International Journal of Statistics and Probability; Vol. 6, No. 3; May 2017
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

# Linear Contrasts Based on an Extension of the Wilcoxon–Mann–Whitney Approach

Rand R Wilcox[1]

[1] Rand R Wilcox, University of Southern California, 3620 S. McClintock Ave Los Angeles, CA 90089-1061, USA

Correspondence: Rand R Wilcox, University of Southern California, 3620 S. McClintock Ave Los Angeles, CA 90089-1061, USA. E-mail: rwilcox@usc.edu

Received: February 10, 2017 Accepted: March 3, 2017 Online Published: April 23, 2017

doi:10.5539/ijsp.v6n3p198 URL: https://doi.org/10.5539/ijsp.v6n3p198

**Abstract**

A well-known approach to comparing two independent groups is to focus on 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 generalization that can be used with any linear contrast based on $J > 2$ independent groups. Roughly, the proposed measure of effect size reflects the probability that among $2K$ random variables, the typical average associated with first $K$ variables is less than the typical average among the other $K$ variables. In effect, it represents a relatively simple measure of effect size that might be used to supplement other measures of effect size when dealing with two-way and higher designs.

**Keywords:** rank-based methods, effect size, Wilcoxon–Mann–Whitney

## 1. Introduction

Let $X_1$ and $X_2$ be two independent random variables. As is evident, one of the best-known approaches to comparing the corresponding groups is the Wilcoxon–Mann–Whitney (WMW) test, which is based on an estimate of $p = P(X_1 < X_2)$, the probability that a randomly sampled observation from the first group is less than a randomly sampled observation from the second group. Certainly, $p$ is a useful and important measure of effect size, it is readily understood by non-statisticians, and additional arguments supporting the use of $p$ are summarized, for example, by Cliff (1996), Ruscio (2008) and Newcombe (2006). But as a method for making inferences about $p$, under general conditions, the WMW test is unsatisfactory. The basic reason is that the standard error of the WMW test statistic was derived assuming that $X_1$ and $X_2$ have identical distributions. When distributions differ, the WMW test uses an incorrect estimate of the standard error. Numerous methods have been proposed for dealing with this issue (e.g., Brunner & Munzel, 2000; Cliff, 1996; Wilcox, 2017), some of which perform reasonably well even with relatively small sample sizes.

Now consider the case of $J$ independent variables having means $\mu_1, \ldots, \mu_J$. From basic principles, a common goal is testing

$$H_0 : \Psi = 0, \tag{1}$$

where

$$\Psi = \sum c_j \mu_j$$

and where the linear contrast coefficients $c_1, \ldots, c_J$ satisfy $\sum c_j = 0$. Roughly, the goal in this paper is to suggest an analog of testing (1) that reduces to an estimate of $p$ when dealing with two independent groups.

To elaborate in a more concrete manner, consider a two-by-two design where, for example, Factor A corresponds to two methods for treating depression and Factor B is gender. The situation can be depicted as follows:

| | | Gender | |
|---|---|---|---|
| | | M | F |
| Method | E | $\mu_1$ | $\mu_2$ |
| | C | $\mu_3$ | $\mu_4$, |

where E and C are the two methods for treating depression. A common way of dealing with main effects for the first factor is to test

$$H_0 : \frac{\mu_1 + \mu_2}{2} = \frac{\mu_3 + \mu_4}{2}.$$

The basic idea here is to focus on

$$p_L = P\left(\frac{X_1 + X_2}{2} < \frac{X_3 + X_4}{2}\right),$$

which generalizes the WMW approach in an obvious way. More broadly, for any $J \geq 2$, the goal is to make inferences about

$$p_L = P\left(\sum c_j X_j < 0\right). \tag{2}$$

An extension of the WMW has already been derived for the particular case where the goal is to deal with an interaction in a 2-by-2 design (e.g., Wilcox, 2017, section 10.6.2). Let

$$p_I = P(X_1 < X_2) - P(X_3 < X_4). \tag{3}$$

As suggested by Patel and Hoel (1973), an analog of no interaction corresponds to the situation where the null hypothesis

$$H_0 : p_I = 0.5, \tag{4}$$

is true. An analog of an ordinal interaction is a situation where both $P(X_1 < X_2)$ and $P(X_3 < X_4)$ are less than 0.5 or both are greater than 0.5. An analog of a disordinal interaction is a situation where these two probabilities are not ordinal. The method for making inferences about $p_I$, described by Wilcox (2017), is based on a simple extension of results by Cliff (1996), which will be called method CPH henceforth. The computational details of method CPH are not provided because they are not directly relevant for the situation at hand. The main point here is that this method is not readily extended to testing for main effects, or more generally, testing (2) for any $J \geq 4$.

