Journal of Data Science **9**(2011), 471-486

Inferences about a Probabilistic Measure of Effect Size When Dealing with More Than Two Groups

Rand R. Wilcox
University of Southern California

Abstract: For two independent random variables, X and Y, let $p = P(X > Y) + 0.5P(X = Y)$, which is sometimes described as a probabilistic measure of effect size. It has been argued that for various reasons, p represents an important and useful way of characterizing how groups differ. In clinical trials, for example, an issue is the likelihood that one method of treatment will be more effective than another. The paper deals with making inferences about p when three or more groups are to be compared. When tied values can occur, the results suggest using a multiple comparison procedure based on an extension of Cliff's method used in conjunction with Hochberg's sequentially rejective technique. If tied values occur with probability zero, an alternative method can be argued to have a practical advantage. As for a global test, extant rank-based methods are unsatisfactory given the goal of comparing groups based on p. The one method that performed well in simulations is based in part on the distribution of the difference between each pair of random variables. A bootstrap method is used where a p-value is based on the projection depth of the null vector relative to the bootstrap cloud. The proposed methods are illustrated using data from an intervention study.

1. Introduction

A fundamental issue is choosing an appropriate method of characterizing how two independent random variables differ. Certainly one of the more obvious approaches is to compare measures of location. A related approach is to use some measure of effect size that is based in part on some measure of scale associated with the groups being compared. One of the better-known and commonly used measures assumes that groups have identical population variances σ^2 and is given by

$$\delta = \frac{\mu_1 - \mu_2}{\sigma},$$

where μ_1 and and μ_2 are the population means. The usual estimate of δ, popularly known as Cohen's d, is

$$d = \frac{\bar{X}_1 - \bar{X}_2}{\sigma},$$

and where $\bar{X}_j$ is the usual sample mean for the jth group ($j = 1, 2$). When dealing with clinical trials, Acion *et al.* (2006) have raised concerns about using δ and related techniques. They go on to argue that often what is needed is some sense of how likely it is that a particular treatment will be beneficial compared to a placebo or some other treatment. In more formal terms, they argue that for two independent random variables, X and Y,

$$p = P(X > Y) + 0.5P(X = Y)$$

is an important and useful way of characterizing how two groups differ. Additional arguments for using p, in a broader context, have been made by Cliff (1996) as well as Vargha and Delaney (2000). This is not to suggest, however, that measures of effect size other than p have no practical value. Even if $p = 0.5$, the means might differ in important ways or the groups might differ in terms of some measure of variation. But the arguments for using p certainly seem to have merit given the goal of gaining perspective on how groups compare.

Given that p is intrinsically interesting and important, there is the issue making inferences about p based on random samples of observations. Various methods have been proposed for testing

$$H_0 : p = 0.5 \tag{1}$$

and computing a confidence interval for p, which are reviewed in Section 2.

As is well known, the Wilcoxon-Mann-Whitney test is based on an estimate of p, namely

$$\hat{p} = \frac{U}{n_1 n_2},$$

where U is the usual Wilcoxon-Mann-Whitney U statistic, and n_1 and n_2 are the samples sizes corresponding to groups 1 and 2, respectively. But for reasons reviewed in Section 2, as a method for testing (1), the Wilcoxon-Mann-Whitney test is unsatisfactory under general conditions.

Although there is an extensive literature regarding methods for testing (1), evidently little is known about how best to proceed when dealing with more than two groups. Accordingly, the goal in this paper is to suggest and compare methods for dealing with $J > 2$ independent groups. Two types of extensions are of interest. The first is aimed at testing the global hypothesis that for all pairs of groups, $p = 0.5$. That is, for J independent groups let p_{jk} be the value of p when comparing groups j and k. The goal is to test

$$H_0 : p_{12} = p_{13} = \cdots = p_{J-1,J} = 0.5. \tag{2}$$

Several methods were considered here, only one of which performed well in simulations.

