TESTING THE EQUALITY OF TWO DISCRETE CONDITIONAL DISTRIBUTIONS

By

VISHNU DAYAL JHA

Punjabi University, Patiala

(Received : January, 1980)

SUMMARY

Wilcoxon Mann–Whitney test has been proposed for testing the equality of purely discrete distributions when the size of each sample is random. Using the mid-rank procedure, the above test has been discussed. Mean and variance of the test statistic (p known) have been computed. The case for unknown p has also been discussed.

1. INTRODUCTION

Let $p(x,y)=P(X=x, Y=y)$ be the joint probability mass function of two discrete variates X and Y with the distribution function

$$P(x,y)=\sum_{i\leqslant x}\sum_{j\leqslant y} p(i,j)$$

and marginals

$$p(x,.)=\sum_{i\leqslant x}\sum_{j} p(i,j) \text{ and } \sum_{i}\sum_{j\leqslant y} p(i,j).$$

Further assume that X takes only two Values 0 and 1 with respective probabilities q and p with $p+q=1$ and Y may take values $0,1,\ldots\ldots$ Our problem of interest is to test $H: P_o(y)=P_1(y)$ against the alternative $A: P_o(y)\neq P_1(y)$, where $P_i(y)=P(Y\leqslant y/X=_i)$ and $P_i(y)=P(Y=y/X=_i)$, the conditional probability mass function of Y given that $X=i$, $i=0, 1$, For notational simplicity we write $P_i(y)=P_i$.

Let $\{Z_i\}=\{(X_i,Y_i)\}$, $(i=1, 2,\ldots,N)$ be a sequence of N independent random variables from a bivariate discrete population $P(x,y)$. We divide the observations $Z_1, Z_2,\ldots Z_N$ into two groups according as the value of X is 1 or 0. Let $U_1, U_2, \ldots, U_n$ $(n>0)$ and $V_1, V_2,\ldots V_{N-n}$ $(N-n=m)$ are independent. Now the problem of testing H is equivalent to testing the hypothesis that the two independentsamples of random sizes are from the same discrete population.

When P_o and P_l are continuous, the probability of getting tied observations is 0 so that this event may be ignored. In the discontinuous case, however ties occur with positive probability and when they do occur, the pooled sample can no longer be uniquely ordered. For the case with a positive probability of ties two procedures have been proposed. One is to order the tied observations randomly, the other is to replace S_{mn} (the sum of ranks of X_i in the combined ordered sample) by

$$S'_{mn} = \sum_{i=1}^{n} R'_i$$

where R'_i = midrank $(X_i)=\frac{1}{2}[N_1(i)+N_2(i)+1]$, $N_1(i)$ is the number of observations (including X_i and $N_2(i)$ the number of observations (including X_i) not larger than X_i.

For the fixed sample sizes, the problem of testing the equality of two discrete distributions using Wilcoxon test [6] has been considered by many authors. Putter [7] has shown under certain regularity conditions that the asymptotic relative efficiency of the randomized test with respect to the midrank test is

$$1-\sum_{k=1}^{r} P_k^3$$

where $P_k=P(Y=\xi_k)$, $(k=1,2,\ldots r)$ and ξ_k are the common discontinuities of P_o and P_l. Buhler [1] has proved the truthness of Putter's argument for infinitely manytvalues of ξ_k. Chanda [2] has worked out the power efficiency of Wilcoxon test for the class of exponential type discrete distributions. Conover [3] developed the theory of rank tests without the assumptions concerning the continuous or discrete nature of the underlying distribution function. He has also discussed, three methods of assigning scores such as the average scores, midrank and randomized rank methods. Conover and Kemp[4] have computed the asymptotic efficiencies for the Wilcoxon test, the Vander Waerden test and the median test when the underlying distributions are Poisson, binomial, discrete uniform and negative binomial. They have also discussed and compared the three methods of handling ties.

For the small samples, the non randomized treatment of ties presents practical difficulties as have been discussed by Putter [7] but the asymptotic problem can be handed. In this paper we propose Wilcoxon Mann-Whitney test for testing the equality of purely

discrete distributions with the common infinitely many discontinuities as 0,1,2,... when the size of each sample is random. We are not supposed to look the small sample properties due to aforesaid difficulties. In section 2, using the midrank procedure we discuss the Wilcoxon Mann-Whitney test for random sample sizes and the mean and variance of the test statistic (p is known) have been computed in section 3. The case for unknown p is discussed in section 4.

