Mann-Whitney Test for Associated Sequences

Isha Dewan
and
B. L. S. Prakasa Rao

April 19, 2002
isid/ms/2002/06

Indian Statistical Institute, Delhi Centre
7, SJSS Marg, New Delhi–110 016, India

MANN-WHITNEY TEST FOR ASSOCIATED SEQUENCES

Isha Dewan and B.L.S. Prakasa Rao
Indian Statistical Institute
New Delhi
India

Abstract

Let $\{X_1, \ldots, X_m\}$ and $\{Y_1, \ldots, Y_n\}$ be two samples independent of each other, but the random variables within each sample are stationary associated with one dimensional marginal distribution functions F and G, respectively. We study the properties of the classical Wilcoxon-Mann-Whitney statistic for testing for stochastic dominance in the above set up.

Key words : U-statistics, Mann-Whitney statistic, Central Limit Theorem, Associated random variables .

1 INTRODUCTION

Suppose that two samples $\{X_1, \ldots, X_m\}$ and $\{Y_1, \ldots, Y_n\}$ are independent of each other, but the random variables within each sample are stationary associated with one dimensional marginal distribution functions F and G respectively. Assume that the density functions f and g of F and G respectively , exist . We wish to test for the equality of the two marginal distribution functions F and G. A commonly used statistic for this nonparametric testing problem is the Wilcoxon Mann-Whitney statistic when the observations $X_i, 1 \leq i \leq m$ are independent and identically distributed (i.i.d.) and $Y_j, 1 \leq j \leq n$ are i.i.d. However, most often the X and the Y observations are not i.i.d. Suppose the samples are from a stationary associated stochastic process.

A finite family $\{X_1, ..., X_n\}$ of random variables is said to be *associated* if

$$\mathrm{Cov}(h_1(X_1, ..., X_n), h_2(X_1, ..., X_n)) \geq 0$$

for any coordinatewise nondecreasing functions h_1, h_2 on R^n such that the covariance exists. An infinite family of random variables is said to be *associated* if every finite subfamily is associated. (cf. Esary, Proschan and Walkup (1967)).

We wish to test the hypothesis that

$$H_0 : F(x) = G(x) \text{ for all } x, \tag{1.1}$$

against the alternative

$$H_1 : F(x) \geq G(x) \text{ for all } x, \tag{1.2}$$

with strict inequality for some x. We can test the above hypothesis conservatively by testing

$$H_0' : \gamma = 0, \tag{1.3}$$

against the alternative

$$H_1' : \gamma > 0, \tag{1.4}$$

where $\gamma = 2P(Y > X) - 1 = P(Y > X) - P(Y < X)$.

Probabilisic aspects of associated random variables have been extensively studied (see, for example, Prakasa Rao and Dewan (2001) and Roussas(1999)). Here we extend the Wilcoxon - Mann - Whitney statistic to stationary sequences of associated variables. Serfling (1980) studied the Wilcoxon statistic when the samples are from stationary mixing processes. Louhichi (2000) gave an example of a sequence of random variables which is associated but not mixing. This shows that tests for samples from stationary associated random sequences need to be studied separately.

In section 2 we state some results that are used to study the properties of Wilcoxon statistic for associated random variables. In section 3 we discuss the asymptotic normality of the Wilcoxon statistic based on independent sequences of stationary associated variables.

2 Preliminaries

We state some theorems that are used in proving the main results in the next section.

Theorem 2.1 : (Bagai and Prakasa Rao(1991)). Suppose X and Y are associated random variables with bounded continuous densities f_X and f_Y, respectively. Then there exists an absolute constant $C > 0$ such that

$$\begin{aligned}&\sup_{x,y} |P[X \leq x, Y \leq y] - P[X \leq x]P[Y \leq y]| \\ &\leq \quad C\{\max(\sup_x f_X(x), \sup_x f_Y(x))\}^{2/3}(\text{Cov}(X,Y))^{1/3}.\end{aligned} \tag{2.1}$$

The following Theorem gives the asymptotic normality of a sequence of associated variables.

