1)	(5)	(4)	(6)
2	6	1	1	133	31	172
	(3)	(1.5)	(1.5)	(5)	(4)	(6)
3	8	0	1	86	13	202
	(3)	(1)	(2)	(5)	(4)	(6)
4	36	15	4	115	69	217
	(3)	(2)	(1)	(5)	(4)	(6)
Sums of ranks	12	6.5	5.5	20	16	24

Computer programs can be used to perform the Friedman test (see **Program Note 9.4** on the website).

Conclusion

In this chapter, we introduced several of the more frequently used nonparametric tests for continuous data. The nonparametric tests are attractive because they do not require an assumption of the normal distribution. Even when the data do come from normal distributions, these nonparametric tests do not sacrifice much power in comparison to tests based on the normality assumption. Although these tests were designed to be used with continuous data, they are often used with ordered data as well. Their use with ordered data can create problems as there are likely to be more ties for ordered data than for continuous data. In the next chapter, we introduce methods for testing hypotheses about ordered or nominal data, as well as about continuous data that are grouped into categories.

EXERCISES

9.1 The following table below shows the annual average fatality rate per 100,000 workers for each state, data originally introduced in Exercise 7.4. A state is placed into one of three groups according to the National Safe Workplace Institute (NSWI) score. Group 1 consists of states whose NSWI score was above 55, group 2 consists of states with scores of 31 to 55, and group 3 consists of states with scores less than or equal to 30. In Exercise 7.4, we examined the correlation between the fatality rates and the NSWI scores. Here we wish to determine whether or not we believe that the median fatality rates for the three groups of states are the same.

State Fatality Rates per 100,000 Workers by the National Safe Workplace Institute Scores

NSWI Groups								
Low (≤ 30)			Middle (31 to 55)			High (> 55)		
State	Rate	Rank	State	Rate	Rank	State	Rate	Rank
AR	12.5	41	LA	11.2	35	NH	4.5	8
WY	29.5	49	KY	11.9	39	WI	6.3	16
NM	12.0	40	GA	10.3	33	RI	3.3	4
KS	9.1	28	VT	6.8	19	AK	33.1	50
ND	13.8	43	AZ	4.1	6	VA	9.9	32
ID	17.1	47	DE	5.8	13	MI	5.3	10.5
TN	8.1	25	MO	5.3	10.5	OR	11.0	34
HI	6.0	14	MD	5.7	12	MN	4.3	7
AL	9.0	27	NC	7.2	21	CT	1.9	1
MS	14.6	44	IN	7.8	23.5	ME	7.8	23.5
SD	14.7	45	WV	16.2	46	TX	11.7	38
SC	6.7	18	FL	9.3	30.5	MA	2.4	2
UT	13.5	42	CO	9.3	30.5	NY	2.5	3
NE	11.3	36	OK	8.7	26	IL	6.9	20
MT	21.6	48	IA	9.2	29	NJ	3.4	5
NV	11.5	37	OH	4.8	9	CA	6.5	17
			PA	6.1	15			
			WA	7.7	22			
Sum of Ranks		584			420			271

Is there any need to use a statistical test of hypothesis to determine whether or not the median fatality rates of these three groups of states are the same? If there is, what test would you use?

9.2 A study was conducted to determine the effect of short-term, low-level exposure of demolition workers to asbestos fibers and silica-containing dusts (Kam 1989). Twenty-three demolition workers were exposed for 26 consecutive days during the destruction of a three-story building. The dependent variable is the percent reduction in the baseline value of the ratio of the forced expiratory volume in the first second to the forced vital capacity (FEV_1/FVC) compared to the same ratio at the end of the demolition project. None of the exposures to asbestos or silica were above the permissible values. The following table shows the data for the 23 workers, grouped according to the level of exposure to asbestos and silica.

Percent Reduction in Pre- and Postproject FEV_1/FVC Values by Level of Exposure to Asbestos and Silica-Containing Dusts

Higer Exposure ($n = 10$)					Lower Exposure ($n = 13$)				
0.73	0.72	0.70	0.33	0.54	0.42	0.70	0.65	0.62	0.81
0.75	0.67	0.73	0.69	0.59	0.64	0.63	0.60	0.66	0.61
					0.68	0.76	0.65		

Test the hypothesis that there is no difference in the median percent reduction for those with the higher level of exposure compared to those with the lower level of exposure. Use a 5 percent significance level.

