Journal of Modern Applied Statistical Methods

Volume 18 | Issue 1 Article 4

2-20-2020

Bivariate Analogs of the Wilcoxon–Mann–Whitney test and the Patel–Hoel Method for Interactions

Rand Wilcox
University of Southern California, rwilcox@usc.edu

Recommended Citation

Wilcox, R. (2019). Bivariate analogs of the Wilcoxon–Mann–Whitney test and the Patel–Hoel method for interactions. Journal of Modern Applied Statistical Methods, 18(1), eP3129. doi: 10.22237/jmasm/1556669880

Journal of Modern Applied Statistical Methods
May 2019, Vol. 18, No. 1, eP3129
doi: 10.22237/jmasm/1556669880

בס"ד

ISSN 1538 − 9472

INVITED ARTICLE

Bivariate Analogs of the Wilcoxon–Mann–Whitney Test and the Patel–Hoel Method for Interactions

Rand Wilcox
University of Southern California
Los Angeles, CA

A fundamental way of characterizing how two independent compares compare is in terms of the probability that a randomly sampled observation from the first group is less than a randomly sampled observation from the second group. The paper suggests a bivariate analog and investigates methods for computing confidence intervals. An interaction for a two-by-two design is investigated as well.

Keywords: Cliff's method, dominance, bootstrap methods

Introduction

Consider two independent random variables, X and Y. One of the more basic methods for comparing the corresponding distributions is in terms of $p = P(X < Y)$, the probability that a randomly sampled observation from the first distribution is less than a randomly sampled observation from the second distribution. The Wilcoxon–Mann–Whitney (WMW) test is based on an estimate of p, but it uses an incorrect estimate of the standard error when distributions differ. That is, inferences about p can be inaccurate regardless of how large the sample size might be. Numerous methods have been derived for dealing with this issue, several of which were compared by Neuhäuser et al. (2007). A method derived by Cliff (1996) was found to perform relatively well.

Now consider the situation where multivariate distributions are to be compared. An ANOVA-type analog of the WMW test was derived by Brunner et al. (2002). Wilcox noted that in the univariate case, inferences about p are related

doi: 10.22237/jmasm/1556669880 | Accepted: January 11, 2019; Published: February 20, 2020.
Correspondence: Rand Wilcox, rwilcox@usc.edu

to how deeply zero is nested in the distribution of $D = X - Y$. Based on this perspective, Brunner et al. (2002) suggested a multivariate technique for the two-sample case. Wilcox (2005) also suggested a projection-type method. Roughly, the data for both groups are projected onto a line connecting the center of the two clouds of data. Then an analog of methods based on p is used. For a description of other rank-based multivariate methods, see for example Puri and Sen (1971), Brunner et al. (2002), Chakraborty and Chaudhuri (2015), as well as Oja and Randles (2004).

The goal is to suggest and investigate an alternative approach that is limited to comparing two independent groups based on bivariate data. A possible appeal of the method is that it is readily interpreted by non-statisticians. As will be seen, it has close ties to p. This is followed by a bivariate generalization of the Patel and Hoel (1973) approach to interactions.

Proposed Measures of Effect Size

For the first of two independent groups, let (X_1,X_2) denote a randomly sampled pair of observations having some unknown bivariate distribution. Let (Y_1,Y_2) denote a randomly sampled pair from the second group. Here, the possible outcomes are broken down into to three categories. The first is $X_1 > Y_1$ and $X_2 > Y_2$, in which case it said that the pair (X_1,X_2) completely dominates (Y_1,Y_2). The second category is when $X_1 < Y_1$ and $X_2 < Y_2$, in which case (Y_1,Y_2) completely dominates (X_1,X_2). The third category is that neither pair completely dominates. Then a way of characterizing the difference between the groups is with

$$\delta = P\left(X_1 < Y_1 \text{ and } X_2 < Y_2\right) - P\left(X_1 > Y_1 \text{ and } X_2 > Y_2\right) \tag{1}$$

the difference between the probability that (Y_1,Y_2) completely dominates minus the probability that (X_1,X_2) completely dominates.

