CHAPTER 14

Nonparametric Hypotheses Tests

Man prefers to believe that which he prefers to be true.

Francis Bacon (Moral: You had better test your hypothesis)

CONTENTS

We consider hypotheses tests in situations where the underlying population distribution is unknown and cannot be reasonably assumed to have any specified parametric form, such as normality. We show how to use the sign test to test

Introductory Statistics, DOI: 10.1016/B978-0-12-374388-6.00014-4

hypotheses concerning the median of the distribution. The signed-rank test for testing that a population distribution is symmetric about 0 is introduced. We present the rank-sum test for testing the equality of two population distributions. Finally, we study the runs test that can be used to test the hypothesis that a sequence of 0s and 1s is a random sequence that does not follow any specified pattern.

14.1 INTRODUCTION

Are we making the earth warmer? More precisely, are humans' actions causing the earth's temperature to rise? Even though data appear to indicate that recent annual average temperatures are among the highest ever recorded, this question is surprisingly difficult to answer. One difficulty involves the change in the geographic locations where measurements are taken over time. For instance, past temperature measurements were usually taken in relatively secluded rural regions, whereas present-day measurements are usually taken near cities having many paved roads (that tend to hold heat). This fact, in itself, will result in higher present-day temperature readings. Another difficulty results from uncertainty concerning the accuracy of measurements from long ago. In addition, there is the statistical question as to whether higher present-day temperatures are due to some real change, such as the burning of carbon-based products that might result in the trapping of the sun's energy in the earth's atmosphere, or whether these higher readings are just the chance fluctuations in random samples.

To get a handle on the statistical part of the problem, we want to be able to test whether a data set of temperatures over time represents a random sample from some fixed probability distribution, or whether the distribution of temperatures is itself changing over time.

In considering this question—Is there a fixed underlying distribution of temperatures that is unchanging over time?—it is important to note that we are not specifying in advance the form of this distribution. In particular, since there is no à priori reason to believe that such an underlying distribution would necessarily be a normal distribution, we certainly do not want to make that supposition. Rather we need to develop a hypothesis test that is valid for any underlying type of distribution. Hypotheses tests that can be used in situations where the underlying distribution of the data is not required to have any particular form will be studied in this chapter. Because the validity of these tests does not rest on the assumption of any particular parametric form (such as normality) for the underlying distribution, these tests are called *nonparametric*.

14.2 SIGN TEST

Consider a large population of elements, each of which has a measurable value. Suppose that the distribution of population values is continuous and that we are interested in testing hypotheses concerning the median, or middle value, of this distribution. If the population distribution is normal, then the median is equal to the mean, and the methods of the previous chapters should be employed. However, we do not make the normality assumption here, but we present tests that can be used for any continuous distribution.

Let η denote the median value of the population. That is, exactly half of the members of the population have values less than η, and half have values greater than η. Equivalently, if X is a randomly chosen member of the population, then

$$P\{X < \eta\} = P\{X > \eta\} = \frac{1}{2}$$

Suppose now that we want to test the null hypothesis that the median is equal to some given value m. To obtain a test of

$$\mathrm{H}_0\colon \eta = m$$

against

$$\mathrm{H}_1\colon \eta \neq m$$

let p denote the proportion of the entire population whose value is less than m. That is,

$$p = P\{X < m\}$$

where X is a randomly chosen member of the population. Now if the null hypothesis is true and m is indeed the median, then p will equal $1/2$. On the other hand, if m is not equal to the median, then p will not equal $1/2$. Therefore, a test of the hypothesis that the median is equal to m is equivalent to a test of the null hypothesis

$$\mathrm{H}_0\colon p = \frac{1}{2}$$

against the alternative

$$\mathrm{H}_1\colon p \neq \frac{1}{2}$$

Thus we see that testing the hypothesis that the median is equal to m is equivalent to testing whether a population proportion is equal to $1/2$. This proportion is, of course, equal to the proportion of the population whose value is less than m.

We can now make use of the results of Sec. 9.5.1 to obtain a test of the null hypothesis H_0 that the median of the population is equal to m. Namely, choose a random sample of n elements of the population, and let TS denote the number of them having values less than m. Note that when H_0 is true, TS will be a binomial random variable with parameters n and $1/2$. The test is to reject the null hypothesis if the value of TS is too large or too small. Specifically, if the observed value of TS is i, then the significance-level-α test calls for rejecting H_0 if either

$$P\{N \geq i\} \leq \frac{\alpha}{2}$$

or

$$P\{N \leq i\} \leq \frac{\alpha}{2}$$

where N is a binomial random variable with parameters $(n, 1/2)$. Figure 14.1 illustrates the test.

Because we have assumed that the population distribution is continuous, there should not, in principle, be any data values exactly equal to m. However, since measurements are recorded to the accuracy of the instrumentation used, this may occur in practice. If there are values equal to m, they should be eliminated and the value of n reduced accordingly.

In terms of the p value, the foregoing can be summed up as follows.

To test

$$H_0\colon \eta = m \quad \text{against} \quad H_1\colon \eta \neq m$$

choose a random sample. Discard any values equal to m. Let n be the number of values that remain. Let the test statistic be the number of values that are less than m. If there are i such values, then the p value is

$$p \text{ value} = 2\,\text{Min}(P\{N \leq i\}, P\{N \geq i\})$$

where N is a binomial random variable with parameters n and $1/2$. The null hypothesis is then rejected at all significance levels greater than or equal to the p value and is not rejected otherwise.

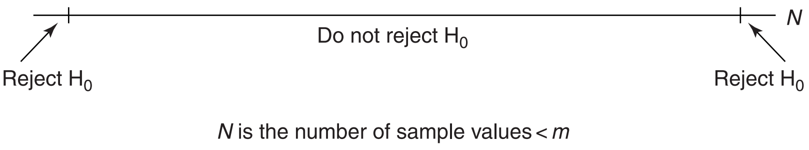


FIGURE 14.1

A test of $H_0\colon \eta = m$ *against* $H_1\colon \eta \neq m$.

To find the p value, it is not necessary to compute both $P\{N \leq i\}$ and $P\{N \geq i\}$. Rather we only need to compute the smaller of these two probabilities. Since $E[N] = n/2$, this will be $P\{N \leq i\}$ when i is small (compared to $n/2$) and $P\{N \geq i\}$ when i is large (compared to $n/2$). When i is near $n/2$, no computations are necessary since both probabilities are close to $1/2$ and so the p value is near 1. Thus, from a practical point of view, the p value can be expressed as

$$p \text{ value} = \begin{cases} 2P\{N \leq i\} & \text{if } i \leq \frac{n}{2} \\ 2P\{N \geq i\} & \text{if } i \geq \frac{n}{2} \end{cases}$$

where N is binomial with parameters n and $1/2$.

■ Example 14.1

The inventory ordering policy of a particular shoe store is partly based on the belief that the median foot size of teenage boys is 10.25 inches. To test this hypothesis, the foot size of each of a random sample of 50 boys was determined. Suppose that 36 boys had sizes in excess of 10.25 inches. Does this disprove the hypothesis that the median size is 10.25?

Solution

Let N be a binomial random variable with parameters $(50, 1/2)$. Since 36 is larger than $50(1/2) = 25$, we see that the p value is

$$p \text{ value} = 2P\{N \geq 36\}$$

We can now use either the normal approximation or Program 5-1 to explicitly compute this probability. Since

$$E[N] = 50 \times \frac{1}{2} = 25 \quad \text{Var}(N) = 50 \times \frac{1}{2} \times \frac{1}{2} = 12.5$$

the normal approximation yields the following:

$$\begin{aligned} p \text{ value} &= 2P\{N \geq 36\} \\ &= 2P\{N \geq 35.5\} \quad \text{(the continuity correction)} \\ &= 2P\left\{\frac{N-25}{\sqrt{12.5}} \geq \frac{35.5-25}{\sqrt{12.5}}\right\} \\ &\approx 2P\{Z \geq 2.97\} \\ &= 0.0030 \quad \text{from Table D.1} \end{aligned}$$

(Program 5-1, which computes binomial probabilities, yields the exact value 0.0026.) Thus the belief that the median shoe size is 10.25 inches is rejected even at the 1 percent level of significance. There appears to be strong evidence that the median shoe size is greater than 10.25. ■

Suppose $X_1, \ldots, X_n$ are the n sample data values. Since the value of the test statistic depends on only the signs, either positive or negative, of the values $X_i - m$, the foregoing test is called the *sign test*.

14.2.1 Testing the Equality of Population Distributions when Samples Are Paired

The sign test can also be used to compare two populations when there is a natural pairing between the elements of their samples. We illustrate this by an example.

■ Example 14.2

An experiment was performed to see if two different sunscreen lotions, both having sun protection factor 15, are equally effective. A group of 12 volunteers exposed their backs to the sun for 1 hour in midday. Each volunteer had brand A sunscreen on one side of his or her spine and brand B on the other side. A measure of the amount of sunburn resulting on both sides of the spine was then determined for each volunteer. If 10 of the volunteers had less of a burn on the side receiving brand A sunscreen than on the side receiving brand B, can we conclude that the brands are not equally effective?

Solution

In this example we can imagine that we have two different populations, the population of backs receiving brand A sunscreen and the population receiving brand B. The "paired" members of the two samples are the two sides of each volunteer's back. Now if the two sunscreens were equally effective, then the median of the difference in sunburn of the two sides of a volunteer's back would equal 0. That is, just by chance, roughly half of the time brand A should perform better than brand B, and vice versa. Thus, we can test for equality of effectiveness by testing the hypothesis that the median of the difference between the brand A and the brand B sunburn of each volunteer is equal to 0.

Since the number of differences whose value is negative is 10, which is greater than $12(1/2) = 6$, we obtain from the sign test that the p value is

$$p \text{ value} = 2P\{N \geq 10\}$$

where N is binomial with parameters $(12, 1/2)$. Since

$$\begin{aligned} P\{N \geq 10\} &= P\{N = 10\} + P\{N = 11\} + P\{N = 12\} \\ &= \frac{12!}{10!\,2!}\left(\frac{1}{2}\right)^{12} + \frac{12!}{11!\,1!}\left(\frac{1}{2}\right)^{12} + \frac{12!}{12!\,0!}\left(\frac{1}{2}\right)^{12} \\ &= \left[\frac{12 \cdot 11}{2 \cdot 1} + 12 + 1\right]\left(\frac{1}{2}\right)^{12} = \frac{79}{4096} \end{aligned}$$

we see that

$$p \text{ value} = \frac{158}{4096} = 0.0386$$

Thus, the null hypothesis of equal effectiveness is rejected at any significance level greater than or equal to 3.86 percent. (For instance, it is rejected at the 5, but not the 1, percent level of significance.) ■

14.2.2 One-Sided Tests

We can also use the sign test to test one-sided hypotheses about a population median. Suppose we want to test

$$H_0: \eta \le m$$

against

$$H_1: \eta > m$$

where η is the population median and m is some specified value. Again, let p denote the proportion of the population whose values are less than m. Now if the null hypothesis is true and so m is at least as large as η, then the proportion of the population whose value is less than m is at least 1/2 (Fig. 14.2). Similarly,