Theorem 2.2 : (Newman (1980, 1984)). Let $\{X_n, n \geq 1\}$ be a stationary associated sequence of random variables with $E[X_1^2] < \infty$ and $0 < \sigma^2 = V(X_1) + 2\sum_{j=2}^{\infty}\text{Cov}(X_1, X_j) < \infty$. Then, $n^{-1/2}(S_n - E(S_n)) \xrightarrow{\mathcal{L}} N(0, \sigma^2)$ as $n \to \infty$.

Assume that

$$\sup_x f(x) < c \qquad \sup_x g(x) < c. \tag{2.2}$$

Further assume that

$$\sum_{j=2}^{\infty} \text{Cov}^{\frac{1}{3}}(X_1, X_j) < \infty, \tag{2.3}$$

and

$$\sum_{j=2}^{\infty} \text{Cov}^{\frac{1}{3}}(Y_1, Y_j) < \infty. \tag{2.4}$$

This would imply

$$\sum_{j=2}^{\infty} \text{Cov}(X_1, X_j) < \infty, \tag{2.5}$$

and

$$\sum_{j=2}^{\infty} \text{Cov}(Y_1, Y_j) < \infty. \tag{2.6}$$

Theorem 2.3 : (Peligard and Suresh (1995)). Let $\{X_n, n \geq 1\}$ be a stationary associated sequence of random variables with $E(X_1) = \mu,\ E(X_1^2) < \infty$. Let $\{\ell_n, n \geq 1\}$ be a sequence of positive integers with $1 \leq \ell_n \leq n$. Let $S_j(k) = \sum_{i=j+1}^{j+k} X_i,\ \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Let $\ell_n = o(n)$ as $n \to \infty$.Assume that (2.5) holds. Then, with $\ell = \ell_n$

$$\begin{aligned}
B_n &= \frac{1}{n-\ell}\left(\sum_{j=0}^{n-\ell} \frac{|S_j(\ell) - \ell\bar{X}_n|}{\sqrt{\ell}}\right) \\
&\to (\mathrm{Var}(X_1) + 2\sum_{i=2}^{\infty} \mathrm{Cov}(X_1, X_i))\sqrt{\frac{2}{\pi}} \quad \text{in } L_2 - \text{mean as } n \to \infty. \qquad (2.7)
\end{aligned}$$

In addition assume that $\ell_n = O(n/(logn)^2)$ as $n \to \infty$, the convergence above holds in the almost sure sense.

Theorem 2.4 : (Roussas (1993)). Let $\{X_n, n \geq 1\}$ be a stationary associated sequence of random variables with bounded one dimensional probability density function. Suppose

$$\begin{aligned}
u(n) &= 2\sum_{j=n+1}^{\infty} \mathrm{Cov}(X_1, X_j) \\
&= O(n^{-(s-2)/2}) \ \text{ for some } \ s > 2. \qquad (2.8)
\end{aligned}$$

Let ψ_n be any positive norming factor. Then, for any bounded interval $I_M = [-M, M]$, we have

$$\sup_{x \in I_M} \psi_n |F_n(x) - F(x)| \to 0, \qquad (2.9)$$

almost surely as $n \to \infty$, provided

$$\sum_{n=1}^{\infty} n^{-s/2} \psi_n^{s+2} < \infty. \qquad (2.10)$$

3 Wilcoxon Statistic

The Wilcoxon two-sample statistic is the U-statistic given by

$$U = \frac{1}{mn} \sum_{i=1}^{m}\sum_{j=1}^{n} \phi(Y_j - X_i), \qquad (3.1)$$

where

$$\phi(u) = \begin{cases} 1 & \text{if } u > 0, \\ 0 & \text{if } u = 0, \\ -1 & \text{if } u < 0. \end{cases}$$

Note that ϕ is a kernel of degree $(1,1)$ with $E\phi(Y - X) = \gamma$. We now obtain the limiting distribution of the statistic U under some conditions.

