Advances in Social Sciences Research Journal – Vol.4, No.20
Publication Date: Oct. 25, 2017
DoI:10.14738/assrj.420.3774.
Multach, M., & Wilcox, R. (2017). Some Results on a Wilcoxon--Mann--Whitney Approach to Interactions in a Two-Way ANOVA Design. *Advances in Social Sciences Research Journal, (420)* 35-39.

Some Results on a Wilcoxon-Mann-Whitney Approach to Interactions in a Two-Way ANOVA Design.

Matthew Multach
Dept. of Psychology
University of Southern California

Rand Wilcox
Dept. of Psychology
University of Southern California

ABSTRACT

There are now several ways of characterizing an interaction in a two-way ANOVA design. For four independent random variables X_j ($j = 1, \cdots, 4$), let $Z = X_1 - X_2$ and $Z^* = X_3 - X_4$. One approach is based on $p = P(Z < Z^*)$, which represents a simple generalization of the Wilcoxon—Mann—Whitney method. Recently, two methods for making inferences about p were derived. One goal in this paper is to report simulation results indicating that both methods can be unsatisfactory when there is heteroscedasticity. The main goal is to describe an alternative approach that performs much better in simulations.

INTRODUCTION

Consider four independent random variables in the context of a two-by-two ANOVA design. There are now a variety of methods aimed at dealing with some notion of an interaction (e.g., De Neve & Thas, 2016; Gao & Alvo, 2005; Wilcox, 2017a). Denoting the random variables by X_j ($j = 1, \cdots, 4$), let $Z = X_1 - X_2$ and $Z^* = X_3 - X_4$. One way of proceeding is to focus on $p = P(Z < Z^*)$. Note that for two groups, the Wilcoxon—Mann—Whitney is based on an estimate of $P(X_1 < X_2)$. So the use of p generalizes the Wilcoxon--Mann—Whitney method in an obvious way.

Consider the goal of testing

$$H_0: p = 1/2. \qquad (1)$$

Recently, two methods for testing (1) were derived and studied via simulations (De Neve and Thas, 2016; Wilcox 2017b). However, extant simulation results do not take into account the possible impact of heteroscedasticity. One goal here is to report new simulation results indicating that both methods can unsatisfactory, in terms of controlling the Type I error probability, when there is heteroscedasticity, particularly when there are unequal sample sizes.

The main goal is to describe an alternative method that performs substantially better in simulations, even when there is homoscedasticity. The methods stemming from both De Neve & Thas (2016) as Wilcox (2017b) are based on the seemingly obvious estimate of p, say $\hat{p}$, which is reviewed in section 2. Wilcox mentioned an alternative estimator ($\tilde{p}$ in section 2), but there are no results on how well this estimator performs for the situation at hand. A seemingly natural speculation is that surely $\hat{p}$ is preferable to $\tilde{p}$, for reasons that will be clear in section 2. But simulation results indicate that $\tilde{p}$ provides adequate control over the Type I error

probability in a range of situations where the methods examined by De Neve and Thas, as well as Wilcox, are unsatisfactory.

The paper is organized as follows. Section 2 describes the methods for testing (1) that are to be compared. Section 3 reports simulation results and section 4 illustrates the new method.

DESCRIPTION OF THE METHODS

Let X_{ij} $(i = 1, \cdots, n_j; j = 1, \cdots, 4)$ be a random sample of size n_j from the jth group. Then an unbiased estimate of p is

$$\hat{p} = \frac{1}{M}\sum\sum\sum\sum I(X_{i1} - X_{j2} < X_{k3} - X_{m4}), \qquad (2)$$

where $M = n_1 n_2 n_3 n_4$, and the indicator function $I(X_{i1} - X_{j2} < X_{k3} - X_{m4})$=1 if $X_{i1} - X_{j2} < X_{k3} - X_{m4}$, otherwise $I(X_{i1} - X_{j2} < X_{k3} - X_{m4})$=1. The De Neve and Thas (2016) method for making inferences about p is based in part on a link function $g(\hat{p})$ that maps the unit interval onto the real line. De Neve and Thas mention two possibilities: $g(x) = x/(1-x)$ and the probit link function $g(x) = \Phi^{-1}(x)$, where $\Phi(x)$ is the standard normal distribution. Let

$$A_1 = \sum_i \left\{ \sum_{j,k,m} I(X_{i1} - X_{j2} < X_{k3} - X_{m4}) - \hat{p} \right\}^2$$

$$A_2 = \sum_j \left\{ \sum_{i,k,m} I(X_{i1} - X_{j2} < X_{k3} - X_{m4}) - \hat{p} \right\}^2$$

$$A_3 = \sum_k \left\{ \sum_{i,j,m} I(X_{i1} - X_{j2} < X_{k3} - X_{m4}) - \hat{p} \right\}^2$$

$$A_4 = \sum_m \left\{ \sum_{i,j,k} I(X_{i1} - X_{j2} < X_{k3} - X_{m4}) - \hat{p} \right\}^2$$

