ISSN: 2469-5831

Wilcox. Int J Clin Biostat Biom 2017, 3:015
DOI: 10.23937/2469-5831/1510015
Volume 3 | Issue 1
Open Access

International Journal of
Clinical Biostatistics and Biometrics

RESEARCH ARTICLE

Analogs of the Wilcoxon-Mann-Whitney Test When There is a Covariate

Rand R Wilcox*

Department of Psychology, University of Southern California, USA

***Corresponding author:** *Rand R Wilcox, Department of Psychology, University of Southern California, USA, E-mail: rwilcox@usc.edu*

Abstract

For two independent random variables $Y_j(j = 1, 2)$, let $p(x) = P(Y1 < Y2 | X = x)$, where X is some covariate of interest. For m values of the covariate, $x_1, \ldots, x_m$, the paper deals with the goal of testing $H_0 : p(x_j) = 0.5\ (j = 1, \ldots, m)$ in a manner that controls Family Wise Error Rate (FWE), the probability of one or more Type I errors. If m is relatively small, extant multiple comparison methods can be used to control FWE. But if m is relatively large, the actual level can be substantially smaller than the nominal level, raising concerns about relatively poor power. The paper describes a method for addressing this issue when $p(x)$ is estimated via a running interval smoother.

Keywords

Effect size, Smoothers, Multiple comparisons

Introduction

For two independent groups, let $Y_j\ (j = 1, 2)$ be some random variable associated with the j^{th} group. As is evident, one approach to comparing these groups is in terms of some measure of location. As is well known, there is vast literature regarding how this might be done. Another approach is to focus on

$$p = P(Y_1 < Y_2) \tag{1}$$

the probability that a randomly sampled observation from the first group is less than a randomly sampled observation from the second group. In the event tied values can occur, (1) is replaced with

$$p = P(Y_1 < Y_2) + 0.5P(Y_1 = Y_2) \tag{2}$$

Cliff [1], Acion, et al. [2], Kraemer and Kupfer [3], and Vargha and Delaney [4], among others, summarize arguments for focusing on p when comparing groups. Indeed, it seems fairly evident that p provides a useful perspective.

Certainly the best-known method for making inferences about p is the Wilcoxon-Mann-Whitney (WMW) test. It is well known, however, that under general conditions, the WMW method uses an incorrect estimate of the standard error. Numerous methods have been derived for dealing with this issue. Wilcox [5] summarizes the relevant literature.

This paper is focused on comparing two independent groups based on p when there is a covariate, say X. Let $p(x)$ denote the value of p given that $X = x$. Here, no particular parametric model is assumed regarding the nature of $p(x)$. Rather, $p(x)$ is estimated with a particular nonparametric regression estimator, which is described in section 3. Let $p_k(x_k) = P(Y_1 < Y_2 | X = x_k)\ (k = 1, \ldots, m)$, where $x_1, \ldots, x_m$ are m covariate values to be determined. The basic goal is to test

$$H_0 : p_k(x_k) = 0.5 \tag{3}$$

for each $k = 1, \ldots, m$ such that the family wise error rate (FWE), meaning the probability of one or more Type I errors, is equal to some specified value, α. If the number of covariate values, m, is reasonably small, one can simply proceed along the lines in Wilcox [5]. A concern, however, is that important differences might be missed due to using too few covariate values. A guess is that a method in Wilcox [5] is readily adapted to the situation at hand when dealing with a large number of covariate values, but preliminary simulations made it clear that this approach is unsatisfactory. The actual FWE can

Citation: Wilcox RR (2017) Analogs of the Wilcoxon-Mann-Whitney Test When There is a Covariate. Int J Clin Biostat Biom 3:015. doi.org/10.23937/2469-5831/1510015

Received: September 19, 2017: **Accepted:** December 20, 2017: **Published:** December 22, 2017

be substantially smaller than the nominal level.

The goal is this paper is suggesting a modification of the method in Wilcox [5] that performs better in simulations. The proposed method is based in part on Cliff's [1] method for testing (1). Cliff's method was chosen based on results in Neuhauser, et al. [6] where several techniques were compared [7]. Particularly important here, for reasons made clear in section 2, is that Cliff's method has been found to perform relatively well even with sample sizes as small as eight.

The paper is organized as follows. Section 2 reviews Cliff's method. Section 3 describes the proposed method and section 4 reports simulation results. Section 5 illustrates the method using data from a study dealing with the emotional and physical well being of older adults.