Theorem 3.1: Le $\{X_i, i \geq 1\}$ and $\{Y_j, j \geq 1\}$ be independent sequences of random variables with one dimensional distribution functions F and G, respectively, such that each sequence is stationary associated satisfying conditions (2.3) to (2.6). Then , as $m, n \rightarrow \infty$ such that $\frac{m}{n} \rightarrow c \in (0, \infty)$, we have

$$\sqrt{m}(U - \gamma) \xrightarrow{\mathcal{L}} N(0, A^2) \text{ as } n \rightarrow \infty,$$

where A^2 is as given by (3.19). If $F = G$, then

$$\begin{aligned} \sigma_X^2 &= \sigma_Y^2 \\ &= 4(\frac{1}{12} + 2\sum_{j=2}^{\infty} \text{Cov}(F(X_1), F(X_j))), \end{aligned} \tag{3.2}$$

so that

$$A^2 = 4(1+c)(\frac{1}{12} + 2\sum_{j=2}^{\infty} \text{Cov}(F(X_1), F(X_j))). \tag{3.3}$$

Proof: Following Hoeffding's decomposition (Lee (1980)), we can write U as

$$U = \gamma + H_{m,n}^{(1,0)} + H_{m,n}^{(0,1)} + H_{m,n}^{(1,1)}, \tag{3.4}$$

where

$$H_{m,n}^{(1,0)} = \frac{1}{m}\sum_{i=1}^{m} h^{(1,0)}(X_i),$$

$$h^{(1,0)}(x) = \phi_{10}(x) - \gamma,\ \phi_{10}(x) = 1 - 2G(x),$$

$$H_{m,n}^{(0,1)} = \frac{1}{n}\sum_{j=1}^{n} h^{(0,1)}(Y_j),$$

$$h^{(0,1)}(y) = \phi_{01}(y) - \gamma,\ \ \phi_{01}(y) = 2F(y) - 1,$$

and

$$H_{m,n}^{(1,1)} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} h^{(1,1)}(X_i, Y_j),$$

where

$$h^{(1,1)}(x, y) = \phi(x - y) - \phi_{10}(x) - \phi_{01}(y) + \gamma.$$

It is easy to see that

$$\begin{aligned} E(\phi_{10}(X)) &= \gamma, \\ E(\phi_{10}^2(X)) &= 4\int_{-\infty}^{\infty} G^2(x)dF(x) - 4\int_{-\infty}^{\infty} G(x)dF(x) + 1, \end{aligned}$$

and

$$\text{Cov}(\phi_{10}(X_i), \phi_{01}(X_j)) = 4\ \text{Cov}(G(X_i), G(X_j)). \tag{3.5}$$

Since the random variables $X_1, \ldots, X_m$ are associated , so are $\phi_{10}(X_1), \ldots, \phi_{10}(X_m)$ since ϕ is monotone (see, Esary, Proschan and Walkup (1967)). Furthermore conditions (2.2), (2.5) and (2.6) imply that

$$\sum_{j=2}^{\infty} \mathrm{Cov}(G(X_1), G(X_j)) < \infty,$$

and

$$\sum_{j=2}^{\infty} \mathrm{Cov}(F(Y_1), F(Y_j) < \infty,$$

since

$$|\mathrm{Cov}(G(X_1), G(X_j))| < (\sup_x g)\mathrm{Cov}(X_1, X_j),$$

and

$$|\mathrm{Cov}(F(Y_1), F(Y_j))| < (\sup_x f)\mathrm{Cov}(Y_1, Y_j),$$

by Newman's inequality (1980). Following Newman (1980,1984), we get that

$$m^{-1/2} \sum_{i=1}^{m} (\phi_{10}(X_i) - \gamma) \xrightarrow{\mathcal{L}} N(0, \sigma_X^2) \text{ as } n \to \infty, \tag{3.6}$$

where

$$\sigma_X^2 = 4\int_{-\infty}^{\infty} G^2(x)dF(x) - 4\int_{-\infty}^{\infty} G(x)dF(x) + 1 + 8\sum_{j=2}^{\infty} \mathrm{Cov}(G(X_1), G(X_j)). \tag{3.7}$$

