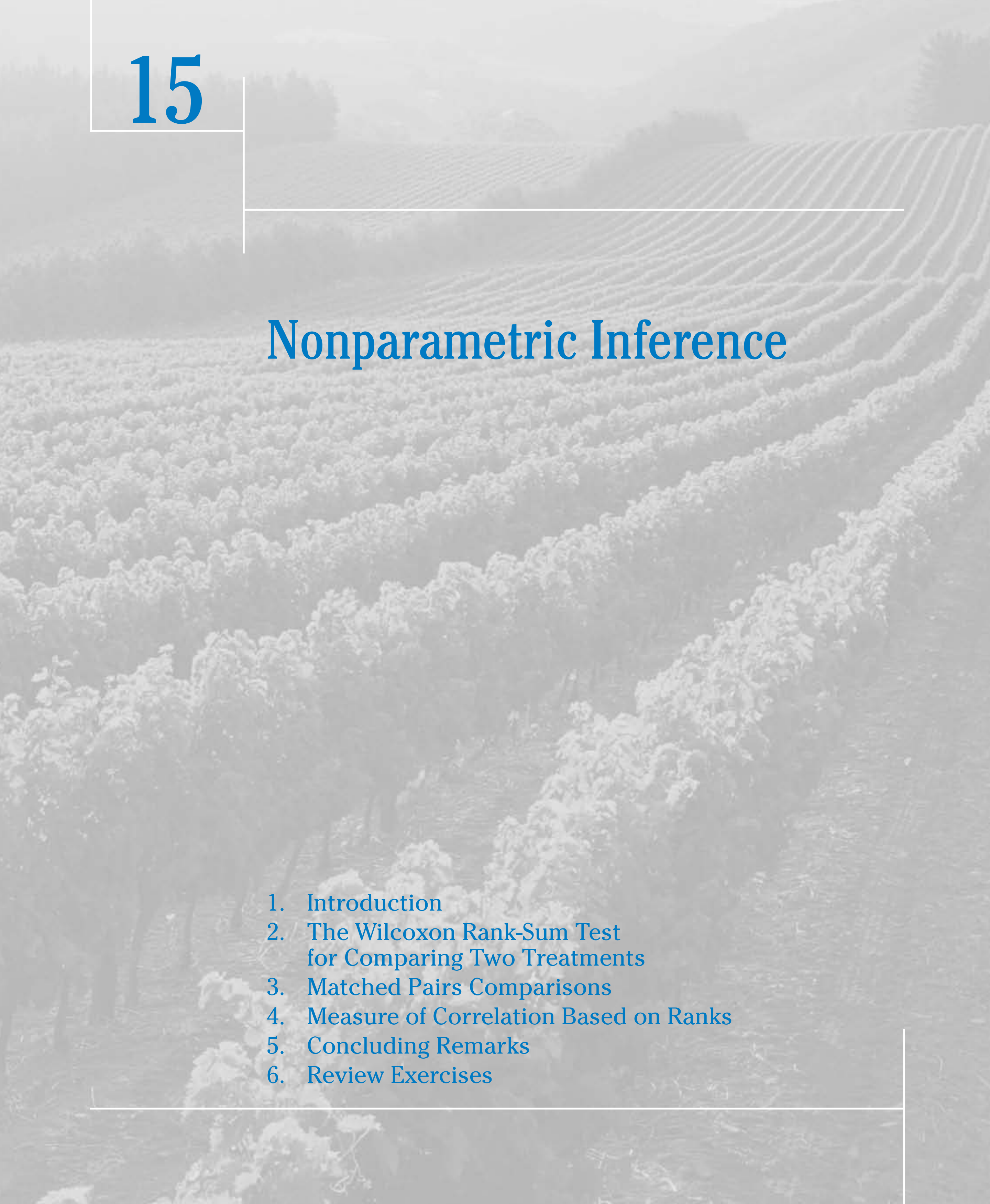


15

Nonparametric Inference

Selecting the Best Vintage

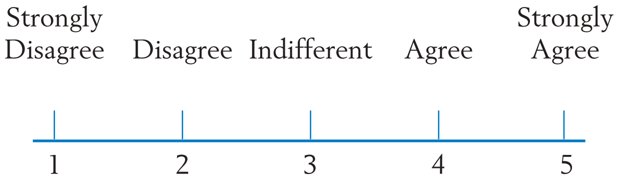
Wines can be ranked without reference to a quantitative scale of measurement. Individuals use non-quantitative characteristics to help select their favorite wines.
© Peter Beck/Corbis Stock Market.

Using their professional judgment, wine critics do give a numeral rating as a guide to overall qualitative placement and wine quality. However, ratings among critics differ on the same bottle of wine and even the meaning of 90 points on one scale is different from that of 90 points on another critic's scale. Ratings can never substitute for your own palate or your own wine tasting rankings.

1. INTRODUCTION

Nonparametric refers to inference procedures that do not require the population distribution to be normal or some other form specified in terms of parameters. Nonparametric procedures continue to gain popularity because they apply to a very wide variety of population distributions. Typically, they utilize simple aspects of the sample data, such as the signs of the measurements, order relationships, or category frequencies. Stretching or compressing the scale of measurement does not alter them. As a consequence, the null distribution of a nonparametric test statistic can be determined without regard to the shape of the underlying population distribution. For this reason, these tests are also called **distribution-free tests.** This distribution-free property is their strongest advantage.

What type of observations are especially suited to a nonparametric analysis? Characteristics like degree of apathy, taste preference, and surface gloss cannot be evaluated on an objective numerical scale, and an assignment of numbers is, therefore, bound to be arbitrary. Also, when people are asked to express their views on a five-point rating scale,



the numbers have little physical meaning beyond the fact that higher scores indicate greater agreement. Data of this type are called **ordinal data,** because only the order of the numbers is meaningful and the distance between the two numbers does not lend itself to practical interpretation. Nonparametric procedures that utilize information only on order or rank are particularly suited to measurements on an ordinal scale.

2. THE WILCOXON RANK-SUM TEST FOR COMPARING TWO TREATMENTS

The problem of comparing two populations based on independent random samples has already been discussed in Section 2 of Chapter 10. Under the assumption of normality and equal standard deviations, the parametric inference procedures were based on Student's t statistic. Here we describe a useful nonparametric procedure named after its proposer F. Wilcoxon (1945). An equivalent alternative version was independently proposed by H. Mann and D. Whitney (1947).

For a comparative study of two treatments A and B, a set of $n = n_A + n_B$ experimental units is randomly divided into two groups of sizes n_A and n_B, respectively. Treatment A is applied to the n_A units, and treatment B to the other n_B units. The response measurements, recorded in a slightly different notation than before, are

Treatment A	X_1,	X_2,	$\ldots, X_{n_A}$
Treatment B	Y_1,	Y_2,	$\ldots, Y_{n_B}$

These data constitute independent random samples from two populations. Assuming that larger responses indicate a better treatment, we wish to test the null hypothesis that there is no difference between the two treatment effects versus the one-sided alternative that treatment A is more effective than treatment B. In the present nonparametric setting, we only assume that the distributions are continuous.

Model: Both Population Distributions Are Continuous

Hypotheses

H_0 : The two population distributions are identical.

H_1 : The distribution of population A is **shifted** to the right of the distribution of population B.

Note that no assumption is made regarding the *shape* of the population distribution. This is in sharp contrast to our t test in Chapter 10, where we assumed that the population distributions were normal with equal standard deviations. Figure 1 illustrates the above hypotheses H_0 and H_1.