Proposed Test Statistics. The test statistic may be defined as

$$U_N=\frac{1}{N(N-1)}\sum_{i\neq j=1}^{n} H(Z_i,Z_j) \tag{2.1}$$

where

$$H(Z_i,Z_j)=\begin{cases}1 & \text{if } X_i=1,\ X_j=0 \text{ and } Y_i<Y_j\\ 1/2 & \text{if } X_i=1,\ X_j=0 \text{ and } Y_i=Y_j\\ 0 & \text{Otherwise}\end{cases}$$

which can be put as

$$U_N=\frac{1}{N(N-1)}\sum_{j=1}^{N-n}\sum_{i=1}^{n}\phi(U_i,U_j) \tag{2.2}$$

where

$$\phi(u,v)=\begin{cases}1 & \text{if } u<v\\ 1/2 & u=v\\ 0 & \text{Otherwise}\end{cases}$$

Then $N(N\text{-}1)U_N$ is the total number of pairs (U_i,V_j) such that $U_i<V_j$: The hypotheses $H : P_o(y)=P_1(y)$ is rejected U_N is either too large or too small.

The statistic U_N is a real-valued symetric in the u's, symmetric in v's with expectation $E(U_N)$ and with finite second moment, and if $n<N\text{-}n$ and $n\to\infty$ such that $\lim\frac{n}{N\text{-}n}$ exists, then $n^{\frac{1}{2}}(U_N\text{-}EU_N)$ has a limiting normal distribution with mean 0 (Fraser (1957), *pp.* 231). Thus if $P_0(y)$ and $P_1(y)$ are both absolutely continuous, the Mann-Whitney test statistic for testing the hypothesis H: $P_o(y)=P_1(y)$ consists of rejecting H when $|U_N\text{-}E(U_N \mid H)|$ is too large, $E(U_N \mid H)$ being the expectation of U_N under H. When $P_0(y)$ and $P_1(y)$ are both pure discrete, a slight modification is necessary because the variance of test statistic under the hypothesis is not distribution-free as will be shown in the later section. So we will develop a new modified test statistic and will be shown asymptotically normally distributed with mean zero and variance unity when p in known.

2. MEAN AND VARIANCE OF THE TEST STATISTIC (p IS KNOWN)

$$B_p(U_N) = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} E\,H(Z_i,Z_j) = E\,H(Z_i,Z_j)$$

$$= p\,q \sum_{j>i=o}^{\infty} p_1(i)\,p_o(j) + \frac{1}{2}\,pq \sum_{i=o}^{\infty} p_1(i)\,p_o(i)$$
$$= \pi \text{ (say)}$$

Hence

$$Ep\,(U_N \mid H) = \frac{pq}{2}$$

To compute the variance we use the statistic (2.2)

Now

$$N^2\,(N\text{-}1)^2\,U_N^2 = \sum_{j=1}^{N-n} \sum_{i=1}^{n} \phi^2\,(u_i,v_j) + \sum_i \sum_{j \neq i} \sum_k \phi\,(u_i,v_k)\,\phi)\,(u_j,v_k)$$

$$+ \sum_i \sum_{j \neq k} \sum_k \phi\,(u_i,v)\,\phi\,u_i,v_k) + \sum_i \sum_{i \neq k, j \neq l} \sum_k \sum_l \phi\,(u_i,v_j)\,\phi\,(u_k,v_l)$$

