Applications of Non-Parametric Statistics

Chan Yiu Man
Department of Mathematics
National University of Singapore

1. Introduction

In the broadest sense a nonparametric statistical method is one that does not rely for its validity or its utility on any assumptions about the form of distribution that is taken to have generated the sample values on the basis of which inferences about the population distribution are to be made. However weaker assumptions are still required in most of the nonparametric statistical methods. Some of the nonparametric tests such as sign test were used as early as in the eighteenth century. Since nonparametric tests are based on weaker assumptions, they have wider applications. Of course if we have more knowledge about the underlying distribution, a more powerful test which depends on how much we know should be used.

We shall discuss through examples some applications of the sign test, the Wilcoxon signed rank test and Spearman's and Kendall's correlation coefficients. Throughout the discussion, we assume that there should be some kind of ordering between observations of the random variables. That is, we assume that the measurement of scale is at least ordinal. Also there should be some natural basis for pairing observations. That is, we limit our discussion in bivariate situations to the matched pairs case.

2. Sign Test

Let (X_i, Y_i), $i = 1, \ldots, n'$ be a bivariate random sample. Define

$$\text{sign}(X_i, Y_i) = \begin{cases} + & \text{if } X_i < Y_i \\ 0 & \text{if } X_i = Y_i \\ - & \text{if } X_i > Y_i. \end{cases}$$

Lecture given at the Workshop on Hypothesis Testing and Non-Parametric Statistics for school teachers, 7 September 1987.

We are interested in whether $P(X_i > Y_i) \geq .5$ for all i or $P(X_i < Y_i) \geq .5$ for all i. This can be translated into a model that $P(+) \geq .5$ or $P(-) \geq .5$. If the tied pairs are omitted, then we would like to test

$$H_0 : P(+) = P(-) \quad \text{against} \quad H_1 : P(+) \neq P(-).$$

Let T = total number of +'s. We disregard all tied pairs and let n = total number of +'s and −'s. Under the null hypothesis, $T \sim Bi(n, .5)$. Hence we reject H_0 at a significant level α if $T \leq t$ or $T \geq n - t$ where t is chosen such that

$$\sum_{j=0}^{t} \binom{n}{j} (.5)^n \approx \frac{\alpha}{2}.$$

Alternatively, if we let t' = observed value of T , then we reject H_0 if

$$P(X \geq t') < \frac{\alpha}{2} \quad \text{or} \quad P(X \leq t') < \frac{\alpha}{2},$$

where $X \sim Bi(n, .5)$. If n is large, then by the Central Limit Theorem, $(T/n) \sim N((1/2), (1/4n))$ under H_0 . Hence we reject H_0 at a significant level α if

$$T \leq \frac{1}{2} - z_{\alpha/2}\sqrt{\frac{1}{4n}} \quad \text{or} \quad T \geq \frac{1}{2} + z_{\alpha/2}\sqrt{\frac{1}{4n}}.$$

Alternatively, we reject H_0 if

$$P\left[Z > \left(\frac{t'}{n} - \frac{1}{2}\right) \Big/ \sqrt{\frac{1}{4n}}\right] < \frac{\alpha}{2} \quad \text{or}$$

$$P\left[Z < \left(\frac{t'}{n} - \frac{1}{2}\right) \Big/ \sqrt{\frac{1}{4n}}\right] < \frac{\alpha}{2}.$$

Example 1

Six students went on a diet in an attempt to lose weight, with the following results.

Student	A	B	C	D	E	F
Weight before diet (X_i kg)	79.0	86.7	85.4	82.6	91.3	85.4
Weight after diet (Y_i kg)	74.9	84.4	83.1	80.8	92.2	82.2
Sign(X_i, Y_i)	−	−	−	−	+	−

Is the diet an effective means of losing weight? We want to test

$$H_0 : P(+) = P(-) \quad \text{against} \quad H_1 : P(+) < P(-).$$

The observed number of positive signs is 1 and $P(T \leq 1|H_0) = .1094 > .05$, therefore we do not reject H_0. We conclude that the diet is not an effective means of losing weight.

Remarks

a. If $Y_i - X_i$ are random variables with a symmetric distribution function, then the Wilcoxon test for matched pairs is more appropriate.
b. If $Y_i - X_i \sim N(\mu, \sigma^2)$, then a matched-pair t-test should be used.

Median Test

The sign test can be used in one sample case to test whether the median θ of the population is equal to a particular value θ_0. We proceed as in the matched pairs case except we replace Y_i's by θ_0.