To provide a brief sketch of the approach used here, consider again the case of two independent random variables, $X_1$ and $X_2$. Let

$$D = X_1 - X_2.$$

Inferences about $p$ are based on a nonparametric estimate of the distribution of $D$. This is evident based on how $p$ is typically estimated. In particular, let $X_{ij}$ ($i = 1, \ldots, n_j$; $j = 1, \ldots,$ J) be a random sample from the $j$th group. Then an estimate of the distribution of $D$ can be based on the $n_1 n_2$ pairwise differences

$$D_{ik} = X_{i1} - X_{k1}$$

($i = 1, \ldots, n_1$; $k = 1, \ldots, n_2$). An estimate of $p$ is simply

$$\hat{p} = \frac{1}{n_1 n_2} \sum \sum I(D_{ik}), \tag{5}$$

where the indicator function $I(D_{ik}) = 1$ if $D_{ik} < 0$; otherwise $I(D_{ik}) = 0$.

For a 2-by-2 design ($J = 4$), the approach is to estimate the distribution of $\sum c_j X_j$ in a similar manner and then consider how a confidence for $p_L$ might be computed. But it is evident that computational issues arise for $J > 4$. Here, a simple approximation of the distribution of $\sum c_j X_j$ is suggested for dealing with the case $J > 4$.

The paper is organized as follows. Section 2 describes two methods when $J = 4$. The focus is on a linear contrast that reflects an interaction, but the results extend in an obvious way to linear contrasts relevant to main effects. Section 3 summarizes simulation results regarding how well these methods control the probability of a Type I error. Motivated in part by results in section 3, section 4 describes an approximate method for dealing with $J > 4$ groups and section 5 reports simulation results on how well this method performs.

## 2. Description of the Methods

This section focuses on $J = 4$ using an estimate of the distribution of $D$ that is an obvious generalization of the method used when $J = 2$. But the method becomes increasingly impractical as $J$ increases. An alternative method must be used that represents an approximation of the method used here when estimating the distribution of $D$. One way of judging the adequacy of the approximate method is to compare it to the more "complete" approximation of the distribution of $D$ that is used here, which is done in section 5.

Given the goal of testing (2), let

$$G_{im} = X_{i1} - X_{m2} \, (i = 1, \ldots, n_1; m = 1 \ldots, n_2)$$

and

$$H_{ac} = X_{a3} - X_{c4} \, (a = 1, \ldots, n_3; c = 1 \ldots, n_4).$$

Then an estimate of $p_L$ is

$$\hat{p}_L = \frac{1}{n_1 n_2 n_3 n_4} \sum_i \sum_m \sum_a \sum_c I(G_{im} < H_{ac}). \tag{6}$$

So $\hat{p}_L$ uses a "complete" estimate of the distribution of $D$ in the sense that it uses all $n_1 n_2 n_3 n_4$ combinations of the $X_{ij}$ values.

Note that a similar approach can be used when dealing with main effects in a 2-by-2 design. For the first factor, for example, now

$$G_{im} = X_{i1} + X_{m2} \, (i = 1, \ldots, n_1; m = 1 \ldots, n_2)$$

and

$$H_{ac} = X_{a3} + X_{c4} \, (a = 1, \ldots, n_3; c = 1 \ldots, n_4).$$

For the second factor, now

$$G_{im} = X_{i1} + X_{m3} \, (i = 1, \ldots, n_1; m = 1 \ldots, n_3)$$

and

$$H_{ac} = X_{a2} + X_{c4} \, (a = 1, \ldots, n_2; c = 1 \ldots, n_4).$$

Observe that inferences about $p_L$ cannot be made by simply applying, for example, the methods derived by Cliff (Cliff, 1996) or Bruner and Munzel (2000) using the variables $G$ and $H$. The reason is that there is dependence among the $G_{im}$ variables ($i = 1, \ldots, n_1; m = 1 \ldots, n_2$) and the same is true for $H_{ac}$ ($i = 1, \ldots, n_1; m = 1 \ldots, n_2$). So the estimate of the standard error of $\hat{p}_I$ used by these methods would be incorrect. Moreover, simulations confirmed that this simple approach does indeed perform poorly. Method CPH avoids this problem, but it does not provide a basis for dealing with main effects and linear contrasts based on more than four groups. Here, two methods for dealing with this issue were considered. The first is to use a percentile bootstrap method and the second is based on a bootstrap estimate of the standard error or $\hat{p}_I$.