The second extension deals with the problem of testing

$$H_0 : p_{jk} = 0.5 \tag{3}$$

for each $j < k$ such that the probability of a least one Type I error, among these $(J^2 - J)/2$ hypotheses, is α. Two methods are proposed, one of which is motivated by results reviewed in Section 2. The relative merits of these methods, when dealing with small sample sizes, are studied via simulations. Wilcox (2003, Section 15.3) describes a simple extension of a method derived by Cliff (1997) that is aimed at controlling the probability of at least one Type I error when comparing all pairs of groups. But in terms of both power and the ability to control the probability of at least one Type I error, an alternative strategy (method CH described in Section 3), is found to be better for general use.

It is noted that there are rank-based methods for comparing more than two groups, but they do not provide a test of (2). For example, Rust and Fligner (1984) as well as Brunner, Dette and Munk (1997) derived a rank-based method for comparing $J > 2$ groups that improves on the Kruskall and Wallis test in terms of handling situations where distributions differ. A rough characterization of these methods is that they are designed to be sensitive to differences among the average ranks associated with the J groups, where the ranks are based on the pooled data. The important point here is that they are not based on estimates of p_{jk} and they do not provide a satisfactory approach to testing (2). This is not to suggest, however, that they have no practical value. But given an interest in making inferences about p_{jk}, a method that deals directly with testing (2), or (3), would seem desirable.

2. Review of Techniques for the Two-Sample Case

The Wilcoxon-Mann-Whitney test statistic can be written as

$$Z = \frac{\hat{p} - 0.5}{\sigma_u/(n_1 n_2)},$$

where

$$\sigma_u^2 = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}.$$

(For results on power and sample sizes, see Rosner and Glenn, 2009.) But as previously noted, as a method for testing (1), or computing a confidence interval for p, the Wilcoxon-Mann-Whitney method is known to be unsatisfactory. The reason is that under general conditions, the standard error of $\hat{p}$ is not $\sigma_u/(n_1 n_2)$. Methods for dealing with this issue have been derived by Brunner and Munzel (2000), Cliff (1996), Fligner and Policello (1981) and Mee (1990). The methods

derived by Cliff, and Brunner and Munzel, seem to be particularly effective in terms of controlling the probability of a Type I error, even when there are tied values (Neuhäuser, Lösch and Jöckel, 2007), with Cliff's method having a slight advantage when the sample sizes are small. Accordingly, one of the methods suggested here for dealing with more than two groups is based in part on a simple generalization of Cliff's technique. For completeness, Reiczigel, Zakariás and Rózsa, (2005) suggest a bootstrap method for making inferences about p and they found that it performed better, in terms of controlling the probability of Type I error, than the method derived by Brunner and Munzel (2000) when sample sizes are small, say less than 30, and tied values do not occur. But when tied values can occur, their bootstrap method can perform poorly.

Cliff's method, which includes the ability to handle tied values, is applied as follows. Let

$$p_1 = P(X > Y),$$
$$p_2 = P(X = Y),$$

and

$$p_3 = P(X < Y).$$

Cliff (1996) focuses on testing

$$H_0 : \delta = p_1 - p_3 = 0, \tag{4}$$

which is readily shown to be the same as testing

$$H_0 : p_3 + 0.5p_2 = 0.5.$$

For convenience, let $P = p_3 + 0.5p_2$, in which case this last equation becomes

$$H_0 : P = 0.5. \tag{5}$$

Of course, when tied values occur with probability zero, $P = p_3 = P(X < Y)$. The parameter δ is related to P in a simple manner:

$$\delta = 1 - 2P, \tag{6}$$

so

$$P = \frac{1 - \delta}{2}. \tag{7}$$

Based on the random samples $X_1, \cdots, X_{n_1}$ and $Y_1, \cdots, Y_{n_2}$, Cliff's confidence interval for δ is computed as follows. Let

$$d_{ih} = \begin{cases} -1, & \text{if } X_i < Y_h, \\ 0, & \text{if } X_i = Y_h, \\ 1, & \text{if } X_i > Y_h. \end{cases}$$