and

$$\hat{\tau}^2 = \left(\frac{g'(\hat{p})}{M}\right)(A_1 + A_2 + A_3 + A_4),$$

where $g'(\hat{p})$ is the derivative of g. Here the focus is on the probit link function, in which case $g'(\hat{p}) = 1/\phi(\Phi^{-1}(\hat{p}))$, where ϕ is the probability density function of a standard normal distribution. When the null hypothesis is true, De Neve and Thas established that asymptotically,

$$T = \frac{g(\hat{p}) - g(0.5)}{\hat{\tau}}$$

has a standard normal distribution. An approximate $1 - \alpha$ confidence interval for p is

$$(g^{-1}\{g(\hat{p}) - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\hat{\tau})\}, g^{-1}\{g(\hat{p}) + \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\hat{\tau})\}).$$

This will be called method NT henceforth.

The basic percentile bootstrap method considered by Wilcox (2017b) is applied as follows. First, generate a bootstrap sample from the jth group by sampling with replacement n_j observations from X_{ij} $(i = 1, \cdots, n_j)$. Based on these bootstrap samples, estimate p using (2) and label the result $\hat{p}^*$. Repeat this process Btimes yielding $\hat{p}_b^*$ $(b = 1, \cdots, B)$. $B = 500$ is used

here, which seems to suffice in many situations in terms of controlling the probability of a Type I error (Wilcox, 2017a). However, a larger choice for B might result in higher power (Racine & MacKinnon, 2007).

Put the $\hat{p}_b^*$ in ascending order yielding $\hat{p}_{(1)}^* \leq \cdots \leq \hat{p}_{(B)}^*$. Let $\ell = \alpha B/2$ and $u = B - \ell$. Then an approximate $1 - \alpha$ confidence interval for p is $\left(\hat{p}_{(\ell+1)}^*, \hat{p}_{(u)}^*\right)$. Let P^* be the proportion of $\hat{p}_b^*$ less than 0.5. Then form Liu & Singh (1997), a p-value when testing (1) is $2\min(P^*, 1 - P^*)$.

As indicated in section 3, NT can be unsatisfactory when the sample sizes are relatively small. Method PB performs reasonably well when the sample sizes are equal, but it can be rather unsatisfactory when the sample sizes are both relatively small and unequal. This suggests a simple modification. Let $N = \min\{n_1, \cdots, n_4\}$. Next, randomly sample, without replacement, N values from each of the four groups and let $\check{p}$ be the resulting estimate of p. Repeat this process L times yielding $\check{p}_1, \cdots, \check{p}_L$ and let

$$\tilde{p} = (\check{p}_1 + \cdots + \check{p}_L)/L.$$

Here, $L = 100$ is used, which was found to generally give an estimate of p that is very similar to $\hat{p}$. Then inferences are made about p using the percentile bootstrap method previously described, except that bootstrap estimates of p are based on $\tilde{p}$ rather than $\hat{p}$. This will be called method PB henceforth.

SIMULATION RESULTS

Simulations were used as a partial check on the small-sample properties of methods NT and PB. Simulation estimates of the actual Type I error probability, when testing at the 0.05 level, are based on 4000 replications. The sample sizes considered were $(n_1, n_2, n_3, n_4) =$ (10,10,10,10), (10,10,20,20), and (20,20,40,40). 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. Specifically, data are generated from g-and-h distributions (Hoaglin, 1985). Let Z be a random variable having a standard normal distribution. If Z has a standard normal distribution, then by definition

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

$$V = Zexp(hZ^2/2), \text{ if g=0}$$

has a g-and-h distribution where g and h are parameters that determine the first four moments. The four distributions used here were the 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). Table 1 shows the skewness (κ_1) and kurtosis (κ_2) for each distribution. Hoaglin (1985) summarizes additional properties of the g-and-h distributions. Once data were generated from one of these four distributions, (1) was tested using $\sigma_j X_{ij}$ $(i = 1, \cdots, n_j; j = 1, \cdots, 4)$. Three choices for the $(\sigma_1, \cdots, \sigma_4)$ were used: (1, 1, 1 ,1), (4, 4, 1, 1) and (1, 1, 4, 4). These three choices are labeled VP 1, VP 2 and VP 3. Simulations indicate that for VP 3, and when the sample sizes are unequal, both NT and PB perform reasonably well. That is, when the distributions with the larger variances are associated with the larger sample sizes, control over the Type I probability is fairly good. But for VP 2, this was no longer the case. So for brevity, only results for VP 1 and VP 2 are reported.

Table 1. Skewness (κ_1) and Kurtosis (κ_2) of the g-and-h distribution

g	h	κ_1	κ_2
0.0	0.0	0.00	3.00
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 summarizes the estimated Type I error probabilities. Although the importance of a Type I error probability depends on the situation, Bradley (1978) suggests 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, method NT is unsatisfactory in nearly all of the situations considered. In contrast, method PB satisfies this criterion for all of the situations considered.