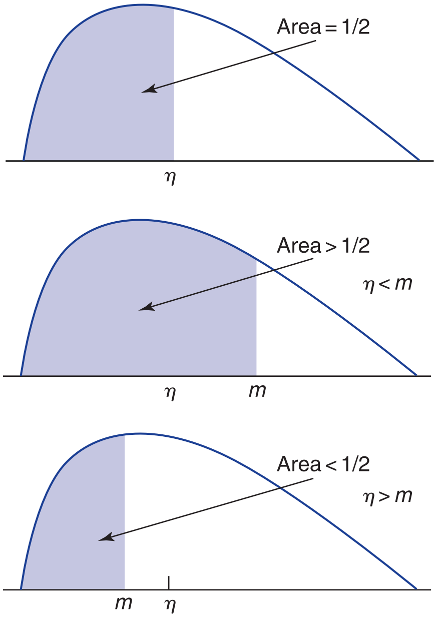


FIGURE 14.2

$P\{X < m\} < 1/2$ *if* $\eta > m$ $P\{X < m\} = 1/2$ *if* $\eta = m$ $P\{X < m\} > 1/2$ *if* $\eta < m$.

if the alternative hypothesis is true and so m is less than η, then the proportion of the population whose value is less than m is less than $1/2$ (see Fig. 14.2). Hence, the preceding is equivalent to testing

$$H_0: p \geq \frac{1}{2}$$

against

$$H_1: p < \frac{1}{2}$$

To use the sign test to test this one-sided hypothesis, choose a random sample of n elements of the population. Suppose that i of them have values that are less than m. The resulting p value is the probability that a value as small as or smaller than i would have occurred by chance if each element had probability $1/2$ of being less than or equal to m. That is, letting N be a binomial random variable with parameters $(n, 1/2)$, we have

$$p \text{ value} = P\{N \leq i\}$$

■ Example 14.3

A bank has decided to build a branch office in a particular community if it can be established that the median annual income of residents of the community is greater than \$40,000. To obtain information, a random sample of 80 families were chosen, and the families were questioned about their income. Of the 80 families, 52 had annual incomes above and 28 had annual incomes below \$40,000. Is this information significant enough, at the 5 percent level of significance, to establish that the median income in the community is greater than \$40,000?

Solution

We need to see if the data are sufficiently strong to reject the null hypothesis when testing

$$H_0: \eta \leq 40 \qquad \text{against} \qquad H_1: \eta > 40$$

If p is the proportion of families in the population with annual incomes below \$40,000, then this is equivalent to testing

$$H_0: p \geq \frac{1}{2} \qquad \text{against} \qquad H_1: p < \frac{1}{2}$$

Since 28 of the 80 sampled families have annual incomes below \$40,000, the p value of the data is

$$p \text{ value} = P\{N \leq 28\}$$

where N is a binomial random variable with parameters $(80, 1/2)$. Using Program 5-1 (or the normal approximation) yields

$$p \text{ value} = 0.0048$$

For such a low p value, the null hypothesis that the median income is less than or equal to \$40,000 is rejected, thus establishing that the median income almost certainly exceeds this value. ■

A test of the one-sided null hypothesis that the median is at least m is similar to the preceding. Thus the one-sided tests are as follows.

One-Sided Hypotheses Tests of the Median

To test either

$$H_0\colon \eta \le m \qquad \text{against} \qquad H_1\colon \eta > m \tag{14.1}$$

or

$$H_0\colon \eta \ge m \qquad \text{against} \qquad H_1\colon \eta < m \tag{14.2}$$

choose a random sample from the population. Remove all values equal to m. Suppose n remain. Let TS denote the number of data values that are less than m. If TS is equal to i, then the p values are

$$p \text{ value} = P\{N \le i\} \quad \text{in case } (14.1)$$
$$p \text{ value} = P\{N \ge i\} \quad \text{in case } (14.2)$$

where N is a binomial random variable with parameters n and $1/2$.

PROBLEMS

1. The published figure for the median systolic blood pressure of middle-aged men is 128. To determine if there has been any change in this value, a random sample of 100 men have been selected. Test the hypothesis that the median is equal to 128 if
   (a) 60 men have readings above 128
   (b) 70 men have readings above 128
   (c) 80 men have readings above 128
   In each case, determine the p value.
2. In 2001, the median household income for the state of Connecticut was \$52,758. A recent survey randomly sampled 250 households and discovered that 42 percent had incomes below the 2001 median and 58 percent

had incomes above it. Does this establish that the median household income in Connecticut is no longer the same as in 2001? What is the p value?

3. Fifty students at the police academy took target practice, using two different types of guns. Each student took half of her or his shots with the less expensive gun and the other half with the more expensive gun. If 29 students had higher scores with the less expensive gun, does this establish that the two guns are not equally effective? Use the 5 percent level of significance.

4. A dermatology clinic wants to compare the effectiveness of a new hand cream and the one it presently recommends to patients suffering from eczema. To gather information, half of its patients are told to put the new skin cream on their left hand and the old cream on their right hand each night for one week; the other half are told to put the new cream on their right hand and the old one on their left. Each patient is examined after one week. Suppose that for 60 percent of the patients the hand receiving the new cream showed greater improvement than the one receiving the old cream.

   When the number of patients involved is equal to

   (a) 10 (b) 20 (c) 50 (d) 100 (e) 500

   do these data prove that the two creams are not equally effective? Use the 5 percent level of significance. Also find the p value in each case.

5. A statistics instructor has made up an examination for a large class of students. She wants the median score on the examination to be at least 72 and thinks that this test will enable her to reach her goal. To be cautious, she has randomly chosen 13 students to take the examination early. If their scores are

   65, 79, 77, 90, 56, 60, 65, 80, 70, 69, 83, 69, 65

   should the hypothesis that the median score will be at least 72 be rejected? Use the 5 percent level of significance.

6. The median selling price of a home in a certain residential community has been steady at \$122,000 for the past 2 years. To determine if the median price has increased, a random sample of 20 recently sold homes were chosen. The selling prices of these homes were (in units of \$1000)

   144, 116, 125, 128, 96, 92, 163, 130, 120, 142, 155,
   133, 110, 105, 136, 140, 124, 130, 88, 146

   Are these data strong enough to establish that the median price has increased? Use the 5 percent level of significance.

7. To test the hypothesis that the median weight of 16-year-old females from Los Angeles is at least 110 pounds, a random sample of 200 such females were chosen. If 120 females weighed less than 110 pounds, does this discredit the hypothesis? Use the 5 percent level of significance. What is the p value?
8. In an attempt to prove that fish oil lowers blood cholesterol levels, a nutritionist instructed 24 volunteers to take a certain fish oil supplement for 3 months. After this time each volunteer had his or her blood cholesterol level checked. Suppose that a comparison with their levels before the beginning of the experiment showed that 16 of the 24 volunteers experienced a reduction in cholesterol levels.
   (a) What are the null and the alternative hypotheses?
   (b) Is the null hypothesis rejected? Use the 5 percent level of significance.
   (c) What is the p value?

14.3 SIGNED-RANK TEST

In Sec. 14.2.1 we saw how the sign test could be used to test the null hypothesis that two populations have the same distribution of values, when the data consisted of paired samples. To test this hypothesis, we considered the differences of the paired-sample values. We noted that if the null hypothesis were true, then the median of these differences would be 0. The sign test was then used to test this latter hypothesis.

The only information needed for the sign test of the equality of two population distributions, when paired samples are used, is the number of times the first data value in a pair is larger than the second. That is, the sign test does not require the actual values of the data pairs, only knowledge of which is larger. However, although it is easy to use, the sign test is not a particularly efficient test of the null hypothesis that the population distributions are the same. For if this null hypothesis is indeed true, then not only will the distribution of paired differences have median zero but it will also have the stronger property of being symmetric about zero. That is, for any number x it will be just as likely for the first value in the pair to be larger than the second by the amount x as for the second value to be larger than the first by this amount (see Fig. 14.3). The sign test, however, does not check for symmetry of the distribution of differences, only that its median value is equal to zero.

For instance, suppose that the data consist of 12 paired values whose differences are as follows:

$$2, 5, -0.1, -0.4, -0.3, 9, 7, 8, 12, -0.5, -1, -0.6$$

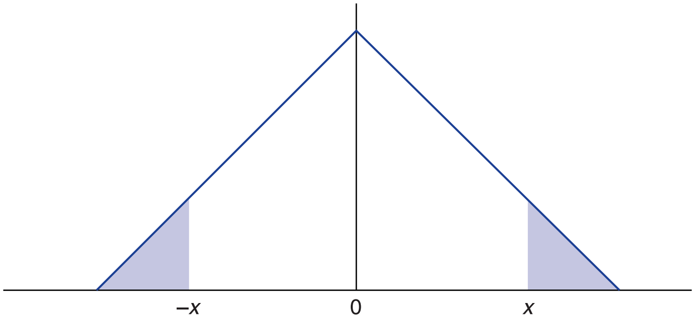


FIGURE 14.3

A density symmetric about 0*. Areas of shaded regions are equal.*

Since six of the differences are positive and six are negative, this data set is perfectly consistent with the hypothesis that the median of the differences is 0. On the other hand, since all the large values are on the positive side, the data do not appear to be consistent with the hypothesis that their distribution is symmetric about 0. Thus, it seems highly unlikely that the population distributions are equal.

Suppose again that we want to use data consisting of paired samples to test the hypothesis that two population distributions are equal. We now present a test that is more sensitive than the sign test. It is called the *signed-rank test*, and it proceeds by testing whether the distribution of the differences of the paired values is symmetric about 0.

Suppose that paired samples of size n are chosen from the two populations. Let D_i denote the difference between the first population value and the second population value of the ith pair, for $i = 1, \dots, n$. Now order these n differences according to their absolute values. That is, the first difference should be the value of D_i having smallest absolute value, and so on. The test statistic for the signed-rank test is the sum of the ranks (or positions) of the negative numbers in the resulting sequence.

■ Example 14.4

Suppose the data consist of the following four paired-sample values:

i	X_i	Y_i
1	4.6	6.2
2	3.8	1.5
3	6.6	11.7
4	6.0	2.1

The differences $X_i - Y_i$ are thus

$$-1.6, 2.3, -5.1, 3.9$$

Ordering these differences according to absolute value, from the smallest to the largest, gives the following:

$$-1.6, 2.3, 3.9, -5.1$$

Since the ranks of the negative values are 1 and 4, the value of the signed-rank test statistic is

$$\text{TS} = 1 + 4 = 5$$

In other words, since the ranked differences in positions 1 and 4 are negative, $\text{TS} = 1 + 4 = 5$. ■

The signed-rank test is like the sign test, in that it considers those data pairs in which the first population value is less than the second. But whereas the sign test gives equal weight to each such pair, the signed-rank test gives larger weights to the pairs whose differences are farthest from zero.

The signed-rank test calls for the rejection of the null hypothesis when we are testing

$$\text{H}_0\text{: two population distributions are equal}$$

against

$$\text{H}_1\text{: two population distributions are not equal}$$

if the test statistic TS is either sufficiently large or sufficiently small. A large value of TS indicates that the majority of the larger values of the differences have negative signs; whereas a small value indicates that the majority have positive signs. Either situation would be evidence against the symmetry of the distribution of differences and thus evidence against H_0.

If the value of the test statistic is t, then the signed-rank test rejects H_0 if either

$$P\{\text{TS} \le t\} \le \frac{\alpha}{2}$$

or

$$P\{\text{TS} \ge t\} \le \frac{\alpha}{2}$$

Here, α is the level of significance, and the probabilities are to be computed under the assumption that H_0 is true. Equivalently, we have the following statement concerning the p value.