An estimate of δ is

$$\hat{\delta} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{h=1}^{n_2} d_{ih}. \tag{8}$$

Let

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

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

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

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

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

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 $1 - \alpha$ confidence interval for δ, Cliff (1996, p. 140) recommends

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

Note that this confidence interval for δ is readily modified to give a confidence for P. Letting

$$C_\ell = \frac{\hat{\delta} - \hat{\delta}^3 - z\hat{\sigma}\sqrt{(1 - \hat{\delta}^2)^2 + z^2\hat{\sigma}^2}}{1 - \hat{\delta}^2 + z^2\hat{\sigma}^2}$$

and

$$C_u = \frac{\hat{\delta} - \hat{\delta}^3 + z\hat{\sigma}\sqrt{(1 - \hat{\delta}^2)^2 + z^2\hat{\sigma}^2}}{1 - \hat{\delta}^2 + z^2\hat{\sigma}^2},$$

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

$$\left(\frac{1 - C_u}{2}, \frac{1 - C_\ell}{2} \right). \tag{9}$$

The strategy suggested by Wilcox (2003), when performing all pairwise comparisons, is to simply replace z with the $1-\alpha$ quantile of a Studentized Maximum modulus distribution with infinite degrees of freedom. This will be called method SMM

3. Two Multiple Comparison Procedures

This section describes the two proposed methods for testing (3). One of the methods for testing (3), which is called method CH, is based on a direct estimate of p_{jk}, but the other method, called method DBH is not.

Method CH

To describe the motivation for first of the new methods of performing all pairwise comparisons, first note that the approach used by Wilcox (2003) is nearly tantamount to using the Bonferroni inequality. For instance, consider $C = 10$ comparisons with the goal that the probability of a Type I error be less than or equal to α. If, for example, $\alpha = 0.05$, the Bonferroni method tests each hypothesis at the 0.05/10=0.005 level. So the resulting value for z, using Cliff's method, is 2.808, the 0.995 quantile of a standard normal distribution. Using instead the method in Wilcox (2003), z would be 2.79. It is known, however, that replacing the Bonferroni method with a sequentially rejective method results in as much or more power. Moreover, assuming that each test is level robust, there are sequentially rejective methods for which the probability of at least one Type I error is less than or equal to the nominal level. Here, Hochberg's (1988) method is used, which is based in part on p-values. There is no explicit expression for a p-value when using Cliff's method, but this is easily addressed with the aid of a computer by determining the smallest α value for which Cliff's method rejects.

Hochberg's method begins by computing a p-value for each of the C tests to be performed, which are labeled $P_1, \ldots, P_C$. Next, put the p-values in descending order yielding $P_{[1]} \geq P_{[2]} \geq \cdots \geq P_{[C]}$. Let $d_k = \alpha/k$ and proceed as follows:

1. Set $k = 1$.
2. If $P_{[k]} \leq d_k$, stop and reject all C hypotheses; otherwise, go to step 3.
3. Increment k by 1. If $P_{[k]} \leq d_k$, stop and reject all hypotheses having a p-value less than or equal d_k.
4. If $P_{[k]} > d_k$, repeat step 3.
5. Continue until a significant result is found or all C hypotheses have been tested.

Note that Hochberg's method will have as much or more power than the method based on the Studentized maximum modulus distribution when $\alpha/C \leq 1 - 2P(Z \leq z)$, where now z is the $1 - \alpha$ quantile of a Studentized maximum modulus distribution with infinite degrees of freedom and Z is a standard normal distribution. The 0.05 quantiles of the Studentized maximum modulus distribution are given in Wilcox (2003) for $2 \leq C \leq 28$. Based on these reported quantiles, for $\alpha = 0.05$, Hochberg's method will have more power when $C \leq 28$. But, for example, with $\alpha = 0.01$ and $C = 28$, $\alpha/C > 1 - 2P(Z \leq z)$, meaning that it is not necessarily true that Hochberg's method will have more power.