Cox and Stuart Test for Trend

The Cox and Stuart test is a modified sign test. Consider a sequence of random variables $X_1, \ldots, X_n$, arranged in a particular order, such as the order in which the random variables are observed. We want to see if a trend exists in the sequence. We group the random variables into pairs $(X_1, X_{1+c}), (X_2, X_{2+c}), \ldots, (X_{n-c}, X_n)$, where $c = n/2$ if n is even and $c = (n+1)/2$ if n is odd. If there is an upward trend, then

$$P(X_i < X_{i+c}) > P(X_i > X_{i+c}) \quad \text{for all } i.$$

Define

$$\text{sign}(X_i, X_{i+c}) = \begin{cases} + & \text{if } X_i < X_{i+c} \\ 0 & \text{if } X_i = X_{i+c} \\ - & \text{if } X_i > X_{i+c}. \end{cases}$$

We can now apply the sign test to test if $P(X_i < X_{i+c}) = P(X_i > X_{i+c})$, or equivalently if $P(+) = P(-)$. The following example shows how this test is applied.

Example 2

The total annual precipitation in a certain city is recorded each year for 19 years, and this record is examined to see if there is any trend in

the amount of precipitation. The amounts of precipitation in centimeters were 114.9, 116.4, 106.1, 92.1, 115.2, 132.7, 89.8, 145.2, 89.8, 148.1, 104.3, 85.6, 116.2, 96.3, 106.0, 91.6, 126.6, 92.0 and 101.3. We want to test

H_0: there is no trend in the precipitation against
H_1: there is a trend in the precipitation.

The numbers are paired as (114.9, 104.3) −, (116.4, 85.6) −, (106.1, 116.2) +, (92.1, 96.3) +, (115.2, 106.0) −, (132.7, 91.6) −, (89.8, 126.6) +, (145.2, 92.0) −, (89.8, 101.3) +, and the middle number 148.1 is omitted. There are no ties, so $n = 9$. Hence $T \sim Bi(9, .5)$ under H_0. The critical region of size .0390 is $\{t : t \leq 1 \text{ or } t \geq 8\}$, where t is the observed value of T. Since the observed value of T is 4, therefore the hypothesis that no trend exists is not rejected.

Cox and Stuart Test for Correlation

The following example illustrates how the Cox and Stuart test can be used in detecting correlation.

Example 3

Cochran (1937) compared the reactions of several patients with each of two drugs, to see if there was a positive correlation between the two reactions for each patient.

Patient	1	2	3	4	5	6	7	8	9	10
Drug 1	.7	−1.6	−.2	−1.2	−.1	3.4	3.7	.8	0.	2.0
Drug 2	1.9	.8	1.1	.1	−.1	4.4	5.5	1.6	4.6	3.4

We want to test

H_0: There is no positive correlation against
H_1: There is a positive correlation.

Ordering the pairs according to the reaction from drug 1 gives:

Patient	2	4	3	5	9	1	8	10	6	7
Drug 1	−1.6	−1.2	−.2	−.1	0.	.7	.8	2.0	3.4	3.7
Drug 2	.8	.1	1.1	−.1	4.6	1.9	1.6	3.4	4.4	5.5

To test if there exists a positive (or negative) correlation is equivalent to testing the presence of an upward (or downward) trend on the newly arranged sequence of observations on drug 2. The pairs are (.8,1.9) +,

(.1,1.6) +, (1.1,3.4) +, (−.1,4.4) +, (4.6,5.5) +. The critical region of size .0312 (obtained from the binomial table with $n = 5$ and $p = .5$) is $\{t : t = 5\}$. Since the observed value of T is 5, therefore we reject H_0 and conclude that there is a positive correlation between reactions to the two drugs.

3. Wilcoxon Matched Pair Signed Rank Test

Suppose we have a set of bivariate random variables $(X_i, Y_i), i = 1, \ldots, n'$. As in the previous section, we are interested in testing whether

$$P(Y_i > X_i) = P(Y_i < X_i) \quad \text{for} \quad \text{all} \quad i.$$

Let us consider the differences

$$D_i = Y_i - X_i, \; i = 1, \ldots, n'.$$

The above test is equivalent to testing if all of the differences D_i have a median which is equal to zero for all i's. That is, we must test

$$H_0 : d_{.5} = 0 \quad \text{against} \quad H_1 : d_{.5} \neq 0,$$

where $d_{.5}$ is the median of D_i's for all i. We delete those pairs with a difference of zero (i.e. $X_i = Y_i$) and then rank the remaining pairs by the relative size of the absolute difference $|D_i|$. In case of several pairs having the same absolute difference, we assign to each of these pairs the average of the ranks that would have otherwise been assigned to them. An additional assumption of symmetry for the distribution of the differences is made. (A distribution is symmetric about $x = c$, for some c, if

$$P(X \leq c - x) = P(X \geq c + x)$$

for each possible value of x. In fact, c is just the median and the mean of the population distribution). The consequence of this additional assumption is that the required scale of measurement is changed from ordinal to interval. (An *interval scale* means that the distance between any two measurements is meaningful.) Let R_i, called *signed rank*, be defined as follows.