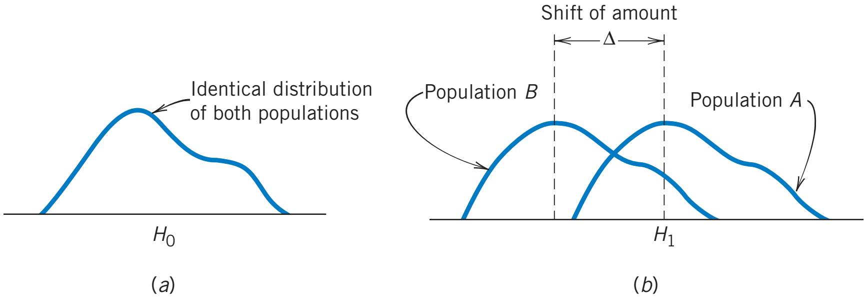


Figure 1 (*a*) Null distribution. (*b*) A shift alternative.

The basic concept underlying the rank-sum test can now be explained by the following intuitive line of reasoning. Suppose that the two sets of observations are plotted on the same diagram using different markings A and B to identify their sources. Under H_0, the samples come from the same population, so that the two sets of points should be well mixed. However, if the larger observations are more often associated with the first sample, for example, we can infer that population A is possibly shifted to the right of population B. These two situations are diagrammed in Figure 2, where the combined set of points in each case is serially numbered from left to right. These numbers are called the **combined sample ranks.** In Figure 2*a*, large as well as small ranks are associated with each sample, whereas in Figure 2*b*, most of the larger ranks are associated with the first sample. Therefore, if we consider the sum of the ranks associated with the first sample as a test statistic, a large value of this statistic should reflect that the first population is located to the right of the second.

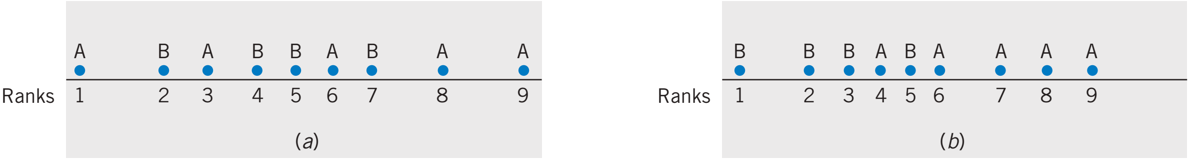


Figure 2 Combined plot of the two samples and the combined sample ranks. (*a*) Mixed ranks. (*b*) Higher ranks are mostly A.

To establish a rejection region with a specified level of significance, we must consider the distribution of the rank-sum statistic under the null hypothesis. This concept is explored in Example 1, where small sample sizes are investigated for easy enumeration.

Example 1 Determining the Null Distribution of the Rank-Sum Statistic

To determine if a new hybrid seedling produces a bushier flowering plant than a currently popular variety, a horticulturist plants 2 new hybrid seedlings and 3 currently popular seedlings in a garden plot. After the plants mature, the following measurements of shrub girth in inches are recorded.

	Shrub Girth (in inches)		
Treatment A (new hybrid)	31.8	39.1	
Treatment B (current variety)	35.5	27.6	21.3

Do these data strongly indicate that the new hybrid produces larger shrubs than the current variety?

SOLUTION We wish to test the null hypothesis

$$H_0 : A \text{ and } B \text{ populations are identical}$$

versus the alternative hypothesis

$$H_1 : \text{Population } A \text{ is shifted from } B \text{ toward larger values}$$

For the rank-sum test, the two samples are placed together and ranked from smallest to largest:

Combined sample ordered observations	21.3	27.6	31.8	35.5	39.1
Ranks	1	2	3	4	5
Treatment	B	B	A	B	A

$$\text{Rank sum for } A \quad W_A = 3 + 5 = 8$$
$$\text{Rank sum for } B \quad W_B = 1 + 2 + 4 = 7$$

Because larger measurements and therefore higher ranks for treatment A tend to support H_1, the rejection region of our test should consist of large values for W_A:

$$\text{Reject } H_0 \quad \text{if} \quad W_A \geq c$$

To determine the critical value c so that the Type I error probability is controlled at a specified level α, we evaluate the probability distribution of W_A under H_0. When the two samples come from the same population, every pair of integers out of $\{1, 2, 3, 4, 5\}$ is equally likely to be the ranks for the two A measurements. There are $\binom{5}{2} = 10$ potential pairs, so that each collection of possible ranks has a probability of $\frac{1}{10} = .1$ under H_0. These rank collections are listed in Table 1 with their corresponding W_A values. The null distribution of W_A can be obtained immediately from Table 1 by collecting the probabilities of identical values (see Table 2). The observed value $W_A = 8$

TABLE 1 Rank Collections for Treatment A with Sample Sizes $n_A = 2$, $n_B = 3$

Ranks of A	Rank Sum W_A	Probability
1,2	3	.1
1,3	4	.1
1,4	5	.1
1,5	6	.1
2,3	5	.1
2,4	6	.1
2,5	7	.1
3,4	7	.1
3,5	8	.1
4,5	9	.1
		Total 1.0

TABLE 2 Distribution of the Rank Sum W_A for Sample Sizes $n_A = 2, n_B = 3$

Value of W_A	3	4	5	6	7	8	9
Probability	.1	.1	.2	.2	.2	.1	.1

has the significance probability $P_{H_0}(W_A \geq 8) = .1 + .1 = .2$. In other words, we must tolerate a Type I error probability of .2 in order to reject H_0. The rank-sum test leads us to conclude that the evidence is not sufficiently strong to reject H_0. Note that even if the A measurements did receive the highest ranks of 4 and 5, a significance level of $\alpha = .1$ would be required to reject H_0.

Guided by Example 1, we now state the rank-sum test procedure in a general setting.

Wilcoxon Rank-Sum Test

Let $X_1, \ldots, X_{n_A}$ and $Y_1, \ldots, Y_{n_B}$ be independent random samples from continuous populations A and B, respectively. To test H_0 : The populations are identical:

1. Rank the combined sample of $n = n_A + n_B$ observations in increasing order of magnitude.
2. Find the rank sum W_A of the first sample.
3. (a) For H_1: Population A is shifted to the right of population B; set the rejection region at the upper tail of W_A.

   (b) For H_1: Population A is shifted to the left of population B; set the rejection region at the lower tail of W_A.

   (c) For H_1: Populations are different; set the rejection region at both tails of W_A having equal probabilities.

A determination of the null distribution of the rank-sum statistic by direct enumeration becomes more tedious as the sample sizes increase. However, tables for the null distribution of this statistic have been prepared for small samples, and an approximation is available for large samples. To explain the use of Appendix B, Table 7, first we note some features of the rank sums W_A and W_B.

The total of the two ranks sums $W_A + W_B$ is a constant, which is the sum of the integers $1, 2, \ldots, n$, where n is the combined sample size. For instance, in Example 1,

$$\begin{aligned} W_A + W_B &= (3 + 5) + (1 + 2 + 4) \\ &= 1 + 2 + 3 + 4 + 5 = 15 \end{aligned}$$

Therefore, a test that rejects H_0 for large values of W_A is equivalent to a test that rejects H_0 for small values of W_B. We can just as easily designate W_B the test statistic and set the rejection region at the lower tail. Consequently, we can always concentrate on the rank sum of the smaller sample and set the rejection region at the lower (or upper) tail, depending on whether the alternative hypothesis states that the corresponding population distribution is shifted to the left (or right).

Second, the distribution of each of the rank-sum statistics W_A and W_B is symmetric. In fact, W_A is symmetric about $n_A(n_A + n_B + 1)/2$ and W_B is symmetric about $n_B(n_A + n_B + 1)/2$. Table 2 illustrates the symmetry of the W_A distribution for the case $n_A = 2, n_B = 3$. This symmetry also holds for the test statistic calculated from the larger sample size.