Suppose the value of TS is t. The p value of the signed-rank test is given by

$$p \text{ value} = 2\,\mathrm{Min}(P\{\mathrm{TS} \le t\}, P\{\mathrm{TS} \ge t\})$$

where the probabilities are to be determined under the assumption that H_0 is true.

To be able to implement the signed-rank test, we need to be able to compute the preceding probabilities. The key to accomplishing this is the fact than when H_0 is true, and so the distribution of differences is symmetric about zero, each of the differences is equally likely to be either positive or negative, independent of the others. Program 14-1 makes use of this fact to explicitly determine the necessary probabilities and the resulting p value. The inputs needed are the sample size n and the value of the test statistic TS.

Example 14.5

A psychology instructor wanted to see if students would perform equally well on two different examinations. He selected 12 students, who all agreed to take part in the experiment. Six of the students were given examination A, and the other six examination B. On the next day the students were tested on the examination they had not yet taken. Thus, each of the 12 students took both examinations. The following pairs of scores were obtained by the students on the two examinations:

	Student											
Examination	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
A	763	419	586	920	881	758	262	332	717	909	940	835
B	797	404	576	855	762	707	195	341	728	817	947	849

Thus, for instance, student 3 scored 586 on examination A and 576 on test B. The paired differences are as follows:

$$-34, 15, 10, 65, 119, 51, 67, -9, -11, 92, -7, -14$$

Ordering these in increasing order of their absolute values gives

$$-7, -9, 10, -11, -14, 15, -34, 51, 65, 67, 92, 119$$

Since the differences in positions 1, 2, 4, 5, and 7 are negative, the value of the test statistic is

$$\mathrm{TS} = 1 + 2 + 4 + 5 + 7 = 19$$

To obtain the p value, we now run Program 14-1, which computes the p value for the signed-rank test that a population distribution is symmetric about 0. Our sample size is 12 and our observed value of the sum of signed ranks is 19. The p value as computed by Program 14-1 is 0.1293945.

Thus, the p value is 0.129, and so the null hypothesis that the distributions of student scores on the two examinations are identical cannot be rejected at the 10 percent level of significance. ■

By making use of the fact that the ordered differences are independent random variables that are each equally likely to be either positive or negative, it can be established that when H_0 is true, the mean and variance of TS are given by, respectively,

$$E[\mathrm{TS}] = \frac{n(n+1)}{4}$$

and

$$\mathrm{Var}(\mathrm{TS}) = \frac{n(n+1)(2n+1)}{24}$$

In addition, it can be shown that for moderately large values of n, TS will have a distribution, when H_0 is true, that is approximately normal with the preceding mean and variance. These facts enable us to approximate the p value when Program 14-1 is not available.

■ Example 14.6

Let us see how well the normal approximation of the p value works for the data of Example 14.5. Since $n = 12$, we obtain from the preceding formulas that, when H_0 is true,

$$E[\mathrm{TS}] = \frac{12 \cdot 13}{4} = 39 \quad \mathrm{Var}(\mathrm{TS}) = \frac{12 \cdot 13 \cdot 25}{24} = 162.5$$

The value of the test statistic is 19. Since this value is less than $E[\mathrm{TS}]$, it is clear that $P\{\mathrm{TS} \le 19\}$ is smaller than $P\{\mathrm{TS} \ge 19\}$. Therefore,

$$\begin{aligned} p \text{ value} &= 2P\{\mathrm{TS} \le 19\} \\ &= 2P\{\mathrm{TS} \le 19.5\} \quad \text{(continuity correction)} \\ &= 2P\left\{\frac{\mathrm{TS} - 39}{\sqrt{162.5}} \le \frac{19.5 - 39}{\sqrt{162.5}}\right\} \\ &\approx 2P\{Z \le -1.530\} \\ &= 0.126 \end{aligned}$$

Thus, the normal approximation yields an approximate p value that is quite close to the actual p value of 0.129 given in Example 14.5. ■

A general rule of thumb is that the normal approximation to the p value will be quite good provided that n, the number of paired-data values, is at least 25. (In fact, this may be rather conservative, since in our example the approximation was quite good and n was only equal to 12.)

■ Example 14.7

Suppose from a sample of 25 paired values the value of TS is 238. Assuming that the distribution of differences is symmetric about 0, we see from the preceding formulas that

$$E[\text{TS}] = \frac{25 \cdot 26}{4} = 162.5$$

$$\sqrt{\text{Var(TS)}} = \sqrt{\frac{25 \cdot 26 \cdot 51}{24}} = 37.165$$

The normal approximation to the p value is thus

$$\begin{aligned} p \text{ value} &= 2P\{\text{TS} \geq 238\} \\ &= 2P\{\text{TS} \geq 237.5\} \\ &= 2P\left\{\frac{\text{TS} - 162.5}{37.165} \geq \frac{237.5 - 162.5}{37.165}\right\} \\ &\approx 2P\{Z \geq 2.018\} \\ &= 0.0436 \end{aligned}$$

On the other hand, using Program 14-1 yields the exact p value:

$$p \text{ value} = 0.0422$$

Thus, once again we see that the approximation is quite close to the exact value. ■

14.3.1 Zero Differences and Ties

If a difference has value 0 (because its paired values are equal), then that data value should be discarded and the value of n should be reduced by 1.

If some of the differences have the same absolute value, then the weight given to a negative difference should be the average of the ranks of all the differences with the same absolute value. For instance, if the differences are

$$-1, 3, 8, -3$$

then the ordered differences are

$$-1, 3, -3, 8$$

Since there is a tie between the second and third differences, the value of the test statistic is

$$\text{TS} = 1 + \frac{2+3}{2} = 3.5$$

Program 14-1 should not be used if there are any ties. Instead the normal approximation should be employed.

Table 14.1 sums up the signed-rank test.

Table 14.1 Signed-Rank Test

It tests the hypothesis that two population distributions are equal by using paired samples, where $X_1, \ldots, X_n$ are the sample from the first population; $Y_1, \ldots, Y_n$ are the sample from the second population; X_i and Y_i are paired; and $D_i = X_i - Y_i$, $i = 1, \ldots, n$.

To test

$$\text{H}_0\text{: distribution of } D_i \text{ is symmetric about } 0$$

against

$$\text{H}_1\text{: distribution of } D_i \text{ is not symmetric about } 0$$

first eliminate any D_i equal to 0. Change the value of n to let it reflect the number of non-zero differences. Take TS equal to the sum of the positions of the negative differences D_i, when the D_i are ranked in increasing order of their absolute values. If two or more of the D_i values are equal, then they are each given a rank equal to the average of their ranks.

If $\text{TS} = t$, then

$$p \text{ value} = 2\,\text{Min}(P\{\text{TS} \le t\}, P\{\text{TS} \ge t\})$$

where the probabilities are to be computed under the assumption that H_0 is true. The probability can be approximated by using the fact than when H_0 is true, TS is approximately a normal random variable with mean and variance, respectively, given by

$$E[\text{TS}] = \frac{n(n+1)}{4} \qquad \text{Var(TS)} = \frac{n(n+1)(2n+1)}{24}$$

If there are no ties, then the exact p value can be obtained by running Program 14-1.

PROBLEMS

1. Determine the value of the signed-rank test statistic if the differences of paired values are as follows.
   (a) −17, 33, 22, −8, 55, −41, −18, 40, 39, 14, −88, 99, 102, −5, 7
   (b) 44, 2, 1, −0.4, −3, −13, 44, 50, 1.1, −2.2, 0.01, −4, −6.6
   (c) 12, 15, 19, 8, −3, −7, −22, −55, 48, 31, 89, 92
2. Assuming that the difference of paired values has a distribution that is symmetric about 0, determine the mean and variance for the signed-rank test statistic for each of the parts of Prob. 1.
3. For each part of Prob. 1, find the p value of the hypothesis that the distribution of the differences is symmetric about 0. Use the normal approximation.
4. Compare the answers obtained in Prob. 3 with the exact p values given by Program 14-1.
5. A history professor wondered if the student graders for her course tended to take into account whether term papers were handwritten or typed. As an experiment, the professor divided up 28 students into 14 pairs. Each pair of students was regarded by the professor to have roughly the same abilities. The professor then assigned a project and asked one member of each pair to turn in a handwritten report and the other member to turn in a typed report. For each pair, the decision as to which student was asked for the handwritten report was based on the flip of a coin. The grades given to the projects were as follows:

Pair	Handwritten	Typed
1	83	88
2	75	91
3	75	72
4	60	70
5	72	80
6	55	65
7	94	90
8	85	89
9	78	85
10	96	93
11	80	86
12	75	79
13	66	64
14	55	68

(a) Would you conclude that how the paper is presented, either typed or handwritten, had an effect on the score given? Use the 5 percent level of significance.

(b) What is the p value?

6. To test the effectiveness of sealants on reducing cavities, half the teeth of 100 children were treated and the other half left untreated. After 6 months the difference between the number of cavities in the treated and untreated teeth of each child was determined. The signed-rank test statistic for these differences was 1830. Can we conclude, at the 5 percent level of significance, that sealants make a difference? What about at the 1 percent level of significance?

7. A consumer organization wanted to determine whether automobile repair shops were giving different estimates to women than to men. It selected two cars having the identical defect and gave one to a man and the other to a woman. Randomly choosing eight repair shops, the organization had the man take his car to four of these shops and the woman go to the other four. One week later they repeated the process, with the man going to the shops previously visited by the woman, and vice versa. The dollar prices quoted were as follows:

Shop	Price quoted to man ($)	Price quoted to woman ($)
1	145	145
2	220	300
3	150	200
4	100	125
5	250	400
6	150	135
7	180	200
8	240	275

Test the hypothesis that the sex of the person bringing the car to the repair shop does not affect the quoted price, using the

(a) Sign test

(b) Signed-rank test

8. Eleven patients having high albumin content in their blood are treated with a medicine. The measured values of their albumin both before and after the medication are as follows:

Blood Content of Albumin (grams per 100 milliliters)

Patient	Before medication	After medication
1	5.04	4.82
2	5.16	5.20
3	4.75	4.30
4	5.25	5.06
5	4.80	5.38
6	5.10	4.89
7	6.05	5.22
8	5.27	4.69
9	4.77	4.52
10	4.86	4.72
11	6.14	6.26

(a) What is the value of the test statistic of the signed-rank test?

(b) What is the p value of the test that the treatment has no effect?

9. An engineer claims that painting the exterior of a particular aircraft will affect its cruising speed. To check this claim, 10 aircraft just off the assembly line were flown to determine cruising speed prior to painting, and they were flown again after being painted. The following data resulted:

	Cruising speed (miles per hour)	
Aircraft	No paint	Paint
1	426.1	416.7
2	438.5	431.0
3	440.6	442.6
4	418.5	423.6
5	441.2	447.5
6	427.5	423.9
7	412.2	412.8
8	421.0	419.8
9	434.7	424.1
10	411.9	418.7

Do the data establish that the engineer is correct? Use the 5 percent level of significance.

10. Let $X_1, \ldots, X_n$ be a random sample of data from a certain population. Suppose we want to test the hypothesis that data from this population

are symmetric about some value v. Explain how we could accomplish this by using the signed-rank test. (*Hint*: Let $D_i = X_i - v$.)

