Uzbekistan Journal of Mathematics and Computer Science
Volume 1 No. 2 Page 79–85 (2025)
DOI: https://doi.org/10.56143/ujmcs.v1i2.9
Research article

Asymptotic Properties of the Wilcoxon-Mann-Whitney Statistics

Shomurodov Nozmbek * To'rayev Alimardon

Abstract

Random variables seen in many practical problems of statistical physics, quantum field theory, and reliability theory are associated connected random variables. This article focuses on nonparametric estimates for statistics constructed by associated random variables. It proves a theorem for a sequence of stationary associated random variables with two identical marginal distributions.

Keywords: *associated; statistics; stationary; sequence.*

1. INTRODUCTION

It is well-known that independent random variables have been extensively studied in science. However, in nature and technology, random variables are often dependent. Therefore, the study of dependent random variables, specifically associated random variables, under certain conditions, and demonstrating their applications in practical problems has been the focus of many prominent experts. In this field, renowned mathematicians such as Newman, Prakasa Rao, Harris, Fortuin, Lebowitz, Hoeffding, Wilcoxon, Mann, Whitney, and their students have achieved fundamental results. Currently, with the development of several directions in mathematical statistics, the importance of the theory of associated random variables has significantly increased, which is well-known among specialists. The topic of this master's thesis is dedicated to gathering future-relevant results on associated random variables, which have been relatively less studied compared to dependent variables, and to studying nonparametric estimators for statistics constructed based on associated variables.

1.1. Preliminaries

Definition 1.1. Let (X, Y) be a random vector with $E[X^2] < \infty$ and $E[Y^2] < \infty$. Define

$$H(x, y) = P(X \leq x, Y \leq y) - P(X \leq x)P(Y \leq y). \tag{1.1}$$

Recall the Hoeffding identity [8]:

$$\operatorname{cov}(X, Y) = \int_{\mathbb{R}^2} H(x, y)\, dx\, dy. \tag{1.2}$$

This identity was extended to the multivariate case by Block and Fang (1988) using the concept of cumulants for random vectors. Yu (1993) generalized Newman's (1984) earlier work by extending the covariance identity to

Received : *4–february–2025,* ***Accepted :*** *20–october–2025*
**** Corresponding author***

absolutely continuous functions of the components of the random vector X. Cuesta-Molina (1992) generalized the Hoeffding identity to semi-monotonic functions $K(\cdot)$ in the following form:

$$K(x', y') - K(x, y') - K(x', y) + K(x, y) \geq 0 \text{ [9]}$$

for all $x \leq x'$ and $y \leq y'$. This was proven as:

$$E[K(X, Y)] - E[K(X^*, Y^*)] = \int_{\mathbb{R}^2} H(x, y) K(dx, dy),$$

where X^* and Y^* are independent random variables with the same marginal distributions as X and Y, respectively. These results were further generalized by Yu (1993), Cuesta-Molina (1992), and Prakasa Rao (1998) to the multivariate case. Cuadras (2002) showed that if $\alpha(x)$ and $\beta(y)$ are functions with finite variation, then:

$$\operatorname{cov}(\alpha(X), \beta(Y)) = \int_{\mathbb{R}^2} H(x, y) \alpha(dx) \beta(dy).$$

This result is a special case of (1.2.). From this, we can see that $\operatorname{cov}(X_1, X_n) \to 0$ as $n \to \infty$. In particular, we have:

$$\sup_n |\operatorname{cov}(X_1, X_n)| < \infty.$$

Using the association property of $X_1, \ldots, X_n$, we observe that $\operatorname{cov}(X_1, X_n) > 0$ and obtain:

$$0 \leq \operatorname{cov}(X_1, X_j) = [\operatorname{cov}(X_1, X_j)]^{2/3} [\operatorname{cov}(X_1, X_j)]^{1/3} \leq [\sup \operatorname{cov}(X_1, X_n)]^{2/3} [\operatorname{cov}(X_1, X_j)]^{1/3}.$$

Therefore,

$$\sum_{j=2}^{n} \operatorname{cov}(X_1, X_j) \leq [\sup \operatorname{cov}(X_1, X_n)]^{2/3} \sum_{j=2}^{n} [\operatorname{cov}(X_1, X_j)]^{1/3} < \infty. \tag{1.3}$$

Let $R_1, R_2, \ldots, R_n$ be the ranks of $X_1, X_2, \ldots, X_n$. The Wilcoxon signed-rank statistic is defined as...