Method DBH

To describe an alternative approach to testing (1), let θ_1 and θ_2 be the population medians associated with two independent groups. It is known that under general conditions, the Wilcoxon-Mann-Whitney test is unsatisfactory for testing

$$H_0 : \theta_1 = \theta_2 \tag{10}$$

(e.g., Hettmansperger, 1984; Fung, 1980). The same is true of the more modern rank-based methods aimed at allowing differences in dispersion. To provide a rough indication why, let $D = X - Y$ and note that under general conditions, $\theta_d \neq \theta_1 - \theta_2$, where θ_d is the population median of D.

The important point here is that the null hypothesis given by (1) is equivalent to

$$H_0 : \theta_d = 0,$$

which is not the same as (10). So for J groups, the goal of testing (3) corresponds to testing

$$H_0 : \theta_{djk} = 0 \tag{11}$$

for each $j < k$, where θ_{djk} is the value of θ_d when comparing groups j and k.

The proposed method for testing (11) is as follows. Let X_{ij} $(i = 1, \cdots, n_j; j = 1, \cdots, J)$ be a random sample of size n_j from the jth group. Generate a bootstrap sample from the jth group by randomly sampling with replacement n_j observations from $X_{1j}, \cdots, X_{n_j j}$, which will be labeled $X^*_{1j}, \cdots, X^*_{n_j j}$. Let M^*_{djk}, $j < k$, be the usual sample median based on the $n_j n_k$ differences $X^*_{ij} - X^*_{\ell k}$ $(i = 1, \cdots, n_j;\ \ell = 1, \cdots, n_k)$. Repeat this process B times yielding M^*_{djkb}, $b = 1, \cdots, B$. Based on general theoretical results in Liu and Singh (1997), in conjunction with a strategy for dealing with tied values suggested in Wilcox (2006), a p-value when testing (11) is readily computed as follows. Let

$$\varrho_{jk} = \frac{1}{B}\left(\sum I(M^*_{djkb} > 0) + 0.5 \sum I(M^*_{djkb} = 0)\right),$$

where the indicator function $I(M^*_{djkb} > 0) = 1$ if $M^*_{djkb} > 0$; otherwise $I(M^*_{djkb} > 0) = 0$. Then a (generalized) p-value when testing H_0: $\theta_{djk} = 0$ is

$$2\min(\varrho_{jk}, 1 - \varrho_{jk}).$$

All indications are that method DBH performs well when dealing with continuous distributions. However, it is not recommended when dealing with discrete distributions where tied values can occur.

4. A Global Test: Method WMWAOV

Now consider the goal of testing (2). The initial strategy considered here was to use a simple generalization of the percentile bootstrap method in Wilcox (2005, pp. 308-310) based on estimates of p_{jk}. Roughly, bootstrap estimates of p_{jk} are obtained, after which one measures how deeply the null vector $(0.5, \cdots, 0.5)$ is nested within the resulting data cloud. If the null hypothesis is true, the null vector should have a reasonably deep location within the bootstrap cloud of points. However, in terms of controlling the probability of a Type I error, the method was found to be unsatisfactory in simulations; the actual level when testing at the 0.05 level was found to be greater than 0.075 with sample sizes of 20. A variation was considered that was based on the test statistic

$$H = \sum_{j<k} (\hat{p}_{jk} - 0.5)^2$$

in conjunction with the bootstrap method in Wilcox (2005, Section 7.6). But this method proved to be unsatisfactory in simulations as well. The only method found to perform reasonably well in simulations is based on an extension of method DBH. That is, the strategy is to test (2) with a method based on the equivalent hypothesis

$$H_0 : \theta_{djk} = 0, \ \forall \ j < k, \tag{12}$$

where θ_{djk} is the value of θ_d when comparing groups j and k.