THE USE OF APPENDIX B, TABLE 7

The **Wilcoxon rank-sum test** statistic is taken as

$$W_S = \text{sum of ranks of the smaller sample in the combined sample ranking}$$

When the sample sizes are equal, take the sum of ranks for either of the samples. Appendix B, Table 7 gives the upper- as well as the lower-tail probabilities:

Upper-tail probability	$P[W_S \geq x]$
Lower-tail probability	$P[W_S \leq x^*]$

By the symmetry of the distribution, these probabilities are equal when x and x^* are at equal distances from the center. The table includes the x^* values corresponding to the x's at the upper tail.

Example 2 Using Table 7 in Appendix B to Set the Rejection Region

Find $P[W_S \geq 25]$ and $P[W_S \leq 8]$ when

$$\text{Smaller sample size} = 3$$
$$\text{Larger sample size} = 7$$

SOLUTION From Table 7, we read $P = P[W_S \geq x]$ opposite the entry $x = 25$, so $.033 = P[W_S \geq 25]$.

The lower tail entry $P[W_S \leq 8]$ is obtained by reading $P[W_S \leq x^*]$ opposite $x^* = 8$. We find $P[W_S \leq 8] = .033$ illustrating the symmetry of W_S.

$$P = P[W_S \geq x] = P[W_S \leq x^*]$$

Smaller Sample Size = 3, Larger Sample Size = 7

	x	P	x^*
	22	—	11
	23	—	10
	24	—	9
→	25	.033	8
	26	—	7
	27	—	6

The steps to follow when using Appendix B, Table 7 in performing a rank-sum test are:

Use the rank-sum W_S of the smaller sample as the test statistic. (If the sample sizes are equal, take either rank sum as W_S.)

1. If H_1 states that the population corresponding to W_S is shifted to the right of the other population, set a rejection region of the form $W_S \geq c$ and take c as the smallest x value for which $P \leq \alpha$.
2. If H_1 states that the population corresponding to W_S is shifted to the left, set a rejection region of the form $W_S \leq c$ and take c as the largest x^* value for which $P \leq \alpha$.
3. If H_1 states that the population corresponding to W_S is shifted in either direction, set a rejection region of the form $W_S \leq c_1$ or $W_S \geq c_2$ and read c_1 from the x^* column and c_2 from the x column, so that $P \leq \alpha/2$.

Example 3 Apply the Rank-Sum Test to Compare Two Geological Formations

Two geological formations are compared with respect to richness of mineral content. The mineral contents of 7 specimens of ore collected from formation 1 and 5 specimens collected from formation 2 are measured by chemical analysis. The following data are obtained:

Mineral Content

Formation 1	7.6	11.1	6.8	9.8	4.9	6.1	15.1
Formation 2	4.7	6.4	4.1	3.7	3.9		

Do the data provide strong evidence that formation 1 has a higher mineral content than formation 2? Test with α near .05.

SOLUTION To use the rank-sum test, first we rank the combined sample and determine the sum of ranks for the second sample, which has the smaller size. The observations from the second sample and their ranks are underlined here for quick identification:

Combined ordered values	$\underline{3.7}$	$\underline{3.9}$	$\underline{4.1}$	$\underline{4.7}$	4.9	6.1	$\underline{6.4}$	6.8	7.6	9.8	11.1	15.1
Ranks	$\underline{1}$	$\underline{2}$	$\underline{3}$	$\underline{4}$	5	6	$\underline{7}$	8	9	10	11	12

The observed value of the rank-sum statistic is

$$W_S = 1 + 2 + 3 + 4 + 7 = 17$$

We wish to test the null hypothesis that the two population distributions are identical versus the alternative hypothesis that the second population, corresponding to W_S, lies to the left of the first. The rejection region is therefore at the lower tail of W_S.

Reading Appendix B, Table 7 with smaller sample size $= 5$ and larger sample size $= 7$, we find $P[W_S \le 21] = .037$ and $P[W_S \le 22] = .053$. Hence, the rejection region with $\alpha = .053$ is established as $W_S \le 22$. Because the observed value falls in this region, the null hypothesis is rejected at $\alpha = .053$. In fact, it would be rejected if α were as low as $P[W_S \le 17] = .009$.

Example 4 Comparing Two Flame-Retardant Materials

Flame-retardant materials are tested by igniting a paper tab on the hem of a dress worn by a mannequin. One response is the vertical length of damage to the fabric measured in inches. The following data (courtesy of B. Joiner) for 5 samples, each taken from two fabrics, were obtained by researchers at the National Bureau of Standards as part of a larger cooperative study.

Fabric *A*	5.7	7.3	7.6	6.0	6.5
Fabric *B*	4.9	7.4	5.3	4.6	6.2

Do the data provide strong evidence that a difference in flammability exists between the two fabrics? Test with α near .05.

SOLUTION The sample sizes are equal, so that we can take the rank sum of either sample as the test statistic. We compute the rank sum for the second sample.

Ordered values	$\underline{4.6}$	$\underline{4.9}$	$\underline{5.3}$	5.7	6.0	$\underline{6.2}$	6.5	7.3	$\underline{7.4}$	7.6
Ranks	$\underline{1}$	$\underline{2}$	$\underline{3}$	4	5	$\underline{6}$	7	8	$\underline{9}$	10

$$W_S = 1 + 2 + 3 + 6 + 9 = 21$$

Because the alternative hypothesis is two-sided, the rejection region includes both tails of W_S. From Appendix B, Table 7, we find that

$$P[W_S \geq 37] = .028 = P[W_S \leq 18]$$

Thus with $\alpha = .056$, the rejection region is $W_S \geq 37$ or $W_S \leq 18$. The observed value does not fall in the rejection region so the null hypothesis is not rejected at $\alpha = .056$.

LARGE SAMPLE APPROXIMATION

When the sample sizes are large, the null distribution of the rank-sum statistic is approximately normal and the test can therefore be performed using the normal table. Specifically, with W_A denoting the rank sum of the sample of size n_A, suppose that both n_A and n_B are large. Then W_A is approximately normally distributed. Under H_0, the distribution of W_A has

$$\text{Mean} = \frac{n_A(n_A + n_B + 1)}{2}$$

$$\text{Variance} = \frac{n_A n_B(n_A + n_B + 1)}{12}$$

Large Sample Approximation to the Rank-Sum Statistic

$$Z = \frac{W_A - n_A(n_A + n_B + 1)/2}{\sqrt{n_A n_B(n_A + n_B + 1)/12}}$$

is approximately $N(0, 1)$ when H_0 is true.

The rejection region for the Z statistic can be determined by using the standard normal table.

Example 5 The Error When Using the Large Sample Approximation

Investigate the amount of error involved in the large sample approximation to the distribution of the rank-sum statistic when $n_A = 9$, $n_B = 10$, and $\alpha = .05$.

SOLUTION The approximate one-sided rejection region is

$$R: \frac{W_A - 9(20)/2}{\sqrt{9 \times 10 \times 20/12}} = \frac{W_A - 90}{12.247} \geq 1.645$$

which simplifies to $R: W_A \geq 110.1$. From Appendix B, Table 7, we find $P[W_S \geq 110] = .056$ and $P[W_S \geq 111] = .047$, which are quite close to $\alpha = .05$. The error decreases with increasing sample sizes.

HANDLING TIED OBSERVATIONS

In the preceding examples, observations in the combined sample are all distinct and therefore the ranks are determined without any ambiguity. Often, however, due to imprecision in the measuring scale or a basic discreteness of the scale, such as a five-point preference rating scale, observed values may be repeated in one or both samples. For example, consider the two samples

Sample 1	20	24	22	24	26
Sample 2	26	28	26	30	18

The ordered combined sample is

$$18 \quad 20 \quad 22 \quad \underbrace{24 \quad 24}_{\text{Tie}} \quad \underbrace{26 \quad 26 \quad 26}_{\text{Tie}} \quad 28 \quad 30$$

Here two ties are present; the first has 2 elements, and the second 3. The two positions occupied by 24 are eligible for the ranks 4 and 5, and we assign the average rank $(4 + 5)/2 = 4.5$ to each of these observations. Similarly, the three tied observations 26, eligible for the ranks 6, 7, and 8, are each assigned the average rank $(6 + 7 + 8)/3 = 7$. After assigning average ranks to the tied observations and usual ranks to the other observations, the rank-sum statistic can then be calculated. When ties are present in small samples, the distribution in Appendix B, Table 7 no longer holds exactly. It is best to calculate the null distribution of W_S under the tie structure or at least to modify the variance in the standardized statistic for use in large samples. See Lehmann [1] for details.