2. Main part

Let $\{X_n, n \geq 1\}$ be a sequence of stationary random variables. We can express T as a linear combination of two U-statistics (Hettmansperger (1984)):

$$T = nU_n^{(1)} + \binom{n}{2} U_n^{(2)}, \tag{2.1}$$

where

$$nU_n^{(1)} = \sum_{i=1}^{n} \phi(X_i),$$

$$\binom{n}{2} U_n^{(2)} = \sum_{1 \leq i < j \leq n} \psi(X_i, X_j), \tag{2.2}$$

and

$$\psi(x, y) = I(x + y > 0). \tag{2.3}$$

For a stationary sequence $\{X_n, n \geq 1\}$, we have:

$$E(U_n^{(2)}) = \frac{1}{\binom{n}{2}} \sum_{1\le i<j\le n} P_{ij} = \frac{1}{\binom{n}{2}} \sum_{j=2}^{n} (n-j+1)p_{1,j},$$

where $p_{ij} = P[X_i + X_j > 0]$. Define:

$$\theta = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \psi(x,y)dF(x)dF(y),$$

$$\theta = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \psi(x,y)dF(x)dF(y) = 1 - \int_{-\infty}^{\infty} F(-x)dF(x),$$

$$\psi_1(x_1) = E(\psi(x_1,x_2)) = \int_{-\infty}^{\infty} \psi(x_1,x_2)dF(x_2) = 1 - F(-x_1).$$

Then,

$$h^{(1)}(x_1) = \psi_1(x_1) - \theta, \tag{2.4}$$

and

$$h^{(2)}(x_1,x_2) = \psi(x_1,x_2) - \psi_1(x_1) - \psi_1(x_2) + \theta = \psi(x_1,x_2) + F(-x_1) + F(-x_2) - 2 + \theta. \tag{2.5}$$

The Hoeffding decomposition (H-decomposition) for $U_n^{(2)}$ is given by (Lee (1990)):

$$U_n^{(2)} = \theta + 2H_n^{(1)} + H_n^{(2)}, \tag{2.6}$$

where $H_n^{(j)}$ is the j-th degree U-statistic based on the kernel $h^{(j)}$, $j = 1,2$:

$$H_n^{(j)} = \frac{1}{\binom{n}{j}} \sum h^{(j)}(X_{i_1},\ldots,X_{i_j}), \tag{2.7}$$

with the sum taken over all subsets $1 \le i_1 < \ldots < i_j \le n$ of $\{1,\ldots,n\}$.

2.1. Variance Decomposition

Here, the sum is taken over all subsets $\{1,\ldots,n\}$ for $1 \le i_1 < \ldots < i_j \le n$. Considering the H-decomposition, we obtain the following:

$$\mathrm{Var}\left(U_n^{(2)}\right) = 4\mathrm{Var}\left(H_n^{(1)}\right) + \mathrm{Var}\left(H_n^{(2)}\right) + 4\mathrm{Cov}\left(H_n^{(1)}, H_n^{(2)}\right). \tag{2.8}$$

Now, consider the following equality (Dewan and Prakasa Rao (2001)):

$$\mathrm{Var}\left(H_n^{(1)}\right) = \frac{1}{n}\left(\sigma_1^2 + 2\sum_{j=2}^{\infty} \sigma_{1j}^2\right) + o\left(\frac{1}{n}\right), \tag{2.9}$$

where

$$\sigma_1^2 = \mathrm{Var}\left(F(-X_1)\right),$$

$$\sigma_{1j}^2 = \mathrm{Cov}\left(F(-X_1), F(-X_{1+j})\right). \tag{2.10}$$

Using Newman's inequality and (1.3), we can write:

$$\sum_{j=2}^{\infty} \sigma_{1j}^2 = \sum_{j=2}^{\infty} \mathrm{Cov}\left(F(-X_1), F(-X_{1+j})\right) < \infty.$$