The percentile bootstrap method is applied as follows. Let $X^*_{ij}$ be a bootstrap sample from the $j$th group, which is obtained by randomly sampling with replacement $n_j$ values from the $j$th group. Let $\hat{p}^*$ be the estimate of $p_I$ based on this bootstrap sample. Repeat this process $B$ times yielding $\hat{p}^*_b$ ($b = 1, \ldots, B$). Let $\hat{p}^*_{(1)} \leq \cdots \leq \hat{p}^*_{(B)}$ be the $\hat{p}^*_b$ values written in ascending order. Here, $B = 500$ is used, which often seems to suffice when using a percentile bootstrap (Wilcox, 2017), However, $B$ greater than 500 might increase power (Racine & MacKinnon, 2007). Let $\ell = \alpha B/2$, rounded to the nearest integer, and let $u = B - \ell$. Then, based on general results in Liu and Singh (1997) an approximate $1 - \alpha$ confidence interval for $p_I$ is

$$(\hat{p}^*_{(\ell+1)}, \, p^*_{(u)}).$$

Let $P^*$ be the proportion of $\hat{p}^*$ values less than 0.5. When testing (4), a p-value is given by $2\min(P^*, 1 - P^*)$. This is called method PB henceforth.

A bootstrap estimate of the squared standard error of $\hat{p}^*_I$ is given by

$$\hat{\tau}^2 = \frac{1}{B-1} \sum (\hat{p}^*_{Ib} - \bar{p}^*_I),^2$$

where $\bar{p}^*_I = \sum \hat{p}^*_{Ib}/B$. Now $B = 100$ is used, which seems to suffice based on results in Efron (1987) and which is further supported by studies summarized by Wilcox (2017). So a reasonable test statistic for testing (4) is

$$T = \frac{\hat{p}_I - 0.5}{\hat{\tau}}. \tag{7}$$

This will be call method BT henceforth. Here, the null distribution of $T$ is approximated with a Student's T distribution with degrees of freedom estimated as described by Brunner and Munzel (2000). This approach is called method BT henceforth. Simulations reported in the next section indicate that the percentile bootstrap method performs better than the method based on $T$, so for brevity further details regarding the degrees of freedom are not provided. (The estimated degrees of freedom were computed via the R function bmp described in Wilcox, 2017, section 5.7.2.)

## 3. Simulation Results

Simulations were used as a partial check on the small-sample properties of methods PB and BT. Simulation estimates of the actual Type I error probability, when testing at the 0.05 level, are based on 2000 replications. (This choice for the number of replications was based in part on an effort to avoid high execution time.) The sample sizes considered were $(n_1, n_2, n_3, n_4) = (10, 10, 10, 10)$, $(20, 20, 20, 20)$ and $(10, 20, 30, 40)$. Unequal sample sizes offered no new insights, so they are not reported. Data were generated from four types of distributions: normal, symmetric and heavy-tailed (roughly meaning that outliers tend to be common), asymmetric and relatively light-tailed, and asymmetric and relatively heavy-tailed. More specifically, data are generated from g-and-h distributions (Hoaglin, 1985), which arise as follows. Let $Z$ be a random variable having a standard normal distribution. Then

$$W = \begin{cases} \frac{\exp(gZ)-1}{g} \exp(hZ^2/2), & \text{if } g > 0 \\ Z\exp(hZ^2/2), & \text{if } g = 0 \end{cases}$$

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

| g | h | $\kappa_1$ | $\kappa_2$ |
|---|---|---|---|
| 0.0 | 0.0 | 0.00 | 3.0 |
| 0.0 | 0.2 | 0.00 | 21.46 |
| 0.2 | 0.0 | 0.61 | 3.68 |
| 0.2 | 0.2 | 2.81 | 155.98 |

Table 2. Estimated Type I Error Probability, $\alpha = 0.05$

| | | | $n = 10$ | | $n = 20$ | |
|---|---|---|---|---|---|---|
| VP | $g$ | $h$ | BT | PB | BT | PB |
| 1 | 0.0 | 0.0 | 0.079 | 0.071 | 0.068 | 0.055 |
| 1 | 0.0 | 0.2 | 0.079 | 0.068 | 0.068 | 0.052 |
| 1 | 0.2 | 0.0 | 0.081 | 0.067 | 0.073 | 0.056 |
| 1 | 0.2 | 0.2 | 0.076 | 0.054 | 0.070 | 0.052 |
| 2 | 0.0 | 0.0 | 0.80 | 0.066 | 0.072 | 0.064 |
| 2 | 0.0 | 0.2 | 0.078 | 0.066 | 0.071 | 0.067 |
| 2 | 0.2 | 0.0 | 0.081 | 0.066 | 0.071 | 0.063 |
| 2 | 0.2 | 0.2 | 0.078 | 0.066 | 0.071 | 0.061 |