To elaborate, again let X_{ij} $(i = 1, \cdots, n_j; \ j = 1, \cdots, J)$ be a random sample of size n_j from the jth group. Generate a bootstrap sample from jth group by randomly sampling with replacement n_j observations from $X_{1j}, \cdots, X_{n_j j}$, which will be labeled $X^*_{1j}, \cdots, X^*_{n_j j}$. Let M^*_{djk}, $j < k$, be the value of M_d based on the bootstrap samples from groups j and k. Repeat this process B times yielding M^*_{djkb}, $b = 1, \cdots, B$. So M^*_{djkb} represents B vectors, each having length $(J^2 - J)/2$. Based on general results in Liu and Singh (1997), a p-value for testing (12) can be obtained by measuring how deeply $\mathbf{0} = (0, \cdots 0)$ is nested within the bootstrap cloud of points.

To avoid certain computational and theoretical issues (to be explained), it is convenient to measure the depth of a point, within a bootstrap data cloud of B points, using a variation of the projection-type technique discussed by Donoho and Gasko (1992). Here, a rough outline of the method is provided. Complete computational details can be found in Wilcox (2005, Section 6.2.5). For notational convenience, write the C bootstrap estimates M^*_{djkb} $(j < k)$, for the bth bootstrap sample, as $\mathbf{M}_b = (M^*_{d12b}, \cdots M^*_{d,J-1,J,b})$. Let $\hat{\eta}$ be an estimate of some robust multivariate measure of location based on the B vectors $\mathbf{M}_1, \cdots \mathbf{M}_B$. Here, the marginal medians are used, but as is well known, various alternatives have been proposed. For fixed i $(i = 1, \cdots, B)$, project all B points onto the line $\mathcal{L}_i$ connecting $\mathbf{M_i}$ and $\hat{\eta}$. For fixed b, let D_{ib} be the distance between $\hat{\eta}$ and the projection of $\mathbf{M_b}$. Let

$$d_{ib} = \frac{D_{ib}}{q_2 - q_1},$$

where q_2 and q_1 are estimates of the upper and lower quartiles, respectively. Here, the estimates of the quartiles are based on the ideal fourths (Frigge *et al.*, 1989). The projection distance of $\mathbf{M_b}$ from the center of the bootstrap data cloud is

$$G_b = \max d_{ib},$$

the maximum taken over $i = 1, \cdots, B$. Let $\mathbf{G_0}$ be the depth of the null vector $(0, \cdots, 0)$. Then, based on general results in Liu and Singh (1997), a p-value for testing (11) is

$$\frac{1}{B}\sum I_b,$$

where the indicator function $I_b = 1$ if $G_0 \leq G_b$; otherwise $I_b = 0$. This will be called method WMWAOV. Like method DBH, WMWAOV can perform poorly when tied values can occur. However, for continuous distributions, it was found to perform well in simulations.

Notice that the covariance matrix associated with the bootstrap data cloud can be singular. This is because the C differences, θ_{djk}, for all $j < k$, can be linearly dependent. Consequently, measuring depth using Mahalanobis distance or some robust analog can fail. A practical advantage of the projection method just described is that no covariance matrix is used, and more generally the inverse of a matrix is not required, so this issue does not arise.

5. Simulation Results

Simulations were used to study the small-sample properties of the proposed methods. Method CH is invariant under order preserving transformations of the data. But for methods DBH and WMWAOV, this is not quite the case, and

so simulations results are reported here for four types of distributions: normal, symmetric and heavy-tailed, asymmetric and relatively light-tailed, and asymmetric with relatively heavy tails. Although Cliff's method is invariant under order preserving transformations, changes in scale can affect the Type I error probability, which is an issue that has not been addressed in extant simulations. So a secondary goal is to report results on this issue when $J = 2$.

Data were generated from one of four g-and-h distributions, one of which was standard normal. If Z has a standard normal distribution, then

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

has a g-and-h distribution where g and h are parameters that determine the first four moments. When $g = h = 0$, X has a standard normal distribution. With $g = 0$, this distribution is symmetric, and it becomes increasingly skewed as g gets large. As h gets large, the g-and-h distribution becomes more heavy-tailed. Table 1 shows the skewness (κ_1) and kurtosis (κ_2) for the distributions used in the simulations, which would seem to cover a range of values often found in practice. They correspond to a standard normal ($g = h = 0$), a symmetric heavy-tailed distribution ($h = 0.2$, $g = 0.0$), an asymmetric distribution with relatively light tails ($h = 0.0$, $g = 0.2$), and an asymmetric distribution with heavy tails ($g = h = 0.2$). (Additional properties of the g-and-h distribution are summarized by Hoaglin, 1985.)