Cliff's Method

Momentarily ignoring the covariate, Cliff's method for making inferences about p is applied as follows. Let $Y_{ij}\left(i = 1, \ldots, n_j; j = 1, 2\right)$ be a random sample of n_j observations from the j^{th} group. Let

$$d_{ih} = \begin{Bmatrix} -1 \; if \; Y_{i1} < Y_{h2} \\ 0 \; if \; Y_{i1} = Y_{h2} \\ 1 \; if \; Y_{i1} > Y_{h2} \end{Bmatrix}$$

An estimate of $\delta = P\left(Y_{i1} > Y_{i2}\right) - P\left(Y_{i1} < Y_{i2}\right)$ is

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

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

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

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} \left(d_{ih} - \hat{\delta}\right)^2,$$

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

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

Then $\hat{\sigma}^2 = \frac{\left(n_1 - 1\right) s_1^2 + \left(n_2 - 1\right) s_2^2 + \tilde{\sigma}^2}{n_1 n_2}$ 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, Cliff [1] recommends

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

Cliff's confidence interval for δ is readily modified to give a confidence for p. Letting

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

and $C_u \frac{\hat{\delta} - \hat{\delta}^3 + z\hat{\sigma}\sqrt{\left(1 - \hat{\delta}_2\right)^2 + z^2\hat{\sigma}^2}}{1 - \hat{\delta}^2 + z^2\hat{\sigma}^2}$

$a1 - \alpha$ confidence interval for p is

$$\left(\frac{1 - C_u}{2}, \frac{1 - C_l}{2}\right).$$

The Proposed Method

Now let $\left(Y_{ij}, X_{ij}\right)$ $\left(i = 1, \ldots, n_j; j = 1, 2\right)$ be a random sample with the covariate, X, included. The proposed method is based in part on a basic component of the running interval smoother, which has been studied extensively [5]. Momentarily focus on a single value of the covariate x, in which case the goal is to test $H_0: p(x) = 0.5$. The basic strategy is quite simple: Compute a confidence interval for p based on the Y_{ij} values such X_{ij} is close to x.

The precise details are as follows. For the j^{th} group, let M_j be the usual sample median and compute the median absolute deviation estimator MAD_j, which is the median based on

$\left|X_{1j} - M_j\right|, \ldots \left|X_{n_j j} - M_j\right|. \; Let \; MADN_j = MAD_j / 0.6745$. Under normality, $MADN_j$ estimates the standard deviation. Let

$$N_j(x) = \left\{i : \left|X_{ij} - x\right| \le f \times MADN_j\right\},$$

where the constant f, called the span, is to be determined. That is, for fixed j, $N_j(x)$ indexes the observed covariate values that are close to x. The X_{ij} values for which $i \in N_j(x)$ are called the nearest neighbors. Let $\hat{p}(x)$ be the estimate of $p(x)$ based on the Y_{ij} values for which $i \in N_j(x)$. Choices for the span that generally perform well are $f = 0.8 \; or \; 1$ (e.g., Wilcox) [5]. Here $f = 1$ is used. As is evident, $H_0: p(x) = 0.5$ can be tested simply by applying Cliff's methods using the Y_{ij} values for which $i \in N_j(x)$.

There remains the problem of choosing the covariate values. In some situations there might be substantive reasons for using particular values. But otherwise it is clearly prudent to choose a reasonably wide range of covariate values. Here, two approaches are described and there relative merits are discussed in section 4.

To describe the first approach, let $N_j(x)$ denote the cardinality of the set $N_j(x)$. The basic strategy is to focus on covariate values where $Nj(x) \ge \eta$, where η is some constant to be determined. That is, focus on situations where the sample sizes are sufficiently large when applying Cliff's method. Given η, let $z_{\ell j}$ be the smallest X_{ij} value such that simultaneously $N_j(X_{i1}) \ge \eta$ *and* $N_j(X_{i2}) \ge \eta$. In a similar manner, let z_{uj} be the largest X_{ij} value such that simultaneously $N_j(X_{i1}) \ge \eta$ *and* $N_j(X_{i2}) \ge \eta$. Let $z_\ell = max\{z_{\ell 1}, z_{\ell 2}\}$ *and* $z_u = min\{z_{u1}, z_{u2}\}$. The covariate values are taken to be the m values evenly spaced between z_ℓ and z_u, inclusive, which are labeled $x_1, \ldots, x_m$ c. Here, $m = 25$ is used. Extant results suggest that Cliff's method performs reasonably well when $\eta \ge 8$, *so* $\eta = 8$ is used henceforth unless stated otherwise. This approach to choosing the covariate values will be called method S henceforth.