A related perspective is based on a generalization of how analogs of the WMW test deal with tied values. Let p_1 denote the probability that (X_1,X_2) completely dominates. Let p_2 denote the probability that neither point dominates, and let p_3 indicate the probability that (Y_1,Y_2) completely dominates. Let $P = p_3 + 0.5p_2$. Note that when $\delta = 0$, $P = 0.5$, in which case there is interest in testing

$$H_0 : P = 0.5 \tag{2}$$

As will be seen, computing a confidence interval for δ plays a role when testing (2).

Next, an analog of the Patel–Hoel approach to interactions is described. Consider a two-by-two design and let δ_1 denote δ for the first level and let δ_2 denote δ for the second level. Generalizing the Patel and Hoel (1973) notion of an interaction an obvious way, no interaction is taken to mean that $\Delta = \delta_1 - \delta_2 = 0$. Alternatively, let P_1 denote P for the first level of the first factor and let P_2 denote the value of P for the second level. No interaction is taken to mean $P_I = P_1 - P_2 = 0$.

Inferences About *P*, *δ* and Δ.

Consider the goal of making inferences about δ, which will yield a method for making inferences about P. Let (X_{i1},X_{i2}) denote a random sample from the first group ($i = 1, \ldots, n_1$). And let (Y_{j1},Y_{j2}) denote a random sample from the second group ($j = 1, \ldots, n_2$). Let $d_{ij} = 1$ if (Y_{j1},Y_{j2}) completely dominates (X_{i1},X_{i2}) and $d_{ij} = -1$ if (X_{j1},X_{j2}) completely dominates (Y_{i1},Y_{i2}). If neither pair completely dominates, $d_{ij} = 0$. An estimate of δ is simply

$$\hat{\delta} = \frac{1}{n_1 n_2} \sum d_{ij}$$

An estimate of P is $\hat{P} = \left(1 - \hat{\delta}\right)/2$.

Now, consider the goal of testing (1) as well as computing a $1 - \alpha$ confidence interval for P. Four strategies are considered. The first approach is based on a simple modification of the method derived by Cliff (1996). Let

$$\bar{d}_{i.} = \frac{1}{n_2} \sum_{h=1}^{n_2} d_{ih},$$

$$\bar{d}_{.h} = \frac{1}{n_1} \sum_{i=1}^{n_1} d_{ih},$$

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} \left(\bar{d}_{.h} - \hat{\delta}\right)^2,$$

$$s_2^2 = \frac{1}{n_2 - 1} \sum_{h=1}^{n_2} \left(\bar{d}_{.h} - \hat{\delta}\right)^2,$$

$$\tilde{\sigma}^2 = \frac{1}{n_1 n_2 - 1} \sum \sum \left(d_{ih} - \hat{\delta}\right)^2.$$

Then

$$\hat{\sigma}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \tilde{\sigma}^2}{n_1 n_2}$$

estimates the squared standard error of $\hat{\delta}$. Let z be the $1 - \alpha/2$ quantile of a standard normal distribution. Rather than use the more obvious confidence interval for δ, results in Cliff (1996, p. 140) suggest using instead

$$\frac{\hat{\delta} - \hat{\delta}^3 \pm z\hat{\sigma}\sqrt{\left(1 - \hat{\delta}^2\right)^2 + z^2\hat{\sigma}^2}}{1 - \hat{\delta}^2 + z^2\hat{\sigma}^2}$$

(Also see Feng & Cliff, 2004.)

The confidence interval for δ is readily modified to give a confidence for P. Letting

$$C_1 = \frac{\hat{\delta} - \hat{\delta}^3 - z\hat{\sigma}\sqrt{\left(1 - \hat{\delta}^2\right)^2 + z^2\hat{\sigma}^2}}{1 - \hat{\delta}^2 + z^2\hat{\sigma}^2}$$

and

$$C_2 = \frac{\hat{\delta} - \hat{\delta}^3 + z\hat{\sigma}\sqrt{\left(1 - \hat{\delta}^2\right)^2 + z^2\hat{\sigma}^2}}{1 - \hat{\delta}^2 + z^2\hat{\sigma}^2}$$

a $1 - \alpha$ confidence interval for P is

$$\left(\frac{1 - C_2}{2}, \frac{1 - C_1}{2}\right)$$

This will be called method C, henceforth.