Additionally,

$$\mathrm{Var}\left(H_n^{(2)}\right) = \binom{n}{2}^{-2} \sum_{1\le i<j\le n} \sum_{1\le k<l\le n} \mathrm{Cov}\left\{h^{(2)}(X_i, X_j), h^{(2)}(X_k, X_l)\right\},$$

where

$$\mathrm{Cov}\left\{h^{(2)}(X_i, X_j), h^{(2)}(X_k, X_l)\right\} = \mathrm{Cov}\left\{\psi(X_i, X_j), \psi(X_k, X_l)\right\} +$$

$$+\mathrm{Cov}\left\{\psi(X_i, X_j), F(-X_k)\right\} + \mathrm{Cov}\left\{\psi(X_i, X_j), F(-X_l)\right\} +$$
$$+\mathrm{Cov}\left\{\psi(X_k, X_l), F(-X_k)\right\} + \mathrm{Cov}\left\{\psi(X_k, X_l), F(-X_l)\right\} +$$

$$+\mathrm{Cov}\left\{F(-X_k), F(-X_l)\right\} + \mathrm{Cov}\left\{F(-X_k), F(-X_l)\right\}.$$

Using Newman's (1980) inequality, we obtain: Using Newman's (1980) inequality, we obtain:

$$\left|\mathrm{Cov}\left(F(-X_k), F(-X_l)\right)\right| \le \sup_x \left(f(x)\right)^2 \mathrm{Cov}\left(X_k, X_l\right). \tag{2.11}$$

Due to the boundedness of the density function, the following result from Bagai and Prakasa Rao (1991) holds:

$$\left|\mathrm{Cov}\left(\psi(X_i, X_j), \psi(X_k, X_l)\right)\right| =$$

$$= \left|P\left[X_i + X_j > 0, X_k + X_l > 0\right] - P\left[X_i + X_j > 0\right] P\left[X_k + X_l > 0\right]\right| \le$$

$$\le C\left[\mathrm{Cov}\left(X_i + X_j, X_k + X_l\right)\right]^{1/3} =$$

$$= C\left[\mathrm{Cov}\left(X_i, X_k\right) + \mathrm{Cov}\left(X_j, X_k\right) + \mathrm{Cov}\left(X_i, X_l\right) + \mathrm{Cov}\left(X_j, X_l\right)\right]^{1/3}. \tag{2.12}$$

Let $Z = X_i + X_j$. Then, $\psi(X_i, X_j) = I(X_i + X_j > 0) = I(Z > 0)$. Note that this function has a jump at $z = 0$. From equation (2.12), we can conclude that:

$$\left|\mathrm{Cov}\left(\psi(X_i, X_j), F(X_k)\right)\right| =$$

$$= \left|\int_{-\infty}^{\infty} \left(P\left[X_i + X_j \le 0, X_k \le x\right] - P\left[X_i + X_j \le 0\right] P\left[X_k \le x\right]\right) dF(x)\right| \le$$
$$\le \int_{-\infty}^{\infty} |P[X_i + X_j \le 0, X_k \le x] - P[X_i + X_j \le 0] P[X_k \le x]| dF(x) \le$$
$$\le C \int_{-\infty}^{\infty} [\mathrm{cov}(X_i + X_j, X_k)]^{1/3} dF(x) = C[\mathrm{cov}(X_i + X_j, X_k)]^{1/3} =$$
$$= C[\mathrm{cov}(X_i, X_k) + \mathrm{cov}(X_j, X_k)]^{1/3}.$$

Using equations (2.11), (2.12), and (2.13) in equation (2.10), we obtain the following:

$$\left|\operatorname{cov}\left\{h^{(2)}(X_i,X_j),h^{(2)}(X_k,X_l)\right\}\right| \leq$$

$$\leq C\left[\operatorname{cov}(X_i,X_k)+\operatorname{cov}\left(X_j,X_k\right)+\operatorname{cov}(X_i,X_l)+\operatorname{cov}\left(X_j,X_l\right)\right]^{\frac{1}{3}}+$$

$$+\left[\operatorname{cov}(X_i,X_k)+\operatorname{cov}\left(X_j,X_k\right)\right]^{\frac{1}{3}}+\left[\operatorname{cov}(X_i,X_l)+\operatorname{cov}\left(X_j,X_l\right)\right]^{\frac{1}{3}}+$$

