Journal of Data Science 13(2015), 43-52

Robust Multiple Comparisons Based On Combined Probabilities From Independent Tests

Rand R. Wilcox[1*] and Florence Clark [2]

[1] *Dept. of Psychology, University of Southern California*
[2] *Division of Occupational Science & Occupational Therapy, University of Southern California*

Abstract: Motivated by a situation encountered in the Well Elderly 2 study, the paper considers the problem of robust multiple comparisons based on K independent tests associated with 2K independent groups. A simple strategy is to use an extension of Dunnett's T3 procedure, which is designed to control the probability of one or more Type I errors. However, this method and related techniques fail to take into account the overall pattern of p-values when making decisions about which hypotheses should be rejected. The paper suggests a multiple comparison procedure that does take the overall pattern into account and then describes general situations where this alternative approach makes a practical difference in terms of both power and the probability of one or more Type I errors. For reasons summarized in the paper, the focus is on 20% trimmed means, but in principle the method considered here is relevant to any situation where the Type I error probability of the individual tests can be controlled reasonably well.

Key words: obust methods, familywise error, meta analysis Fisher's method, Dunnett's T3, Yuen's method

1. Introduction

Considers the situation where K independent tests are performed yielding the p-values $p_1, \ldots, p_K$. In the meta analysis literature, there are well-known methods for testing the assumption that all K hypotheses are true based on these K p-values (e.g., Hedges & Olkin, 1985; Cousins, 2008). Perhaps the best-known method for accomplishing this goal is a technique derived by Fisher (1932), which is reviewed in section 2. But a limitation of all these methods is that they do not indicate which are significant given the goal of controlling the familywise error rate (FWE), meaning the probability of one or more Type I errors. That is, rejecting via Fisher's method, for example, indicates that not all K hypotheses are true, but of course it can be of interest to determine which of the K hypotheses should be rejected. One goal in the paper is to suggest a generalization of Fisher's method, called method F here, which is aimed at dealing with this issue.

* Corresponding author.

Recently, Chen and Nadarajah (2014) suggested an alternative to Fisher's method. It will be evident that the Chen and Nadarajah method can replace Fisher's method when applying method F. The resulting multiple comparisons procedure will be called method CN.

The motivation for this paper stems from a situation encountered in the Well Elderly 2 study (Jackson et al., 2009; Clark, et al., 1997). A portion of the study dealt with the impact of an intervention program aimed at reducing depressive symptoms in older adults. One goal was to compare a control group to an experimental group with a separate analysis done for three ethnic groups: Caucasian, Black/African American and Hispanic/Latino. For convenience, these three ethnic groups are labeled A, B and C, respectively. The data were skewed with outliers, so the control group was compared to the experimental group using 20% trimmed means in conjunction with the method derived by Yuen (1974). The resulting p-values were .022, .126 and .096, respectively. With FWE set at .05, none of the three tests are significant using a simple generalization of Dunnett's (1980) T3 procedure; see Wilcox (2012, section 7.4.1), which will be called method YSM henceforth. (Details of this method are summarized in section 2.) Using instead the improvement on the Bonferroni method derived by Hochberg (1988), or the method derived by Benjamini and Hochberg (1995) aimed at controlling the false discovery rate, again all three tests are not significant. However, Fisher's (1932) method for testing the hypothesis that all three hypotheses are true, based on these three p-values, yields a significant result at the .011 level. But what, if anything, can be said about which tests should be rejected? More generally, how might global tests based on p-values be extended to multiple comparisons and under what circumstances, if any, would such a method offer a practical advantage in terms of power?

A simple step-down technique is suggested for performing multiple tests based on p-values. A criticism of the new method is that it does not provide confidence intervals for the differences between the measures of location. However, certainly there is some interest regarding the strength of the empirical evidence that the typical response in say an experimental group is greater than the typical response for participants in a control group. In an 2 exploratory study, particularly when the sample sizes are relatively small, if there is evidence that an experimental treatment has practical value, this might help motivate further study with the goal of getting a reasonably precise estimate of some appropriate measure of effect size.

