

Nonparametric Tests

INTRODUCTION

Most tests of hypotheses and significance (or decision rules) considered in previous chapters require various assumptions about the distribution of the population from which the samples are drawn. For example, the one-way classification of Chapter 16 requires that the populations be normally distributed and have equal standard deviations.

Situations arise in practice in which such assumptions may not be justified or in which there is doubt that they apply, as in the case where a population may be highly skewed. Because of this, statisticians have devised various tests and methods that are independent of population distributions and associated parameters. These are called *nonparametric tests*.

Nonparametric tests can be used as shortcut replacements for more complicated tests. They are especially valuable in dealing with nonnumerical data, such as arise when consumers rank cereals or other products in order of preference.

THE SIGN TEST

Consider Table 17.1, which shows the numbers of defective bolts produced by two different types of machines (I and II) on 12 consecutive days and which assumes that the machines have the same total output per day. We wish to test the hypothesis H_0 that there is no difference between the machines: that the observed differences between the machines in terms of the numbers of defective bolts they produce are merely the result of chance, which is to say that the samples come from the same population.

Table 17.1

Day	1	2	3	4	5	6	7	8	9	10	11	12
Machine I	47	56	54	49	36	48	51	38	61	49	56	52
Machine II	71	63	45	64	50	55	42	46	53	57	75	60

A simple nonparametric test in the case of such paired samples is provided by the *sign test*. This test consists of taking the difference between the numbers of defective bolts for each day and writing only the *sign* of the difference; for instance, for day 1 we have $47-71$, which is negative. In this way we obtain from Table 17.1 the sequence of signs

$$- \quad - \quad + \quad - \quad - \quad - \quad + \quad - \quad + \quad - \quad - \quad - \quad (1)$$

(i.e., 3 pluses and 9 minuses). Now if it is just as likely to get a + as a −, we would expect to get 6 of each. The test of H_0 is thus equivalent to that of whether a coin is fair if 12 tosses result in 3 heads (+) and 9 tails (−). This involves the binomial distribution of Chapter 7. Problem 17.1 shows that by using a two-tailed test of this distribution at the 0.05 significance level, we cannot reject H_0 ; that is, there is no difference between the machines at this level.

Remark 1: If on some day the machines produced the same number of defective bolts, a difference of *zero* would appear in sequence (I). In such case we can omit these sample values and use 11 instead of 12 observations.

Remark 2: A normal approximation to the binomial distribution, using a correction for continuity, can also be used (see Problem 17.2).

Although the sign test is particularly useful for paired samples, as in Table 17.1, it can also be used for problems involving single samples (see Problems 17.3 and 17.4).

THE MANN–WHITNEY U TEST

Consider Table 17.2, which shows the strengths of cables made from two different alloys, I and II. In this table we have two samples: 8 cables of alloy I and 10 cables of alloy II. We would like to decide whether or not there is a difference between the samples or, equivalently, whether or not they come from the same population. Although this problem can be worked by using the t test of Chapter 11, a non-parametric test called the *Mann–Whitney U test*, or briefly the *U test*, is useful. This test consists of the following steps:

Table 17.2

Alloy I				Alloy II				
18.3	16.4	22.7	17.8	12.6	14.1	20.5	10.7	15.9
18.9	25.3	16.1	24.2	19.6	12.9	15.2	11.8	14.7

Step 1. Combine all sample values in an array from the smallest to the largest, and assign ranks (in this case from 1 to 18) to all these values. If two or more sample values are identical (i.e., there are *tie scores*, or briefly *ties*), the sample values are each assigned a rank equal to the *mean* of the ranks that would otherwise be assigned. If the entry 18.9 in Table 17.2 were 18.3, two identical values 18.3 would occupy ranks 12 and 13 in the array so that the rank assigned to each would be $\frac{1}{2}(12 + 13) = 12.5$.

Step 2. Find the sum of the ranks for each of the samples. Denote these sums by R_1 , and R_2 , where N_1 and N_2 are the respective sample sizes. For convenience, choose N_1 as the smaller size if they are unequal, so that $N_1 \leq N_2$. A significant difference between the rank sums R_1 and R_2 implies a significant difference between the samples.