Therefore

$$N^2(N-1)^2\,E\left(U_N^2 \mid n\right)$$

$$= n(N-n)\left[\frac{1}{4} \sum_{i=0}^{\infty} p_1(i)p_0(i) + \sum_{j>i=0}^{\infty} p_1(j)p_0(i)\right]$$

$$+ (N-n)n(n-1)\left[\frac{1}{4} \sum_{i=0}^{\infty} p_1^2(i)p_0(i) = \sum_{j>i=0}^{\infty} p_0(i)p_1(i)p_0(j)\right.$$

$$\left. + \sum_{i,j>k=0}^{\infty} p_0(i)p_0(j)p_1(k)\right]$$

$$+ n(N-n)(N-n-1)\left[\frac{1}{4} \sum_{i=0}^{\infty} p_0(i)\,p_1^2(i)\right.$$

$$\left. + \sum_{i>k,\,j=0} p_0(i)p_1(k)p_1(j) + \sum_{i>j=0} p_0(i)p_1(i)p_1(j)\right.$$

$$+n(n-1)(N-n)(N-n-1)\left[\frac{1}{4}\sum_{i=0} p_1(i)p_0(i) + \sum_{i>j=0} p_0(i)p_1(j)\right]$$

Since n is a binomial random variate $b(N, p)$. If we wrire

$$N^{(r)}=N(N-1)\ldots(N-r+1)$$

then

$$E[n^{(r)}\,(N-n)^{(s)}]=p^r\,q^s\,N^{(r+s)}$$

Let us put

$$A_1=\sum_{i=0}^{\infty} p_1(i)p_0(i)+\sum_{j>i=0}^{\infty} p_1(j)p_0(i)$$

$$A_2=\tfrac{1}{4}\sum_{i=0}^{\infty} p_1^2(i)p_0(i)+\sum_{j>i=0}^{\infty} p_0(i)p_1(i)p_0(j) + \sum_{i,j>k=0}^{\infty} p_0(i)p_0(j)p_1(k)$$

$$A_3=\frac{1}{4}\sum_{i=0}^{\infty} p_0(i)p_1^2(i)+\sum_{i>j=0}^{\infty} p_0(i)p_1(i)p_1(\mathrm{j}) + \sum_{i>j,k=0}^{\infty} p_0(i)p_1(j)p_1(k)$$

$$A_4=\tfrac{1}{4}\sum_{i=0}^{\infty} p_1(i)p_0(i)+\sum_{i>j=0}^{\infty} p_0(i)p_1(j)$$

Hence

$$N^2(N-1)^2\;E(U_N^2)=pq\,\mathrm{N}(N-1)\,A_1+p^2q\,N(N-1)(N-2)A_2 + pq^2N(N-1)(N-2)\,A_3+ p^2q^2\,N(N-1)\,(N-2)\,(N-3)A_4^2$$

$$\mathrm{Var}_p(U_N)=\frac{pq}{N(N-1)}A_1[+p(N-2)\,A_2+q(N-2)A_3 + pq(N-2)\,(N-3)\,A_4^2\,]-p^2q^2\,A_4^2$$

$$= \frac{pq}{N(N-1)}[A_1 + p(N-2)\,A_2 + q(N-2)\,A_3 - 2pq(2N-3)\,A_4^2]$$

$$= \frac{s^2(U_N)}{N(N-1)}$$

Hence by the central limit theorem, the variate $\sqrt{N(N-1)}\,(U_N-\mu)/s(U_N)$ approximates the standard normal distribution.

Here we are to note that $P_0=P_1$ *i.e.* $p_1(i)=p_0(i)=p(i)$ for all $i=0,1,2,\ldots$ $\mu_H=pq/2$ and

$$s_H^2\,(U_N)=\frac{pq}{12}\{(N-2)\,[4-\sum_{i=0}^{\infty} p^3(i)]+6-\sum_{i=0}^{\infty} p^2(i)-6pq(2N-3)\}$$

It follows therefore that the variance of U_N depends on the common unknown distribution function $P(y)$. Let us define $\binom{N}{3}$ R as the number of triplets (Y_i, Y_j, Y_k) $(i\neq j\neq k)$ such that $Y_i=Y_j=Y_k$. R can be written in the form

$$R=\frac{1}{M}\sum_{i=0}^{M} R_i$$

where $M=\binom{N}{3}$ and R_i is a random variable associated with the i-th triplet taking the value 1 if the Y's are all equal and O otherwise.