The second approach to choosing the covariate values is as follows. Let $\hat{\gamma}q_j$ be an estimate of the q^{th} quantile of the covariate associated with the j^{th} group, 0.5. Let $w_\ell = max\{\hat{\gamma}_{q1}, \hat{\gamma}_{q2}\}$ and $w_u = min\{\hat{\gamma}_1 - q_1, \hat{\gamma}_1 - q_2\}$. The covariate values are taken to be m values evenly spaced between w_ℓ and w_u. Here, both q = 0.1 and 0.05 are considered. This alternative approach to choosing the covariate values will be called method Q henceforth. The relative merits of methods S and Q are summarized in section 6.

Controlling FWE

There remains the issue of controlling FWE. There is a wide range of techniques that might be used (e.g., Wilcox [5]). One possibility is to use the sequentially rejective technique derived by Hochberg [8], which has been found to perform well in simulations when m = 5. But as m increases, this approach results in FWE levels well below the nominal level. A similar concern arises using a critical value based on Studentized maximum modulus distribution (with infinite degrees of freedom) as well the method derived by Hommel [9]. This is not surprising for the following reason. Note that in various situations, the set $N_j(x_k) \cap N_j(x_\ell)$, $k \ne \ell$, will not be empty, in which case the test statistics corresponding to these two covariate values will be correlated. In terms of controlling FWE, what is needed is a method that takes this into account. Results in Wilcox [5] suggest how to proceed. The basic idea is to determine a critical p-value, p_α, when both Y and X have a normal distribution and there is no association between Y and X. Then simulations are used to check on the impact of non-normality as well as situations where there is an association between Y and X.

To be a bit more precise, let p_k be the p-value when testing $H_0: p(x_k) = 0.5$ $(k = 1, \ldots, m)$. Let p_{min} denote $min\{p1, \ldots, pm\}$ and let p_α denote the α quantile of p_{min}. So for any k, $H_0: p(x_k) = 0.5$ is rejected when $p_k \le p_\alpha$, in which case FWE will be equal to α. So the strategy is to use simulations to determine p_α when both X and Y are independent and both have a standard normal distribution. Then simulations are used to determine the impact on FWE when X and Y are dependent and when Y has a non-normal distribution.

Table 1 shows estimates of p_α when using method S, m = 25 and α = 0.05 and there is a common sample size n ranging between 30 and 800. (The same p_α values are used by method Q). These estimates are based on 2000 replications. When n is small, execution time is reasonably low, but as n increases, execution time becomes an issue. Note that generally the estimates of p_α decrease as n increases Overall, the rate of the decrease is very small, particularly for $n \ge 200$. Here, for sample sizes not included in Table 1, Cleveland's [10] smoother is used to estimate p_α based on 1/n and the values in Table 1. (The R function l plot. pred in Wilcox, 2017, is used here) [5] For unequal sample sizes, p_α is determined for both n_1 and n_2 and the results are averaged.

Simulation Results

This section reports simulation results on the ability of the methods in the previous section to control FWE under non-normality and when there is an association between Y and X.

Data were generated based on

$$Y = X^a + \in, \tag{6}$$

where is some random variable having a median of zero and $a = 1$ *or* 2. The distribution for the error term, $\in$, was taken to be a one of four g-and-h distributions [11] one of which is the standard normal distribution. If Z has a standard normal distribution, then by definition

$$V = \begin{cases} \dfrac{(exp(g Z) - 1)}{g} exp\left(hZ^2/2\right), & \text{if } g > 0 \\ Zexp\left(hZ^2/2\right), & \text{if } g = 0 \end{cases}$$

has a g-and-h distribution where g and h are parameters that determine the first four moments. The four distributions used here were the standard normal $(g = h = 0.0)$, a symmetric heavy-tailed distribution $(h = 0.2, g = 0.0)$, an asymmetric distribution with rel-

Table 1: Estimates of p_α, α = 0.05, when m = 25.

n:	30	50	60	70	80	100	150
$p_{0.05}$:	0.008566	0.008385	0.006758	0.006871	0.006157	0.006629	0.006629
n:	200	300	400	500	600	800	
$p_{0.05}$:	0.00468	0.004536	0.004953	0.004294	0.004288	0.004148	

N = Sample Size; $p_{0.05}$: Critical 0.05 p-value.