$$R_i = \begin{cases} \text{Rank}(|D_i|) & \text{if } D_i > 0 \\ -\text{Rank}(|D_i|) & \text{if } D_i < 0. \end{cases}$$

If $d_{.5} = 0$, then the size of each difference and the resulting rank are independent of whether the difference is above or below the median. Let

$$T = \Sigma R_i / (\Sigma R_i^2)^{.5}.$$

In case of no ties,

$$\Sigma R_i^2 = \Sigma i^2 = \frac{n(n+1)(2n+1)}{6}.$$

For the case of no ties, it is more convenient to use only the positive signed ranks. Let

$$T^+ = \Sigma(R_i \text{ where } D_i \text{ is positive}).$$

Notice that

$$\Sigma R_i = 2T^+ - n(n+1)/2.$$

The exact distribution of T^+ under H_0 can be derived in the following way. Firstly, we count the number of sign combinations for ranks $|D_i|$'s that will give a particular value of T^+. For example, both $(1, -2, -3, 4)$ and $(-1, 2, 3, -4)$ give $T^+ = 5$ for $n = 4$. Then the probability that T^+ equals that particular value is obtained by dividing that number by 2^n. Under H_0, we have

$$P(R_i = a_i) = P(R_i = -a_i) = \frac{1}{2},$$

where $a_i \in \{1, 2, \ldots, n\}$. Therefore $E(R_i) = 0$ and $Var(R_i) = a_i^2$. Hence when n is large, T is approximately distributed as $N(0, 1)$ and T^+ is approximately distributed as

$$N\left(\frac{n(n+1)}{4}, \quad \frac{n(n+1)(2n+1)}{24}\right).$$

Quantiles w_p of the exact distribution of T^+ can be found in most of the nonparametric textbooks. The cumulative probability function $P(T^+ \leq t)$ is given in Lehmann's book up to $n = 20$. For the two sided hypothesis, we reject H_0 if

$$T < w_{\alpha/2} \quad \text{or} \quad T > w_{1-\alpha/2}.$$

Alternatively, we reject H_0 if

$$P(T^+ \leq \text{observed value of } T^+) < \frac{\alpha}{2} \quad \text{or}$$

$$P(T^+ \geq \text{observed value of } T^+) < \frac{\alpha}{2}.$$

The following example illustrates how this test can be applied.

Example 4

To test whether a new fertilizer will give a higher yield than the fertilizer that has been used in the past, eight fields in a farm are each divided into two parts which are assigned to the two fertilizers at random. The yields in kilogram are given below.

Field i	1	2	3	4	5	6	7	8
Old fertilizer X_i	88	77	76	64	96	72	65	65
New fertilizer Y_i	86	71	77	68	91	72	77	70
Difference D_i	−2	−6	1	4	−5	0	12	5
Rank($\lvert D_i \rvert$)	2	6	1	3	4.5	–	7	4.5

We want to test

$$H_0 : d_{.5} = 0 \quad \text{against} \quad H_1 : d_{.5} > 0.$$

From the table showing quantiles of the Wilcoxon signed rank test statistic (p.460-1 Conover(1980)) with $n = 7, \alpha = .05$, we have $w_{.05} = 4$. Hence $w_{.95} = 24$ $(w_p = \frac{n(n+1)}{2} - w_{1-p})$ and the critical region is $\{t : t \geq 24\}$. Since $T^+ = 15.5$, we do not reject H_0. Alternatively, from the table in Lehmann's book (p.418-21), $P(T^+ > 15.5) \approx .44$. Hence H_0 is not rejected.

4. Measures on rank correlation

Suppose we have a random sample $(x_i, y_i), i = 1, 2, \ldots, n$, of n independent observations on the bivariate random variable (X, Y). In many applications the question of possible independence of X and Y is considered and the statistics related to the correlation coefficient are often used

in this connection. Technically a test of zero correlation is not, of course, a test of independence, but where such a test is used the primary concern is often to establish dependence, rather than independence.

The most commonly used measure of correlation is Pearson's product moment correlation coefficient, denoted by r and defined as

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{[\sum(x_i - \bar{x})^2]^{.5}[\sum(y_i - \bar{y})^2]^{.5}}$$

This measure of correlation may be used without any requirements concerning the type of underlying distribution. However, the distribution of r depends on the distribution of (X, Y). Therefore r cannot be used as a test statistic unless the distribution of (X, Y) is known.