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 ($h = 0.2$, $g = 0$), an asymmetric distribution with relatively light tails ($h = 0$, $g = 0.2$), and an asymmetric distribution with heavy tails ($g = h = 0.2$). Table 1 summarizes the skewness ($\gamma_1$) and kurtosis ($\gamma_2$) of these distributions.

The estimated Type I error probabilities are summarized in Table 2. Bradley (1978) suggests that in general, when testing at the 0.05 level, the actual level should be between 0.025 and 0.075. Based on this criterion, method BT is unsatisfactory when $n = 10$, while method PB satisfies this criterion for all of the situations considered.

A Welch-type method can be used to test (1), which allows heteroscedasticity (e.g., Wilcox, 2017, section 7.4.1). It is evident that it is sensitive to different features of the distribution compared to method PB. So at some level power comparisons are meaningless. However, to provide at least some perspective, consider testing both (1) and (2) using the contrast coefficients 1, 1, −1, −1 (main effects associated with the first factor) when $\delta$ is added to the first group. For symmetric distributions, estimated power for the Welch and PB methods differed by about two units in the second decimal place. It is when distributions differ in skewness that the choice of method might make a difference in terms of power. Consider, for example, the situation where the first three groups have standard normal distributions and the fourth group has a lognormal distribution that has been shifted to have a median of 0.8. For $n = 30$ and $\delta = 0.5$, the estimated power was 0.62 and 0.84 for the Welch and PB methods, respectively. This is not to suggest that method PB has, in general, more power. The only point is that the choice of method can make a substantial difference.

## 4. Dealing with More Than Four Groups

Now consider the case of $J \geq 4$ independent variables. The goal in this section is to suggest a method for testing (2) using an approximation of the complete estimate of the distribution of $D$. The approximate method is applied as follows. Let $m = \min\{n_1, \ldots, n_J\}$. For each $j$, randomly sample without replacement $m$ values from $X_{ij}$ yielding say $Y_{ij}$ ($i = 1, \ldots, m$; $j = 1, \ldots, J$). Let

$$M_i = \sum_{j=1}^{J} c_j Y_{ij}$$

($i = 1, \ldots, m$) in which case an estimate of $p_L$ is

$$\hat{p}_L = \frac{1}{m} \sum I(M_i < 0). \tag{8}$$

Now repeat this process $N$ times yielding $\hat{p}_{1L}, \ldots, \hat{p}_{NL}$. Then the final estimate of $p_L$ is taken to be

$$\tilde{p}_L = \frac{1}{N} \sum_{k=1}^{N} \hat{p}_{kL}. \tag{9}$$

Inferences based on $\tilde{p}_L$ used in conjunction with a percentile bootstrap method, are henceforth called method APB.

Table 3. Estimated Type I Error Probability for Method APB,$n = N = 10$, $\alpha = 0.05$

| VP | $g$ | $h$ | $J = 4$ | $J = 6$ |
|---|---|---|---|---|
| 1 | 0.0 | 0.0 | 0.062 | 0.060 |
| 1 | 0.0 | 0.2 | 0.062 | 0.062 |
| 1 | 0.2 | 0.0 | 0.059 | 0.065 |
| 1 | 0.2 | 0.2 | 0.062 | 0.064 |
| 2 | 0.0 | 0.0 | 0.066 | 0.064 |
| 2 | 0.0 | 0.2 | 0.064 | 0.060 |
| 2 | 0.2 | 0.0 | 0.064 | 0.069 |
| 2 | 0.2 | 0.2 | 0.064 | 0.063 |

To provide some perspective on the choice of $N$, consider the case where $J = 4$ and $p_L$ is estimated with $\hat{p}_L$ given by (6) as well as $\tilde{p}_L$. The issue is how well $\tilde{p}_L$ approximates the value returned by $\hat{p}_L$. Generating data from standard normal distributions, with $n_1 = n_2 = n_3 = n_4 = 40$, the following results were obtained based on 5000 replications. With $N = 10$, there is a 0.95 probability that $|\hat{p}_L - \tilde{p}_L| \leq 0.03$. For $N = 20$ this probability is 0.996. For $N = 50$, now $|\hat{p}_L - \tilde{p}_L| \leq 0.02$ with probability 0.997. For $N = 100$ $|\hat{p}_L - \tilde{p}_L| \leq 0.01$ with probability 0.969, and for $N = 200$, this probability is now 0.997.