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

g	**h**	$\boldsymbol{\kappa_1}$	$\boldsymbol{\kappa_2}$
0.0	0.0	0.00	3.0
0.0	0.2	0.00	21.46
0.2	0.0	1.75	8.9
0.2	0.2	2.81	155.99

To gain some information about the effect of different scales, simulations were run where observations in the first group were multiplied by $\sigma_1 = 4$. For brevity, this will be called the heteroscedastic case. Note than when dealing with skewed distributions, changing the scale in this manner alters, for example, the value of the parameter θ_{d12}. In particular, the null hypothesis given by (11) is no longer true. Accordingly, observations in the first group were shifted so that the null hypothesis is true. This was done by first estimating the actual value using simulations based on 20,000 replications. That is, compute $\hat{\theta}_{d12}$ based on the sample size being used, repeat this 20,000 times, and take the mean of the

results, say $\tilde{\theta}_{d12}$, as the true value of θ_{d12}. Then, after generating data from a particular g-and-h distribution, replace X_{i1} with $X_{i1} - \tilde{\theta}_{d12}$.

Table 2 shows the estimated Type I error probability when testing at the 0.05 level, based on 2,000 replications, for the two sample case with sample sizes $n_1 = n_2 = 20$. As is evident, $\sigma_1 = 1$ corresponds to the homoscedastic case and $\sigma_1 = 4$ is the heteroscedastic case. As can be seen, there is little separating the two methods. The main difference is that in all cases, the level of method WMWAOV is greater than the level of Cliff's methods. For the heteroscedastic case where $g = h = 0.2$, the estimated level of method WMWAOV is 0.60 and for Cliff's method it is 0.032. This suggests that WMWAOV has a power advantage in this particular case. To explore the extent this is true, simulations were run with 1.5 subtracted from each observation in the first group. The power of WMWAVO and Cliff's method were estimated to be 0.336 and 0.268, respectively.

Table 2: Estimated Type I error probabilities, $J = 2$, $n_1 = n_2 = 20$

g	h	σ_1	WMWAOV	CLIFF
0.0	0.0	1	0.054	0.042
		4	0.066	0.048
0.0	0.2	1	0.051	0.042
		4	0.059	0.055
0.2	0.0	1	0.053	0.042
		4	0.064	0.039
0.2	0.2	1	0.050	0.042
		4	0.060	0.032

Table 3 shows $\hat{\alpha}$, an estimate of the Type I error probability when $J = 4$ groups are compared with methods DBH, CH and WMWAOV. Three variance patterns were considered where the observations in the jth groups are multiplied by σ_j. The three choices were $(\sigma_1, \cdots, \sigma_4) = (1, 1, 1, 1)$, (1, 1, 1, 4) and (4, 1, 1, 1). For convenience, these variance patterns are labeled VP 1, VP 2 and VP 3, respectively. Of course, for equal sample sizes, there is no practical difference between the latter two variance patterns, and so for this special case, only results for VP 2 are reported. Table 4 reports the results when unequal sample sizes are used.

Clark *et al.* (2009) report results on a study dealing with the effectiveness of lifestyle intervention strategies for adults aged 60 and older. A portion of the study dealt with comparing groups of participants based on measures of physical wellbeing. Here, results for the physical composite variable are reported, which was chosen in part because there are no tied values. The three groups that are compared differ in terms of amount of treatment received. The sample sizes

Table 3: Estimated Type I error probabilities, $J = 4$ groups, $(n_1, n_2, n_3, n_4) = (20, 20, 20, 20)$, $\alpha = 0.05$

g	h	VP	DBH	CH	WMWAOV
0.0	0.0	1	0.060	0.029	0.048
		2	0.061	0.026	0.051
0.0	0.2	1	0.060	0.029	0.034
		2	0.058	0.025	0.037
0.2	0.0	1	0.061	0.029	0.046
		2	0.059	0.027	0.053
0.2	0.2	1	0.060	0.029	0.044
		2	0.059	0.025	0.033