Table 2: Some properties of the g-and-h distribution.

g	h	κ_1	κ_2
0	0	0	3
0	0.2	0	21.46
0.2	0	0.61	3.68
0.2	0.2	2.81	155.98

κ_1 = kappa_1 = skewness; κ_2 = kappa_2 = kurtosis.

Table 3: Estimates of FWE when testing at the α = 0.05 and using Method S to choose the covariate values.

g	h	n_1	n_2	a = 1	a = 2
0	0	30	30	0.048	0.067
0	0.2	30	30	0.042	0.063
0.2	0	30	30	0.047	0.063
0.2	0.2	30	30	0.052	0.061
0	0	30	60	0.048	0.079
0	0.2	30	60	0.051	0.057
0.2	0	30	60	0.053	0.06
0.2	0.2	30	60	0.052	0.047
0	0	150	200	0.065	0.11
0	0.2	150	200	0.062	0.097
0.2	0	150	200	0.065	0.106
0.2	0.2	150	200	0.064	0.099

Table 4: Estimates of FWE when testing at the α = 0.05 and using Method Q to choose the covariate values.

g	h	n_1	n_2	a = 1	a = 2
0	0	30	30	0.038	0.066
0	0.2	30	30	0.042	0.063
0.2	0	30	30	0.04	0.063
0.2	0.2	30	30	0.039	0.062
0	0	30	60	0.036	0.048
0	0.2	30	60	0.034	0.047
0.2	0	30	60	0.035	0.047
0.2	0.2	30	60	0.034	0.047
0	0	150	200	0.021	0.036
0	0.2	150	200	0.02	0.031
0.2	0	150	200	0.021	0.034
0.2	0.2	150	200	0.021	0.03

g = skewness parameter for the g-and-h distribution; h = kurtosis parameter for the g-and-h distribution; n_1 = first sample size; n_2 = second sample size; a = 1 = a straight regression line, a = 2 = a quadratic regression line.

atively light tails $(h = 0.0,\ g = 0.2)$, and an asymmetric distribution with heavy tails $(g = h = 0.2)$. Table 2 shows the skewness (κ_1) and kurtosis (κ_2) for each distribution. Figure 1 shows plots of these distributions. Additional properties of the g-and-h distribution are summarized by Hoaglin [11]. The results reported here are for situations where the distribution of X was taken to be standard normal. A few simulations were run where X and $\in$ have the same g-and-h distribution, no new insights were found, so the results are not reported.

It is noted that if there is no covariate, transforming to some g-and-h distribution does not alter the results based on Cliff's method because it depends only on the ranks of the data. But when there is a covariate and there is an association between Y and X, this is no longer the case. So an issue understands the impact on FWE when there is an association.

Table 3 shows the results for a = 1 and 2 and sample sizes (n_1, n_2) = (30, 30), (30, 60) and (150, 200), where the covariate values were chosen based on method S. Again 2000 replications were used. Although the importance of a Type I error depends on the situation, Bradley [12] suggested that as a general guide, when testing at the 0.05 level, the actual level should be between 0.025 and 0.075. Based on this criterion, the proposed method is satisfactory for all of the situations considered when a = 1. However, when a = 2, this is no longer the case. For (n_1, n_2) = (30, 60) and $g = h = 0$, the estimate is 0.079. For (n_1, n_2) = (150, 200), estimates are approximately equal to 0.1. For sample sizes (n_1, n_2) = (40, 100), and (n_1, n_2) = (50, 100), not shown in Table 3, again estimates exceeded 0.075. Lowering the span to 0.8 and even 0.6 does not correct this problem. Using η = 14 was found to be unsatisfactory as well. A closer examination of the simulation results revealed that when a = 2, inferences based on the more extreme covariate values result in FWE values greater than the nominal level. This motivated the second approach to choosing the covariate values, method Q in the previous section.

Some additional simulations were run where $Y = I_{x<0}\ X + \in$, where the indicator function $I_{X>0} = 1\ if\ X > 0$; otherwise $I_{X>0} = 0$. The results were very similar to those where a = 1. So there are indications that method S can perform reasonably well when the regression line is not straight. But if there is sufficient curvature for the more extreme covariate values, this is no longer the case.