Exercises

15.1 Independent random samples of sizes $n_A = 4$ and $n_B = 2$ are taken from two continuous populations.

(a) Enumerate all possible collections of ranks associated with the smaller sample in the combined sample ranking. Attach probabilities to these rank collections under the null hypothesis that the populations are identical.

(b) Obtain the null distribution of W_S = sum of ranks of the smaller sample. Verify that the tail probabilities agree with the tabulated values.

15.2 Independent samples of sizes $n_A = 2$ and $n_B = 2$ are taken from two continuous populations.

(a) Enumerate all possible collections of ranks associated with population A. Also attach probabilities to these rank collections assuming that the populations are identical.

(b) Obtain the null distribution of W_A.

15.3 Using Appendix B, Table 7, find:

(a) $P[W_S \geq 39]$ when $n_A = 5$, $n_B = 6$.

(b) $P[W_S \leq 15]$ when $n_A = 6$, $n_B = 4$.

(c) The point c such that $P[W_S \geq c]$ is close to .05 when $n_A = 7, n_B = 7$.

15.4 Using Appendix B, Table 7, find:

(a) $P[W_S \geq 57]$ when $n_A = 6$, $n_B = 8$.

(b) $P[W_S \leq 31]$ when $n_A = 8$, $n_B = 6$.

(c) $P[W_S \geq 38$ or $W_S \leq 22]$ when $n_A = 5$ and $n_B = 6$.

(d) The point c such that $P[W_S \leq c]$ is close to .05 when $n_A = 4, n_B = 7$.

(e) The points c_1 and c_2 such that $P[W_S \leq c_1] = P[W_S \geq c_2]$ is about .025 when $n_A = 7, n_B = 9$.

15.5 See Table D. 10 in the data bank. The number of breathing pauses per hour(BPH) helps determine a sleeping disorder. We took a random sample of 3 males, population A, and a random sample of 2 females and obtained the values of BPH

Males	10.39	7.61	2.42
Females	2.58	.41	

(a) Evaluate W_A.

(b) Evaluate W_S.

15.6 The following data pertain to the serum calcium measurements in units of IU/L and the serum alkaline phosphate measurements in units of μg/ml for two breeds of pigs, Chester White and Hampshire:

Chester White

Calcium	116	112	82	63	117	69	79	87
Phosphate	47	48	57	75	65	99	97	110

Hampshire

Calcium	62	59	80	105	60	71	103	100
Phosphate	230	182	162	78	220	172	79	58

Using the Wilcoxon rank-sum procedure, test if the serum calcium level is different for the two breeds.

15.7 Referring to the data in Exercise 15.6, is there strong evidence of a difference in the serum phosphate level between the two breeds?

15.8 A project (courtesy of Howard Garber) is constructed to prevent the decline of intellectual performance in children who have a high risk of the most common type of mental retardation, called cultural-familial. It is believed that this can be accomplished by a comprehensive family intervention program. Seventeen children in the high-risk category are chosen in early childhood and given special schooling until the age of $4\frac{1}{2}$. Another 17 children in the same high-risk category form the control group. Measurements of the psycholinguistic quotient (PLQ) are recorded for the control and the experimental groups at the age of $4\frac{1}{2}$ years.

Do the data at the bottom of the page strongly indicate improved PLQs for the children who received special schooling? Use the Wilcoxon rank-sum test with a large sample approximation: Use $\alpha = .05$.

15.9 The possible synergetic effect of insecticides and herbicides is a matter of concern to many environmentalists. It is feared that farmers who apply both herbicides and insecticides to a crop may enhance the toxicity of the insecticide beyond the desired level. An experiment is conducted with a particular insecticide and herbicide to determine the toxicity of the treatments.

Treatment 1: A concentration of .25 μg per gram of soil of insecticide with no herbicide.

Treatment 2: Same dosage of insecticide used in treatment 1 plus 100 μg of herbicide per gram of soil.

PLQ at Age $4\frac{1}{2}$ Years

Experimental group	105.4	118.1	127.2	110.9	109.3	121.8	112.7	120.3	
Control group	79.6	87.3	79.6	76.8	79.6	98.2	88.9	70.9	

Experimental group	110.9	120.0	100.0	122.8	121.8	112.9	107.0	113.7	103.6
Control group	87.0	77.0	96.4	100.0	103.7	61.2	91.1	87.0	76.4

Several batches of fruit flies are exposed to each treatment, and the mortality percent is recorded as a measure of toxicity. The following data are obtained:

Treatment 1	Treatment 2
40	36
28	49
31	56
38	25
43	37
46	30
29	41
18	

Determine if the data strongly indicate different toxicity levels among the treatments.

15.10 Morphologic measurements of a particular type of fossil excavated from two geological sites provided the following data:

Site A	Site B
1.49	1.31
1.32	1.46
2.01	1.86
1.59	1.58
1.76	1.64

Do the data strongly indicate that fossils at the sites differ with respect to the particular morphology measured?

15.11 If $n_A = 1$ and $n_B = 9$, find

(a) The rank configuration that most strongly supports H_1: Population A is shifted to the right of population B.

(b) The null probability of $W_A = 10$.

(c) Is it possible to have $\alpha = .05$ with these sample sizes?

15.12 One aspect of a study of gender differences involves the play behavior of monkeys during the first year of life (courtesy of H. Harlow, U. W. Primate Laboratory). Six male and six female monkeys are observed in groups of four families during several ten-minute test sessions. The mean total number of times each monkey initiates play with another age mate is recorded.

Males	3.64	3.11	3.80	3.58	4.55	3.92
Females	1.91	2.06	1.78	2.00	1.30	2.32

(a) Plot the observations.

(b) Test for equality using the Wilcoxon rank-sum test with α approximately .05.

(c) Determine the significance probability.

3. MATCHED PAIRS COMPARISONS

In the presence of extensive dissimilarity in the experimental units, two treatments can be compared more efficiently if alike units are paired and the two treatments applied one to each member of the pair. In this section, we discuss two nonparametric tests, the **sign test** and the **Wilcoxon signed-rank test,** that can be safely applied to paired differences when the assumption of normality is suspect. The data structure of a matched pairs experiment is given in Table 3, where the observations on the ith pair are denoted by (X_i, Y_i). The null hypothesis of primary interest is that there is no difference, or

$$H_0 : \text{No difference in the treatment effects}$$

TABLE 3 Data Structure of Matched Pairs Sampling

Pair	Treatment A	Treatment B	Difference $A - B$
1	X_1	Y_1	D_1
2	X_2	Y_2	D_2
.	.	.	.
.	.	.	.
.	.	.	.
n	X_n	Y_n	D_n

THE SIGN TEST

This nonparametric test is notable for its intuitive appeal and ease of application. As its name suggests, the sign test is based on the signs of the response differences D_i. The test statistic is

$$\begin{aligned} S &= \text{number of pairs in which treatment } A \text{ has a higher response than treatment } B \\ &= \text{number of positive signs among the differences } D_1, \ldots, D_n \end{aligned}$$

When the two treatment effects are actually alike, the response difference D_i in each pair is as likely to be positive as it is to be negative. Moreover, if measurements are made on a continuous scale, the possibility of identical responses in a pair can be neglected. The null hypothesis is then formulated as

$$H_0: P[+] = .5 = P[-]$$

If we identify a plus sign as a success, the test statistic S is simply the number of successes in n trials and therefore has a binomial distribution with $p = .5$ under H_0. If the alternative hypothesis states that treatment A has higher responses than treatment B, which is translated $P[+] > .5$, then large values of S should be in the rejection region. For two-sided alternatives $H_1: P[+] \neq .5$, a two-tailed test should be employed.