14.4 RANK-SUM TEST FOR COMPARING TWO POPULATIONS

Consider two populations having a certain measurable characteristic, and suppose we are interested in testing the hypothesis that the two population distributions of this characteristic are the same. To test this hypothesis, suppose that independent samples of sizes n and m are drawn from the two populations.

If we were willing to assume that the underlying probability distributions were both normal, then we would apply the two-sample tests developed in Chap. 10. However, instead we will develop a nonparametric test that does not require the assumption of normality.

To begin, rank the $n + m$ data values from the two samples from smallest to largest. That is, give rank 1 to the smallest data value, rank 2 to the second smallest, and so on. For the time being we will assume that the $n + m$ values are all distinct, so there are no ties. Designate one of the samples (it makes no difference which one) as the first sample. The test we will consider makes use of the test statistic TS, defined to equal the sum of the ranks of the first sample. That is,

$$\text{TS} = \text{sum of ranks of data in first sample}$$

■ Example 14.8

To determine if reflex reaction time is age dependent, a sample of eight 20-year-old men and an independent sample of nine 50-year-old men were chosen. The following represents their reaction times (in seconds) to a given stimulus.

20-year-olds: 4.22, 5.13, 1.80, 3.34, 2.72, 2.80, 4.33, 3.60

50-year-olds: 5.42, 3.39, 2.55, 4.45, 5.55, 4.96, 5.88, 6.30, 5.10

Putting these 17 values in increasing order gives the following:

$$1.80,{}^{*}\ 2.55,\ 2.72,{}^{*}\ 2.80,{}^{*}\ 3.34,{}^{*}\ 3.39,\ 3.60,{}^{*}\ 4.22,{}^{*}\ 4.33,{}^{*}$$
$$4.45,\ 4.96,\ 5.10,\ 5.13,{}^{*}5.42,\ 5.55,\ 5.88,\ 6.30$$

We have put a star next to the data values that come from the 20-year-olds (which we are taking to be the first sample). Hence, the value of the sum of the ranks of the first sample is

$$\text{TS} = 1 + 3 + 4 + 5 + 7 + 8 + 9 + 13 = 50$$ ■

Let H_0 be the hypothesis that the two population distributions are identical, and suppose that the value of the test statistic TS is t. Since we want to reject H_0 if the value of TS is either significantly large or significantly small, it follows that the significance-level-α test will call for rejection of H_0 if either

$$P\{\text{TS} \leq t\} \leq \frac{\alpha}{2}$$

or

$$P\{\text{TS} \geq t\} \leq \frac{\alpha}{2}$$

where both of the preceding probabilities are to be computed under the assumption that H_0 is true. In other words, the null hypothesis will be rejected if the sum of the ranks from the first sample is either too small or too large to be explained by chance. As a result, it follows that the significance-level-α test will call for rejection of H_0 if the p value of the data set, given by

$$p \text{ value} = 2\,\text{Min}(P\{\text{TS} \leq t\}, P\{\text{TS} \geq t\})$$

is less than or equal to α.

To determine the probabilities, we need to know the distribution of TS when H_0 is true. To begin, suppose that the first sample is the one of size n. Now, when H_0 is true and so all the $n + m$ data values come from the same distribution, it follows that the set of ranks of the first sample will have the same distribution as a random selection of n of the values $1, 2, \ldots, n + m$. Using this, we can show that when H_0 is true, the mean and variance of TS are given by the following formulas.

When H_0 is true,

$$E[\text{TS}] = \frac{n(n + m + 1)}{2}$$

$$\text{Var}(\text{TS}) = \frac{nm(n + m + 1)}{12}$$

In addition, it can be shown that when n and m are both of at least moderate size (both being larger than 7 should suffice), TS will, when H_0 is true, have an approximately normal distribution. Hence, if the sample sizes are not too small, it follows that TS will be approximately normal with a mean and variance as stated in the preceding.

■ Example 14.9

In Example 14.8, $n = 8$, $m = 9$, and the value of the sum of the ranks of the first sample was TS = 50. Now,

$$E[\text{TS}] = \frac{n(n+m+1)}{2} = 72 \quad \text{Var(TS)} = \frac{nm(n+m+1)}{12} = 108$$

Since the observed value of TS was less than its mean, we have

$$\begin{aligned} p \text{ value} &= 2P\{\text{TS} \le 50\} \\ &= 2P\{\text{TS} \le 50.5\} \\ &= 2P\left\{\frac{\text{TS}-72}{\sqrt{108}} \le \frac{50.5-72}{\sqrt{108}}\right\} \\ &\approx 2P\{Z \le -2.069\} \\ &= 0.0385 \end{aligned}$$

Therefore, the null hypothesis that the two population distributions are identical would be rejected at the 5 percent level of significance. ■

If there are any ties, then the rank of a data value should be the average of the ranks of all those with the same value. For instance, if the first-sample data are 2, 4, 4, 6 and the second-sample data are 5, 6, 7, then the sum of the ranks of the data of sample 1 is $1 + 2.5 + 2.5 + 5.5 = 11.5$. The test should then be run exactly as before. (It turns out that the possibility of ties has the effect of reducing the variance of TS when the null hypothesis is true. As a result, the p value previously given will be larger than the actual p value that takes into account the tied values. In consequence, the test presented will be conservative, in that whenever ties are present and it calls for rejection, then a more sophisticated test that takes the ties into account will also call for rejection of the null hypothesis.)

Statistics in Perspective

The test developed in this section is called the *two-sample rank-sum test*. Aside from the sign test, which goes back to Arbuthnot in 1710 (see Sec. 9.5), it was one of the first nonparametric tests to be developed. It was jointly and independently discovered in the mid-1940s by Wilcoxon and the team of Mann and Whitney. Because of this, the test is sometimes called the *Wilcoxon sum-of-ranks test* and sometimes the *Mann–Whitney test*. The publications of Wilcoxon and Mann–Whitney were the beginning of a wave of research on nonparametric tests, one that has not yet abated.

■ Example 14.10

In an attempt to determine if the vocabulary skills of two different students are similar, an English teacher had each of them write a short essay on the same topic. The teacher then counted the number of times each student used words having four or more letters. The following data resulted:

	Number of words used having i letters	
i	Student 1	Student 2
4	44	49
5	16	11
6	8	5
7	7	4
8	4	1
9	2	1
10	3	0

Thus, for instance, 8 out of the 84 words (having four or more letters) written by student 1 and 5 of the 71 words used by student 2 were six-letter words. Use these data to test the hypothesis that the word-length frequency distributions of the two students are the same.

Solution

The data consist of one sample of 84 words and another sample of 71 words. Since in the combined samples of 155 words the data value 4 appears 93 times, each of these 93 data values is given a rank equal to the average of the rank numbers 1 through 93. That is, each is given rank

$$\frac{1+2+\cdots+93}{93}=\frac{1+93}{2}=47$$

Also, since the next-smallest data value (the value 5) occurs 27 times, each of these values shares the ranks from 94 through 120. Therefore, each of the data values 5 is given rank

$$\frac{94+120}{2}=107$$

Similarly, the data values 6, 7, 8, 9, and 10 are given rank values as follows:

Data value	4	5	6	7	8	9	10
Rank	47	107	127	139	147	151	154

The sum of the ranks of the sample of 71 words has value

$$TS = 47 \times 49 + 107 \times 11 + 127 \times 5 + 139 \times 4 + 147 + 151 = 4969$$

Since $n = 71$ and $m = 84$, we see that

$$\frac{n(n+m+1)}{2} = 5538 \qquad \frac{nm(n+m+1)}{12} = 77{,}532$$

Thus, the approximate p value is

$$\begin{aligned} p \text{ value} &= 2P\{TS \le 4969\} \\ &= 2P\{TS \le 4969.5\} \\ &= 2P\left\{\frac{TS - 5538}{\sqrt{77{,}532}} \le \frac{4969.5 - 5538}{\sqrt{77{,}532}}\right\} \\ &\approx 2P\{Z \le -2.04\} \\ &= 0.041 \end{aligned}$$

and so the hypothesis that the word-length distributions of the two students are identical is rejected at the 5 percent level of significance. ■

In cases where the sample sizes are not large, say, when either is less than 8, we can no longer assume that the distribution of the sum of the ranks is approximately normal. However, we can still employ the rank-sum test by directly computing the exact p value. To do so, we use the fact that when H_0 is true, the set of ranks of the first sample has the same distribution as a random selection of n of the values $1, 2, \ldots, n + m$. By using this, it is possible (with the help of a computer) to explicitly determine the p value. Program 14-2 computes the exact p value for the rank-sum test. The inputs needed are the sizes of the first and second samples and the sum of the ranks of the elements of the first sample. Whereas either sample can be designated as the first sample, the program will run fastest if the first sample is the one whose sum of ranks is smaller. In addition, since this program implicitly assumes there are no ties, it can be used only when the value of TS is an integer.

Historical Perspective

The use of statistical techniques to make literary comparisons goes back a long way. In 1901, Thomas Mendenhall, who had been a professor of physics at Ohio State University, published a comparison of the frequencies of the number of letters in the words used by Shakespeare and other authors. Mendenhall noted that nearly all Shakespeare's plays had approximately the same frequency distribution. He showed that Shakespeare used a higher proportion of words with one, two, four, or five letters and a lesser proportion of the others when compared

Thomas Mendenhall

(Ohio State University Photo Archive, Columbus) with either Dickens or Thackeray. Francis Bacon's distribution was also found to be quite different from Shakespeare's. Excitement was stirred, however, because an analysis of the plays of Christopher Marlowe produced a word-size frequency distribution almost identical to that of Shakespeare's.

More recently, statistical analysis has been employed to decide the authorship of 12 *Federalist* papers. These papers, consisting of 77 letters, appeared anonymously in New York State newspapers between 1787 and 1788. The letters tried to persuade the citizens of New York to ratify the Constitution. Although it was generally known that the authors of the papers were Alexander Hamilton, John Jay, and James Madison, it was not known which of them was responsible for each specific paper. As of 1964, authorship of most of the papers had been determined. However, a long-standing dispute remained concerning the authorship of 12 of them. In a book published in 1964, Harvard statisticians Frederic Mosteller and David Wallace used a statistical analysis to conclude that all 12 papers had been written by Madison alone. Their analysis considered such things as the frequency distributions of each author's use of such words as *by*, *from*, *to*, and *upon*.

■ Example 14.11

Let us reconsider Example 14.9, this time using Program 14-2 to compute the p value. This program runs best if you designate the sample having the smaller sum of ranks as the first sample. The size of the first sample is 8. The size of the second sample is 9. The sum of the ranks of the first sample is 50. Program 14-2 computes the p value as 3.595229E-02.

Thus the exact p value is 0.0359, which is reasonably close to the approximate value of 0.0385 obtained by using the normal approximation in Example 14.9. ■

14.4.1 Comparing Nonparametric Tests with Tests that Assume Normal Distributions

The strength of nonparametric tests is that they can be used without making any assumptions about the form of the underlying distributions. The price that one pays for using a nonparametric test is that it will not be as effective in cases where the distributions are normal or approximately normal as would a test that starts out by assuming normality. Somewhat surprisingly, the loss in effectiveness is relatively small. For instance, it can be shown that when sample sizes n and m are large, the efficiency of the nonparametric rank-sum test is approximately 95 percent of that of the two-sample t test when the distributions are indeed normal. By this, we loosely mean that when the population distributions are normal but unequal, then the chance of rejection with the nonparametric test with samples of size n is roughly the same as with the normal-based t test with samples of size $0.95n$. This is an impressive result, and it might easily lead one to conclude that the nonparametric test is superior if one is not absolutely certain that the distributions are close to normal. For if the distributions are not normal, then the normal test is based on a false assumption; and even if the distributions are normal, the nonparametric one is almost as good. However, even when the underlying distributions are not normal, the normal test will be a good one when sample sizes n and m are large. This is so because this test is based on a test statistic that will be approximately normal even when the population distributions are not. We can thus conclude that, for large sample sizes, the normal-based t test will be an effective test.