Simulations reported below indicate that method C performs well when the sample sizes are equal, including situations where there is heteroscedasticity, meaning that the marginal distributions have different variances. That is, $\text{VAR}(X_1) \neq \text{VAR}(Y_1)$ and $\text{VAR}(X_2) \neq \text{VAR}(Y_2)$. However, when $n_1 \neq n_2$ and

simultaneously the marginal variances differ, method *C* can be unsatisfactory in terms of controlling the Type I error probability. A closer look at the simulation results revealed that the estimate of the standard error of $\hat{\delta}$ can be inaccurate.

Three methods were examined that were aimed at dealing with unequal sample sizes. The first was to use a bootstrap estimate of the standard error, which is computed as follows. Begin by generating a bootstrap sample from the first group. That is, randomly sample with replacement n_1 pairs of values from (X_{i1},X_{i2}), $i = 1, \ldots, n_1$. Next, generate a bootstrap sample from the second group, compute $\hat{\delta}$ based on these two bootstrap samples and label the result δ^*. Repeat this B times yielding $\delta_1^*, \ldots, \delta_B^*$. The estimate of squared standard error of $\hat{\delta}$ is

$$\tilde{\sigma}^2 = \frac{1}{B-1}\sum\left(\delta_b^* - \bar{\delta}^*\right)^2$$

where $\bar{\delta}^* = \sum \delta_b^* / B$. Here, $B = 100$ was used, which has been found to suffice for a range of other situations (Wilcox, 2017). Using method *C*, but with $\hat{\sigma}^2$ replaced by $\tilde{\sigma}^2$, was found to improve the control over the Type I error probability, but situations were found where it was unsatisfactory.

The second approach was to use a basic percentile bootstrap method, which does not use an estimate of the standard error. Based on a bootstrap sample from each group compute an estimate of P and lable the result P^*. Repeat this process B times yielding $P_1^*, \ldots, P_B^*$, only now $B = 599$ bootstrap samples are used, which is motivated by results in Wilcox (2017). Put these B bootstrap estimates in ascending order and label the result $P_{(1)}^* \leq \cdots \leq P_{(B)}^*$. Let $\ell = \alpha B/2$, rounded to the nearest integer, and $B - \ell$. Then an approximate $1 - \alpha$ confidence interval for P is $\left(P_{(\ell+1)}^*, P_{(u)}^*\right)$. Let $A = \sum I(P^* > 0.5)$, where the indicator function $I(P^* > 0.5) = 1$ if $P^* > 0.5$; otherwise $I(P^* > 0.5) = 0$. Let $p^* = A/B$. A (generalized) p-value is $2\min(p^*, 1 - p^*)$, which is called method *PB* henceforth. But this approach proved to be unsatisfactory as well in some of the situations described here.

The finding that method *C* performs well in simulations when the sample sizes are equal motivated the third approach. Let $n = \min\{n_1,n_2\}$. Next, generate a bootstrap sample of size n from each group and compute a $1 - \alpha$ confidence interval for δ using method *C* based on these bootstrap samples. Let L denote the lower end of the confidence interval and U the upper end. Repeat this process B times yielding $L_1, \ldots, L_B$ and $U_1, \ldots, U_B$. The final confidence interval for δ is taken to be $(\bar{L},\bar{U})$, where $\bar{L} = \sum L_b/B$ and $\bar{U} = \sum U_b/B$. This will be called method *CPB* henceforth.

This is the only method found to perform reasonably well in simulations when there is both unequal sample sizes and heteroscedasticity. A confidence interval for P can be computed in a similar manner.