The method studied here is based in part on 20% trimmed means. As is well known, heavy-tailed distributions, roughly meaning distributions for which outliers are likely to occur, appear to be quite common based on modern outlier detection techniques, as predicted by Tukey (1960). Consequently, when comparing groups using the usual sample mean, power can be relatively poor. In addition, skewed distributions can result in poor control over the Type I error probability and inaccurate confidence intervals (e.g., Wilcox, 2012a, b). These practical concerns are reduced substantially using Yuen's (1974) method (e.g. Wilcox, 2012b). In the event sampling is from normal distributions, power is nearly as high as methods based on means. This follows almost immediately from results reported by Rosenberger and Gasko (1983) who studied the efficiency of trimmed means. In terms of maximizing power when dealing with data from actual studies, there is some evidence that a 20% trimmed mean is usually preferable (e.g., Wu, 2002.) This is not to suggest, however, that a 20% trimmed mean is always optimal-it is not. For example, when

dealing with sufficiently heavy-tailed distributions, the usual sample median can have a substantially smaller standard error than the mean or 20% trimmed mean, which might translate into substantially higher power and shorter confidence intervals. But a concern about the usual sample median is that for normal distributions, it can have a relatively large standard error, roughly because it trims too many observations. As will be made evident, the methods studied here are relevant when using any method for comparing measures of location that provide reasonably good control over the Type I error probability.

Section 2 describes the details of the proposed method. Section 3 reports simulation results on the ability of the new method to control FWE and how it compares to a generalization of Dunnett's T3 procedure in terms of both Type errors and power.

2. Description of the Method

Momentarily consider a single random sample: $X_1, \ldots, X_n$. Let $X_{(1)} \leq \cdots \leq X_{(n)}$ be the values written in ascending order and let $g = \lfloor \gamma n \rfloor$, where $0 \leq \gamma \leq .5$, where $\lfloor \gamma n \rfloor$ is the greatest integer less than or equal to γn. Then the γ-trimmed mean is

$$\bar{X}_t = \frac{1}{n-2g} \sum_{i=g+1}^{n-g} X_{(i)}.$$

For reasons already explained, the focus is on γ =.2, the 20% trimmed mean. The sample Winsorized mean is

$$\bar{W} = \frac{1}{n} \sum_{i=1}^{n} W_i$$

and the sample Winsorized variance is

$$s^2_{\{w\}} = \frac{1}{n-1} \sum (W_i - \bar{W})^2,$$

where for $i = 1, \ldots, n$,

$$W_i = \begin{cases} X_{(g+1)}, \text{if } X_i \leq X_{(g+1)} \\ X_i, \quad \text{if } X_{(g+1)} < X_i < X_{(n-g)} \\ X_{(n-g)}, \text{if } X_i \geq X_{(n-g)}. \end{cases} \quad (1)$$

Method YSM

For two independent groups let μ_{t1} and μ_{t2} denote the population trimmed means, which are estimated by $\bar{X}_{t1}$and $\bar{X}_{t2}$, respectively. Let $g_j = \lfloor .2n_j \rfloor$ where n_j is the sample size associated with the *j*th group (j = 1, 2). Let

$$d_j = \frac{(n_j-1)s^2_{wj}}{h_j(h_j-1)}, \quad (2)$$

Where S_{wj}^2 is the Winsorized variance for the *jth* group and $h_j = n_j - 2g_j$. Yuen's methodfor testing H_0: μ_{t1} and μ_{t2} is based on the test statistic

$$T_y = \frac{\bar{X}_{t1} - \bar{X}_{t2}}{\sqrt{d_1 + d_2}} \quad (3)$$

The null distribution is approximated with a Student's t distribution having degrees of freedom

$$\hat{v} = \frac{(d_1 + d_2)^2}{\frac{d_1^2}{h_1 - 1} + \frac{d_2^2}{h_2 - 1}}.$$

When Yuen's method is applied *K* times, method YSM consists of controlling FWE using a critical value based on the Studentized maximum modulus distribution $\hat{\boldsymbol{v}}_{\boldsymbol{jk}}$ degrees of freedom, where $\hat{\boldsymbol{v}}_{\boldsymbol{jk}}$ is the value of $\hat{\boldsymbol{v}}$ based on the data associated with groups *j* and *k*.

2.1 Description of the Method

Now consider a situation where Yuen's method is applied *K* times where all *K* tests are independent and p-values are based on Student's t distribution with $\hat{\boldsymbol{v}}$ degrees of freedom.