Probably the best we can say is that if one is not certain that the underlying distribution is at least approximately normal, then for moderate sample sizes the rank-sum nonparametric test is preferred to the two-sample t test. On the other hand, in cases of large sample sizes, either test type can be used. A key difference, however, that can be useful in deciding which type of test to use is that the t test is designed to detect differences in the population means, whereas the rank-sum test is designed to detect any difference in the population distributions.

PROBLEMS

1. The following data are from independent samples from two populations.

   Sample 1: 142, 155, 237, 244, 202, 111, 326, 334, 350, 247

   Sample 2: 212, 277, 175, 138, 341, 255, 303, 188

   (a) Determine the sum of the ranks of the data from sample 1.
   (b) Determine the sum of the ranks of the data from sample 2.

2. There is an algebraic identity stating that the sum of the first k positive integers is equal to $k(k+1)/2$. That is,

$$\sum_{i=1}^{k} i = \frac{k(k+1)}{2}$$

   Use this identity to determine the relationship between the sum of the ranks of the sample of size n and the sum of the ranks of the sample of size m. Use the results of Prob. 1 to check your result. Assume that all $n+m$ data values are different.

3. A study was carried out to determine if educational opportunities in rural and urban California counties are the same. Two counties of roughly the same socioeconomic makeup, one in an urban area and the other in a rural area, were chosen. The Scholastic Aptitude Test (SAT)

scores of a random sample of high school graduates were obtained in both counties. The results were as follows:

Rural	Urban
544	610
567	498
475	505
658	711
590	545
602	613
571	509
502	514
578	609

Find the p value of the test of the hypothesis that the distributions of scores in both counties are identical. Use the normal approximation.

4. A group of 16 volunteers were randomly divided into two subgroups of 8 each. Members of the first subgroup were given daily tablets containing 5 grams of vitamin C, and members of the second subgroup were given a placebo. After 1 month the blood cholesterol levels of the 16 individuals were measured and compared with their levels at the beginning of the experiment. The reductions in blood cholesterol levels for the two subgroups were as follows:

Vitamin C	Placebo
6	9
12	−3
14	0
2	−1
7	5
7	3
1	−4
8	−1

Test the null hypothesis, at the 5 percent level of significance, that vitamin C and the placebo are equally effective in reducing cholesterol. Assume that the distribution of the test statistic is approximately normal when the null hypothesis is true. (A negative data value means that the blood cholesterol level increased. For instance, the data value −4 indicates an increase of 4 in the blood cholesterol reading.)

5. A study was conducted to test the hypothesis that the starting salary distribution of seniors graduating from Stanford University with a degree in Computer Science (CS) was the same as the one for CS graduates of the University of California at Berkeley. A random sample of recently graduating students yielded the following yearly salaries (in units of $1000):

Stanford	Berkeley
57.8	52.6
60.4	56.6
71.2	61.0
52.5	47.9
68.0	55.0
69.6	62.5
70.0	66.4
54.0	57.5
48.8	56.5
57.6	49.8

What conclusion would you draw at the 5 percent level of significance?

6. An experiment designed to determine the effectiveness of vitamin B1 in stimulating the growth of mushrooms was performed. The vitamin was applied to 9 mushrooms, selected at random from a set of 17. The remaining 8 mushrooms were left untreated. The weights (in grams) of all 17 mushrooms at the end of the experiment were as follows:

   Untreated mushrooms: 18, 12.4, 13.5, 14.6, 24, 21, 23, 17.5

   Vitamin B1 mushrooms: 34, 27, 21.2, 29, 20.5, 19.6, 28, 33, 19

   Test, at the 5 percent level, the hypothesis that the vitamin B1 treatment had no effect.

7. Twenty-four workers were randomly divided into two sets of 12 each. Each set of workers was put through a 2-week training program. However, the first set of workers spent an additional day on “motivational” material. At the end of the training session the workers were given a series of tests and then ranked according to their performances. If the sum of the ranks of the workers who went through the motivational material was 136, what is the p value of the test of the hypothesis that the motivational material has no effect?

8. Use Program 14-2 to find the exact p value in Prob. 4.

9. Use Program 14-2 to find the exact p value in Prob. 5.

10. Redo Prob. 1 in Sec. 10.4, this time using a nonparametric test.

14.5 RUNS TEST FOR RANDOMNESS

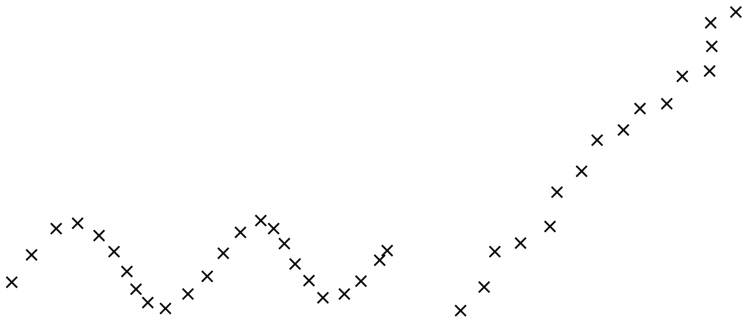
A basic assumption in much of data analysis is that a set of data constitutes a random sample from some population. However, sometimes the data set is not actually a random sample from a population but rather is one that has some internal pattern. For instance, the data might tend to be increasing or decreasing over time, or they may follow some cyclical pattern where they increase and then decrease in a cyclic manner (see Fig. 14.4). In this section we will develop a test of the hypothesis that a given data set constitutes a random sample.

To test the hypothesis that a given sequence of data values constitutes a random sample, suppose initially that each datum can take on only two possible values, which we designate 0 and 1. Consider any data set of 0s and 1s, and call any consecutive sequence of either 0s or 1s a run. For instance, the data set

$$0, 1, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0$$

contains a total of nine runs: 5 runs of 0s and four runs of 1s. The first run consists of the single value 0; the next run consists of the three values 1, 1, 1; the next one consists of the two values 0, 0; and so on.

Suppose that the data set consists of a total of $n + m$ values, of which n are equal to 1 and m are equal to 0. Let R denote the number of runs in the data set. Now if the data set were a random sample from some population, then all possible orderings of the $n + m$ values (consisting of n 1s and m 0s) would be equally likely. By using this result it is possible to determine the probability distribution of R and thus to test the null hypothesis H_0 that the data set is a random sample by rejecting H_0 if the value of R is either too small or too large to be explained by chance. Specifically, if the value of R is r, then the significance-level-α test calls for



Data follow a cyclic trend.

Data follow an increasing trend.

FIGURE 14.4

Nonrandom data sets.

rejecting H_0 if either

$$P\{R \le r\} \le \frac{\alpha}{2}$$

or

$$P\{R \ge r\} \le \frac{\alpha}{2}$$

where these probabilities are computed under the assumption that the null hypothesis is true. The resulting test is called the *runs test*.

We can also perform the runs test by observing the value of R, say it is r, and then calculating the resulting p value:

$$p \text{ value} = 2\,\text{Min}(P\{R \le r\}, P\{R \ge r\})$$

Program 14-3 determines this p value by calculating the relevant probabilities.

■ Example 14.12

The following are the outcomes of the last 24 games played by a local softball team. The letter W signifies a win and L a loss.

$$W,\ L,\ L,\ L,\ W,\ L,\ L,\ W,\ L,\ L,\ W,\ L,\ L,\ W,\ L,\ W,\ L,\ L,\ L,\ L,\ W,\ L,\ W,\ L$$

Is this data set consistent with randomness?

Solution

To test the hypothesis of randomness, note that the data set of 8 W's and 16 L's contains a total of 16 runs. To see whether this justifies a rejection of the hypothesis of randomness, we run Program 14-3.

The number of ones is 8. The number of zeros is 16. The number of runs is 16. The p value as computed by Program 14-3 is 0.052497.

Thus, since the p value is 0.0525, it follows that the hypothesis of randomness cannot be rejected at the 5 percent level of significance. That is, although the evidence of the data is against the hypothesis of randomness, it is not quite strong enough to cause us to reject that hypothesis at the 5 percent level of significance. ■

If Program 14-3 is not readily available to be run, then we can obtain an approximate p value by making use of a result that states that when the null hypothesis is true, R will have an approximately normal distribution with mean and variance given, respectively, by

$$\mu = \frac{2nm}{n+m} + 1$$

and

$$\sigma^2 = \frac{2nm(2nm - n - m)}{(n+m)^2(n+m-1)}$$

This will lead to a good approximation for the distribution of R provided that n and m, the numbers of 1s and 0s in the data set, are both of at least moderate size. (Both being at least 20 should suffice.)

If we observe a total of r runs, then the p value is given by

$$p \text{ value} = 2\,\text{Min}(P\{R \leq r\}, P\{R \geq r\})$$

We can now use the preceding normal approximation to compute the relevant probability concerning R.

■ Example 14.13

Let us repeat Example 14.12, this time determining the approximate p value by using the preceding normal approximation. Since $n = 8$ and $m = 16$, we see that

$$\mu = \frac{2 \cdot 8 \cdot 16}{24} + 1 = 11.667$$

$$\sigma^2 = \frac{256(256 - 24)}{24 \cdot 24 \cdot 23} = 4.4831$$

Since there are a total of 16 runs, the p value is given by

$$p \text{ value} = 2\,\text{Min}(P\{R \leq 16\}, P\{R \geq 16\})$$

Since $E[R] = 11.667$ is less than 16, it follows that $P\{R \geq 16\}$ is smaller than $P\{R \leq 16\}$. Thus,

$$\begin{aligned}
p \text{ value} &= 2P\{R \geq 16\} \\
&= 2P\{R \geq 15.5\} \quad \text{(continuity correction)} \\
&= 2P\left\{\frac{R - 11.667}{\sqrt{4.4831}} \geq \frac{15.5 - 11.667}{\sqrt{4.4831}}\right\} \\
&\approx 2P\{Z \geq 1.81\} \\
&= 0.07
\end{aligned}$$

Since this approximate p value is greater than 0.05, the null hypothesis is not rejected at the 5 percent level of significance. Therefore, the approximate p value leads us to the same conclusion, at the 5 percent level of significance, as does the exact p value. However, the approximate value of 0.07 is not that close to the actual p value of 0.0525. Of course, in this example, the values of n and m (namely, 8 and 16) do not meet the rule of thumb that both should be at least 20 for the approximation to be accurate. ■

Example 14.14

Consider a sequence that contains 20 zeros and 20 ones. Suppose that there are 27 runs in this sequence. Compare the actual p value of the test of the hypothesis that the sequence is random with the approximate p value obtained by using the normal approximation.