9.3 A study was conducted to compare the effectiveness of the applied relaxation method and the applied relaxation method with biofeedback in patients with chronic low back pain (Strong, Cramond, and Mass 1989). Twenty female patients were randomly assigned to each treatment group, and the treatments were then provided. One of the dependent variables studied was the change in the pain rating index — based on the McGill Pain Questionnaire — between pre- and posttreatment. Patients were also followed for a longer period, but those results are not used in this exercise. The actual change data were not shown in the article, but the following table contains hypothetical changes for the two groups.

Hypothetical Data Showing the Changes in Pre- and Posttreatment Values of the McGill Pain Questionnaire for 40 Women Randomly Assigned to the Different Treatments

Relaxation Only										Relaxation with Biofeedback									
10	11	21	18	16	16	15	9	2	19	9	12	7	14	4	2	11	8	9	11
5	18	16	14	12	13	11	13	14	20	6	10	9	7	8	10	6	13	7	8

Use the appropriate one or two-sided test for the null hypothesis of no difference in the median changes in pain rating between the two groups at the 0.10 significance level. Provide the rationale for your choice of either the one-sided or two-sided test.

9.4 The following data are from the 1971 census for Hull, England (Goldstein 1982). The data show by ward, roughly equivalent to a census tract, the number of households per 1000 without a toilet and the corresponding incidence of

infectious jaundice per 100,000 population reported between 1968 and 1973. Group the ward into three groups based on the rate of households without a toilet. Use the Kruskal-Wallis test to determine whether or not there is a difference in the median incidence of jaundice for the three groups at the 0.05 significance level.

Ward	Number of Toilets	Jaundice	Ward	Number of Toilets	Jaundice
1	222	139	12	1	128
2	258	479	13	276	263
3	39	88	14	466	469
4	389	589	15	443	339
5	46	498	16	186	189
6	385	400	17	54	198
7	241	80	18	749	401
8	629	286	19	133	317
9	24	108	20	25	201
10	5	389	21	36	419
11	61	252			

9.5 Exercise 9.4 provides an example of ecological data, data aggregated for a group of subjects. Care must be taken in the use of this type of data (Piantadosi, Byar, and Green 1988). For example, suppose in Exercise 9.4 there was a statistically significant difference in the median incidence of jaundice for the three groups of wards. Is it appropriate to conclude that there is an association between the presence or absence of a toilet in a household and the occurrence of jaundice? Provide the rationale for your answer.

9.6 In the study on Ramipril introduced in Chapter 7, there was a four-week baseline period during which patients took placebo tablets (Walter, Forthofer, and Witte 1987). Of the 160 patients involved in the study, 24 had previously taken medication for high blood pressure, but it had been greater than seven days since they had last taken their medication. These 24 patients had some expectation that medication works. We will examine hypothetical data based on the summary statistics reported to determine whether or not there is a placebo effect — a reduction in blood pressure values associated with taking the placebo — here. The hypothetical systolic blood pressure (SBP in mmHg) values are the following:

Patient Number	Week 0 SBP	Week 4 SBP	Patient Number	Week 0 SBP	Week 4 SBP
1	171	182	13	148	178
2	172	167	14	182	166
3	166	186	15	210	183
4	181	175	16	171	164
5	194	177	17	165	163
6	200	200	18	201	175
7	200	168	19	189	165
8	181	178	20	197	174
9	173	189	21	187	167
10	178	189	22	174	180
11	206	167	23	197	185
12	199	185	24	169	149

Use the sign test to test the hypothesis that the proportion of decreases in SBP between week 0 and week 4 is equal to 0.50 versus the alternative that the proportion of decreases in SBP is greater than 0.50. Use the 0.05 significance level. If there were reversion or regression to the mean here, would that affect our conclusion about the placebo effect? Test the hypothesis of no reversion to the mean at the 0.05 level.

9.7 Use the Wilcoxon Signed Rank test to test the hypothesis that the median change in SBP in Exercise 9.6 is zero versus the alternative hypothesis that the median change is greater than zero. Perform the test at the 0.05 level. Compare your results to those of the sign test. Do you think that there is a placebo effect here?

9.8 As an extension of Example 9.2, of the 13 measurements of lead concentration in the blood, 6 measurements were done in the morning, and the remaining 7 measurements were done in the afternoon. The measurements were as follows:

Mornings:	43	44	51	50	41	47	
Afternoons:	45	40	49	35	46	36	38

Is there any evidence that the morning measurements are different from the afternoon measurements in the differences from the true value of 41? Test the null hypothesis of no difference at the 0.05 significance level using the Wilcoxon Rank Sum test. Would you use a two-sample t test for this data? Why or why not?