As for the Patel—Hoel analog of an interaction, a simple extension of method C can be used to compute a confidence for Δ when $n_1 = n_2$. Let δ_1 represent δ when focusing on level one of Factor A with level two ignored, and let $\hat{\delta}_1$ be the estimate of δ given by (1). An estimate of the squared standard error of $\hat{\delta}_1$ is now denoted by $\hat{\sigma}_1^2$. Similarly, let δ_2 be the estimate of δ_2 when focusing on level two of Factor A, with level one ignored, and denote its estimate with $\hat{\delta}_2$. The estimated squared standard error of $\hat{\delta}_2$ is denoted by $\hat{\sigma}_2^2$. Then an approximate $1-\alpha$ confidence interval for Δ is simply $\hat{\Delta} \pm z_{1-\alpha/2}\sqrt{\hat{\sigma}_1^2+\hat{\sigma}_2^2}$, where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of a standard normal distribution. It can be seen that

$$P_I = \frac{\delta_2 - \delta_1}{2}$$

An estimate of P_I is

$$\hat{P}_I = \frac{\hat{\delta}_2 - \hat{\delta}_1}{2}$$

an estimate of the squared standard error is

$$S^2 = \frac{1}{4}\left(\hat{\sigma}_1^2 + \hat{\sigma}_2^2\right)$$

and a $1-\alpha$ confidence interval for is $\hat{P}_I \pm z_{1-\alpha/2}S$. This will be called method *CPH*. As for $n_1 \neq n_2$, a percentile bootstrap method can be used. Simply proceed in the same manner as method *PB*. For each group generate a bootstrap sample as previously described. Compute an estimate of Δ based on these bootstrap samples and label the result Δ^*. Repeat this process B times yielding $\Delta_b^*\left(b=1,\ldots,B\right)$. Put these B values in ascending order and label the results $\Delta_{(1)}^* \leq \ldots \leq \Delta_{(B)}^*$. Then an approximate $1-\alpha$ confidence interval for Δ is $\left(\Delta_{(\ell+1)}^*, \Delta_{(u)}^*\right)$, where l and u are defined as before. A confidence interval for P_I can be computed in a similar manner.

This will be called method *IPB* henceforth. Although method *PB* was found to be unsatisfactory in some situations, simulations indicate that method *IPB* performs reasonably well.

Simulation Results

Simulations were used to investigate the small-sample properties of methods *C*, *CPB* and *CPH*. Data were generated from one of four bivariate distributions where the marginal distributions were taken to be *g*-and-*h* distributions, which contain the standard normal distribution as a special case. Let Z be a random variable having a standard normal distribution. Then

$$V = \left(\exp\left(gZ\right) - 1\right)\ \exp\left(hZ^2/2\right)/g, \text{ if } g > 0$$
$$V = Z\exp\left(hZ^2/2\right), \text{ if } g = 0$$

has a *g*-and-*h* distribution, where g and h are parameters that determine the first four moments. The four distributions used here are the standard normal ($g = h = 0$), a symmetric heavy-tailed distribution ($h = 0.5$, $g = 0$), an asymmetric distribution with relatively light tails ($h = 0$, $g = 0.5$), and an asymmetric distribution with heavy tails ($g = h = 0.5$). It is noted that in theory, when $h = 0.5$, kurtosis is not defined for a *g*-and-*h* distribution. That is, kurtosis is infinitely large. When $g = h = 0.5$, skewness is not defined as well. But of course, when generating data on a computer, values are in effect generated from a bounded distribution, in which case data are being generated from a distribution with a finite level of skewness and kurtosis. Table 1 summarizes the skewness and kurtosis values used here, where estimates of the skewness and kurtosis, based on one million observations generated from the *g*-and-*h* distribution, are used when they are not defined.

Data for each group were generated by first generating data from a bivariate normal distribution having correlation ρ and where the marginal distributions have mean zero and variance one. Three choices for ρ were used: 0.0, 0.8 and –0.8. Then the marginal distributions were transformed to one of the four g-and-h distributions in Table 1. Heteroscedasticity was considered by multiplying all values in group two by k, where k was taken to be one (homoscedasticity) or four. For $n_1 \neq n_2$, $k = 1/4$ was used as well. The probability of a Type I error, when testing at the 0.05 level, was estimated with 5000 replications except when using a bootstrap method. Now 2000 replications were used due to the increased execution time.