$$+\left[\operatorname{cov}(X_k,X_i)+\operatorname{cov}(X_l,X_i)\right]^{\frac{1}{3}}+\left[\operatorname{cov}\left(X_k,X_j\right)+\operatorname{cov}\left(X_l,X_j\right)\right]^{\frac{1}{3}}+$$

$$+\operatorname{cov}(X_i,X_k)+\operatorname{cov}\left(X_j,X_k\right)+\operatorname{cov}(X_i,X_l)+\operatorname{cov}\left(X_j,X_l\right)\leq$$

$$\leq C\left[\operatorname{cov}(X_i,X_k)+\operatorname{cov}\left(X_j,X_k\right)+\operatorname{cov}(X_i,X_l)+\operatorname{cov}\left(X_j,X_l\right)\right]^{\frac{1}{3}}+$$

$$+\left[\operatorname{cov}(X_k,X_i)+\operatorname{cov}(X_l,X_i)\right]^{\frac{1}{3}}+\left[\operatorname{cov}\left(X_k,X_j\right)+\operatorname{cov}\left(X_l,X_j\right)\right]^{\frac{1}{3}}+$$

$$= \ C\left[\operatorname{cov}(X_i,X_k)+\operatorname{cov}\left(X_j,X_k\right)+\operatorname{cov}(X_i,X_l)+\operatorname{cov}\left(X_j,X_l\right)\right]^{\frac{1}{3}}+$$

$$=\left[\operatorname{cov}(X_k,X_i)+\operatorname{cov}(X_l,X_i)\right]^{\frac{1}{3}}+\left[\operatorname{cov}\left(X_k,X_j\right)+\operatorname{cov}\left(X_l,X_j\right)\right]^{\frac{1}{3}}$$

$$r=(|i-k|])+r(|j-k|])+r(|i-l|])+r[|j-l|]$$
$$\sum r(k)<\infty.$$

Therefore, from Serfling's (1968) theorem, we obtain as $n\to\infty$

$$operatornamevar\left(H_n^{(2)}\right)=o\left(\frac{1}{n}\right). \tag{2.13}$$

Using the Cauchy-Schwarz inequality, the following follows

$$\operatorname{cov}\left(H_n^{(1)},H_n^{(2)}\right)=o\left(\frac{1}{n}\right).$$

Using equations (2.8), (2.9), (2.16), and (2.17), we can write

$$\operatorname{var}\left(U_n^{(2)}\right)=4\left[\sigma_1^2+2\sum_{j=1}^{\infty}\sigma_{1j}^2\right]+o\left(\frac{1}{n}\right).$$

To obtain the limit distribution of the U-statistic, we introduce the following theorem.

Theorem 2.1. *Let $\{X_n,n\geq 1\}$ be a sequence of associated random variables. Suppose*

$$\sum_{k=1}^{\infty}r(k)<\infty \text{ holds. Then, as } n\to\infty,$$

$$n^{1/2}\left(U_n^{(2)}-\theta\right)\xrightarrow{e}N(0,1)$$

where $\sigma_U^2=\sigma_1^2+2\sum_{j=1}^{\infty}\sigma_{1j}^2$.

Proof. Here, we also use relations (2.2)–(2.12), and make appropriate modifications in the remaining relations: Using Newman's (1980) inequality, we obtain

$$|\operatorname{cov}(F(-X_i), F(-X_k))| \leq \sup_x (f(x))^2 \operatorname{cov}(X_i, X_k). \tag{2.14}$$

Due to the boundedness of the density function, the following result follows from Bagai and Prakasa Rao's (1991) theorem:

$$\left|\operatorname{cov}\left(\psi(X_i, X_j), \psi(X_k, X_l)\right)\right| =$$

$$= \left|P\left[X_i + X_j > 0, X_i + X_k > 0\right] - P\left[X_i + X_j > 0\right] P\left[X_i + X_k > 0\right]\right| \leq$$

$$\leq Cr\left(|i - l|\right).$$

Let $Z = X_i + X_j$. Note that the function $\psi(X_i, X_j) = I(X_i + X_j > 0) = I(z > 0)$ has a discontinuity at $z = 0$. Now, from equation (2.12), it follows that