Example 6 Applying the Sign Test to Compare Two Types of Spark Plugs

Mileage tests are conducted to compare a new versus a conventional spark plug. A sample of 12 cars ranging from subcompacts to full-sized sedans are included in the study. The gasoline mileage for each car is recorded, once with the conventional plug and once with the new plug. The results are given in Table 4. Test the null hypothesis of no difference versus the one-sided alternative that the new plug is better. Use the sign test and take $\alpha \leq .05$.

TABLE 4 Mileage Data

Car Number	New A	Conventional B	Difference $A - B$
1	26.4	24.3	$+2.1$
2	20.3	19.8	$+.5$
3	25.8	26.9	-1.1
4	26.5	27.2	$-.7$
5	32.5	30.5	$+2.0$
6	38.3	37.9	$+.4$
7	22.1	22.4	$-.3$
8	30.1	28.6	$+1.5$
9	22.9	23.1	$-.2$
10	32.6	31.6	$+1.0$
11	27.3	25.5	$+1.8$
12	29.4	28.6	$+.8$

SOLUTION We are to test

$$H_0: \text{No difference between } A \text{ and } B, \text{ or } P[+] = .5$$

versus the one-sided alternative

$$H_1: \text{The new plug } A \text{ is better than the conventional plug } B, \text{ or } P[+] > .5$$

Looking at the differences $A - B$, we can see that there are 8 plus signs in the sample of size $n = 12$. Thus, the observed value of the sign test statistic is $S = 8$. We will reject H_0 for large values of S. Consulting the binomial table for $n = 12$ and $p = .5$, we find $P[S \geq 9] = .073$ and $P[S \geq 10] = .019$. If we wish to control α below .05, the rejection region should be established at $S \geq 10$. The observed value $S = 8$ is too low to be in the rejection region, so that at the level of significance $\alpha = .019$, the data do not sustain the claim of mileage improvement.

The significance probability of the observed value is $P[S \geq 8] = .194$.

An application of the sign test does not require the numerical values of the differences to be calculated. The number of positive signs can be obtained by glancing at the data. Even when a response cannot be measured on a well-defined numerical scale, we can often determine which of the two responses in a pair is better. This is the only information that is required to conduct a sign test.

For large samples, the sign test can be performed by using the normal approximation to the binomial distribution. With large n, the binomial distribution with $p = .5$ is close to the normal distribution with mean $n/2$ and standard deviation $\sqrt{n/4}$.

Large Sample Approximation to the Sign Test Statistic

Under H_0,

$$Z = \frac{S - n/2}{\sqrt{n/4}}$$

is approximately distributed as $N(0, 1)$.

Example 7 Applying the Sign Test to a Large Sample of Beer Preferences

In a TV commercial filmed live, 100 persons tasted two beers A and B and each selected their favorite. A total of $S = 57$ preferred beer A. Does this provide strong evidence that A is more popular?

SOLUTION According to the large sample approximation,

$$Z = \frac{S - n/2}{\sqrt{n/4}} = \frac{57 - 50}{\sqrt{25}} = 1.4$$

The significance probability $P[Z > 1.4] = .0808$ is not small enough to provide strong support to the claim that beer A is more popular.

HANDLING TIES

When the two responses in a pair are exactly equal, we say that there is a **tie.** Because a tied pair has zero difference, it does not have a positive or a negative sign. In the presence of ties, the sign test is performed by discarding the tied pairs, thereby reducing the sample size. For instance, when a sample of $n = 20$ pairs has 10 plus signs, 6 minus signs, and 4 ties, the sign test is performed with the effective sample size $n = 20 - 4 = 16$ and $S = 10$.

THE WILCOXON SIGNED-RANK TEST

We have already noted that the sign test extends to ordinal data for which the responses in a pair can be compared without being measured on a numerical scale. However, when numerical measurements are available, the sign test may result in a considerable loss of information because it includes only the signs of the differences and disregards their magnitudes. Compare the two sets of paired differences plotted in the dot diagrams in Figure 3. In both cases, there are $n = 6$ data points with 4 positive signs, so that the sign test will lead to identical

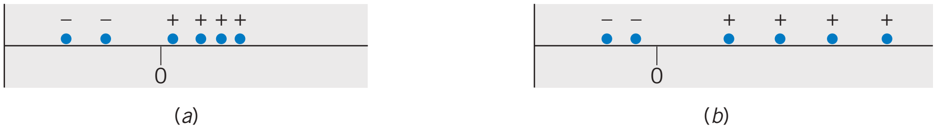


Figure 3 Two plots of paired differences with the same number of + signs but with different locations for the distributions.

conclusions. However, the plot in Figure 3*b* exhibits more of a shift toward the positive side, because the positive differences are farther away from zero than the negative differences. Instead of attaching equal weights to all the positive signs, as is done in the sign test, we should attach larger weights to the plus signs that are farther away from zero. This is precisely the concept underlying the signed-rank test.

In the signed-rank test, the paired differences are ordered according to their numerical values without regard to signs, and then the ranks associated with the positive observations are added to form the test statistic. To illustrate, we refer to the mileage data given in Example 6 where the paired differences appear in the last column of Table 4. We attach ranks by arranging these differences in increasing order of their **absolute** values and record the corresponding signs.

Paired differences	2.1	.5	−1.1	−.7	2.0	.4	−.3	1.5	−.2	1.0	1.8	.8
Ordered absolute values	.2	.3	.4	.5	.7	.8	1.0	1.1	1.5	1.8	2.0	2.1
Ranks	1	2	3	4	5	6	7	8	9	10	11	12
Signs	−	−	+	+	−	+	+	−	+	+	+	+

The signed-rank statistic T^+ is then calculated as

$$\begin{aligned} T^+ &= \text{sum of the ranks associated with positive observations} \\ &= 3 + 4 + 6 + 7 + 9 + 10 + 11 + 12 \\ &= 62 \end{aligned}$$

If the null hypothesis of no difference in treatment effects is true, then the paired differences $D_1, D_2, \ldots, D_n$ constitute a random sample from a population that is symmetric about zero. On the other hand, the alternate hypothesis that treatment A is better asserts that the distribution is shifted from zero toward positive values. Under H_1, not only are more plus signs anticipated, but the positive signs are also likely to be associated with larger ranks. Consequently, T^+ is expected to be large under the one-sided alternative, and we select a rejection region in the upper tail of T^+.

Steps in the Signed-Rank Test

1. Calculate the differences $D_i = X_i - Y_i, i = 1, \ldots, n$.
2. Assign ranks by arranging the absolute values of the D_i in increasing order; also record the corresponding signs.
3. Calculate the signed-rank statistic T^+ = sum of ranks of positive differences D_i.
4. Set the rejection region at the upper tail, lower tail, or at both tails of T^+, according to whether treatment A is stated to have a higher, lower, or different response than treatment B under the alternative hypothesis.

Selected tail probabilities of the null distribution of T^+ are given in Appendix B, Table 8 for $n = 3$ to $n = 15$.

USING APPENDIX B, TABLE 8

By symmetry of the distribution around $n(n+1)/4$, we obtain

$$P[T^+ \geq x] = P[T^+ \leq x^*]$$

when $x^* = n(n+1)/2 - x$. The x and x^* values in Appendix B, Table 8 satisfy this relation. To illustrate the use of this table, we refer once again to the mileage data given in Example 6. There, $n = 12$ and the observed value of T^+ is found to be 62. From the table, we find $P[T^+ \geq 61] = .046$. Thus, the null hypothesis is rejected at the level of significance $\alpha = .046$, and a significant mileage improvement using the new type of spark plug is indicated.

$$P = P[T^+ \geq x] = P[T^+ \leq x^*]$$
$$n = 12$$

	x	P	x^*
	56	·	22
	·	·	·
	·	·	·
	·	·	·
→	61	.046	17
	62		16
	·	·	·
	·	·	·
	·	·	·
	68	·	10

With increasing sample size n, the null distribution of T^+ is closely approximated by a normal curve, with mean $n(n+1)/4$ and variance $n(n+1)(2n+1)/24$.