Table 1. Some properties of the g-and-h distribution.

g	h	K_1	K_2
0.0	0.0	0.0	3.0
0.0	0.5	0.0	11,986.2
0.5	0.0	1.8	8.9
0.5	0.5	126.1	24,711.9

Reported in Table 2 are the estimated Type I error probabilities using method *C* when $n_1 = n_2 = 10$. When $k = 1$, for fixed ρ, altering g or h does not change the results because the ranks of the values remain the same. Although the importance of a Type I error can depend on the situation, Bradley (1978) suggested as a general guide, when testing at the 0.05 level, the actual level should be between 0.025 and 0.075. Method C satisfies this criterion. The largest estimate is 0.056 and the lowest is 0.025. But when $n_1 \neq n_2$, and $k \neq 1$, the estimate exceeds 0.080 in some situations.

Table 2. Estimates of the actual Type I error probability using method *C*, $\alpha = 0.05$, $n_1 = n_2 = 10$

g	h	p	k=1	k=4
0.0	0.0	0.0	0.046	0.053
0.0	0.0	0.8	0.043	0.051
0.0	0.0	-0.8	0.025	0.025
0.0	0.5	0.0	0.046	0.055
0.0	0.5	-0.8	0.025	0.025
0.0	0.5	0.8	0.043	0.048
0.5	0.0	0.0	0.046	0.056
0.5	0.0	0.8	0.043	0.050
0.5	0.0	-0.8	0.025	0.025
0.5	0.5	0.0	0.046	0.052
0.5	0.5	0.8	0.043	0.054
0.5	0.5	-0.8	0.025	0.025

Shown in Table 3 are the estimated Type I error probabilities using method *CPB* when $n_1 = 10$ and $n_2 = 40$. As can be seen *CPB* generally performs reasonably well but situations are encountered where it does not satisfy Bradley's criterion; estimates less than 0.025 occur, the lowest estimate being 0.016 when $\rho = -0.8$ and $k = 1$. Increasing the first sample size to $n_1 = 20$, now the estimate is 0.033. For $n_1 = 20$ and $n_2 = 50$ the estimate is 0.034.

Table 3. Estimates of the actual Type I error probability, $\alpha = 0.05$, $n_1 = 10$, $n_2 = 40$ using method *CPB*.

g	*h*	*p*	*k*=1	*k*=4	*k*=1/4
0.0	0.0	0.0	0.053	0.056	0.065
0.0	0.0	0.8	0.024	0.015	0.024
0.0	0.0	-0.8	0.016	0.027	0.031
0.0	0.5	0.0	0.053	0.055	0.061
0.0	0.5	0.8	0.024	0.019	0.031
0.0	0.5	-0.8	0.016	0.022	0.023
0.5	0.0	0.0	0.053	0.052	0.068
0.5	0.0	0.8	0.024	0.019	0.028
0.5	0.0	-0.8	0.016	0.027	0.041
0.5	0.5	0.0	0.053	0.061	0.059
0.5	0.5	0.8	0.024	0.018	0.029
0.5	0.5	-0.8	0.016	0.026	0.029

Reported in Tables 4 and 5 are interaction results. Now heteroscedasticity was introduced by multiplying the values in the second of the four groups by *k*. Table 4 shows the results using method CPH when $n_1 = n_2 = n_3 = n_4 = 10$. As can be seen, all of the estimates are reasonably close to the nominal level; the lowest estimate is 0.048 and the highest is 0.064. Using instead method IPB, the estimates (not shown in Table 4) were a bit higher than the estimates using CPH but always less than 0.07. However, when there is both unequal sample sizes and heteroscedasticity, estimates using CPH can exceed 0.075. Results in Table 5 are based on method IPB when there are unequal sample sizes. Now the lowest estimate is 0.022 and the highest is 0.071. So Bradley's criterion is met in all situations except when $k = 1$ and $\rho = 0.8$.

