THE EGYPTIAN STATISTICAL JOURNAL
RSSR, CAIRO UNIV., VOL., 45, NO.1, 2001

The Possession of the Markov Chain Property by the Wilcoxon-Mann-Whitney Statistic

Ali S. Barakat

Department of Statistics, An-Najah National University, Nablus, West Bank, Palestine. Fax. +(972)9-2387982, E-mail: barakat@najah.edu

Abstract. The Wilcoxon-Mann-Whitney statistic is rewritten using nearest neighbors techniques and a characterization in terms of a Markov chain is established here for the first time.

Keywords. Nearest neighbors, Wilcoxon, Markov chain

1. Introduction

Let $X_1, X_2, \ldots\ldots, X_n$ and $Y_1, Y_2, \ldots\ldots, Y_m$ be two independent random samples of observations in the Euclidean space R^d from unknown distribution function F_i, $i = 1,2$, respectively.

The two samples are combined into a single sample $Z_1, Z_2, \ldots, Z_N$ of size N=n+m, , such that

$$Z_i = \begin{cases} X_i & , \quad i = 1,2,\ldots,n \\ Y_{i-n} & , \quad i = n+1, n+2, \ldots\ldots, N \end{cases}$$

Let $R_j(X_i)$ = rank of observation X_i with respect to distance from X_j. Then we define X_i as the k^{th} nearest neighbor of X_j if $R_j(X_i) = k$ and as a k-nearest neighbor if $R_j(X_i) \le k$. Assume that there will be no ties.

Let $I(i,j) = I\{Z_i$ and its j^{th} nearest neighbor are from different samples$\}$ where $I\{E\}$ is the indicator function of the event E.

For i = 1,2,....,N and k=1,2,.....,N-1, define

$$B_i = \sum_{k=1}^{N-1} B_{i,k} = \sum_{k=1}^{N-1} \sum_{j=1}^{k} I(i,j)$$

The object of the present investigation is to rewrite B_i as a linear function of the Wilcoxon-Mann-Whitney statistic and to characterize a Markovian structure for the Wilcoxon test.

2. A representation for B_i

Choose the first variable to be Z_i (fixed), $i = 1,2,\ldots,N$ and calculate $\|Z_j - Z_i\|$, $j = 1,2,\ldots,N, j \neq i$. The combined ordered arrangement of the two samples can be denoted by a vector of indicator random variables $Z_{i,k}$, where $Z_{i,k} = 1$ if the point Z_i and its k^{th} nearest neighbor, Z_j, belong to different samples and $Z_{i,k} = 0$ if both points belong to the same sample, $k=1,2,\ldots,N-1$. The rank of the observation for which $Z_{i,k}$ is an indicator is k, and therefore the vector Z_i indicates the rank-order statistic of the combined ordered arrangement of the two samples and in addition identifies the sample to which each observation belongs. B_i can be expressed in terms of this notation. This kind of statistic is called a linear rank statistic which is defined as

$$B_{N-1}(Z_i) = \sum_{k=1}^{N-1} a_k Z_{i,k}$$

where a_k are given numbers.

Lemma

For each $i(1 \leq i \leq N)$, a linear relationship exists between the statistic B_i and the Wilcoxon-Mann-Whitney statistic.

Proof:

$$B_{i,k} = \sum_{j=1}^{k} I(i,j) = \sum_{j=1}^{k} Z_{i,j}$$

$$B_i = \sum_{k=1}^{N-1} \sum_{j=1}^{k} Z_{i,j}$$

If we take $a_j = (N-j)$ then $B_{N-1}(Z_i)$ can be written in terms of $B_{i,k}$ as follows:

$$B_{N-1}(Z_i) = \sum_{k=1}^{N-1} (N-k) Z_{i,j}$$

$$= \sum_{k=1}^{N-1}\sum_{j=1}^{k} Z_{i,j}$$

$$= \sum_{k=1}^{N-1} B_{i,k}$$

$$= B_i$$

Therefore,

$$B_i = \sum_{j=1}^{N-1}(N-j)Z_{i,j}$$

$$= \begin{cases} mN - W_i & , \quad i = 1,2,....,n \\ nN - W_i & , \quad i = n+1, N+2,....,N \end{cases}$$

where

$W_i = \sum_{j=1}^{N-1} jZ_{i,j}$ is the wilcoxon rank sum statistic.

So, B_i has a Wilcoxon-Mann-Whitney distribution.

Thus B_i is actually the same as the Wilcoxon-Mann-Whitney rank sum test, since a linear relationship exists between the two test statistics. Therefore, all the properties of the tests are the same. One of the important properties is a Markovian property which is proved in the following section.