Solution

Let us start with the normal approximation. Since $n = m = 20$, the mean and standard deviation of the number of runs are, respectively,

$$\mu = \frac{2 \cdot 20 \cdot 20}{40} + 1 = 21 \qquad \sigma = \sqrt{\frac{2 \cdot 20 \cdot 20 \cdot 760}{40 \cdot 40 \cdot 39}} = 3.121$$

Therefore, since the observed number of runs is 27, the normal approximation gives the following:

$$\begin{aligned} p \text{ value} &= P\{R \geq 27\} \\ &= 2P\{R \geq 26.5\} \\ &= 2P\left\{\frac{R - 21}{3.121} \geq \frac{26.5 - 21}{3.121}\right\} \\ &\approx 2P\{Z \geq 1.762\} \\ &= 0.078 \end{aligned}$$

On the other hand, running Program 14-3 gives the exact p value:

The number of ones is 20. The number of zeros is 20. The number of runs is 27. The p value is computed by program 14-3 as 0.075996.

Therefore, in this example (where both n and m are equal to 20) the approximate p value of 0.078 is quite close to the actual p value of 0.076. ■

We can also use the runs test to test for randomness when the sequence of data values does not comprise just 0s and 1s. To test whether a given sequence of data $X_1, X_2, \ldots, X_n$ constitutes a random sample from some population, let s_m denote the sample median of this data set. Now, for each data value determine whether it is less than or equal to s_m or whether it is greater than s_m. Put a 0 in position i if X_i is less than or equal to s_m, and put a 1 otherwise. If the original data set constituted a random sample from some distribution, then the sequence of 0s and 1s will also constitute a random sample. Therefore, we can test whether the original data set is a random sample by using the runs test on the resulting sequence of 0s and 1s.

■ Example 14.15

The average summer temperatures in degrees Fahrenheit for 20 successive years from 1971 to 1990 in a given west coast city are

$$72, 71, 70, 82, 80, 77, 71, 85, 75, 80, 82, 81, 83, 82, 85, 86, 83, 81, 82, 84$$

Test the hypothesis that the data constitute a random sample.

Solution

The sample median m is the average of the 10th and the 11th smallest values. Therefore,

$$s_m = \frac{81 + 82}{2} = 81.5$$

The data of 0s and 1s that indicate whether each value is less than or equal to or greater than 81.5 are as follows:

$$0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,1\,0\,1\,1\,1\,1\,1\,0\,1\,1$$

Thus, the sequence of 0s and 1s consists of 10 of each and has a total of 10 runs. To determine if this value is significantly greater or larger than could be expected by chance if the data were truly random, we run Program 14-3.

The number of ones is 10. The number of zeros is 10. The number of runs is 10. The p value is computed as 0.8281409.

For such a large p value, the hypothesis of randomness is not rejected. That is, the data give no evidence of not being a random sample.

If we had used the normal approximation, then we would have first computed μ and σ, the null hypothesis mean and standard deviation of the total number of runs. Since $n = m = 10$,

$$\mu = \frac{200}{20} + 1 = 11 \qquad \sigma = \sqrt{\frac{200(180)}{400 \cdot 19}} = 2.176$$

Since the observed number of runs is 10, the p value given by the normal approximation is

$$\begin{aligned} p \text{ value} &= 2P\{R \le 10\} \\ &= 2P\{R \le 10.5\} \\ &= 2P\left\{\frac{R-11}{2.176} \le \frac{10.5-11}{2.176}\right\} \\ &\approx 2P\{Z \le -0.23\} \\ &= 0.818 \end{aligned}$$

Thus, the normal approximation is quite accurate in this example. ■

■ Example 14.16

The following are the successive numbers of points scored by a certain high school basketball team in the 23 games it played in the 1994–1995 season. Is it reasonable to suppose that the scores constitute a random sample?

77, 62, 58, 64, 66, 72, 59, 69, 80, 74, 72, 69, 74, 83, 85, 87, 80, 88, 76,

77, 82, 85, 83

Solution

The sample median is the 12th-smallest score, namely, 76. The sequence of 0s and 1s indicating whether each value is less than or equal to or greater than 76 is as follows:

1 0 0 0 0 0 0 0 1 0 0 0 0 1 1 1 1 1 0 1 1 1 1

Thus, this sequence consists of twelve 0s and eleven 1s, and has seven runs. From Program 14-3, we see that the p value $= 0.02997$, and thus the hypothesis that the data constitute a random sample is rejected at the 5 percent level of significance. ■

PROBLEMS

Unless otherwise stated, use either Program 14-3 or the normal approximation, whichever is more convenient, in answering the following questions.

1. Suppose a sequence of 0s and 1s contains twenty 0s and thirty 1s. Let R denote the total number of runs. What are the (a) largest and (b) smallest possible values of R?

2. Determine the number of runs for the following data sets of 0s and 1s.
   (a) 1 1 0 0 0 0 1 1 0 1 0 1 1 0 0 0 1 1
   (b) 0 1 1 0 0 0 0 1 1 1 0 1 1 0 1 1 1 1
   (c) 0 0 0 1 1 0 0 1 1 0 1 1 1 1 1 0 1 0
3. The following data relate to the acceptability of the 26 most recently produced watches at a Swiss watch factory. The value 1 signifies that the watch is acceptable and the value 0 that it is unacceptable.

   1 1 1 0 1 1 1 1 1 1 1 1 1 1 0 0 0 0 1 1 1 1 0 0 1 1

   Test the hypothesis, at the 5 percent level of significance, that the data constitute a random sample.
4. A production run of 60 items resulted in 12 defectives. The defectives are item numbers 9, 14, 15, 26, 30, 36, 37, 44, 45, 46, 59, and 60.
   (a) What is the value of R, the total number of runs?
   (b) Can we conclude, at the 5 percent level of significance, that the successive items do not constitute a random sample?
5. A total of 25 people, 10 of whom are women, are to be interviewed. The interviewer is told to interview them in a randomly chosen order. Suppose that the sequence of sexes of the successively interviewed people is as follows:

   F F M F F F F M M F F M F F M M M M M M M M M M M

   Did the interviewer follow instructions? Explain, and give the relevant p value.
6. Over the last 50 days the Dow Jones industrial average increased on 32 days and decreased on the other 18 days. If the total number of runs (of increasing or decreasing Dow Jones average) was 22, what is the p value of the test of the hypothesis that the increases and decreases constituted a random sample?
7. The lifetimes, in hours, of 30 successively produced storage batteries are as follows:

   148, 152, 155, 147, 176, 170, 165, 149, 138, 155, 160, 153, 162, 155, 159,

   174, 168, 149, 182, 177, 191, 185, 178, 176, 182, 184, 181, 177, 160, 154

   (a) What is the sample median?
   (b) What is the value of R, the number of runs in the corresponding data that details for each data value whether it is less than (or equal to) or greater than the sample median?
   (c) Do these data disprove the hypothesis that the sequence of values constitutes a random sample?

8. The following data represent end-of-year Dow Jones averages for a sequence of 10 consecutive years:

   910, 890, 1010, 1033, 1080, 1275, 1288, 1553, 1980, 2702

   Test the hypothesis, at the 5 percent level of significance, that these data can be thought of as constituting a random sample.

14.6 TESTING THE EQUALITY OF MULTIPLE PROBABILITY DISTRIBUTIONS

Whereas in Section 14.4 we showed how to test the hypothesis that population distributions are identical when there are two such populations, we are sometimes faced with the situation where there are more than two populations. So suppose there are k populations and that F_i is the distribution function of some measurable value of the elements of population i. We are interested in testing the null hypothesis

$$\mathrm{H}_0\colon F_1 = F_2 = \cdots = F_k$$

against the alternative

$$\mathrm{H}_1\colon \text{not all of the } F_i \text{ are equal}$$

To test the null hypothesis, suppose that independent samples are drawn from each of the k populations. Let n_i denote the size of the sample chosen from population $i, i = 1, \ldots, k$, and let $N = \sum_{i=1}^{k} n_i$ denote the total number of data values obtained. Now rank these N data values from smallest to largest, and let R_i denote the sum of the ranks of the n_i data values from population $i, i = 1, \ldots, k$.

Noting that when H_0 is true, the rank of any individual data value is equally likely to be any of the ranks $1, 2, \ldots, N$, it follows that the expected rank of any individual data is $\frac{1+2+\cdots+N}{N} = \frac{N+1}{2}$. Consequently, with $\bar{n} = \frac{N+1}{2}$, it follows that when H_0 is true the expected value of the sum of the ranks of the n_i data values from population i is $n_i(N+1)/2 = n_i\bar{n}$. That is, if R_i is the sum of the ranks of the n_i data values from population i then, when H_0 is true,

$$E[R_i] = n_i\bar{n}$$

Drawing our inspiration from the goodness-of-fit test, let us consider the test statistic

$$T = \sum_{i=1}^{k} \frac{(R_i - n_i\bar{n})^2}{n_i\bar{n}}$$

and use a test that rejects the null hypothesis when T is large. Now,

$$\begin{aligned} T &= \frac{1}{\bar{n}} \sum_{i=1}^{k} \left(\frac{R_i^2 - 2R_i n_i \bar{n} + n_i^2 \bar{n}^2}{n_i} \right) \\ &= \frac{1}{\bar{n}} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 2 \sum_{i=1}^{k} R_i + \bar{n} \sum_{i=1}^{k} n_i \\ &= \frac{1}{\bar{n}} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 2 \sum_{i=1}^{k} R_i + N\bar{n} \end{aligned}$$

Because $\sum_{i=1}^{k} R_i$ is the sum of the ranks of all N data values,

$$\sum_{i=1}^{k} R_i = 1 + 2 + \cdots + N = \frac{N(N+1)}{2} = N\bar{n}$$

and so

$$T = \frac{1}{\bar{n}} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - N\bar{n}$$

Hence, rejecting H_0 when T is large is equivalent to rejecting H_0 when $\sum_{i=1}^{k} R_i^2/n_i$ is large. So, we might as well let the test statistic be

$$\text{TS} = \sum_{i=1}^{k} \frac{R_i^2}{n_i}$$

To determine the appropriate significance-level-α test, we need the distribution of TS when H_0 is true. While its exact distribution is rather complicated, we can use the result that, when H_0 is true and all n_i are at least 5, the distribution of

$$\frac{12}{N(N+1)} \text{TS} - 3(N+1)$$

is approximately that of a chi-squared random variable with $k-1$ degrees of freedom. Using this, we thus see an approximate significance-level-α test of the null hypothesis that all distributions are identical is to

$$\text{reject} \quad H_0 \quad \text{if} \quad \frac{12}{N(N+1)} \text{TS} - 3(N+1) \geq \chi^2_{k-1,\alpha}$$

The preceding is known as the *Kruskal-Wallis test*.

■ Example 14.17

Suppose that 10 experts are to analyze three different wines, giving numerical scores ranging from 0 (terrible wine) to 10 (exceptional wine) to each wine. Use the resulting scores given below to test, at the 5 percent level of significance, the null hypothesis that the wines are of identical quality.