Denote the resulting p-values by $p_1, \ldots, p_K$. As previously noted, there are several methods for testing the global hypothesis that all *K* tests are not significant based on these *K* independent p-values. The method derived by Fisher (1932) is based on the test statistic

$$F = -2\sum \log(p_i),$$

which has a chi-squared distribution with 2*K* degrees of freedom when all *K* hypotheses are true. That is, reject at the α level if *F* exceeds the $1 - \alpha$ quantile of a chi-squared distribution with 2K degrees of freedom.

Recently, Chen and Nadarajah (2014) suggested an alternative to Fisher's method. The test statistic (for two-sided hypotheses) is

$$C = \sum \Phi^{-1}(p_i/2),$$

where $\Phi^{-1}(pi/2)$ is the $pi/2$ quantile of a standard normal distribution. The null distribution of *C* is a chi-squared distribution with K degrees of freedom. Simulation results reported by Chen and Nadarajah (2014) indicate that their method and Fisher's method tend to have higher power compared to several other test statistics that have been proposed. Regarding Fisher's test, this is consistent with other studies summarized by Cousins (2008). However, it is also known that the power of alternative methods can depend on the pattern of the p-values.

Consider, for example, the largest of the *K* p-values, say p_M. If all *K* hypotheses are true, then a simple modification of the method derived by Tippett (1931) would be to conclude that at least

one of the *K* hypotheses is false if p_M is less than or equal to the α quantile of a beta distribution with parameters *K* and 1. If, for example, six independent tests are performed and all six p-values are equal to .2, both Fisher's method and the method derived by Chen and Nadarajah are not significant at the .05 level. In contrast, p_M rejects at the .001 level. For this reason, p_M was considered in the context of the multiple comparison technique described in section 2.1, but simulations indicated that generally its power does not compare well to the Fisher or Chen-Nadarajah techniques, so p_M is not discussed further.

Momentarily focus on Fisher's method. The multiple comparison technique studied here is based on a simple step-down strategy. An approach that is not quite satisfactory is described first, followed by a modification aimed at improving the control over the probability of one or more Type I errors.

Let $p_{(1)} \leq \cdots \leq p_{(K)}$ denote the ordered p-values and suppose the goal is to have the probability of one or more Type I errors equal to α. First, perform Fisher's global test based on all *K* p-values. If not significant at the α level, stop. If significant, reject the hypothesis associated with the smallest p-value and then apply Fisher's method based on $p_{(2)} \leq \cdots \leq p_{(K)}$ again at the α level. If not significant, stop. If significant, reject the hypothesis associated $p_{(2)}$ and then apply Fisher's method based on $p_{(3)} \leq \cdots \leq p_{(K)}$ at the α level. Continue in this fashion until a non-significant is obtained or all K hypotheses are rejected.

A rough argument that the step-down procedure just described will provide reasonably good control over the probability of one or more Type I errors is as follows. First consider the case where all *K* hypotheses are true. Then when using the step-down approach, the probability of one or more Type I errors is controlled at the α level simply because the probability of rejecting at the first step is α. Consider the case where exactly one of the *K* hypotheses is false. Further imagine that for the one false hypothesis, power is close to one, in which case, with near certainty, step 2 will be reached and Fisher's method would be applied based on $K - 1$ true hypotheses. Because, Fisher's method controls the probability of one or more Type I errors based on the $K - 1$ true hypotheses, the probability of one or more Type I errors will be equal to α. In a similar manner, if two hypotheses are false and there is a high probability that they are rejected, then in step 3, again the probability of rejecting one or more of the $K - 2$ true hypotheses is α.

Method F

However, this argument for using the step-down method is not completely satisfactory. Consider a situation where some of the hypotheses are true but some are false. For the true hypotheses, the goal is to have the probability of one or more Type I errors approximately equal to some specified α value. A speculation was that in situations where the power associated with the false hypotheses is relatively low, it might be possible that the probability of one or more Type I errors exceeds α for the true hypotheses. This was found to be the case in simulations. The probability of one or more Type I errors was found to be as high as .075, and in one case .08, when testing at the .05 level. To guard against this possibility, the step-down method was modified to terminate as follows. At step *k*, if Fisher's method is significant at the α/k level, reject the hypothesis associated with $p_{(k)}$ and continue to the next step, otherwise stop. In effect,

a novel application of the sequentially rejective method derived by Hochberg (1988) is used to control the probability of one or more Type I errors. This will be called method F henceforth. Simulations in section 3 strongly indicate that the probability of one or more Type I errors will not exceed the nominal level under normality. Indeed, the probability of one or more Type I errors was found to be very close to the nominal level for K = 4, 10 and 20 and when sampling from a normal distribution. For non-normal distributions, simulations indicate that FWE is again controlled well, even when there is heteroscedasticity.