Table 4. Interaction, estimates of the actual Type I error probability using method *CPH*, $\alpha = 0.05$, $n_1 = n_2 = n_3 = n_4 = 10$

g	*h*	*p*	*k*=1	*k*=4
0.0	0.0	0.0	0.048	0.053
0.0	0.0	0.8	0.060	0.064
0.0	0.5	0.0	0.048	0.052
0.0	0.5	0.8	0.060	0.062
0.5	0.0	0.0	0.048	0.052
0.5	0.0	0.8	0.060	0.062
0.5	0.5	0.0	0.048	0.052
0.5	0.5	0.8	0.060	0.063

Table 5. Interaction, estimates of the actual Type I error probability using method *IPB*, $\alpha = 0.05$, $n_1 = 10$, $n_2 = 40$, $n_3 = 10$, $n_4 = 40$

g	*h*	*p*	*k*=1	*k*=4	*k*=1/4
0.0	0.0	0.0	0.058	0.067	0.069
0.0	0.0	0.8	0.022	0.026	0.041
0.0	0.0	-0.8	0.052	0.069	0.059
0.0	0.5	0.0	0.054	0.064	0.061
0.0	0.5	0.8	0.022	0.028	0.030
0.0	0.5	-0.8	0.052	0.069	0.061
0.5	0.0	0.0	0.054	0.066	0.071
0.5	0.0	0.8	0.022	0.028	0.039
0.5	0.0	-0.8	0.052	0.071	0.056
0.5	0.5	0.0	0.054	0.066	0.063
0.5	0.5	0.8	0.022	0.024	0.032
0.5	0.5	-0.8	0.052	0.071	0.057

Illustrations.

Data from Thomson and Randall-Maciver (1905) are used to illustrate the proposed methods. They report four measurements for male Egyptian skulls from five different time periods: 4,000 BC, 3,300 BC, 1,850 BC, 200 BC and 150 AD. There are thirty skulls from each time period. Here the focus is on the first and last time periods and two of the measures: skull height and skull length. The probability that a randomly sampled pair of observations from 4,000 BC completely dominates a randomly sampled pair from 150 AD was estimated to be 0.52. The probability that a randomly sampled pair from 150 AD dominates was estimated to be 0.09 and the estimate of δ is 0.43. The estimate of P is 0.28 and the p-value based on method C is less than 0.001. That is, the results indicate that skull height and length, taken together, tend to be smaller in 150 AD.

A second illustration is based on data from the Well Elderly 2 study (Clark, et al., 2011). The general goal was to assess the impact of an intervention program aimed at improving the health and wellbeing of older adults. The sample sizes for the control group and the experimental group were 227 and 187, respectively. Included were two measures of meaningful activities. Comparing the control group to the experimental group, Cliff's method indicated that based on the first measure, meaningful activities are more likely after intervention. The p-value is 0.04. For the second measure the p-value was 0.05. Of interest is whether these two measures, taken together, again indicate that meaningful activities tend to be higher in the experimental group. The probability that the control group dominates was

estimated to be only 0.29. The probability that the experimental group dominates was estimated to 0.40. The estimate of P was 0.56, the 0.05 confidence interval was (0.52, 0.62), and the p-value was 0.02. So the results provide further evidence that meaningful activities tend to be more likely after intervention.

Conclusion

All indications are that method *C* provides reasonably accurate control over the Type I error probability when the sample sizes are equal. However, it is unclear when or how this method can be justified when dealing with unequal sample sizes. Based on extant results, the best advice is to always use method *CPH* instead. At some level this issue does not seem to raise any practical concerns. In situations where method *C* performs reasonably well, it does not appear to have any practical advantage over method *CPH* in terms of Type I errors and power. The one feature of method *C* that might make it more appealing is that it does not use bootstrap samples. That is, changing the seed in the random number generator can have some impact on the results using method *CPH*. Using a relatively high number of bootstrap samples can minimize this concern.