Similarly, we see that

$$n^{-1/2} \sum_{j=1}^{n} (\phi_{01}(Y_j) - \gamma) \xrightarrow{\mathcal{L}} N(0, \sigma_Y^2) \text{ as } n \to \infty, \tag{3.8}$$

where

$$\sigma_Y^2 = 4\int_{-\infty}^{\infty} F^2(x)dG(x) - 4\int_{-\infty}^{\infty} F(x)dG(x) + 1 + 8\sum_{j=2}^{\infty} \mathrm{Cov}(F(Y_i), F(Y_j)). \tag{3.9}$$

Note that $E(H_{m,n}^{(1,1)}) = 0$. Consider

$$\begin{aligned} \mathrm{Var}(H_{m,n}^{(1,1)}) &= E(H_{m,n}^{(1,1)})^2 \\ &= \frac{\Delta}{m^2 n^2}, \end{aligned} \tag{3.10}$$

where

$$\Delta = \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{i'=1}^{m}\sum_{j'=1}^{n} \Delta(i, j; i', j'), \tag{3.11}$$

and

$$\Delta(i, j; i', j') = \mathrm{Cov}(h^{(1,1)}(X_i, Y_j), h^{(1,1)}(X_{i'}, Y_{j'})). \tag{3.12}$$

Following Serfling (1980),

$$\begin{aligned}\Delta(i,j;i',j') &= 4(E(F_{i,i'}(Y_j,Y_{j'})-F(Y_j)F(Y_{j'}))\\ &\quad -\mathrm{Cov}(G(X_i,X_{i'})))\\ &= 4(E(G_{j,j'}(X_i,X_{i'})-G(X_i)G(X_{i'}))\\ &\quad -\mathrm{Cov}(F(Y_j,Y_{j'})), \end{aligned} \tag{3.13}$$

where $F_{i,i'}$ is the joint distribution function of $(X_i, X_{i'})$ and $G_{j,j'}$ is the joint distribution function of $(Y_j, Y_{j'})$.

Then, by Theorem 2.1, there exists a constant $C > 0$ such that

$$\begin{aligned}\Delta(i,j;i',j') &\leq C[\mathrm{Cov}^{\frac{1}{3}}(X_i,X_{i'})+\mathrm{Cov}(X_i,X_{i'})]\\ &= r_1(|i-i'|)\ \text{(say)}, \end{aligned} \tag{3.14}$$

by stationarity and

$$\begin{aligned}\Delta(i,j;i',j') &\leq C[\mathrm{Cov}^{\frac{1}{3}}(Y_j,Y_{j'})+\mathrm{Cov}(Y_j,Y_{j'})]\\ &= r_2(|j-j'|)\ \text{(say)}, \end{aligned} \tag{3.15}$$

by stationarity. Note that

$$\sum_{k=1}^{\infty} r_1(k) < \infty, \qquad \sum_{k=1}^{\infty} r_2(k) < \infty. \tag{3.16}$$

by (2.3) - (2.6). Then, following Serfling (1980), we have

$$\Delta = o(mn^2) \tag{3.17}$$

as m and $n \to \infty$ such that $\frac{m}{n}$ has a limit $c \in (0, \infty)$.

Hence, from (3.4), we have

$$\begin{aligned}\sqrt{m}(U-\gamma) &= \sqrt{m}\frac{1}{m}\sum_{i=1}^{m} h^{(1,0)}(X_i)+\sqrt{\frac{m}{n}}\frac{1}{\sqrt{n}}\sum_{j=1}^{n} h^{(0,1)}(Y_j)+\sqrt{m}H_{m,n}^{(1,1)}\\ &\xrightarrow{\mathcal{L}} N(0,A^2), \end{aligned} \tag{3.18}$$

where

$$A^2 = \sigma_X^2 + c\sigma_Y^2, \tag{3.19}$$

since $E(H_{m,n}^{(1,1)}) = 0$ and $\mathrm{Var}(\sqrt{m}H_{m,n}^{(1,1)}) \to 0$ as $m, n \to \infty$ such that $\frac{m}{n} \to c \in (0, \infty)$. This completes the proof of the theorem.