Method CN

Method F was described in terms of Fisher's test, but of course the argument for considering the proposed approach applies to any reasonable alternative to Fisher's test such as the Chen-Nadarajah method. Method CN refers to method F but with Fisher's test replaced by the Chen-Nadarajah test.

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

g	h	κ_1	κ_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

3. Description of the Method

Simulations were used to check the properties of the step-down method and how it compares to method YSM. As a partial check on the impact of non-normality, observations were generated from g-and-h distributions, which contain normal distributions as a special case.

To elaborate, let Z be a standard normal random variable. Then

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

has a g-and-h distribution where g and h are parameters that determine the first four moments (Hoaglin, 1985). The standard normal distribution corresponds to $g = h = 0$. The case g = 0 corresponds to a symmetric distribution, and as g increases, skewness increases as well. The parameter h determines heavy tailedness. As h increases, heavy tailedness increases. The six marginal distributions considered here are (g, h) = (0,0), (0, 2), (.2, 0) and (2., .2). So a standard normal distribution is included and corresponds to $g = h = 0$. Table 1 summarizes the skewness (κ_1) and kurtosis (κ_2) for the g-and-h distributions used in the simulations.

Let X_{ik} and Y_{ik} $(i = 1, \ldots, n;\ k = 1, \ldots, K)$ denote the data generated from 2K independent groups. For each k the hypothesis of equal trimmed means is tested via Yuen's method using the

X_{ik} and Y_{ik} values. The sample sizes used were $n_1 = n_2$ = 20 and 200. A few simulations were run with additional sample sizes as will be indicated. No new insights were achieved with $n_1 = n_2$ = 200, so for brevity they are not reported. To assess power, the impact of heteroscedasticity, as well as FWE when fewer than *K* of the hypotheses are true, Y_{ik} was replaced by $\lambda Y_{ik} + \delta_k$. For *K* = 6 the choices for the δ_k values included δ = (0,0,0,0,0,0), meaning that all six hypotheses are true, (.1,.2,.2,0,0,0), (.1,.2,.2,0,0,0),(5,.5,.5,0,0,0), (1.5,1.5,1.5,0,0,0), in which case three hypotheses are true and three are false, and finally (1,1,1,1,1,1) and (4,.4,.4,.4,.4,.4). The choices for λ were 1 (homoscedasticity) and 4. Skewed distributions were shifted so that the population trimmed mean is zero, in which case the population trimmed mean associated with λY_{ik} is not altered when $\lambda \neq 1$.

Table 2 reports the estimated Type I error probabilities for the situations where all hypotheses are true, and where three hypotheses are true. For the case where three hypotheses are true, results are reported for $\delta = (.1, .2, .2, 0, 0, 0)$ and $\delta = (.5, .5, .5, 0, 0, 0)$. So the last six columns of Table 2 indicate the estimate of FEW among the three true hypotheses. Both *X* and *Y* have the g-and-h distribution indicated by column 1 and 2. Estimated Type I error probabilities were based on 4000 replications. If, for example, the actual level is .05, then the standard error of the estimated level is $\sqrt{.05(.95)/4000} = .003$. Results using Fisher's method, the Chen-Nadarajah method, and the extension of Yuen's method (Wilcox, 2012, section 7.4.1) are labeled F, CN and YSM, respectively. As can be seen, the highest estimate using F is .060 and .061 when using CN, which occurred $g = .2, h = 0$ and λ = 4. Increasing *n* to 30 the estimates are .0.57, 0.61 and 0.61 for methods F, CN and YSM, respectively. For *n*=50 the estimates are .052, .051 and .054. The lowest estimates are .027, .028 and .019 for F, CN and YSM, respectively. So all indications are that generally F and CN have actual levels about as close or closer to nominal level than YSM.