The situation is similar when dealing with an interaction. A simple extension of method *C*, method *CPH*, performed well in simulations. But for unequal sample sizes, method *IPB* should be used. When method *CPH* performs well, it provides a slight advantage in terms of controlling the Type I error probability. More precisely, when the actual level using *CPH* exceeds the nominal level, the level using method *IPB* was estimated to be higher by a few units in the third decimal place. A positive feature of *IPB* is that, among all of the situations considered, the actual level was found to be less than 0.07 when testing at the 0.05 level. A possible appeal of method *CPH* is that it does not use bootstrap samples in contrast to *IPB*.

Finally, R functions for applying the methods in this paper are stored in the file `Rallfun-v35`, which can be downloaded from https://dornsife.usc.edu/labs/rwilcox/software/. The function `MULNC` applies method *C* when the sample sizes are equal and method *CPB* otherwise. As for interactions, the function `MULNC.int` performs method *CPH* and `MULNCpb.int` performs *IPB*.

References

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*(2), 144–152. doi: 10.1111/j.2044-8317.1978.tb00581.x

Brunner, E., Munzel, U. & Puri, M. L. (2002). The multivariate nonparametric Behrens–Fisher problem. *Journal of Statistical Planning and Inference, 108*(1-2), 37–53. doi: 10.1016/s0378-3758(02)00269-0

Chakraborty, A. & Chaudhuri, P. (2015). A Wilcoxon–Mann–Whitney-type test for infinite-dimensional data. *Biometrika, 102*(1), 239–246. doi: 10.1093/biomet/asu072

Clark, F., Jackson, J., Carlson, M., Chou, C.-P., Cherry, B. J., Jordan-Marsh M., Knight, B. G., Mandel, D., Blanchard, J., Granger, D. A., Wilcox, R. R., Lai, M. Y., White, B., Hay, J., Lam, C., Marterella, A. & Azen, S. P. (2011). Effectiveness of a lifestyle intervention in promoting the well-being of independently living older people: results of the Well Elderly 2 Randomised Controlled Trial. *Journal of Epidemiology and Community Health, 66*(9), 782-790. doi: 10.1136/jech.2009.099754.

Cliff, N. (1996). *Ordinal Methods for Behavioral Data Analysis*. Mahwah, NJ: Erlbaum. doi: 10.4324/9781315806730

Feng, D. & Cliff, N. (2004). Monte Carlo evaluation of ordinal *d* with improved confidence interval. *Journal of Modern and Applied Statistical Methods, 3*(2), 322-332. doi: 10.22237/jmasm/1099267560

Neuhäuser M., Lösch C. & Jöckel K-H. (2007). The Chen-Luo test in case of heteroscedasticity. *Computational Statistics & Data Analysis, 51*(10), 5055–5060. doi: 10.1016/j.csda.2006.04.025

Oja, H. & Randles, R. H. (2004). Multivariate Nonparametric Tests. *Statistical Science, 19*(4), 598–605. doi: 10.1214/088342304000000558

Patel, K. M. & Hoel, D. G. (1973). A nonparametric test for interaction in factorial experiments. *Journal of the American Statistical Association, 68*(343), 615–620. doi: 10.1080/01621459.1973.10481394

Puri, M. L. & Sen, P. K. (1971). *Nonparametric Methods in Multivariate Analysis*. New York: Wiley.

Thomson, A. & Randall-Maciver, R. (1905). *Ancient Races of the Thebaid*. Oxford, UK: Oxford University Press.

Wilcox, R. R. (2004). A multivariate projection-type analogue of the Wilcoxon-Mann-Whitney test. *British Journal of Mathematical and Statistical Psychology, 57*(2), 205–213. doi: 10.1348/0007110042307212

Wilcox, R. R. (2005). Depth and a multivariate generalization of the Wilcoxon-Mann-Whitney test. *American Journal of Mathematical and Management Sciences, 25*(3-4), 343–364. doi: 10.1080/01966324.2005.10737655

Wilcox, R. R. (2017). *Introduction to Robust Estimation and Hypothesis Testing* (4th ed.). San Diego, CA: Academic Press.