Large Sample Approximation to Signed-Rank Statistic

$$Z = \frac{T^+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}$$

is approximately distributed as $N(0, 1)$.

This result can be used to perform the signed-rank test with large samples.

Example 8 Applying the Signed-rank Test to Compare Spark Plugs

Refer to the mileage data in Example 6. Obtain the significance probability for the signed-rank test using (a) the exact distribution in Appendix B, Table 8 and (b) the large sample approximation.

SOLUTION

(a) For the mileage data, $T^+ = 62$ and $n = 12$. From Appendix B, Table 8, the exact significance probability is $P[T^+ \geq 62] = .039$.

(b) The normal approximation to this probability uses

$$z = \frac{62 - 12(13)/4}{\sqrt{12(13)(25)/24}} = \frac{23}{12.75} = 1.804$$

From the normal table, we approximate $P[T^+ \geq 62]$ by $P[Z \geq 1.804] = .036$.

The normal approximation improves with increasing sample size.

*HANDLING TIES

In computing the signed-rank statistic, ties may occur in two ways: Some of the differences D_i may be zero or some nonzero differences D_i may have the same absolute value. The first type of tie is handled by discarding the zero values after ranking. The second type of tie is handled by assigning the average rank to each observation in a group of tied observations with nonzero differences D_i.

See Lehmann [1] for instructions on how to modify the critical values to adjust for ties.

Exercises

15.13 In a taste test of two chocolate chip cookie recipes, 13 out of 18 subjects favored recipe A. Using the sign test, find the significance probability when H_1 states that recipe A is preferable.

15.14 Two critics rate the service at six award-winning restaurants on a continuous 0-to-10 scale. Is there a difference between the critics' ratings?

(a) Use the sign test with α below .05.

(b) Find the significance probability.

	Service Rating	
Restaurant	Critic 1	Critic 2
1	6.1	7.3
2	5.2	5.5
3	8.9	9.1
4	7.4	7.0
5	4.3	5.1
6	9.7	9.8

15.15 A social researcher interviews 25 newly married couples. Each husband and wife are independently asked the question: "How many children would you like to have?" The following data are obtained.

	Answer of	
Couple	Husband	Wife
1	3	2
2	1	1
3	2	1
4	2	3
5	5	1
6	0	1
7	0	2
8	1	3
9	2	2
10	3	1
11	4	2
12	1	2
13	3	3
14	2	1
15	3	2
16	2	2
17	0	0
18	1	2
19	2	1
20	3	2
21	4	3
22	3	1
23	0	0
24	1	2
25	1	1

Do the data show a significant difference of opinion between husbands and wives regarding an ideal family size? Use the sign test with α close to .05.

15.16 Use Appendix B, Table 8, to find:

(a) $P[T^+ \geq 54]$ when $n = 11$.

(b) $P[T^+ \leq 32]$ when $n = 15$.

(c) The value of c so that $P[T^+ \geq c]$ is nearly .05 when $n = 14$.

15.17 Use Appendix B, Table 8, to find:

(a) $P[T^+ \geq 65]$ when $n = 12$.

(b) $P[T^+ \leq 10]$ when $n = 10$.

(c) The value c such that $P[T^+ \geq c] = .039$ when $n = 8$.

(d) The values c_1 and c_2 such that $P[T^+ \leq c_1] = P[T^+ \geq c_2] = .027$ when $n = 11$.

15.18 Referring to Exercise 15.14, apply the Wilcoxon signed-rank test with α near .05.

15.19 The null distribution of the Wilcoxon signed-rank statistic T^+ is determined from the fact that under the null hypothesis of a symmetric distribution about zero, each of the ranks $1, 2, \ldots, n$ is equally likely to be associated with a positive sign or a negative sign. Moreover, the signs are independent of the ranks.

(a) Considering the case $n = 3$, identify all $2^3 = 8$ possible associations of signs with the ranks 1, 2, and 3, and determine the value of T^+ for each association.

(b) Assigning the equal probability of $\frac{1}{8}$ to each case, obtain the distribution of T^+ and verify that the tail probabilities agree with the tabulated values.

15.20 A married couple's monthly credit charges are divided into his and hers and the difference, husband's minus wife's, is calculated. A random sample of 30 married couples yielded the Wilcoxon signed-rank statistic $T^+ = 325$. What is the significance probability if the alternative is two-sided?

15.21 In Example 14 of Chapter 10, we presented data on the blood pressure of 15 persons before and after they took a pill.

Before	After	Difference
70	68	2
80	72	8
72	62	10
76	70	6
76	58	18
76	66	10
72	68	4
78	52	26
82	64	18
64	72	− 8
74	74	0
92	60	32
74	74	0
68	72	− 4
84	74	10

(a) Perform a sign test, with α near .05, to determine if blood pressure has decreased after taking the pill.

(b) Perform a Wilcoxon signed-rank test to determine if blood pressure has decreased after taking the pill.

15.22 Charles Darwin performed an experiment to determine if self-fertilized and cross-fertilized plants have different growth rates. Pairs of *Zea mays* plants, one self- and the other cross-fertilized, were planted in pots, and their heights were measured after a specified period of time. The data Darwin obtained were

	Plant height (in $\frac{1}{8}$ inches)	
Pair	Cross-	Self-
1	188	139
2	96	163
3	168	160
4	176	160
5	153	147
6	172	149
7	177	149
8	163	122
9	146	132
10	173	144
11	186	130
12	168	144
13	177	102
14	184	124
15	96	144

Source: C. Darwin, *The Effects of Cross- and Self-Fertilization in the Vegetable Kingdom*, D. Appleton and Co., New York, 1902.

(a) Calculate the paired differences and plot a dot diagram for the data. Does the assumption of normality seem plausible?

(b) Perform the Wilcoxon signed-rank test to determine if cross-fertilized plants have a higher growth rate than self-fertilized plants.

4. MEASURE OF CORRELATION BASED ON RANKS

Ranks may also be employed to determine the degree of association between two random variables. These two variables could be mathematical ability and musical aptitude or the aggressiveness scores of first- and second-born sons on a psychological test. We encountered this same general problem in Chapter 3, where we introduced Pearson's product moment correlation coefficient

$$r = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2} \sqrt{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}}$$

as a measure of association between X and Y. Serving as a descriptive statistic, r provides a numerical value for the amount of linear dependence between X and Y.

Structure of the Observations

The n pairs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ are independent, and each pair has the same continuous bivariate distribution. The $X_1, \ldots, X_n$ are then ranked among *themselves*, and the $Y_1, \ldots, Y_n$ are ranked among *themselves:*

Pair no.	1	2	$\cdots$	n
Ranks of X_i	R_1	R_2	$\cdots$	R_n
Ranks of Y_i	S_1	S_2	$\cdots$	S_n

Before we present a measure of association, we note a few simplifying properties. Because each of the ranks, 1, 2, ..., n must occur exactly once in the set $R_1, R_2, \ldots, R_n$, it can be shown that

$$\overline{R} = \frac{1 + 2 + \cdots + n}{n} = \frac{n+1}{2}$$

$$\sum_{i=1}^{n} (R_i - \overline{R})^2 = \frac{n(n^2 - 1)}{12}$$

for all possible outcomes. Similarly,

$$\bar{S} = \frac{n+1}{2} \qquad \text{and} \qquad \sum_{i=1}^{n} (S_i - \bar{S})^2 = \frac{n(n^2-1)}{12}$$

A measure of correlation is defined by C. Spearman that is analogous to Pearson's correlation r, except that Spearman replaces the observations with their ranks. Spearman's rank correlation r_{Sp} is defined by

$$r_{\text{Sp}} = \frac{\sum_{i=1}^{n} (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^{n} (R_i - \bar{R})^2}\sqrt{\sum_{i=1}^{n} (S_i - \bar{S})^2}} = \frac{\sum_{i=1}^{n}\left(R_i - \frac{n+1}{2}\right)\left(S_i - \frac{n+1}{2}\right)}{n(n^2-1)/12}$$

This rank correlation shares the properties of r that $-1 \leq r_{\text{Sp}} \leq 1$ and that values near $+1$ indicate a tendency for the larger values of X to be paired with the larger values of Y. However, the rank correlation is more meaningful, because its interpretation does not require the relationship to be linear.