Table 4 shows the simulation results when using method Q. As can be seen, now the largest estimate is 0.066. Note that given the sample sizes, altering the distribution of the error term has an even smaller impact on the estimate of FWE compared to method S. The main difficulty is that the estimate drops below 0.025 in some situations, the lowest estimate being 0.020. For a = 2 and (n_1, n_2) = (600, 600), the estimates are nearly identical to those when (n_1, n_2) = (150, 200). For (n_1, n_2) = (50, 600), $g = h = 0$ and a = 2 the estimate is 0.050. Simulations were also run using q = 0.05. The estimates differed from those in Table 4 by at most three units in the third decimal place.

An Illustration

Data from the Well Elderly 2 study [13] are used to illustrate the proposed method. A general goal in the Well Elderly 2 study was to assess the efficacy of an intervention strategy aimed at improving the physical and emotional health of older adults. (The data are available at http://www.icpsr.umich.edu/icpsrweb/landing.

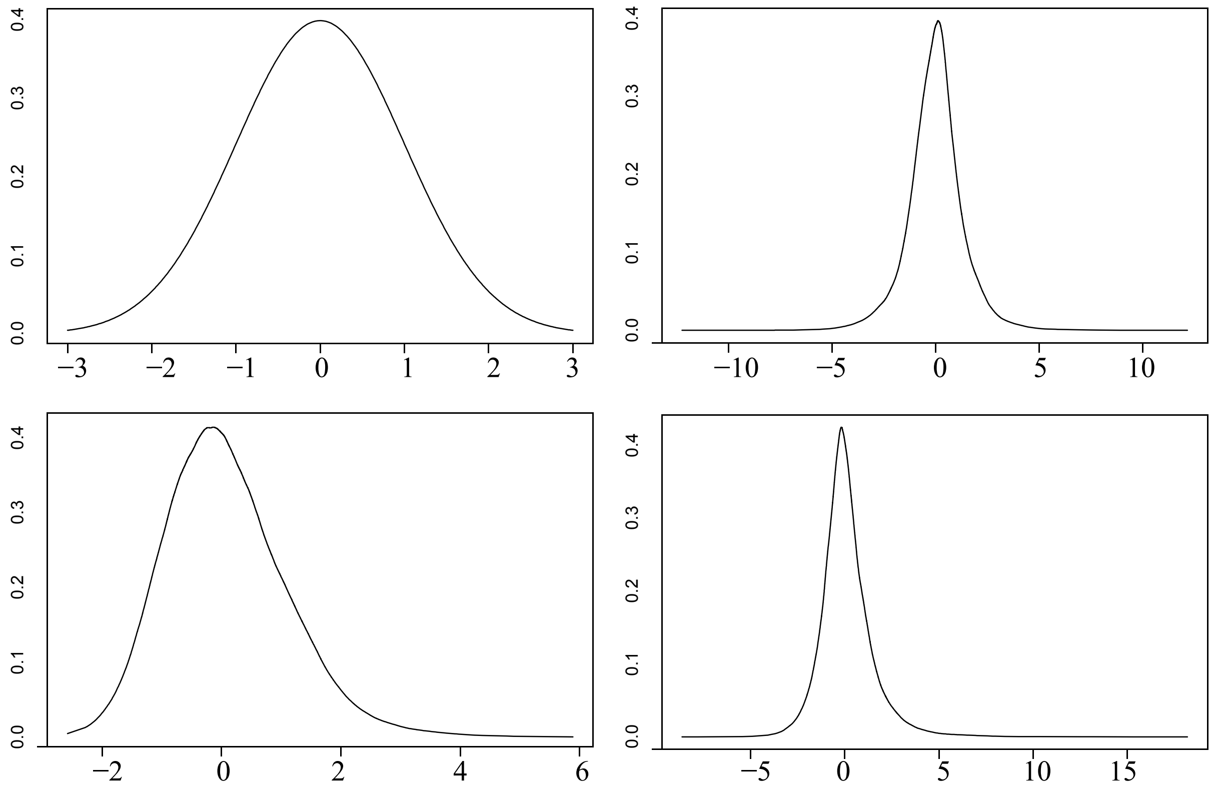

Figure 1: The four distributions used in the simulations. The upper left is (g, h) = (0.0, 0.0) (standard normal), the upper right is (g, h) = (0.0, 0.2), the lower left is (g, h) = (0.2, 0.0) and the lower right is (g, h) = (0.2, 0.2).