$$\left|\operatorname{cov}\left(\psi(X_i, X_j), F(X_k)\right)\right| =$$

$$= \left|\int_{-\infty}^{\infty} \left(P\left[X_i + X_j \leq 0, X_k \leq x\right] - P\left[X_i + X_j \leq 0\right] P\left[X_k \leq x\right]\right) dF(x)\right| \leq$$

$$\leq \int_{-\infty}^{\infty} \left|P\left[X_i + X_j \leq 0, X_k \leq x\right] - P\left[X_i + X_j \leq 0\right] P\left[X_k \leq x\right]\right| dF(x) \leq$$

$$= C\left[r\left(|i - k|\right) + r\left(|j - k|\right)\right]. \tag{2.15}$$

Using equations (2.14) and (2.15) in equation (2.12)we obtain

$$\left|\operatorname{cov}\left(h^{(2)}\left(X_i, X_j\right), h^{(2)}\left(X_k, X_l\right)\right)\right| \leq$$

$$= r\left(|i - k|\right) + r\left(|j - k|\right) + r\left(|i - l|\right) + r\left(|j - l|\right).$$

$$\sum_{k=1}^{\infty} r(k) < \infty.$$

From this and Serfling's (1968) theorem, we obtain as $n \to \infty$

$$\operatorname{var}\left(H_n^{(2)}\right) = o\left(\frac{1}{n}\right). \tag{2.16}$$

Using the Cauchy-Schwarz inequality, the following follows

$$\operatorname{cov}\left(H_n^{(1)}, H_n^{(2)}\right) = o\left(\frac{1}{n}\right). \tag{2.17}$$

Using equations (2.8), (2.9), (2.16), and (2.17), we can write

$$\operatorname{var}\left(U_n^{(2)}\right) = 4\left[\sigma_1^2 + 2\sum_{j=1}^{\infty} \sigma_{1j}^2\right] + o\left(\frac{1}{n}\right).$$

3. Conclusion

It is well-known that independent random variables have been sufficiently studied in science. However, in nature and technology, random variables are often dependent. Therefore, the study of dependent variables, specifically associated random variables, under certain conditions, and demonstrating their applications in practical problems has been the focus of many prominent experts. The topic of this master's thesis is dedicated to gathering future-relevant results on associated random variables, which have been relatively less studied compared to dependent variables, and to studying nonparametric estimators for statistics constructed based on associated variables.

References

[1] Bagai, I. and Prakasa Rao, B.L.S. Estimation of the survival function for stationary associated processes, *Statist. Probab. Lett.*, 12, 385-391.(1991)

[2] Baek, Jong Li., Park, Sung Tae., Chung, Sung Mo. and Seo, Hye Young. On the almost sure convergence of weighted sums of negatively associated random variables, *Commun. Korean Math. Soc.*, 20, 539-546. (2005).

[3] Bagai, I. and Prakasa Rao, B.L.S. Kernel-type density and failure rate estimation for associated sequences, *Ann. Inst. Statist. Math.*, 47, 253-266.(1995)

[4] Birkel, T. On the convergence rate in the central limit theorem for associated processes, *Ann. Probab.*, 16, 1689-1698.(1988b)

[5] Bulinski, A.V. Rates of convergence in the central limit theorem for fields of associated random variables, *Theor. Probab. Appl.*, 40, 136-144.(1995)

[6] Matula, P. On almost sure limit theorems for positively dependent random variables, *Statist. Probab. Lett.*, 74, 59-66.(2005)

[7] Newman, C.M. and Wright, A.L. Associated random variables and martingale inequalities, *Z. Wahrsch. Theorie und Verw. Gebiete*, 59, 361-371. (1982)

[8] Hoeffding, W. Masstabinvariante Korrelations-theorie, Schr. Math. Inst.*University Berlin*, 5, 181-233. (1940)

[9] Queseda- Molina, J.J. A generalization of an identity of Hoeffding and some applications, *J. Ital. Statist. Soc.*, 3, 405-411.(1992)

Affiliations

Shomurodov Nozimbek
Address: Tashkent State Transport University, Tashkent.
e-mail: nozimshoh007@gmail.com
ORCID ID: https://orcid.org/0009-0005-2748-4484

Turaev Alimardon
Address: Tashkent State Transport University, Tashkent.
e-mail: alimardontoxirovich0413@gmail.com
ORCID ID: https://orcid.org/0000-0002-0584-9738