Spearman's Rank Correlation

$$r_{\text{Sp}} = \frac{\sum_{i=1}^{n}\left(R_i - \frac{n+1}{2}\right)\left(S_i - \frac{n+1}{2}\right)}{n(n^2-1)/12}$$

1. $-1 \leq r_{\text{Sp}} \leq 1$.
2. r_{Sp} near $+1$ indicates a tendency for the larger values of X to be associated with the larger values of Y. Values near -1 indicate the opposite relationship.
3. The association need not be linear; only an increasing/decreasing relationship is required.

Example 9 Calculating Spearman's Rank Correlation

An interviewer in charge of hiring large numbers of data entry persons wishes to determine the strength of the relationship between ranks given on the basis of an interview and scores on an aptitude test. The data for six applicants are

Interview rank	5	2	3	1	6	4
Aptitude score	47	32	29	28	56	38

Calculate r_{Sp}.

SOLUTION There are 6 ranks, so that $\overline{R} = (n + 1)/2 = 7/2 = 3.5$ and $n(n^2 - 1)/12 = 35/2 = 17.5$. Ranking the aptitude scores, we obtain

Interview R_i	5	2	3	1	6	4
Aptitude S_i	5	3	2	1	6	4

Thus,

$$\begin{aligned}\sum_{i=1}^{n}\left(R_i - \frac{n+1}{2}\right)\left(S_i - \frac{n+1}{2}\right)\\ &= (5 - 3.5)(5 - 3.5) + (2 - 3.5)(3 - 3.5) + \cdots + (4 - 3.5)(4 - 3.5)\\ &= 1.5(1.5) + (-1.5)(-.5) + \cdots + (.5)(.5)\\ &= 16.5\end{aligned}$$

and

$$r_{\text{Sp}} = \frac{16.5}{17.5} = .943$$

The relationship between interview rank and aptitude score appears to be quite strong.

Figure 4 helps to stress the point that r_{Sp} is a measure of any monotone relationship, not merely a linear relation.

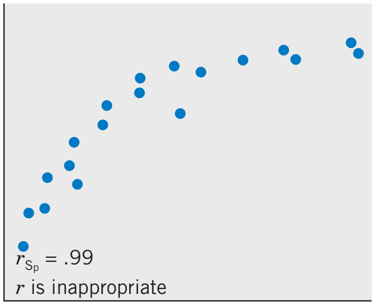


Figure 4 r_{Sp} is a measure of any monotone relationship.

A large sample approximation to the distribution of r_{Sp} is available.

If X and Y are independent,

$$\sqrt{n - 1}\; r_{\text{Sp}} \text{ is approximately distributed as } N(0, 1)$$

provided that the sample size is large.

This approximation leads to a convenient form of a test for independence. Reject

$$H_0 : X \text{ and } Y \text{ are independent}$$

in favor of

$$H_1 : \text{Large values of } X \text{ and } Y \text{ tend to occur together and small values tend to occur together}$$

if

$$\sqrt{n - 1}\, r_{\text{Sp}} \ \geq \ z_{\text{a}}$$

Recall that z_α is the upper α point of a standard normal distribution. Two-tailed tests can also be conducted.

Example 10 Establishing Dependence When Large *X* and *Y* Tend to Occur Together and so Do Small *X* and *Y*

The grade point average (GPA) and Scholastic Achievement Test (SAT) scores for 40 applicants yielded $r_{\text{Sp}} = .4$. Do large values of GPA and SAT tend to occur together? That is, test for lack of independence using $\alpha = .05$.

SOLUTION For $\alpha = .05$, the rejection region is $\sqrt{n - 1}\, r_{\text{Sp}} \geq z_{.05} = 1.96$. Since

$$\sqrt{n - 1}\, r_{\text{Sp}} = \sqrt{39}(.4) = 2.498$$

we reject $H_0 : X$ and Y are independent at level $\alpha = .05$. Large values of GPA and SAT tend to occur together and so do small values.

Exercises

15.23 Refer to Exercise 11.31 and the first four countries in Table 8. The number of Internet users per one hundred residents and the human development index (HDI) are

Internet/100	21.3	26.2	14.3	20.6
HDI	.866	.870	.852	.824

Calculate Spearmann's rank correlation.

15.24 Refer to Exercise 11.68 and the height and speed of the four tallest roller coasters. Given the data

Height	456	420	415	377
Speed	128	120	100	107

Evaluate r_{Sp}

15.25 The following scores are obtained on a test of dexterity and aggression administered to a random sample of 10 high-school seniors.

Student	1	2	3	4	5	6	7	8	9	10
Dexterity	23	29	45	36	49	41	30	15	42	38
Aggression	45	48	16	28	38	21	36	18	31	37

Evaluate Spearman's statistic.

15.26 Referring to Example 10, determine the significance probability of $r_{\text{Sp}} = .4$, using the one-sided test, when $n = 40$.

5. CONCLUDING REMARKS

In contrast to nonparametric procedures, Student's t and the chi-square statistic $(n - 1)S^2/\sigma^2$ were developed to make inferences about the parameters μ and σ of a normal population. These *normal-theory parametric* procedures can be seriously affected when the sample size is small and the underlying distribution is not normal. Drastic departures from normality can occur in the forms of conspicuous asymmetry, sharp peaks, or heavy tails. For instance, a t test with an intended level of significance of $\alpha = .05$ may have an actual Type I error probability far in excess of this value. These effects are most pronounced for small or moderate samples sizes precisely when it is most difficult to assess the shape of the population. The selection of a parametric procedure leaves the data analyst with the question: Does my normality assumption make sense in the present situation? To avoid this risk, a nonparametric method could be used in which inferences rest on the safer ground of distribution-free properties.

When the data constitute measurements on a meaningful numerical scale and the assumption of normality holds, parametric procedures are certainly more efficient in the sense that tests have higher power than their nonparametric counterparts. This brings to mind the old adage, "You get what you pay for." A willingness to assume more about the population form leads to improved inference procedures. However, trying to get too much for your money by assuming more about the population than is reasonable can lead to the "purchase" of invalid conclusions. A choice between the parametric and nonparametric approach should be guided by a consideration of loss of efficiency and the degree of protection desired against possible violations of the assumptions.

Tests are judged by two criteria: control of the Type I error probability and the power to detect alternatives. Nonparametric tests guarantee the desired control of the Type I error probability α, whatever the form of the distribution. However, a parametric test established at $\alpha = .05$ for a normal distribution may suffer a much larger α when a departure from normality occurs. This is particularly true with small sample sizes. To achieve universal protection, nonparametric tests, quite expectedly, must forfeit some power to detect alternatives when normality actually prevails. As plausible as this argument sounds, it is rather surprising that the loss in power is often marginal with such simple procedures as the Wilcoxon rank-sum test and the signed-rank test.

Finally, the presence of dependence among the observations affects the usefulness of nonparametric and parametric methods in much the same manner. When either method is used, the level of significance of the test may seriously differ from the nominal value selected by the analyst.