Step 3. To test the difference between the rank sums, use the statistic

$$U = N_1N_2 + \frac{N_1(N_1 + 1)}{2} - R_1 \tag{2}$$

corresponding to sample 1. The sampling distribution of U is symmetrical and has a mean and variance given, respectively, by the formulas

$$\mu_U = \frac{N_1N_2}{2} \quad \sigma_U^2 = \frac{N_1N_2(N_1 + N_2 + 1)}{12} \tag{3}$$

If N_1 and N_2 are both at least equal to 8, it turns out that the distribution of U is nearly normal, so that

$$z = \frac{U - \mu_U}{\sigma_U} \quad (4)$$

is normally distributed with mean 0 and variance 1. Using Appendix II, we can then decide whether the samples are significantly different. Problem 17.5 shows that there is a significant difference between the cables at the 0.05 level.

Remark 3: A value corresponding to sample 2 is given by the statistic

$$U = N_1 N_2 + \frac{N_2(N_2 + 1)}{2} - R_2 \quad (5)$$

and has the same sampling distribution as statistic (2), with the mean and variance of formulas (3). Statistic (5) is related to statistic (2), for if U_1 and U_2 are the values corresponding to statistics (2) and (5), respectively, then we have the result

$$U_1 + U_2 = N_1 N_2 \quad (6)$$

We also have

$$R_1 + R_2 = \frac{N(N + 1)}{2} \quad (7)$$

where $N = N_1 + N_2$. Result (7) can provide a check for calculations.

Remark 4: The statistic U in equation (2) is the total number of times that sample 1 values precede sample 2 values when all sample values are arranged in increasing order of magnitude. This provides an alternative *counting method* for finding U .

THE KRUSKAL–WALLIS H TEST

The U test is a nonparametric test for deciding whether or not two samples come from the same population. A generalization of this for k samples is provided by the *Kruskal–Wallis H test*, or briefly the *H test*.

This test may be described thus: Suppose that we have k samples of sizes N_1, N_2, \dots, N_k , with the total size of all samples taken together being given by $N = N_1 + N_2 + \dots + N_k$. Suppose further that the data from all the samples taken together are ranked and that the sums of the ranks for the k samples are R_1, R_2, \dots, R_k , respectively. If we define the statistic

$$H = \frac{12}{N(N + 1)} \sum_{j=1}^k \frac{R_j^2}{N_j} - 3(N + 1) \quad (8)$$

then it can be shown that the sampling distribution of H is very nearly a *chi-square distribution* with $k - 1$ degrees of freedom, provided that N_1, N_2, \dots, N_k are all at least 5.

The H test provides a nonparametric method in the *analysis of variance* for one-way classification, or one-factor experiments, and generalizations can be made.

THE H TEST CORRECTED FOR TIES

In case there are too many ties among the observations in the sample data, the value of H given by statistic (8) is smaller than it should be. The corrected value of H , denoted by H_c , is obtained by dividing

the value given in statistic (8) by the correction factor

$$1 - \frac{\sum (T^3 - T)}{N^3 - N} \tag{9}$$

where T is the number of ties corresponding to each observation and where the sum is taken over all the observations. If there are no ties, then $T = 0$ and factor (9) reduces to 1, so that no correction is needed. In practice, the correction is usually negligible (i.e., it is not enough to warrant a change in the decision).

THE RUNS TEST FOR RANDOMNESS

Although the word “random” has been used many times in this book (such as in “random sampling” and “tossing a coin at random”), no previous chapter has given any test for randomness. A nonparametric test for randomness is provided by the *theory of runs*.

To understand what a run is, consider a sequence made up of two symbols, a and b , such as

$$a a | b b b | a | b b | a a a a | b b b | a a a a | \tag{10}$$

In tossing a coin, for example, a could represent “heads” and b could represent “tails.” Or in sampling the bolts produced by a machine, a could represent “defective” and b could represent “nondefective.”

A *run* is defined as a set of identical (or related) symbols contained between two different symbols or no symbol (such as at the beginning or end of the sequence). Proceeding from left to right in sequence (10), the first run, indicated by a vertical bar, consists of two a ’s; similarly, the second run consists of three b ’s, the third run consists of one a , etc. There are seven runs in all.

It seems clear that some relationship exists between randomness and the number of runs. Thus for the sequence

$$a | b | a | b | a | b | a | b | a | b | a | b | \tag{11}$$

there is a *cyclic pattern*, in which we go from a to b , back to a again, etc., which we could hardly believe to be random. In such case we have *too many* runs (in fact, we have the maximum number possible for the given number of a ’s and b ’s).