Numerical Scores given to
Wines A, B, C by Ten Experts

A	B	C
7.21	6.04	6.42
6.60	6.26	4.80
6.22	7.44	7.05
7.38	8.02	5.84
8.20	6.91	7.13
7.07	6.65	6.54
6.72	7.11	7.04
5.89	7.15	5.22
9.02	6.61	6.83
6.88	7.29	7.08

Solution

Ordering the 30 rankings gives that the sum of the ranks of wines A, B, and C are

$$R_A = 176, \quad R_B = 175, \quad R_C = 114$$

yielding that

$$\frac{12}{N(N+1)}\text{TS} - 3(N+1) = \frac{12}{30(31)} \frac{(176)^2 + (175)^2 + (114)^2}{10} - 93 = 3.254$$

Because $\chi^2_{2,.05} = 5.99$ it follows that the null hypothesis can not be rejected. Indeed, the resulting p value is

$$p \text{ value} \approx P\{\chi^2_2 > 3.254\} = .1965 \quad ■$$

14.6.1 When the Data Are a Set of Comparison Rankings

Before giving the general problem, let us illustrate by again considering Example 14.17. However, suppose now that the 10 wine experts each compare and rank the 3 wines, with each expert ranking the wines from best (score equal to 1) to worst (score equal to 3). Suppose we want to use the results of these rankings to test the null hypothesis that the wines are identical in quality and thus the ranking by each expert is equally likely to be any of the $3! = 6$ possible orderings. We could test this null hypothesis by viewing the data as the results of 10 trials, where each trial corresponds to the rankings of an expert. That is, we could take the result of an expert as a permutation of the numbers $1, 2, 3$, where the result i, j, k would mean that the expert gave rank i to wine A, rank j to wine B, and rank k to wine C. The null hypothesis would be that all 6 possible outcomes of a trial are equally likely, and this could be tested by a goodness-of-fit test. However, a total of 10 trial results in a situation where each trial has 6 possible outcomes is not nearly enough trials to obtain meaningful test results. Thus, we consider a different test.

Let R_{A}, R_{B}, and R_{C} be the sum of the ranks given by the experts to the wines A, B, and C, respectively. Now, if the null hypothesis is true and the wines are of identical quality then each expert would be equally likely to give wine A any of the rankings $1, 2$, or 3. Thus, the expected rank given to wine A by each expert is $1(1/3) + 2(1/3) + 3(1/3) = 2$. As there are 10 experts, it follows that when H_0 is true

$$E[R_{\mathrm{A}}] = 10 \cdot 2 = 20$$

Indeed the same reasoning shows that, when H_0 is true,

$$E[R_{\mathrm{A}}] = E[R_{\mathrm{B}}] = E[R_{\mathrm{C}}] = 20$$

So, again inspired by the goodness-of-fit test statistic, it seems reasonable to reject the null hypothesis when

$$\frac{(R_{\mathrm{A}} - 20)^2}{20} + \frac{(R_{\mathrm{B}} - 20)^2}{20} + \frac{(R_{\mathrm{C}} - 20)^2}{20}$$

is sufficiently large. Equivalently, we would want to reject when

$$(R_{\mathrm{A}} - 20)^2 + (R_{\mathrm{B}} - 20)^2 + (R_{\mathrm{C}} - 20)^2$$

is sufficiently large.

In the general case, let us suppose that k items are to be ranked by each of n experts. (In the wine example, $k = 3$, $n = 10$.) To test the null hypothesis H_0 that all k items are of identical quality, we let R_i denote the sum of the rankings given to item i by all n experts. Using that, under H_0, each expert is equally likely to give wine i any of the rankings $1, 2, \ldots, k$, it follows that the expected ranking given to wine i by each expert is

$$\frac{1 + 2 + \cdots + k}{k} = \frac{k + 1}{2}$$

As there are n experts, it follows, when H_0 is true, that

$$E[R_i] = n(k + 1)/2$$

Because $E[R_i]$ is the same for all i, it follows that to reject H_0 when $\sum_{i=1}^{n} \frac{(R_i - E[R_i])^2}{E[R_i]}$ is large is equivalent to letting the test statistic be

$$\mathrm{TS} = \sum_{i=1}^{n} (R_i - n(k+1)/2)^2$$

and rejecting H_0 when TS is sufficiently large.

To determine how large TS need be to justify rejecting H_0, we make use of the result that, when H_0 is true,

$$\frac{12}{nk(k+1)}\mathrm{TS} = \frac{12}{nk(k+1)} \sum_{i=1}^{n} (R_i - n(k+1)/2)^2$$

has approximately a chi-squared distribution with $k-1$ degrees of freedom.

■ Example 14.18

If the 10 experts in Example 14.17 had used comparative ranking of wines A, B, C, the following data would have resulted. Use them to test the null hypothesis that the wines are all of equal quality.

Rankings of Wines A, B, C by Ten Experts

A	B	C
1	3	2
1	2	3
3	1	2
2	1	3
1	3	2
1	2	3
3	1	2
2	1	3
1	3	2
3	1	2

Solution

We see by the preceding that $R_{\mathrm{A}} = 18$, $R_{\mathrm{B}} = 18$, $R_{\mathrm{C}} = 24$. Hence,

$$\mathrm{TS} = (18-20)^2 + (18-20)^2 + (24-20)^2 = 24$$

Because $\frac{12}{10 \cdot 3 \cdot 4}\mathrm{TS} = \frac{\mathrm{TS}}{10}$ is, when H_0 is true, approximately a chi-squared random variable with 2 degrees of freedom, and because the data give that $\frac{\mathrm{TS}}{10} = 2.4$, the

p value of the test of H_0 is

$$p \text{ value} \approx P(\chi^2_2 \geq 2.4) = .3012$$

showing that the data are not inconsistent with the hypothesis that the wines are of identical quality. ■

The preceding test for the equality of multiple population distributions when the data consists of a set of comparison rankings is known as the *Freedman test*.

PROBLEMS

1. Use the data of Prob. 1 of Sec. 11.2 to test the hypothesis that the three population distributions are identical.
2. Redo Prob. 11 of Sec. 11.2, this time using a nonparametric test.
3. The following are the weights of random samples of adult males from three different political affiliations. Use them to test the null hypothesis that the weight of a randomly chosen man is independent of his political affiliation.

   Weights of Republicans: 204, 178, 195, 187, 152, 166, 240, 182

   Weights of Democrats: 200, 168, 175, 192, 156, 164, 180, 166

   Weights of Independents: 172, 177, 168, 183, 159, 172, 192, 165

4. Nine members of a population were randomly chosen and asked to give a comparison ranking of 3 different possible styles that an automobile company could use in a new car it is developing. Their rankings of the styles A, B and C were as follows:

1	2	3
A	C	B
A	B	C
C	B	A
B	A	C
A	B	C
A	C	B
C	B	A
B	C	A
C	A	B

Use these rankings to test the hypothesis that all three styles would be equally favored by the entire population.

14.7 PERMUTATION TESTS

Until now all the nonparametric tests we have considered use test statistics based solely on the ranks of the data values. Permutation tests, however, are nonparametric tests that use the precise data values obtained. For an example, suppose we want to to test the null hypothesis H_0 that the data $X_1, \ldots, X_N$ is a sample from some population having an unknown population distribution. In a permutation test, the data is observed, a test statistic is derived, and the p value is then computed conditional on knowing the set S of data values observed. For instance if $N = 3$ and $X_1 = 5$, $X_2 = 7$, $X_3 = 2$, then the p value is computed conditional on the information that the set of data values is $S = \{2, 5, 7\}$. The computation of the p value makes use of the fact that, conditional on the set of data values S, each of the $N!$ possible ways of assigning these N values to the original data is equally likely when the null hypothesis is true. That is, suppose that $N = 3$ and the set of data values is, as in the preceding, $S = \{2, 5, 7\}$. Now the null hypothesis H_0 states that X_1, X_2, X_3 are independent and identically distributed. Consequently, if H_0 is true then, given the data set S, it follows that the vector (X_1, X_2, X_3) is equally likely to equal any of the $3!$ permutations of the values $2, 5, 7$.

The implementation of a permutation test is as follows. Depending on the alternative hypothesis, a test statistic TS is chosen. Suppose, for the moment, that large values of the test statistic are evidence for the alternative hypothesis. The data values are then observed, say that $X_i = x_i$, $i = 1, \ldots, N$, and the value of TS is calculated. If the value of the test statistic is $\text{TS} = t$, the resulting p value of the null hypothesis that results from these data is the probability that TS would be at least as large as t when all possible assignments of the N data values to the variables $X_1, \ldots, X_N$ are equally likely.

For an illustration, suppose we are to observe data over N weeks, with X_i being the data value observed in week i, $i = 1, \ldots, N$, and that we want to use these data to test the null hypothesis

$$\text{H}_0\text{: } X_1, \ldots, X_N \text{ are independent and identically distributed}$$

against

$$\text{H}_1\text{: } X_i \text{ tends to increase as } i \text{ increases}$$

Now if the null hypothesis is true and the data are independent and identically distributed then, conditional on knowing the set of values $\{x_1, \ldots, x_N\}$, but not knowing which value corresponds to X_1 or which corresponds to X_2 and so on, the statistic $\sum_{j=1}^{N} j\, X_j$ would be distributed as if we randomly paired up the two data sets $\{1, \ldots, N\}$ and $\{x_1, \ldots, x_N\}$ and then summed the products of the N paired values. On the other hand, if the alternative hypothesis were true, then $\sum_{j=1}^{N} j\, X_j$ would tend to be larger than if we just randomly paired the values $1, \ldots, N$ with the values in the set $\{x_1, \ldots, x_N\}$, and then summed the products of the N pairs. This is because the sum of the paired values of two sets of equal size is largest when the largest values are paired with each other, the second largest are paired with each other, and so on. (In statistical terms the correlation coefficient of data pairs $(j, X_j), j = 1, \ldots, N$ is large when the X_j tend to increase as j increases.) Consequently, one possible permutation test of H_0 versus H_1 is to use the test statistic

$$\mathrm{TS} = \sum_{j=1}^{N} j\, X_j$$

and then reject the null hypothesis when TS is sufficiently large.

The test is performed as follows.

1. Observe the data values—say $X_j = x_j, j = 1, \ldots, N$.
2. Let $t = \sum_{j=1}^{N} j\, x_j$.
3. Compute the p value, equal to the probability that a random pairing of the numbers $1, 2, \ldots, N$ with the values $x_1, x_2, \ldots, x_N$ would result in the sum of the products of the paired terms being at least t.

To compute the p value we use the result that, for a given set of data values $S = \{x_1, \ldots, x_N\}$, when the null hypothesis that the data values are independent and identically distributed is true then $\mathrm{TS} = \sum_{j=1}^{N} j\, X_j$ is approximately a normal random variable with mean

$$E[\mathrm{TS}] = \frac{N(N+1)}{2}\bar{x}$$

and variance

$$\mathrm{Var}(\mathrm{TS}) = \frac{N(N+1)(2N+1)}{6(N-1)}(ss - N\bar{x}^2) + \frac{N^2(N+1)^2}{4(N-1)}\left(\bar{x}^2 - \frac{ss}{N}\right)$$

where

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \text{and} \quad ss = \sum_{i=1}^{N} x_i^2$$

That is, $\bar{x}$ is the average of the data values observed, and ss is the sum of the squares of these values.

■ Example 14.19

To determine if the weekly sales of DVD players is on a downward trend the manager of a large electronics store has been tracking such sales for the past

12 weeks, with the following sales figures from week 1 to week 12 (the current week) resulting:

$$22, 24, 20, 18, 16, 14, 15, 15, 13, 17, 12, 14$$

Are the data strong enough to reject the null hypothesis that the distribution of sales is unchanging in time, and so enable the manager to conclude that there is a downward trend in sales?