Table 2: Estimated probability of one or more Type I errors, α=.05,n_1=n_2=20,K=6

			$\delta = (0,0,0,0,0,0)$			$\delta = (.1,.2,.2,0,0,0)$			$\delta = (.5,.5,.5,0,0,0)$		
g	h	λ	F	CN	YSM	F	CN	YSM	F	CN	YSM
0.0	0.0	1	.050	.048	.050	.044	.044	.024	.056	.054	.024
0.0	0.2	1	.044	.042	.041	.042	.036	.020	.052	.045	.020
0.2	0.0	1	.048	.047	.049	.046	.045	.024	.058	.055	.023
0.2	0.2	1	.042	.042	.041	.041	.037	.019	.050	.044	.019
0.0	0.0	4	.058	.056	.066	.033	.033	.035	.035	.035	.035
0.0	0.2	4	.059	.061	.050	.027	.028	.027	.029	.029	.027
0.2	0.0	4	.060	.061	.068	.035	.035	.037	.036	.036	.037
0.2	0.2	4	.052	.051	.055	.028	.029	.029	.030	.030	.029

F=Fisher, CN=Chen–Nadarajah, YSM=Yuen

A few additional simulations were run with *n* = 400, $g = h$ = 0 and λ =1 as a partial check on the software and the large sample properties of method F and CN. For situations where three of the six hypotheses are true, again the estimated FWE was close to the nominal level. For the situation where one hypothesis is true, the expectation was that the probability of a Type I error would be less than or equal to the nominal level, particularly when power is relatively low for

the hypotheses that are false, simply because the likelihood of reaching the true hypothesis is relatively low. For the situation where the difference between the population trimmed means is .2 for five of the hypotheses, the probability of rejecting the one true hypothesis was estimated to be .039 using F. (For n = 20 the estimate is .019.) (A similar result was obtained with CN.) Increasing the differences in the population trimmed means to .5, the probability of rejecting the one true hypothesis was estimated to be .010. So even with a fairly large sample size there is room for improvement.

Table 3 reports estimated power, meaning the probability of detecting one or more false hypotheses. As can be seen, all indications are that F and CN have more power than YSM, with F seeming to have a slight advantage over CN.

Table 3: Estimated power, the probability of detecting one or more true difference, α=.05, $n_1 = n_2 = 20, K = 6$

			$\delta = (.4, .4, .4, .4, .4, .4)$			$\delta = (.5, .5, .5, 0, 0, 0)$			$\delta = (.1, .2, .2, 0, 0, 0)$		
g	h	λ	F	CN	YSM	F	CN	YSM	F	CN	YSM
0.0	0.0	1	.521	.510	.334	.397	.396	.306	.077	.078	.070
0.0	0.2	1	.458	.443	.286	.338	.334	.262	.064	.066	.057
0.2	0.0	1	.516	.506	.335	.393	.394	.303	.074	.074	.067
0.2	0.2	1	.451	.436	.279	.332	.334	.256	.063	.065	.054
			$\delta = (1, 1, 1, 1, 1, 1)$			$\delta = (1.5, 1.5, 1.5, 0, 0, 0)$			$\delta = (.1, .2, .2, 0, 0, 0)$		
0.0	0.0	4	.388	.378	.259	.413	.409	.314	.064	.063	.059
0.0	0.2	4	.339	.328	.220	.357	.355	.271	.051	.050	.050
0.2	0.0	4	.350	.331	.205	.386	.381	.270	.059	.060	.068
0.2	0.2	4	.298	.283	.168	.328	.322	.227	.049	.048	.051

F=Fisher, CN=Chen–Nadarajah, YSM=Yuen

Additional simulations were run with K=10 and 20 yielding results that were very similar to those reported in Table 2 and 3. For example, when K=20, $g = h = 0, \lambda = 1$ and all K hypothese are true, FEW was estimated to be .055, .052 and .058 for methods F, CN and YSM, repectively.

4. Concluding Remarks

As previously noted, a step-down variation of method YSM was considered in the simulations in section 3 that is based on a generalization of the technique derived by Holm (1979). Simulations indicate that it performs in a very similar manner to YSM in terms of both Type I errors and power. That is, it offers no advantage and does not perform as well as methods F and CN, so details about the method were not provided.

