THE EFFECT OF TIES ON CRITICAL VALUES OF SOME TWO-SAMPLE RANK TESTS

BU-258-M D. S. Robson August, 1968

Abstract

When ties occur among the observations in two samples the effect upon the usual rank-test statistic is to reduce its variance, and the use of tabled critical values then results in a loss of power. We consider here the type of adjustments to be made to the tabled critical values, with special reference to a two-sample test of data falling on the perimeter of a circle -- or where observations are compass directions.

Biometrics Unit, Department of Plant Breeding and Biometry, Cornell University.

THE EFFECT OF TIES ON CRITICAL VALUES OF SOME TWO-SAMPLE RANK TESTS

BU-258-M D. S. Robson August, 1968

A rank test of the hypothesis that $X_1, \cdots, X_a$ and $Y_1, \cdots, Y_b$ constitute two independent simple random samples from the same population is usually based upon a test statistic of the form

$$T = \sum_{i=1}^{n} \delta_i f(i)$$

where

$$\delta_i = \begin{cases} 1 \text{ if the } i^{th} \text{ ranking of the } n = a + b \text{ observations is an X} \\ 0 \text{ if the } i^{th} \text{ ranking observation is a Y} \end{cases}$$

and where $f(i)$ is some specified function of the ranks. In the Wilcoxon rank-sum test, for example, $f(i) = i$. When tables of critical values of T are available they apply only to the continuous case where no ties occur among the n observations. When ties do occur the common practice in the case of the rank-sum test is to assign the average of the two (or more) ranks to each of the two (or more) tied observations, and in general we shall suppose that the function $f(i)$ is similarly averaged when ties occur. Such a modification produces a test statistic T^* having the same mean value as T but a smaller variance than T, so use of the tabled critical values of T when ties occur results in a loss in power.

To express this reduction in variance algebraically we denote the number of distinct values among the n observations by the symbol k ($k \leq n$) and, after ordering these k values, let n_ν be the frequency of the ν^{th} value in the combined sample, with $a_\nu(b_\nu)$ denoting its frequency in the X-sample (Y-sample); $a_\nu + b_\nu = n_\nu$. We then have

---

Biometrics Unit, Department of Plant Breeding and Biometry, Cornell University.

$$(1) \qquad T^* = \sum_{\nu=1}^{k} \frac{a_\nu}{n_\nu} \sum_{n_1+\cdots+n_{\nu-1}+1}^{n_1+\cdots+n_\nu} f(i)$$

$$= \sum_{\nu=1}^{k} a_\nu \bar{f}_\nu$$

and

$$E(T^*) = \sum_{\nu=1}^{k} \frac{an_\nu}{n} \bar{f}_\nu = \frac{a}{n} \sum_{1}^{k} n_\nu \bar{f}_\nu = a\bar{f} = E(T) \ ;$$

in fact,

$$T^* = E(T | n_1, \cdots, n_k, a_1, \cdots, a_k) \ .$$

The variance of T under this null hypothesis is seen to be

$$\mathrm{var}(T) = \frac{ab}{n(n-1)} \sum_{i=1}^{n} [f(i) - \bar{f}]^2$$

while, for fixed $n_1, \cdots, n_k$,

$$(2) \qquad \mathrm{var}(T^*) = \frac{ab}{n(n-1)} \sum_{\nu=1}^{k} n_\nu (\bar{f}_\nu - \bar{f})^2 \ .$$

The reduction in variance is therefore

$$\mathrm{var}(T) - \mathrm{var}(T^*) = \frac{ab}{n(n-1)} \sum_{\nu=1}^{k} \sum_{i=n_1+\cdots+n_{\nu-1}+1}^{n_1+\cdots+n_\nu} [f(i) - \bar{f}_\nu]^2 \ .$$

For large samples the distribution of T^* is approximately normal, and hence $\mathrm{var}(T^*)$ may be employed to determine the critical region of the test. To achieve

the potential increase in power for small samples the critical values of T^* would need to be tabulated for each configuration of $n_1, \cdots, n_k$ -- an impossible task. It is noteworthy, however, that the approach to normality is rapid in most cases. For small samples it would appear that if T_c is a tabulated critical value of T for sample sizes $\underline{a}$ and $\underline{b}$ then a reasonable approximation to the corresponding critical value of T^* would be

$$T_c^* \doteq T_c \sqrt{\frac{\text{var}(T^*)}{\text{var}(T)}} \ .$$

Similar arguments apply in the multivariate case where the observations are ranked according to some index function and, in the p-variate case, the test statistic is

$$T_j = \sum_{i=1}^{n} \delta_i f_j(i) \quad \text{for } j = 1, \cdots, p \ .$$

As an example we consider the case where the observations lie on the perimeter of a circle -- representing, for example, the directions selected by two samples of animals in an orientation experiment. One rank test used in this situation is based on the circular order of the observations with

$$T_1 = \sum_{i=1}^{n} \delta_i \sin \frac{2\pi i}{n} \qquad T_2 = \sum_{i=1}^{n} \delta_i \cos \frac{2\pi i}{n}$$

and critical values are tabulated for

$$R^2 = T_1^2 + T_2^2 \ .$$

Since

$$E[T_1] = E[T_2] = 0$$

$$\operatorname{var}(T_1) = \operatorname{var}(T_2) = \frac{ab}{2(n-1)}$$

$$\operatorname{cov}(T_1, T_2) = 0$$

then for large samples the statistic $2(n-1)R^2/ab$ is approximately distributed as chi-square with 2 degrees of freedom. When ties occur and we define T_1^* and T_2^* as in (1) then the corresponding chi-square statistic becomes

$$\frac{[T_1^*]^2 \operatorname{var}(T_2^*) + [T_2^*]^2 \operatorname{var}(T_1^*) - 2T_1^* T_2^* \operatorname{cov}(T_1^*, T_2^*)}{\operatorname{var}(T_1^*) \operatorname{var}(T_2^*) - [\operatorname{cov}(T_1^*, T_2^*)]^2}$$

where the variances are given by (2) and the covariance is given by the corresponding formula

$$(3) \qquad \operatorname{cov}(T_1^*, T_2^*) = \frac{ab}{n(n-1)} \sum_{\nu=1}^{k} n_\nu (\bar{f}_{1\nu} - \bar{f}_1)(\bar{f}_{2\nu} - \bar{f}_2) .$$

Note that (2) and (3) simplify somewhat in this case where $\bar{f}_1 = \bar{f}_2 = 0$. Again, tabulation of critical values for small samples would be a formidable task. If R_c is a tabulated critical value of R then a reasonable approximation to the critical value of the above test statistic would be $2(n-1)R_c^2/ab$ rather than the corresponding critical value from a chi-square table.