Y-axis = likelihood; X-axis = f(x). f (x) is the probability density function.

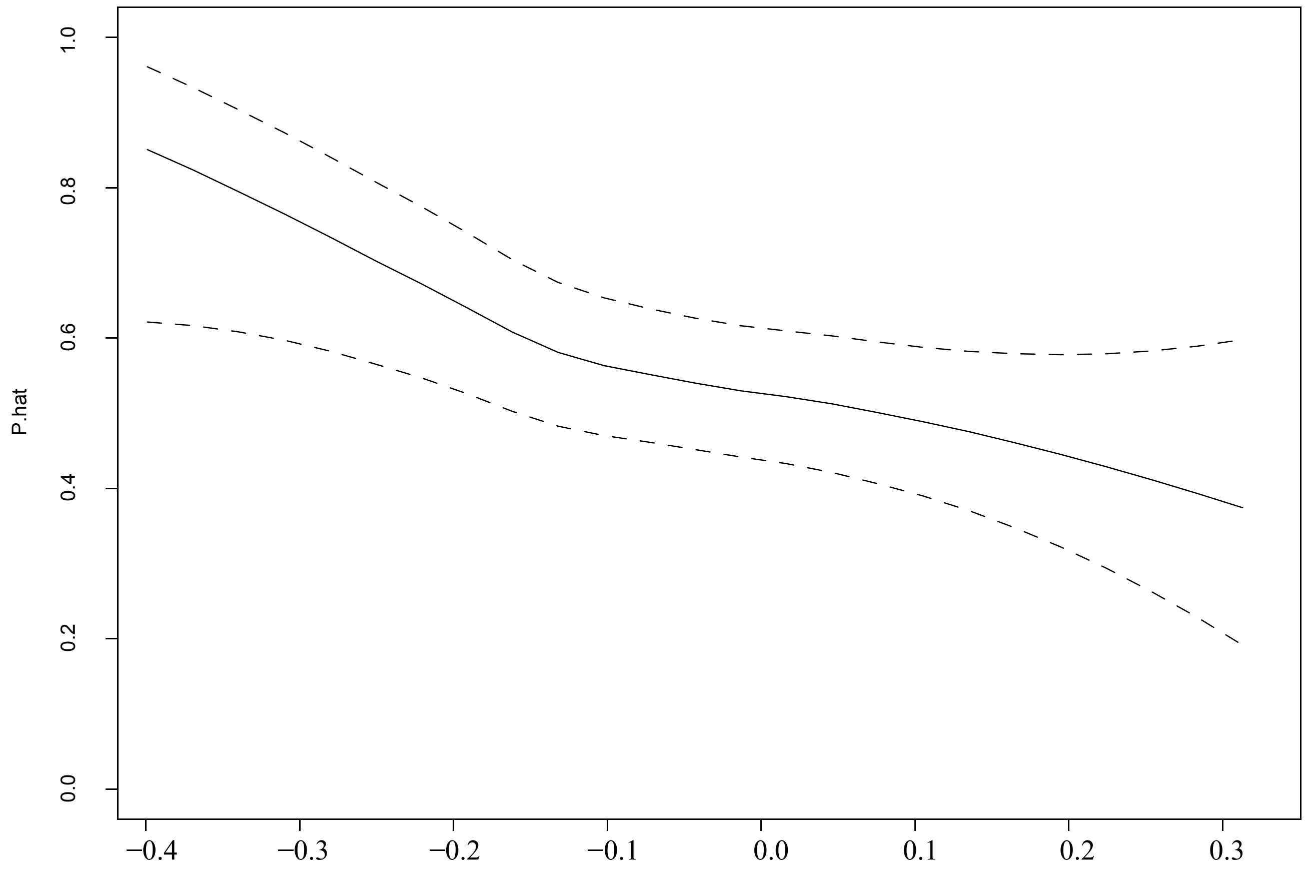

Figure 2: The solid line indicates the estimated probability that a randomly sampled individual from the control group will have a lower MAPA score than a randomly sampled participant in the experimental group. The dashed lines indicate a confidence band having, approximately, simultaneous probability coverage 0.95.