All indications are that the step-down method studied here controls FWE fairly well, better than the extension of Yuen's method, and simultaneously it provides better power, sometimes substantially so. But as previously noted, a limitation of the method is that it does not provide confidence intervals having some specified simultaneous probability coverage. Also, while all of the simulation results indicate that the step-down method performs well in terms of controlling FWE, it is evident that this does not constitute a proof that this will always be the case. The only

point is that the empirical evidence indicates that it has practical value in terms of both FWE and power.

References

[1] Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, B, 57, 289-300.

[2] Chen, Z. & Nadarajah, S. (2014). On the optimally weighted z-test for combining probabilities from independent studies. Computational Statistics and Data Analysis, 70. 387-394.

[3] Clark, F., Azen, S. P., Zemke, R., Jackson, J, Carlson, M., Mandel, D. et al. (1997). Occupational therapy for independent-living older adults. A randomized controlled trial. Journal of the American Medical Association, 278, 1321-1326.

[4] Cousins, R. D. (2008). Annotated bibliography of some papers on combining significances or p-values. arXiv:0705.2209v2.

[5] Dunnett, C. W. (1980). Pairwise multiple comparisons in the unequal variance case. Journal of the American Statistical Association, 75, 796-800.

[6] Fisher, R. A. (1932). Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh.

[7] Hedges, L. & Olkin, I. (1985). Statistical Methods for Meta-Analysis. San Diego, CA Academic Press.

[8] Hoaglin, D. C. (1985) Summarizing shape numerically: The g-and-h distributions. In D. Hoaglin, F. Mosteller and J. Tukey (Eds.) Exploring data tables, trends, and shapes. (pp. 461{515). New York: Wiley.

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

[10] Hochberg, Y. & Tamhane, A. C. (2009). Multiple Comparison Procedures. New York: Wiley.

[11] Holm S. (1979). A simple sequentially rejective multiple test procedure. Scandanavian Journal of Statistics, 6, 65-70.

[12] Jackson, J., Mandel, D., Blanchard, J., Carlson, M., Cherry, B., Azen, S., Chou, C.-P., Jordan-Marsh, M., Forman, T., White, B., Granger, D., Knight, B., Clark, F. (2009).

Confronting challenges in intervention research with ethnically diverse older adults: the USC Well Elderly II trial. Clinical Trials, 6, 90-101.

[13] Rosenberger, J. L. & Gasko, M. (1983). Comparing location estimators: Trimmed means, medians, and trimean. In D. Hoaglin, F. Mosteller and J. Tukey (Eds.) Understanding Robust and exploratory data analysis. (pp. 297-336). New York: Wiley.

[14] Staudte, R. G. and Sheather, S. J. (1990). Robust Estimation and Testing New York: Wiley.

[15] Tippett, L. (1931). The Methods of Statistics. Williams and Norgate Ltd. London

[16] Tukey, J. W. (1960). A survey of sampling from contaminated normal distributions. In I. Olkin et al. (Eds.) Contributions to Probability and Statistics. Stanford, CA: Stanford University Press.

[17] Wilcox, R. R. (2012a). Modern Statistics for the Social and Behavioral Sciences: A Practical Introduction. New York: Chapman & Hall/CRC press

[18] Wilcox, R. R. (2012b). Introduction to Robust Estimation and Hypothesis Testing, 2nd Ed. San Diego, CA: Academic Press.

[19] Wu, P.-C. (2002). Central limit theorem and comparing means, trimmed means one-step M-estimators and modified one-step M-estimators under non-normality. Unpublished doctoral dissertation, Dept. of Education, University of Southern California.

[20] Yuen, K. K. (1974). The two sample trimmed t for unequal population variances. Biometrika, 61, 165-170.

Received December 12, 2013; accepted July 26, 2014.

Rand R. Wilcox
Dept. of Psychology
University of Southern California
3620 McClintock Ave, University of Southern California, Los Angeles, CA 90089-1061
rwilcox@usc.edu

Florence Clark
Division of Occupational Science & Occupational Therapy
University of Southern California
1540 Alcazar Street, CHP 133, Los Angeles, CA 90089-9003
fclark@chan.usc.edu