9.9 A psychologist investigated the effect of three different patterns of reward (RR = full reinforcement, RU = reinforcement trial followed by unreinforcement trial, UR = unreinforcement trial followed by reinforcement trial) upon the extent of learning an opposing habit (Siegel 1956). Eighteen litters of rats, three in each litter, were trained under the three patterns of reward, and the three rats in each litter were randomly assigned to the three reinforcement patterns. The extent of learning was measured by counting the number of errors made in the trials to compare the three reward patterns. Because the count of errors is probably not an interval measure and the count data exhibits possible lack of homogeneity of variance, the error counts are ranked within each litter:

Letters:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**Sum**
RR:	1	2	1	1	3	2	3	1	3	3	2	2	3	2	2.5	3	3	2	39.5
RU:	3	3	3	2	1	3	2	3	1	1	3	3	2	3	2.5	2	2	3	42.5
UR:	2	1	2	3	2	1	1	2	2	2	1	1	1	1	1	1	1	1	26

What nonparametric test would you use for these data? Perform the test at the 0.05 significance level and interpret the test results

9.10 A group of researchers investigated the effectiveness of an educational intervention designed to improve physicians' knowledge of the costs of common medications and willingness to consider costs when prescribing (Korn et al. 2003). The researchers administered a written survey before and six months after the intervention. Physicians were asked to agree or disagree with the statements about the relevance of cost for prescribing, using a 5-point Likert scale (1 = strongly agree; 2 = somewhat agree; 3 = neutral; 4 = somewhat disagree;

5 = strongly disagree). They used "the Wilcoxon matched-pairs signed-rank test" which measures the effect of intervention considering the nature of the data. A total of 109 pairs of pre- and postsurvey responses were analyzed and *p*-values were reported for selected questions.

Discuss how the Wilcoxon signed-rank test could have been applied to the data. How many zeros do you think the investigators had in the differences in the 5-point scale? How many tied ranks do you think they had? Do you think the test was appropriate for the data?

REFERENCES

Davis, C. E. "The Effect of Regression to the Mean in Epidemiologic and Clinical Studies." *American Journal of Epidemiology* 104:493–498, 1976.

Friedman, M. "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance." *Journal of American Statistical Association* 32:675–701, 1937.

Goldstein, M. "Preliminary Inspection of Multivariate Data." *The American Statistician* 36:358–362, 1982.

Hollander, M., and D. Wolfe. *Nonparametric Statistical Methods*. New York: Wiley, 1973.

Hunter, J. S. "The National System of Scientific Measurement." *Science* 210:869–874, 1980.

Kam, J. K. "Demolition Worker Hazard: The Effect of Short-Term, Low-Level Combined Exposures." *Journal of Environmental Health* 52:162–163, 1989.

Korn, L. M., S. Reichert, T. Simon, and E. A. Halm. "Improving Physicians' Knowledge of the Costs of Common Medications and Willingness to Consider Costs When Prescribing." *Journal of General Internal Medicine* 18:31–37, 2003.

McPherson, R. S., M. Z. Nichaman, H. W. Kohl, D. B. Reed, and D. R. Labarthe. "Intake and Food Sources of Dietary Fat among Schoolchildren in The Woodlands, Texas." *Pediatrics* 86(4):520–526, 1990.

Nesselroade, J. R., S. M. Stigler, and P. B. Baltes. "Regression Toward the Mean and the Study of Change." *Psychological Bulletin* 88:622–637, 1980.

Piantadosi, S., D. P. Byar, and S. B. Green. "The Ecological Fallacy." *American Journal of Epidemiology* 127:893–904, 1988.

Samuels, M. L. "Statistical Reversion Toward the Mean: More Universal Than Regression Toward the Mean." *The American Statistician* 45:344–346, 1991.

Reisin, E., R. Abel, M. Modan, et al. "Effect of Weight Loss without Salt Restriction on the Reduction of Blood Pressure in Overweight Hypertensive Patients." *The New England Journal of Medicine* 298:1–6, 1978.

Siegel, S. *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill, 1956, p. 169–171.

Steel, R. G. D., and J. H. Torrie. *Principles and Procedures of Statistics*. New York: McGraw-Hill, 1960, p. 158–159.

Stigler, S. M. *The History of Statistics: The Measurement of Uncertainty before 1900*. Cambridge, MA: Belknap Press of Harvard University Press, 1986.

Strong, J., T. Cramond, and F. Mass. "The Effectiveness of Relaxation Techniques with Patients Who Have Chronic Low Back Pain." *The Occupational Therapy Journal of Research* 9:184–192, 1989.

Walter U., R. Forthofer, and R. U. Witte. "Dose-Response Relation of Amgiotensin Converting Enzyme Inhibitor Ramipril in Mild to Moderate Essential Hypertension." *American Journal of Cardiology* 59:125D–132D, 1987.