Y-axis = likelihood; X-axis = f(x). f (x) is the probability density function.

jsp). A portion of the study was aimed at understanding the impact of intervention on a Measure of Meaningful Activities (MAPA). A covariate of interest was the Cortisol Awakening Response (CAR), which is defined as the change in cortisol concentration that occurs during the first 30-45 minutes after waking from sleep. (CAR is taken to be the cortisol level upon awakening minus the level of cortisol after the participants were awake). Extant studies [14,15] indicate that the CAR is associated with various measures of stress. Here, a control group is compared to a group that received intervention.

Figure 2 shows the estimate of $P(Y_1 < Y_2 | X)$, where Y_1 is the MAPA score before intervention, Y_2 is MAPA after intervention, and X is CAR. (Leverage points, outliers among the CAR values, were removed). Selecting covariate points via method S, significant differences are found for CAR ranging between -0.36 and -0.27. Plots of the regression lines, using Cleveland's [10] method, suggested that there is little or no curvature, which in turn suggests that method S controls FWE reasonably well. So there is an indication that when the CAR is sufficiently negative (cortisol increases after awakening), MAPA scores tend to be higher for the group receiving intervention. The covariate values used by method S range between -0.38 and 0.32. In contrast, when using method Q with q = 0.1, they range between -0.19 and 0.15 and no significant results are found. For q = 0.05 the covariate values range between -0.27 and 0.22 with a single significant result when the CAR is equal to -0.24.

Concluding Remarks

In summary, all indications are that method S for choosing the covariate values performs reasonably well except when there is a sufficient amount of curvature near the extreme ends of the covariate. Method Q avoids FWE well above the nominal level in situations where method S breaks down. But method S has the potential of providing comparisons for a wider range of covariate values. As was illustrated, this can make practical difference.

A criticism of method S is that there is no formal method for justifying the assumption that curvature among the more extreme covariate values is not an issue. For now, the best that can be done is to inspect the plot returned by some nonparametric regression estimator. So an argument for using method Q might be that it is safer in terms of controlling FWE.

Finally, R functions for applying methods S and Q are available at Dornsife.usc.edu/cf/labs/wilcox/wilcox-faculty-display.cfm and are stored in the file Rallfun- v34. The function ancdetwmw applies method S and ancdetwmwQ applies method Q.

References

1. Cliff N (1996) Ordinal Methods for Behavioral Data Analysis. Mahwah, NJ: Erlbaum.
2. Acion L, Peterson JJ, Temple S, Arndt S (2006) Probabilistic index: An intuitive non-parametric approach to measuring the size of treatment effects. Stat Med 25: 591-602.
3. Kraemer HC, Kupfer DJ (2006) Size of treatment effects and their importance to clinical research and practice. Biol Psychiatry 59: 990-996.
4. Vargha A, Delaney HD (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.
5. Wilcox RR (2017) Introduction to Robust Estimation and Hypothesis Testing. (4th edn), Academic Press, San Diego, CA.
6. Neuhauser M, Laosch C, Jockel KH (2007) The Chen-Luo test in case of het-eroscedasticity. Computational Statistics & Data Analysis 51: 5055-5060.
7. Ruscio J, Mullen T (2012) Confidence intervals for the probability of superiority effect size measure and the area under a receiver operating characteristic curve. Multivariate Behav Res 47: 201-223.
8. Hochberg Y (1988) A sharper Bonferroni procedure for multiple tests of significance. Biometrika 75: 800-802.
9. Hommel G (1988) A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika 75: 383-386.
10. Cleveland WS (1979) Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association 74: 829-836.
11. Hoaglin DC (1985) Summarizing shape numerically: The g-and-h distribution. In: D Hoaglin, F Mosteller, J Tukey, Exploring Data Tables Trends and Shapes. Wiley, New York, 461-515.
12. Bradley JV (1978) Robustness? British Journal of Mathematical and Statistical Psychology 31: 144-151.
13. Clark F, Jackson J, Carlson M, Chou CP, Cherry BJ, et al. (2011) Effectiveness of a lifestyle intervention in promoting the well-being of independently living older people: Results of the Well Elderly 2 Randomise Controlled Trial. J Epidemiol Community Health 66: 782-790.
14. Chida Y, Steptoe A (2009) Cortisol awakening response and psychosocial factors: A systematic review and meta-analysis. Biol Psychol 80: 265-278.
15. Clow A, Thorn L, Evans P, Hucklebridge F (2004) The awakening cortisol response: Methodological issues and significance. Stress 7: 29-37.