Solution

Let the null hypothesis be that the distribution of sales is unchanged over time, and let the alternative hypothesis be that there is a downward trend in sales. Thus, if the alternative hypothesis is true then there would be a negative correlation between X_j, the sales during week j, and j. So a relatively small value of $\sum_{j=1}^{12} j\, X_j$ would be evidence in favor of the alternative hypothesis. Now, with x_j equal to the observed value of X_j, the sales data gives that

$$\sum_{j=1}^{12} j\, x_j = 1178$$

A calculation yields that when H_0 is true,

$$E[\mathrm{TS}] = 1300 \quad \mathrm{Var(TS)} = 1958.81$$

Because we want to reject the null hypothesis when TS is small, the normal approximation yields that

$$\begin{aligned} p \text{ value} &= P_{\mathrm{H}_0}(TS \le 1178) \\ &= P_{\mathrm{H}_0}\left(\frac{TS - 1300}{\sqrt{1958.81}} \le \frac{1178 - 1300}{\sqrt{1958.81}}\right) \\ &\approx P(Z < -2.757) \\ &= .0029 \end{aligned}$$

Consequently, we reject the null hypothesis that the distribution of sales has not changed as time has passed.

Let us now suppose that whereas the set of 12 data values was as before, they now appeared in the order

$$22, 14, 14, 16, 24, 20, 18, 15, 17, 15, 12, 13$$

With these data, the value of the test statistic is $\sum_{j=1}^{12} j\,X_j = 1233$, and the normal approximation yields that

$$\begin{aligned} p \text{ value} &= P_{H_0}(TS \leq 1233) \\ &= P_{H_0}\left(\frac{TS - 1300}{\sqrt{1958.81}} \leq \frac{1233 - 1300}{\sqrt{1958.81}}\right) \\ &\approx \Phi(-1.514) \\ &= .065 \end{aligned}$$

Consequently, in this case the data would not enable us to reject the null hypothesis at, say, the 5 percent level of significance though it would lead to a rejection at the 10 percent level of significance.

Finally, suppose again that the set of 12 data values was as before, but suppose that they now appeared in the order

22, 14, 14, 16, 24, 13, 18, 15, 17, 15, 12, 20

In this case, the value of the test statistic is $\sum_{j=1}^{12} j\,X_j = 1275$. Thus, the normal approximation yields that

$$\begin{aligned} p \text{ value} &= P_{H_0}(TS \leq 1275) \\ &= P_{H_0}\left(\frac{TS - 1300}{\sqrt{1958.81}} \leq \frac{1275 - 1300}{\sqrt{1958.81}}\right) \\ &\approx \Phi(-.565) \\ &\approx .286 \end{aligned}$$

Consequently, in this case the data would not enable us to reject the null hypothesis at any reasonable level of significance.

PROBLEMS

1. The following are a student's weekly exam scores. Do they prove that the student improved (as far as exam score) as the semester progressed?

   68, 64, 72, 80, 72, 84, 76, 86, 94, 92

2. A baseball player has the reputation of starting slowly at the beginning of a season but then continually improving as the season progresses. Do the following data, which indicate the number of hits he

has in consecutive five-game strings of the season, strongly validate the player's reputation?

8, 3, 7, 12, 4, 7, 13, 6, 0, 9, 12, 4, 4, 6, 10

KEY TERMS

Nonparametric hypotheses tests: A class of hypotheses tests about a population that do not assume that the population distribution is a specified type.

Sign test: A nonparametric test concerning the median of a population. The test statistic counts the number of data values less than the hypothesized median.

Signed-rank test: A nonparametric test of the null hypothesis that a population distribution is symmetric about a specified value.

Rank-sum test: A nonparametric test of the equality of two population distributions. It uses independent samples from the populations and then ranks the combined data from the two samples. The sum of the ranks of (either) one of the samples is the test statistic.

Runs test: A nonparametric test of the hypothesis that an ordered data sequence constitutes a random sample from some population.

SUMMARY

In this chapter we learned how to test a statistical hypothesis without making any assumptions about the form of the underlying probability distributions. Such tests are called *nonparametric*.

Sign Test

The sign test can be used to test hypotheses concerning the median of a distribution. Suppose that for a specified value m we want to test

$$H_0\colon \eta = m$$

against

$$H_1\colon \eta \neq m$$

where η is the median of the population distribution. To obtain a test, choose a sample of elements of the population, discarding any data values exactly equal to m. Suppose n data values remain. The test statistic of the sign test is the number of remaining values that are less than m. If there are i such values, then the p value of the sign test is given by

$$p \text{ value} = \begin{cases} 2P\{N \leq i\} & \text{if } i \leq \frac{n}{2} \\ 2P\{N \geq i\} & \text{if } i \geq \frac{n}{2} \end{cases}$$

where N is a binomial random variable with parameters n and $p = 1/2$. The computation of the binomial probability can be done either by running Program 5-1 or by using the normal approximation to the binomial.

The sign test can also be used to test the one-sided hypothesis

$$\text{H}_0\colon \eta \leq m \quad \text{against} \quad \text{H}_1\colon \eta > m$$

It uses the same test statistic as earlier, namely, the number of data values that are less than m. If the value of the test statistic is i, then the p value is given by

$$p \text{ value} = P\{N \leq i\}$$

where again N is binomial with parameters n and $p = 1/2$.

If the one-sided hypothesis to be tested is

$$\text{H}_0\colon \eta \geq m \quad \text{against} \quad \text{H}_1\colon \eta < m$$

then the p value, when there are i values less than m, is

$$p \text{ value} = P\{N \geq i\}$$

where N is binomial with parameters n and $p = 1/2$.

As in all hypothesis testing, the null hypothesis is rejected at any significance level greater than or equal to the p value.

Signed-Rank Test

The signed-rank test is used to test the hypothesis that a population distribution is symmetric about the value 0. In applications, the population often consists of the differences of paired data. The signed-rank test calls for choosing a random sample from the population, discarding any data values equal to 0. It then ranks the remaining nonzero values, say there are n of them, in increasing order of their absolute values. The test statistic is equal to the sum of the rankings of the negative data values. If the value of the test statistic TS is equal to t, then the p value is

$$p \text{ value} = 2\,\text{Min}(P\{\text{TS} \leq t\}, P\{\text{TS} \geq t\})$$

where the probabilities are to be computed under the assumption that the null hypothesis is true. The p value can be found either by using Program 14-1 or by using the fact that TS will have approximately, when the null hypothesis is true and n is of least moderate size, a normal distribution with mean and variance, respectively, given by

$$E[\text{TS}] = \frac{n(n+1)}{4} \qquad \text{Var(TS)} = \frac{n(n+1)(2n+1)}{24}$$

Rank-Sum Test

The rank-sum test can be used to test the null hypothesis that two population distributions are identical, when the data consist of independent samples from these populations. Arbitrarily designate one of the samples as the first sample. Suppose that the size of this sample is n and that of the other sample is m. Now rank the combined samples. The test statistic TS of the rank-sum test is the sum of the ranks of the first sample. The rank-sum test calls for rejecting the null hypothesis when the value of the test statistic is either significantly large or significantly small.

When n and m are both greater than 7, the test statistic TS will, when H_0 is true, have an approximately normal distribution with mean and variance given by, respectively,

$$E[\text{TS}] = \frac{n(n+m+1)}{2} \qquad \text{Var(TS)} = \frac{nm(n+m+1)}{12}$$

This enables us to approximate the p value, which when $\text{TS} = t$ is given by

$$p \text{ value} \approx \begin{cases} 2P\left\{Z \le \dfrac{t + 0.5 - n(n+m+1)/2}{\sqrt{nm(n+m+1)/12}}\right\} & \text{if } t < \dfrac{n(n+m+1)}{2} \\ 2P\left\{Z \ge \dfrac{t - 0.5 - n(n+m+1)/2}{\sqrt{nm(n+m+1)/12}}\right\} & \text{if } t > \dfrac{n(n+m+1)}{2} \end{cases}$$

For values of t near $n(n+m+1)/2$, the p value is close to 1, and so the null hypothesis would not be rejected (and the preceding probability need not be calculated).

For small values of n and m the exact p value can be obtained by running Program 14-2.

Runs Test

The runs test can be used to test the null hypothesis that a given sequence of data constitutes a random sample from some population. It supposes that each datum is either a 0 or a 1. Any consecutive sequence of either 0s or 1s is called a *run*. The test statistic for the runs test is R, the total number of runs. If the observed value of R is r, then the p value of the runs test is given by

$$p \text{ value} = 2\,\text{Min}(P\{R \le r\}, P\{R \ge r\})$$

The probabilities here are to be computed under the assumption that the null hypothesis is true.

Program 14-3 can be used to determine this p value. If Program 14–3 is not available, we can approximate the p value by making use of the fact that when the null hypothesis is true, R will have an approximately normal distribution. The mean and variance, respectively, of this distribution are

$$\mu = \frac{2nm}{n+m} + 1 \qquad \sigma^2 = \frac{2nm(2nm - n - m)}{(n+m)^2(n+m-1)}$$

REVIEW PROBLEMS

1. Use a nonparametric test to solve Prob. 2 in Sec. 10.4.
2. Use a nonparametric test to solve Prob. 3 in Sec. 10.4.
3. According to the *Federal Reserve Bulletin* of January 1992, in 1989 the sample median net worth of all 55-year-olds in the labor force was $104,500. (The sample mean was $438,300.) Suppose that a random sample of 1000 such workers today yielded the result that 421 had a family net worth (in 1989 dollars) of over $104,500. Can we conclude, at the 5 percent level of significance, that the median net worth has decreased?
4. An experiment was initiated to study the effect of a newly developed gasoline detergent on automobile mileage. The following data, representing mileage per gallon before and after the detergent was added for each of eight cars, resulted:

Car	Mileage without additive	Mileage with additive
1	24.2	23.5
2	30.4	29.6
3	32.7	32.3
4	19.8	17.6
5	25.0	25.3
6	24.9	25.4
7	22.2	20.6
8	21.5	20.7

Find the p value of the test of the hypothesis that mileage is not affected by the additive when using

(a) The sign test

(b) The signed-rank test

Compare your results with each other and with Example 10.8 of Sec. 10.5.

5. Test the hypothesis that the weights of the students given in App. A constitute a random sample. Use the first 40 data values to test this hypothesis. What is the p value?
6. Choose a random sample of 30 of the students in App. A. Use those data to test the hypothesis that the median weight of all the students listed is less than or equal to 130 pounds.
7. Choose a random sample of 40 students from App. A, and use this sample to test the hypothesis that the distribution of blood cholesterol readings is the same for both sexes.
8. A chemist tests a variety of blood samples for a certain virus. The successive results are as follows, with P meaning that the virus is present and A that it is absent.

$$A\ A\ P\ P\ P\ A\ A\ A\ A\ P\ P\ A\ A\ A\ A\ A\ A\ P\ P\ A\ A\ P\ P\ P\ P$$

Test the hypothesis that the chemist tested the samples in a random order. Use the 5 percent level of significance. Also determine the p value.

9. Repeat Example 10.9, this time using a nonparametric test.
10. Explain how we could have used a contingency table analysis to test the hypothesis in Example 14.10. Do this test, find the p value, and compare it with the one obtained in Example 14.10. Since the contingency table test is different from the one used in Example 14.10, the two p values need not be equal.
11. Consider Prob. 7 of Sec. 14.3. Suppose now that the same car had been brought to 16 different automobile repair shops, with the woman bringing it into 8 of them and the man to the other 8. Suppose the data on the quoted repair prices were as given in that problem. Test the hypothesis, at the 5 percent level of significance, that the distributions of price quotes received by the man and by the woman are the same.