Now

$$E(R \mid n)=P(Y_1=Y_2=Y_3,\ X_1,\ X_2,\ X_3,\ X\varepsilon P_0)+P(Y_1=Y_2=Y_3,\ X_1,\ X_2,\ X_3,\ X\varepsilon p_1)+3P(Y_1=Y_2=Y_3,\ X_1\ X_2\varepsilon P_0,\ X_3\varepsilon P_1)+3P(Y_1=Y_2=Y_3,\ X_1\varepsilon P_0,\ X_2, X_3\varepsilon P_1)$$

$$= \frac{n(n-1)\,(n-2)}{N(N-1)\,(N-2)}\sum_{i=0}^{\infty} p_0^3(i)$$

$$+ \frac{(N-n)\,(N-n-1)(N-n-2)}{N(N-1)\,(N-2)}\sum_{i=0}^{\infty} p_1^3(i)$$

$$+ \frac{3n(n-1)\,(N-n)}{N(N-1)\,(N-2)}\sum_{i=0}^{\infty} p_0^2(i)p_1(i)$$

$$+ \frac{3n(N-n)\,(N-n-1)}{N(N-1)\,(N-2)}\sum_{i=0}^{\infty} p_0(i)p_1^2(i)$$

$$E(R)=En(E(R \mid n))=p^3 \sum_{i=1}^{\infty} p_0^3(i)+ q^3 \sum_{i=0}^{\infty} p_1^3(i)$$

$$+ 3p^2q \sum_{i=1}^{\infty} p_0^2(i)p_1(i) + 3pq^2 \sum_{i=0}^{\infty} p_1^2(i)p_0(i)$$

$$= \sum_{i=0}^{\infty} \{p\, p_0(i)+q\, p_1(i)\}^3$$

Under the null hypothesis $H: P_0(i)=p_1(i)=p(i)$, we have

$$E(R) = \sum_{i=0}^{\infty} p^3(i)$$

Hence R is an unbiased estimate of $\sum_{i=0}^{\infty} p^3(i)$.

Similarly we define $\binom{N}{2} S$ as the number of pairs (Y_i, Y_j) $(i \neq j)$ such that $Y_i=Y_j$ and it can easily be shown that under H, S is an unbiased estimator of $\sum_{i=0}^{\infty} p^2(i)$. So under H

$$U_N'= \frac{\sqrt{N(N-1)}\ (U_N-pq)}{s_1(U_N)} \xrightarrow{\alpha} N(0,1)$$

where

$$s_1^2(U_N) = \frac{pq}{12} \{(N-2)(4-R)+6-3S-6pq(2N-3)\}$$

Hence for given α, we can find C_α such that

$$P(\mid U_N' \mid > C_\alpha)=\alpha$$

Case when p is Unknown

$\hat{p}=\frac{n}{N}$ as an usual estimate can be used and consider the test based on the statistic $(U_N-E_{\hat{p}}(U_N))/s_{\hat{p}}(U_N)$ where $E_{\hat{p}}(U_N)$ and $s_{\hat{p}}(U_N)$ are obtained by replacing p and q by $\hat{p}$ and $\hat{q}$ respectively. It can readily be shown that asymptotic distribution of this modified test statistic is normal with mean 0 and finite variance.

ACKNOWLEDGEMENT

I am very grateful to the referees for helpful suggestions and comments.

REFERENCES

[1] Buhler, W.J. (1967) : 'The Treatment of Ties in the Wilcoxon test.' *Ann. Math. Statist.* 38, 519-22.

[2] Chanda, K.C. (1963) : 'On the Efficiency of Two Sample Mann-Whitney Test for Discrete Populations.' *Ann. Math. Statist.* 34, 612-17

[3] Conover, W.J. (1973) : 'Rank Tests for One Sample, Two Samples and k Samples without the Assumption of a Continuous Distribution Function.' *Ann. Statist.* 1, 1105-25.

[4] Conover, W.J. and Kemp, K.E. (1976) : 'Comparisons of the Asymptotic Efficiencies of Two Sample Tests for Discrete Distributions'. Comm. Statist.—*Theory and Methods.* A5(1), 1-15.

[5] Fraser, D.A.S. (1957) : *Nonparametric Methods in Statistics.* John Wiley and Sons. Inc. New York.

[6] Mann, H.B. and Whitney, D.R, (1947) : 'On a Test whether on of the Two Variables is Stochastically larger than the Other'. *IBID*, 18, 50-60.

[7] Putter, J. (1955) : 'The Treatment of Ties in some Nonparametric Tests.' *Ann. Math. Statist.* 26, 368-86.