Caution: When successive observations are dependent, nonparametric test procedures lose their distribution-free property, and conclusions drawn from them can be seriously misleading.

Reference

1. E. L. Lehmann, *Nonparametrics: Statistical Methods Based on Ranks*, Springer, 2007.

USING STATISTICS WISELY

1. A one-sample nonparametric test will provide valid inferences with a small sample size where it may not be possible to check the assumption of normality. Of course, the power of the rank test will generally be less than the normal theory paired *t* test when normality holds.
2. When the two sample sizes are small, it is a good idea to conduct a two-sample Wilcoxon rank-sum test. If software is available, also obtain the corresponding confidence interval for the difference in location. These provide a baseline comparison for the result based on the *t* distribution.
3. If the sample sizes are large enough so the dot diagrams reveal a difference in both location and spread, the Wilcoxon rank-sum test is not appropriate.
4. Remember that nonparametric tests can produce invalid inferences if there is time dependence between the observations.

KEY IDEAS AND FORMULAS

Nonparametric tests obtain their **distribution-free** character because rank orders of the observations do not depend on the shape of the population distribution.

The **Wilcoxon rank-sum test,** based on the test statistic

$$W_A = \text{sum of ranks of the } n_A \text{ observations from population } A, \text{ among all } n_A + n_B \text{ observations}$$

applies to the comparison of two populations. It uses the **combined sample ranks.**

In the paired-sample situation, equality of treatments can be tested using either the **sign test** based on the statistic

$$S = \text{No. of positive differences}$$

or the **Wilcoxon signed-rank** based on the statistic

$$T^+ = \text{sum, over positive differences, of the ranks of their absolute values}$$

The level of a nonparametric test holds whatever the form of the (continuous) population distribution.

Any tie in the observations requires specific handling.

TECHNOLOGY

Nonparametric tests and confidence intervals

MINITAB

One sample—inference about median

Start with the data in *C1*. To find a 95% confidence interval for the median using the sign test:

Stat > Nonparametrics > 1-Sample Sign.
Type *C1* in **Variables.** Click **Confidence interval** and type *0.95* in **Level.**
Click **OK.**

To test a hypothesis concerning the median, instead of **Confidence interval:**

Click **Test Median** and choose the form of the **Alternative** hypothesis. You cannot set the level.

To find a 95% confidence interval for the median using the Wilcoxon signed-rank test:

Stat > Nonparametrics > 1-Sample Wilcoxon.
Type *C1* in **Variables.** Click **Confidence interval** and type *0.95* in **Level.**
Click **OK.**

To test a hypothesis concerning the median, instead of **Confidence interval:**

Click **Test Median** and choose the form of the **Alternative** hypothesis.

Two-sample Wilcoxon test for equality of populations

Start with the data from the first population in *C1* and the data from the second in *C2*. To test at level $\alpha = .05$:

Stat > Nonparametrics > Mann-Whitney.
Type *C1* and *C2* in **Variables.** Type *0.95* in **Confidence level** and select the form of the **Alternative.** Click **OK.**

The output includes a confidence interval for the difference in locations.

6. REVIEW EXERCISES

15.27 From the campus crime statistics in Chapter 11, Table 5, the number of burglaries at the three universities in Florida, population A, and the three in California are

Florida	43	69	90
California	61	74	42

Evaluate W_A

15.28 Using Appendix B, Table 7, find:

(a) $P[W_S \geq 42]$ when $n_1 = 5$, $n_2 = 7$.
(b) $P[W_S \leq 25]$ when $n_1 = 6$, $n_2 = 6$.
(c) $P[W_S \geq 81 \text{ or } W_S \leq 45]$ when $n_1 = 10$, $n_2 = 7$.
(d) The point c such that $P[W_S \geq c] = .036$ when $n_1 = 8$, $n_2 = 4$.

(*continued*)

(e) The points c_1 and c_2 such that $P[W_S \geq c_2] = P[W_S \leq c_1] = .05$ when $n_1 = 3$, $n_2 = 9$.

15.29 (a) Evaluate all possible rank configurations associated with treatment A when $n_A = 3$ and $n_B = 2$.

(b) Determine the null distribution of W_A.

15.30 Five finalists in a figure-skating contest are rated by two judges on a 10-point scale as follows:

Contestants	*A*	*B*	C	*D*	*E*
Judge 1	6	9	2	8	5
Judge 2	8	10	4	7	3

Calculate the Spearman's rank correlation r_{Sp} between the two ratings.

15.31 Using Appendix B, Table 8, find:

(a) $P[T^+ \geq 28]$ when $n = 8$.

(b) $P[T^+ \leq 5]$ when $n = 9$.

(c) The point c such that $P[T^+ \leq c]$ is approximately .05 when $n = 13$.

15.32 Referring to Exercise 15.30, calculate:

(a) The sign test statistic.

(b) The significance probability when the alternative is that Judge 2 gives higher scores than Judge 1.

15.33 In a study of the cognitive capacities of nonhuman primates, 19 monkeys of the same age are randomly divided into two groups of 10 and 9. The groups are trained by two different teaching methods to recollect an acoustic stimulus. The monkeys' scores on a subsequent test are seen below.

Do the data strongly indicate a difference in the recollection abilities of monkeys trained by the two methods? Use the Wilcoxon rank-sum test with α close to .10.

Memory Scores

Method 1	167	149	137	178	179	155	164	104	151	150
Method 2	98	127	140	103	116	105	100	95	131	

15.34 A mixture of compounds called phenolics occurs in wood waste products. It has been found that when phenolics are present in large quantities, the waste becomes unsuitable for use as a livestock feed. To compare two species of wood, a dairy scientist measures the percentage content of phenolics from 6 batches of waste of species A and 7 batches of waste of species B. The following data are obtained.

Percentage of Phenolics

Species *A*	2.38	4.19	1.39	3.73	2.86	1.21	
Species *B*	4.67	5.38	3.89	4.67	3.58	4.96	3.98

Use the Wilcoxon rank-sum test to determine if the phenolics content of species B is significantly higher than that of species A. Use α close to .05.

15.35 (a) Calculate Spearman's rank correlation for the data on Chester Whites in Exercise 15.6.

(b) Test for independence of calcium and phosphate levels using the rejection region

$$\sqrt{n-1}\, r_{\text{Sp}} \geq 1.96 \quad \text{or} \quad \leq -1.96$$

(c) What is the approximate level of significance?

15.36 In the study described in Exercise 1.5, golfers were asked to estimate the size of the hole (cm) in the green by selecting among nine in a board. The size selected and their score for the day for three golfers are

Hole size	11.5	10.0	10.5
Score	84	104	94

Evaluate r_{Sp}.

15.37 Refer to Exercise 10.76. Evaluate:

(a) Sign test statistic.

(b) Signed-rank statistic.

*15.38 ***Confidence interval for median using the sign test.*** Let $X_1, \ldots, X_n$ be a random sample from a continuous population whose median is denoted by M. For testing $H_0\colon M = M_0$, we can use the sign test statistic $S =$ No. of $X_i > M_0$, $i = 1, \ldots, n$. H_0 is rejected at level α in favor of $H_1\colon M \neq M_0$ if $S \leq r$ or $S \geq n - r + 1$, where r is the largest integer satisfying

$$\sum_{x=0}^{r} b(x;\ n,\ .5) \leq \alpha/2$$

If we repeat this test procedure for all possible values of M_0, a $100(1 - \alpha)\%$ confidence interval for M is then the range of values M_0 so that S is in the acceptance region. Ordering the observations from smallest to largest, verify that this confidence interval becomes

$(r + 1)$st smallest to $(r + 1)$st largest observation

(a) Refer to Example 6. Using the sign test, construct a confidence interval for the population median of the differences $A - B$, with a level of confidence close to 95%.

(b) Repeat part (a) using Darwin's data given in Exercise 15.22.