However, with unequal sample sizes and the minimum sample size equal to 10, there is less agreement. Consider, for example, the case where $n_1 = 10$, $n_2 = 20$, $n_3 = 30$ and $n_4 = 40$. For $N = 10$, there a 0.77 probability that $|\hat{p}_I - \tilde{p}_L| \leq 0.05$. For $N = 20$, this probability is increased to 0.92. For $N = 50$, now $|\hat{p}_I - \tilde{p}_L| \leq 0.03$ with probability 0.90. For $N = 100$, $|\hat{p}_I - \tilde{p}_L| \leq 0.03$ occurs with probability 0.98. For $N = 200$, $|\hat{p}_I - \tilde{p}_L| \leq 0.02$ occurs with probability 0.97.

Now consider $n_1 = 20$, $n_2 = 30$, $n_3 = 40$ and $n_4 = 50$. For $N = 10$, there is a 0.94 probability that $|\hat{p}_I - \tilde{p}_L| \leq 0.05$. For $N = 20$, this probability is increased to 0.99. For $N = 50$, now $|\hat{p}_I - \tilde{p}_L| \leq 0.03$ with probability 0.99. For $N = 100$ $|\hat{p}_I - \tilde{p}_L| \leq 0.02$ occurs with probability 0.98. For $N = 200$ $|\hat{p}_I - \tilde{p}_L| \leq 0.02$ occurs with probability 0.999.

So, suppose agreement is deemed acceptable if there is agreement within three units in the second decimal place or less with probability 0.95 or higher. With equal sample sizes, $N = 50$ suffices. For unequal sample sizes, a crude rule is that $N = 50$ suffices provided the minimum sample size is at least 20. If the minimum sample size is 10, $N = 100$ is a better choice. Of course, one could simply use $N = 200$ or larger to be safe. The only concern is that as $N$ increases, execution time increases substantially when testing hypotheses with a percentile bootstrap method, at least based on the R functions described in the final section of this paper. (Additional results regarding the choice $N = 10$ are given in the next section.)

## 5. Simulation Results

This section reports estimated Type I error probabilities when using the method described in the previous section. The number of groups was taken to be 4 or 6. Data were generated as described in section 3. For $J = 6$ groups, now the linear contrast coefficients were taken to be $c_1 = c_2 = c_3 = 1$ and $c_4 = c_5 = c_6 = -1$. Again VP1 refers to homoscedasticity. Now VP 2 means that $\sigma_1 = \sigma_2 = \sigma_3 = 1$ and $\sigma_4 = \sigma_5 = \sigma_6 = 4$. The results are reported in Table 3 for $n = N = 10$. Based on Bradley's criterion, all indications are that method APB is satisfactory even with $n = N = 10$. Note that for $J = 4$, the results reported in Table 2 using the complete method for estimating the distribution of $D$, are very similar to those in Table 3, which were based on the incomplete estimate of the distribution of $D$ described in section 4.

## 6. Concluding Remarks

Of course, despite the simulations reported here, perhaps situations can be found where the method in section 4 breaks down. The main point is that, at least for the situations considered, the proposed method performs reasonably well. Moreover, there is no known alternative method that can deal with the case $J > 4$ in a reasonably accurate manner.

Finally, the R function linWMW computes $\hat{p}_L$ and the R function linWMWpb computes a confidence interval using the percentile bootstrap method in section 4, both of which are available from the author upon request.

## References

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

Brunner, E., & Munzel, U. (2000). The nonparametric Behrens–Fisher problem: asymptotic theory and small-sample approximation. *Biometrical Journal, 42,* 17–25. https/doi.org/10.1002/(SICI)1521-4036(200001)42:1<17::AID-BIMJ17

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

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association, 82*, 171–185.

Newcombe R. G. (2006a). Confidence intervals for an effect size measure based on the Mann-Whitney statistic. Part 1: General issues and tail-area-based methods. *Statistics in Medicine, 25*, 543–557. https/doi.org/10.1002/sim.2323.

Racine, J., & MacKinnon, J. G. (2007). Simulation-based tests than can use any number of simulations. *Communications in Statistics–Simulation and Computation, 36*, 357–365. https://doi.org/10.1080/03610910601161256

Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods, 13*, 19–30. https/doi.org/10.1037/1082-989x.13.1.19

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

## Copyrights