Estimation of the limiting variance

Note that the limiting variance A^2 depends on the unknown distribution F even under the null hypothesis. We need to estimate it so that the proposed test statistic can be used for testing purposes. The unknown variance A^2 can be estimated using the estimators given by Peligard and Suresh (1995). We now give a consistent estimator of the unknown variance A^2 under some conditions.

Let $N = m+n$. Under the hypothesis $F = G$, the random variables $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ are associated with the one-dimensional marginal distribution function F. Denote $Y_1, \ldots, Y_n$ as $X_{m+1}, \ldots, X_N$. Then $X_1, \ldots, X_N$ are associated as independent sets of associated random variables are associated (cf. Esary, Proschan and Walkup (1967)).

Let $\{\ell_N, N \geq 1\}$ be a sequence of positive integers with $1 \leq \ell_N \leq N$. Let $S_j(k) = \sum_{i=j+1}^{j+k} \phi_{10}(X_i), \bar{\phi_N} = \frac{1}{N}\sum_{i=1}^{N} \phi_{10}(X_i)$. Define $\ell = \ell_N$ and

$$B_N = \frac{1}{N-\ell}[\sum_{j=0}^{N-\ell} \frac{|S_j(\ell) - \ell\bar{\phi_N}|}{\sqrt{\ell}}]. \tag{3.20}$$

Note that B_N depends on the unknown function F. Let $\hat{\phi}_{10}(x) = 1-2F_N(x)$ where F_N is the empirical distribution function corresponding to F based on the associated random variables $X_1, \ldots, X_N$. Let $\hat{S}_j(k)$, $\hat{\bar{\phi_N}}$ and $\hat{B_N}$ be expressions analogous to $S_j(k)$, $\bar{\phi}_N$ and B_N with ϕ_{10} replaced by $\hat{\phi}_{10}$. Let $Z_i = \phi_{10}(X_i) - \hat{\phi}_{10}(X_i)$. Then

$$\begin{aligned}
&|B_N - \hat{B}_N| \\
&\quad = |\frac{1}{N-\ell}\sum_{j=0}^{N-\ell} \frac{|S_j(\ell) - \ell\bar{\phi}|}{\sqrt{\ell}} - \frac{1}{N-\ell}\sum_{j=0}^{N-\ell} \frac{|\hat{S}_j(\ell) - \ell\hat{\bar{\phi}}|}{\sqrt{\ell}}| \\
&\quad \leq \frac{1}{(N-\ell)\sqrt{\ell}} \sum_{j=0}^{N-\ell} |S_j(\ell) - \hat{S}_j(\ell) - \ell(\bar{\phi} - \hat{\bar{\phi}})| \\
&\quad = \frac{1}{(N-\ell)\sqrt{\ell}} \sum_{j=0}^{N-\ell} |\sum_{i=j+1}^{j+\ell} Z_i - \ell\frac{1}{N}\sum_{i=1}^{N} Z_i| \\
&\quad \leq \frac{1}{(N-\ell)\sqrt{\ell}} \sum_{j=0}^{N-\ell} \{\sum_{i=j+1}^{j+\ell} |Z_i| + \ell\frac{1}{N}\sum_{i=1}^{N} |Z_i|\}.
\end{aligned} \tag{3.21}$$

Note that

$$|Z_i| = 2|F_N(X_i) - F(X_i)|.$$

Suppose that the density function corresponding to F has a bounded support. Then, for sufficiently large $M > 0$, with probability 1,

$$\begin{aligned}
\sup_{x \in R} |F_N(x) - F(x)| &= \max\{\sup_{x \in [-M,M]} |F_N(x) - F(x)|, \sup_{x \in [-M,M]^c} |F_N(x) - F(x)|\} \\
&= \sup_{x \in [-M,M]} |F_N(x) - F(x)|.
\end{aligned} \tag{3.22}$$