On the other hand, for the sequence

$$a a a a a | b b b b | a a a a | b b b | \tag{12}$$

there seems to be a *trend pattern*, in which the a ’s and b ’s are grouped (or clustered) together. In such case there are *too few* runs, and we would not consider the sequence to be random.

Thus a sequence would be considered nonrandom if there are either too many or too few runs, and random otherwise. To quantify this idea, suppose that we form all possible sequences consisting of N_1 a ’s and N_2 b ’s, for a total of N symbols in all ($N_1 + N_2 = N$). The collection of all these sequences provides us with a sampling distribution: Each sequence has an associated number of runs, denoted by V . In this way we are led to the sampling distribution of the statistic V . It can be shown that this sampling distribution has a mean and variance given, respectively, by the formulas

$$\mu_V = \frac{2N_1N_2}{N_1 + N_2} + 1 \quad \sigma_V^2 = \frac{2N_1N_2(2N_1N_2 - N_1 - N_2)}{(N_1 + N_2)^2(N_1 + N_2 - 1)} \tag{13}$$

By using formulas (13), we can test the hypothesis of randomness at appropriate levels of significance. It turns out that if both N_1 and N_2 are at least equal to 8, then the sampling distribution of V is very nearly a normal distribution. Thus

$$z = \frac{V - \mu_V}{\sigma_V} \tag{14}$$

is normally distributed with mean 0 and variance 1, and thus Appendix II can be used.

FURTHER APPLICATIONS OF THE RUNS TEST

The following are other applications of the runs test to statistical problems:

1. **Above- and Below-Median Test for Randomness of Numerical Data.** To determine whether numerical data (such as collected in a sample) are random, first place the data in the *same order* in which they were collected. Then find the median of the data and replace each entry with the letter *a* or *b* according to whether its value is *above* or *below* the median. If a value is the same as the median, omit it from the sample. The sample is random or not according to whether the sequence of *a*'s and *b*'s is random or not. (See Problem 17.20.)
2. **Differences in Populations from Which Samples Are Drawn.** Suppose that two samples of sizes m and n are denoted by a_1, a_2, \dots, a_m and b_1, b_2, \dots, b_n , respectively. To decide whether the samples do or do not come from the same population, first arrange all $m + n$ sample values in a sequence of increasing values. If some values are the same, they should be ordered by a random process (such as by using random numbers). If the resulting sequence is random, we can conclude that the samples are not really different and thus come from the same population; if the sequence is not random, no such conclusion can be drawn. This test can provide an alternative to the Mann–Whitney U test. (See Problem 17.21.)

SPEARMAN'S RANK CORRELATION

Nonparametric methods can also be used to measure the correlation of two variables, X and Y . Instead of using precise values of the variables, or when such precision is unavailable, the data may be ranked from 1 to N in order of size, importance, etc. If X and Y are ranked in such a manner, the *coefficient of rank correlation*, or *Spearman's formula for rank correlation* (as it is often called), is given by

$$r_s = 1 - \frac{6 \sum D^2}{N(N^2 - 1)} \quad (15)$$

where D denotes the differences between the ranks of corresponding values of X and Y , and where N is the number of pairs of values (X, Y) in the data.

Solved Problems

THE SIGN TEST

- 17.1** Referring to Table 17.1, test the hypothesis H_0 that there is no difference between machines I and II against the alternative hypothesis H_1 that there is a difference at the 0.05 significance level.

SOLUTION

Figure 17-1 shows the binomial distribution of the probabilities of X heads in 12 tosses of a coin as areas under rectangles at $X=0, 1, \dots, 12$. Superimposed on the binomial distribution is the normal distribution, shown as a dashed curve. The mean of the binomial distribution is $\mu = np = 12(0.5) = 6$. The standard deviation is $\sigma = \sqrt{npq} = \sqrt{12(0.5)(0.5)} = \sqrt{3} = 1.73$. The normal curve also has mean = 6 and standard deviation = 1.73. From chapter 7 the binomial probability of X heads is

$$\Pr\{X\} = \binom{12}{X} \left(\frac{1}{2}\right)^X \left(\frac{1}{2}\right)^{12-X} = \binom{12}{X} \left(\frac{1}{2}\right)^{12}$$