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

			n= (10,10,10,10)		n= (10,10,20,20)		n= (20,20,40,40)	
VP	g	h	NT	PB	NT	PB	NT	PB
1	0.0	0.0	0.079	0.067	0.075	0.069	0.074	0.061
1	0.0	0.2	0.077	0.060	0.076	0.067	0.080	0.064
1	0.2	0.0	0.078	0.061	0.075	0.068	0.077	0.064
1	0.2	0.2	0.078	0.059	0.078	0.064	0.080	0.065
2	0.0	0.0	0.089	0.060	0.099	0.071	0.091	0.060
2	0.0	0.2	0.090	0.059	0.092	0.068	0.090	0.061
2	0.2	0.0	0.090	0.060	0.088	0.067	0.090	0.061
2	0.2	0.2	0.086	0.066	0.089	0.069	0.090	0.061

Regarding method NT, it is noted that under normality and homoscedasticity, with n= (20,20,20,20), the estimated Type I error probability is 0.066. For n= (40,40,40,40) the estimate is 0.057. Also, simulation results reported by De Neve and Thas (2016) indicate better control over the Type I error probability than indicated by Table 2 under normality, homoscedasticity and n= (10,10,10,10). It is unclear, however, which link function they used. Switching to the link function $g(x) = x/(1-x)$, simulation estimates were more consistent with their results. That is, apparently, the choice for the link function is important. However, when dealing with unequal sample sizes and heteroscedasticity, estimated Type I error probabilities were consistent with those in Table 2. For example, for VP 2, $g = 0.2$, $h = 0.0$ and n= (10,10,20,20), the estimate was 0.09. Increasing the sample sizes to n= (30,50,70,90), again control over the Type I error probability is unsatisfactory. If instead all of the sample sizes are equal to 40, the estimate is 0.076, and for a common sample size of 50 the estimate is 0.067.

AN ILLUSTRATION

Method PB is illustrated with data from the Well Elderly 2 study (Clark, e.g., 2012). Generally, the Well Elderly 2 study was designed to assess the effectiveness of an intervention program aimed at improving the physical and emotional wellbeing of older adults. A portion of the study focused on measures of depressive symptoms. Here we compare measures for a control group to a group that received intervention while taking into account a second factor: participants who identified themselves as White, versus those who did not. First it is noted that when dealing with 20% trimmed means, a significant interaction is not found at the 0.05 level using the method in Wilcox (2017a, section 7.4.1); the R function bbtrim was used. The p-value is 0.18. With no trimming, the p-value is 0.077. Comparing medians via a percentile bootstrap method (using the R function med2mcp), the p-value is 0.264. Using method PB, the p-value is 0.12. So, of course, the choice of method can make a practical difference. The suggestion is that by considering multiple notions of interactions, this provides a deeper and more nuanced

understanding regarding the nature of the interaction. Even if say the 20% trimmed means had been significant, method PB provides a useful perspective: for randomly sampled participants from each of the four groups, there is an estimated 0.435 probability that the decrease in depressive symptoms, between the participants who describe themselves as White, is less than the decrease between participants who do not identify as White.

CONCLUDING REMARKS

In summary, with equal sample sizes of at least 50, method NT performed fairly well. But otherwise, it can be unsatisfactory with respect to the Type I error probability. Of particular concern are situations where the sample sizes are unequal and there is heteroscedasticity. In contrast, method PB performed reasonably well in all of the situations considered, so it is recommended for general use.

It is noted that method PB can be extended to testing linear contrasts when there are more than four groups (Wilcox, 2017b). The method used by Wilcox is based in part on a simple extension of $\tilde{p}$, which was motivated by computational issues related to estimating the distribution of a linear combination of independent random variables. Here, this computational issue does not arise when using $\hat{p}$ unless the product of the sample sizes exceeds the capacity of the computer being used. As previously noted, a seemingly natural speculation is that $\hat{p}$ is preferable to $\tilde{p}$ for the situation at hand, but the simulation results reported here indicate the opposite conclusion.

Finally, the R functions WMWinterci and interWMWAP perform methods NT and PB, respectively. Both are being added to the R package WRS.

References

Bradley, J. V. (1978) Robustness? British Journal of Mathematical and Statistical Psychology, 31 144--152.

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. (2012). 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, 782--790. doi:10.1136/jech.2009.099754

De Neve, J. & Thas, O. (2016): A Mann–Whitney type effect measure of interaction for factorial designs, Communications in Statistics - Theory and Methods, DOI: 10.1080/03610926.2016.1263739

Gao, X. & Alvo, M. (2005). A nonparametric test for interaction in two-way layouts. The Canadian Journal of Statistics, 33,529--543.

Hoaglin, D. C. (1985). Summarizing shape numerically: The g-and-h distribution. In D. Hoaglin, F. Mosteller & J. Tukey (Eds.) Exploring Data Tables Trends and Shapes. New York: Wiley, pp. 461-511.

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

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

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

Wilcox, R. R. (2017b). Linear Contrasts Based on an Extension of the

Wilcoxon–Mann–Whitney Approach. International Journal of Statistics and Probability, 6, No. 3. doi:10.5539/ ijsp.v6n3p198.