Spearman's rank correlation coefficient ρ

If we transform the x_i values and y_i values to their respective ranks $R(x_i)$ and $R(y_i)$ and then calculate r as above, we obtain the Spearman's rank correlation coefficient ρ, i.e.

$$\rho = \frac{\sum[R(X_i) - \frac{n+1}{2}][R(Y_i) - \frac{n+1}{2}]}{[\sum R(X_i)^2 - \frac{n(n+1)^2}{4}]^{.5}[\sum R(Y_i)^2 - \frac{n(n+1)^2}{4}]^{.5}}.$$

If there are no ties,

$$\rho = 1 - 6\sum \frac{[R(X_i) - R(Y_i)]^2}{n(n^2 - 1)}.$$

Since ρ is based on the ranks, its distribution does not depend on the form of the population distribution. Hence ρ can be used as a test statistic. Under the assumption that X_i and Y_i are independent identically distributed, then each of the $n!$ arrangements of the ranks of the X_i's paired with the ranks of Y_i's is equally likely. Therefore the distribution of ρ is obtained by counting the number of arrangement that give a particular value of ρ and dividing that by $n!$. When n is large, ρ is asymptotically distributed as $N(0, \frac{1}{n-1})$. We illustrate how to use ρ as a test statistic using the data in Example 4.

Rank(X_i)	7	6	5	1	8	4	2.5	2.5
Rank(Y_i)	7	3	5.5	1	8	4	5.5	2

We want to test

H_0: correlation coefficient = 0 against
H_1: correlation coefficient $\neq 0$.

Notice that if we reject H_0, then we also reject the hypothesis that X_i and Y_i are mutually independent. With $n = 8, \alpha = .05$, we obtain $w_{.975} = .7143$ from the table (p.456 Conover). Hence the critical region = $\{\rho : \rho < -.7143 \quad \text{or} \quad \rho > .7143\}$. Since the observed value of ρ is .7530, therefore we reject H_0.

Kendall's Correlation Coefficient τ

Consider a bivariate random sample $(X_i, Y_i), i = 1, 2, \ldots, n$. Two observations (X_i, Y_i) and (X_j, Y_j) are called *concordant* if

$$(X_j - X_i)(Y_j - Y_i) > 0$$

and *disconcordant* if

$$(X_j - X_i)(Y_j - Y_i) < 0,$$

for all $i, j = 1, 2, \ldots, n$ and $i \neq j$. Pairs with ties are neither concordant nor disconcordant. Let N_c and N_d denote the number of concordant and disconcordant pairs respectively. We have

$$0 \leq N_c \leq \frac{n(n-1)}{2} \quad \text{and} \quad 0 \leq N_d \leq \frac{n(n-1)}{2}.$$

Therefore

$$\frac{-n(n-1)}{2} \leq N_c - N_d \leq \frac{n(n-1)}{2}.$$

Kendall's τ is defined as

$$\tau = \frac{(N_c - N_d)}{n(n-1)/2}.$$

If we have arranged the x values in increasing order of magnitude so that $x_1 < x_2 < \ldots < x_n$, then

$$N_c - N_d = \sum_{i<j} \text{sign}(y_j - y_i).$$

The exact distribution of τ can be found in a way similar to the one used in finding the distribution of ρ. When n is large, τ is approximately

distributed as $N(0, \quad [2(2n+5)] / [9n(n-1)])$. Let us consider Example 4 again. (X_i, Y_i): (64,68), (65,70), (65,77), (72,72), (76,77), (77,71), (88,86), (96,91). $N_c = 22$. $N_d = 4$. Hence $\tau = .6429$. The critical region for testing H_0: correlation coefficient = 0 against H_1: correlation coefficient $\neq 0$ is

$$\{\tau : \tau < -\frac{w_p}{28} \quad \text{or} \quad \tau > \frac{w_p}{28}\},$$

where w_p is the p^{th} quantile of $N_c - N_d$. With $n = 8, \alpha = .05$, we have $w_{.975} = 16$ from the table showing quantiles of the Kendall's test statistic (p.458 Conover). Hence we reject H_0 if $|\tau| > .5714$. Since $\tau = .6429$, therefore we reject H_0.

Remarks

a. Spearman's ρ and Kendall's τ can be used to test the trend by pairing the measurements with the time (or order) in which the measurements were taken.

b. We notice that Spearman's $\rho(\rho = .7530)$ is larger than Kendall's $\tau(\tau = .6429)$ based on the same data set. In many situations, ρ tends to be larger than τ, in absolute value. However, as a test of significance, both ρ and τ will give nearly same results in most cases.

References

[1] Cochran, W.G., The efficiencies of the binomial series tests of significance of a mean and a correlation coefficient, *Journal of Royal Statistical Society*, **100** (1937), 69-73.

[2] Conover, W.J., *Practical Nonparametric Statistics*, John Wiley & Sons, 1980.

[3] Lehmann, E.L., *Nonparametrics: Statistical Methods Based on Ranks*, Holden-Day, San Francisco, 1975.