3. A Markovian Property of the $B_{i,k}$

The main result of this paper is the following

Theorem

For every i ($1 \leq i \leq n$), the sequence $\{B_{i,k};\ k = 1,2,.....,N-1\}$ is a Markov chain, i.e., for every k ($\leq N-1$) and max $(0,k-n+1) \leq r_1 \leq r_2 \leq \ldots \leq r_k \leq r_{k+1} \leq \min(k+1,m)$

$P(B_{i,k+1} = r_{k+1} \mid B_{i,j} = r_j;\ j \leq k) = P(B_{i,k+1} = r_{k+1} \mid B_{i,k} = r_k)$

Proof:

Let S be the set of all permutations of $(\{1,2,\ldots,N\}-\{i\})$ satisfying the condition $\{B_{i,1}=r_1, B_{i,2}=r_2,\ldots, B_{i,k}=r_k\}$. It is clear that for any $s\in S$, $B_{i,k+1}$ can assume only the values r_k and r_k+1. If $\{a_1,a_2,\ldots,a_N\}\in S$, then in the set $\{a_1,a_2,\ldots,a_k\}$, we have r_k elements of the set $\{n+1, n+2,\ldots,N\}$ and $(k-r_k)$ elements of the set $\{1,2,\ldots,n\}$. Then we may have either of the following:

أ. $k+1\in\{n+1, n+2, \ldots, N\}$. This happens with the (conditional) probability

$$\frac{m-r_k}{N-1-k}$$

ب. $k+1\notin\{n+1, n+2,\ldots,N\}$. This happens with (conditional) probability

$$1-\frac{m-r_k}{N-1-k}=\frac{n-1-k+r_k}{N-1-k}$$

In case (i), B_{ik+1} can assume only the value r_k+1 with probability $\frac{m-r_k}{N-1-k}$, while in case (ii), $B_{i,k+1}$ can assume only the value r_k with probability $\frac{n-1-k+r_k}{N-1-k}$.

Thus, the assumable values of $B_{i,k+1}$(viz. r_k, r_k+1) and their respective (conditional) probabilities (given the $B_{i,j}$, $j\le k$) depend only on the values r_k assumed by $B_{i,k}$.

Corollary:

For every k $(1\le k\le N-1)$ and r_k, we have

$$p(B_{i,k}=r_k)=\binom{N-1}{k}^{-1}\binom{m}{r_k}\binom{n-1}{k-r_k} \qquad \ldots.(1)$$

for $\max(0,k-n+1)\le r_k\le\min(k,m)$, $i=1,2,\ldots,n$,
and for every $\ell>k$, $r_\ell\ge r_k$

$$p(B_{i,k}=r_k, B_{i,\ell}=r_\ell)=\frac{\binom{m}{r_k}\binom{n-1}{k-r_k}\binom{m-r_k}{r_\ell-r_k}\binom{n-1-k+r_k}{\ell-k-r_\ell+r_k}}{(N-1)!\{(k)!(\ell-k)!(N-1-\ell)!\}^{-1}} \qquad \ldots.(2)$$

for max(0,k-n+1) $\le r_k \le r_\ell \le$ min (ℓ ,m); i = 1,2, ……, n; and $1 \le k \le \ell \le N-1$.

By (1) and (2), we may note that

$$p(B_{i,k+1} = s / B_{i,k} = r) = \binom{N-2-k}{m-s}\binom{N-1-k}{m-r}^{-1} \quad \dots(3)$$

for s ≥ r (and 0 for s < r), so that

$$p(B_{i,k+1} = S / B_{i,k} = r) = \begin{cases} \dfrac{m-r}{N-1-k} & , \quad s = r+1 \\ \dfrac{n-1-k+r}{N-1-k} & , \quad s = r \\ 0 & , \quad s \ge r+2 \, or \, s < r. \end{cases} \quad \dots(4)$$

Hence, from (4) we have

$$E(B_{i,k+1} / B_{i,k}) = \frac{(N-k-2)}{(N-1-k)} B_{i,k} + \frac{m}{(N-1-k)} \quad \dots(5)$$

for k = 1,2,….., N-2.

Also,

$$E(B^2_{i,k+1} / B_{i,k}) = \frac{(N-3-k)}{(N-1-k)} B^2_{i,k} + \frac{2m-1}{(N-1-k)} B_{i,k} + \frac{m}{(N-1-k)} \quad \dots(6)$$

Conclusion:

We conclude that the Wilicoxon-Mann-Whitney test can be characterized by a Markovian property. Furthermore, such Markovian property is essential for a new proof of the normality of the Wilcoxon-Mann-Whitney test via Martingale limit theorem.

Note: For $i = n+1, \ldots, N$, the same resents are obtained but n and m are interchanged.

References:

Bailey, N.T.J. (1964) The Elements of Stochastic Processes. Wiely, New York.

Barakat, A.S., Quade, D., and Salama, I.A. (1996) "Multivariate homogeneity testing using an extended concept of nearest neighbors. "Biometrical Journal, 38, 5, 605-612.

Bishop, Y.M.N., Fienberg, W.E., and Holland, P.W. (1976) Discrete Multivariate Analaysis: Theory and Practice. MIT Press, Massachusetts.

Gibbons, J.D. (1985) Nonparametric statistical inference. Marcel Dekker, New York.

Sen, P.K. and Salama, I.A. (1983) "The Spearman footrule and a Markov chain property". Statistics and Probability Letters, 1, 285-289.