Table 4: Estimated Type I error probabilities, $J = 4$ groups, $(n_1, n_2, n_3, n_4) = (20, 20, 30, 40)$

g	h	VP	DBH	CH	WMWAOV
0.0	0.0	1	0.059	0.036	0.055
		2	0.057	0.030	0.062
		3	0.060	0.026	0.055
0.0	0.2	1	0.061	0.036	0.055
		2	0.058	0.030	0.051
		3	0.061	0.026	0.060
0.2	0.0	1	0.061	0.036	0.060
		2	0.060	0.030	0.063
		3	0.058	0.029	0.064
0.2	0.2	1	0.061	0.029	0.056
		2	0.060	0.030	0.049
		3	0.061	0.026	0.052

were 67, 64, and 41. The output from the R function cidM, designed to perform method DBH, using the default $\alpha = 0.05$ for the probability of one or more Type I errors, was

```
$test
    Group Group p-value     p.crit    P(X<Y) P(X=Y)    P(X>Y)     p.hat
        1     2   0.006 0.01666667 0.6408434      0 0.3591566 0.6408434
        1     3   0.018 0.02500000 0.6286858      0 0.3713142 0.6286858
        2     3   0.490 0.05000000 0.4630275      0 0.5369725 0.4630275
```

The column headed by p.hat indicates the estimate of p. So at the 0.05 level, for groups 1 and 2, as well as 1 and 3, H_0: $p = 0.5$ is rejected. (The R functions used here are available from the author upon request.)

A portion of the output from the R function cidmulv2 (method CH), again testing at the 0.05 level, is

Group	Group	p.hat	p.ci.lower	p.ci.uppper	p-value	p.crit
1	2	0.6408434	0.5413586	0.7295302	0.006	0.01666667
1	3	0.6286858	0.5171831	0.7279818	0.024	0.02500000
2	3	0.4630275	0.3536995	0.5760297	0.530	0.05000000

So again a significant difference is found for groups 1 and 2, as well as 1 and 3. Note, however, that for $\alpha = 0.04$, method CH no longer rejects when comparing groups 1 and 3, but again method DBH rejects, the only point being that the choice of method can make a difference. In fairness, although DBH tends to have more power when there are no tied values, situations can be constructed where CH rejects and DBH does not.

Another portion of the study compared five groups based on a physical composite score. No differences were found using any of the methods in this paper. But it is interesting to note that if 2.5 is added to every observation in the first group, method WMWAOV rejects at the 0.05 level, but methods CH and DBH find no differences. This merely illustrates that it is possible for the global test to find a true difference that is missed using methods CH and DBH. (The R function wmwaov, available from the author, performs method WMWAOV.)

6. Concluding Remarks

In summary, the results reported here strongly indicate that method SMM should be abandoned in favor of method CH. When dealing with continuous distributions, the results indicate that method DBH has a practical advantage over method CH in terms of power. Perhaps this is because the actual Type I error probability associated with DBH is a bit higher than it is with method CH. That is, if an adjustment could be made so that both methods are more level robust, the choice of method might be immaterial in terms of power. When dealing with tied values, all indications are that method CH is best for general use.

Although situations can be found where method WMWAOV rejects when methods CH and DBH do not reject, perhaps such situations are rare in practice. However, it seems prudent to consider the possibility that this event can occur, which is the main reason for including it in this study.

Finally, to stress a point made in the introduction, it is not being suggested that p be used exclusively when studying the differences among groups. Indeed, it seems that multiple perspectives can be useful. For example, there might be situations where the lower quantiles of two distributions differ substantially while the upper quantiles do not. That is, subsets of participants might respond differently to a particular treatment, an issue that can be addressed using the shift function developed by Doksum and Sievers (1976).