Hence, from (3.21) and Theorem 2.4 we get

$$\begin{aligned} |B_N - \hat{B}_N| &\leq \frac{2}{(N-\ell)\sqrt{\ell}} (N-\ell)\,\ell \sup_x |F_N(x) - F(x)| \\ &= 2\sqrt{\ell}\ \psi_N^{-1} \sup_x\ \psi_N\ |F_N(x) - F(x)| \\ &\rightarrow 0 \quad \text{as} \quad N \rightarrow \infty \end{aligned} \tag{3.23}$$

provided $\sqrt{\ell}\ \ \psi_N^{-1} = O(1)$ or $\ell_N = O(\psi_N^2)$. Therefore we get,

$$|B_N - \hat{B}_N| \rightarrow 0 \quad \text{a.s. as} \quad n \rightarrow \infty. \tag{3.24}$$

Hence, from Theorem 2.3,

$$\frac{\pi}{2}\hat{B}_N^2 \rightarrow 4(\frac{1}{12} + 2\sum_{j=2}^{\infty} \mathrm{Cov}(F(X_1), F(X_j))) \tag{3.25}$$

as $n \rightarrow \infty$. Define $J_N^2 = (1+c)\frac{\pi}{2}\hat{B}_N^2$.

Then,
$\frac{\sqrt{N}(U-\gamma)}{J_N} \xrightarrow{\mathcal{L}} N(0,1)$ as $m, n \rightarrow \infty$ such that $\frac{m}{n} \rightarrow c \in (0,\infty)$; as$n \rightarrow \infty$.

Hence the statistic $\frac{\sqrt{N}(U-\gamma)}{J_N}$ can be used as a test statistic for testing $H_0' : \gamma = 0$ against $H_1' = \gamma > 0$.

On the other hand, by using Newman's inequality, one could obtain an upper bound on A^2 given by

$$4(1+c)(\frac{1}{12} + 2\sum_{j=2}^{\infty} \mathrm{Cov}(X_1, X_j)) \tag{3.26}$$

and we can have conservative tests and estimates of power based on (3.27).

Acknowledgements

We thank the referees for their suggestions.

References

Esary, J., Proschan, F. and Walkup, D. (1967). Association of random variables with applications, *Ann. Math. Statist.*, **38**, 1466-1474.

Lee, A.J. (1990). *U-Statistics.* Marcel Dekker, New York.

Louhichi, S. (2000). Weak convergence for empirical processes of associated sequences, *Ann. Inst. Henri Poincare*, **36**, 547-567.

Newman, C.M. (1980). Normal fluctuations and the FKG inequalities, *Comm. Math. Phys.*, **74**, 119-128.

Newman, C.M. (1984). Asymptotic independence and limit theorems for positively and negatively dependent random variables. *Inequalities in Statistics and Probability*, (ed. Y.L. Tong), 127-140, IMS, Hayward.

Peligard, M. and Suresh, R. (1995). Estimation of variance of partial sums of an associated sequence of random variables, *Stoch. Process. Appl.* , **56**, 307-319.

Prakasa Rao. B.L.S. and Dewan, I. (2001). Associated sequences and related inference problems, *Handbook of Statistics, 19 , Stochastic Processes: Theory and Methods*, (eds. C.R. Rao and D.N. Shanbag), 693-728, North Holland, Amsterdam.

Roussas, G.G. (1993) . Curve estimation in random field of associated processes, *J. Nonparametric Statist.*, **2**, 215-224.

Roussas, G.G. (1999) . Positive and negative dependence with some statistical applications, *Asymptotics, nonparametrics and time series* (ed. S. Ghosh), 757-788, Marcel Dekker, New York.

Serfling , R.J. (1968). The Wilcoxon two-sample statistic on strongly mixing processes, *Ann. Math. Statist.*, **39**, 1202-1209.