References

Acion, L., Peterson, J. J., Temple, S. and Arndt, S. (2006). Probabilistic index: An intuitive non-parametric approach to measuring the size of treatment effects. *Statistics in Medicine* **25**, 591-602.

Brunner, E., Dette, H. and Munk, A. (1997). Box-type approximations in nonparametric factorial designs. *Journal of the American Statistical Association* **92**, 1494-1502.

Brunner, E. and Munzel, U. (2000). The nonparametric Behrens-Fisher problem: asymptotic theory and small-sample approximation. *Biometrical Journal* **42**, 17-25.

Clark, F., Jackson, J., Mandel, D., Blanchard, J., Carlson, M., Azen, S. *et al.* (2009). Confronting challenges in intervention research with ethnically diverse older adults: the USC Well Elderly II trial. *Clinical Trials* **6**, 90-101.

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

Doksum K. A. and Sievers G.L. (1976). Plotting with confidence: graphical comparisons of two populations. *Biometrika* **63**, 421-434.

Donoho, D. L. and Gasko, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Annals of Statistics* **20**, 1803-1827.

Fligner, M. A. and Policello II, G. E. (1981). Robust rank procedures for the Behrens-Fisher problem. *Journal of the American Statistical Association* **76**, 162-168.

Frigge, M., Hoaglin, D. C. and Iglewicz, B. (1989). Some implementations of the Boxplot. *American Statistician* **43**, 50-54.

Grissom, R. J. (1994). Probability of the superior outcome of one treatment over another. *Journal of Applied Psychology* **79**, 314-316.

Grissom, R. J. and Kim, J. J. (2005). *Effect Sizes for Research: A Broad Practical Approach.* Erlbaum, Mahwah.

Hettmansperger, T. P. (1984). *Statistical Inference Based on Ranks.* Wiley, New York.

Hoaglin, D. C. (1985). Summarizing shape numerically: The g-and-h distribution. In D. Hoaglin, F. Mosteller and J. Tukey (Eds.) *Exploring Data Tables Trends and Shapes*. Wiley, New York.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75**, 800-802.

Kraemer, H. C. and Kupfer, D. J. (2006). Size of treatment effects and their importance to clinical research and practice. *Biological Psychiatry* **59**, 990-996.

Liu, R. G. and Singh, K. (1997). Notions of limiting P values based on data depth and bootstrap. *Journal of the American Statistical Association* **92**, 266-277.

Mee, R. W. (1990). Confidence intervals for probabilities and tolerance regions based on a generalization of the Mann-Whitney statistic. *Journal of the American Statistical Association* **85**, 793-800.

Neuäuser M, Lösch, C. and Jöckel, K-H. (2007). The Chen-Luo test in case of heteroscedasticity. *Computational Statistics and Data Analysis* **51**, 5055-5060.

Reiczigel, J. Zakariás, I. and Rózsa, L. (2005). A bootstrap test of stochastic equality of two populations. *American Statistician* **59**, 156-161.

Rosner, B. and Glynn, R. J. (2009). Power and sample size estimation for the Wilcoxon rank sum test with application to comparisons of C statistics from alternative prediction models. *Biometrics* **65**, 188-197.

Rust, S. W. and Fligner, M. A. (1984). A modification of the Kruskal-Wallis statistic for the generalized Behrens-Fisher problem. *Communications in Statistics - Theory and Methods* **13**, 2013-2027.

Vargha, A. and Delaney, H. D. (2000). A critique and improvement of the CL common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics* **25**, 101-132.

Wilcox, R. R. (2003). *Applying Contemporary Statistical Techniques*. Academic Press, New York.

Wilcox, R. R. (2005) *Introduction to Robust Estimation and Hypothesis Testing*. Academic Press, New York.

Wilcox R. R. (2006) Comparing medians. *Computational Statistics and Data Analysis* **51**, 1934-1943.

Received June 22, 2010; accepted December 20, 2010.

Rand R. Wilcox
Department of Psychology
University of Southern California
Los Angeles, CA 90089-1061, USA
rwilcox@usc.edu