# Pairwise Multiple Comparison Tests when Data are Nonnormal

**H. J. Keselman, Robert A. Cribbie**
**University of Manitoba**

**and**

**Rand R. Wilcox**
**University of Southern California**

**Keywords: Pairwise multiple comparisons, bootstrap, trimmed means, nonnormality**

**Authors Note:** This research was supported by a Natural Sciences and Engineering Research Council of Canada grant (#OGP0015855).

Abstract

Numerous authors suggest that the data gathered by investigators are not normal in shape. Accordingly, methods for assessing pairwise multiple comparisons of means with traditional statistics will frequently result in biased rates of Type I error and depressed power to detect effects. One solution is to obtain a critical value to assess statistical significance through bootstrap methods. The SAS system can be used to conduct step-down bootstrapped tests. We investigated this approach when data were neither normal in form nor equal in variability in balanced and unbalanced designs. We found that the step-down bootstrap method resulted in substantially inflated rates of error when variances and group sizes were negatively paired. Based on our results, and those reported elsewhere, we recommend that researchers should use trimmed means and Winsorized variances with a heteroscedastic test statistic. When group sizes are equal, the bootstrap procedure effectively controlled Type I error rates.

**Pairwise Multiple Comparison Tests when Data are Nonnormal**

An underlying assumption of most pairwise multiple comparison procedures (MCPs) (e.g., the methods due to Tukey, 1953, Scheffe, 1959, and other procedures available through the major statistical packages) is that the populations from which the data are sampled are normal in form. Although it may be convenient (both practically and statistically) for researchers to assume that their samples are obtained from normally distributed populations, this assumption may rarely be accurate (Micceri, 1989; Pearson, 1931; Wilcox, 1990). Tukey (1960) suggested that outliers should be a common occurrence in distributions and others (e.g. Miller, 1988; Zumbo & Coulombe, 1997) have indicated that skewed distributions frequently depict psychological (reaction time) data. Researchers falsely assuming normally distributed data risk obtaining unsatisfactory Type I and/or Type II error rates for many patterns of nonnormality, especially when other assumptions are also not satisfied (e.g., variance homogeneity) (See Wilcox, 1997).

One potential solution to the problem of nonnormal data is to use bootstrap sampling methods to obtain an empirically determined critical value to assess statistical significance rather than using critical values that are based on the presumption of normally distributed data (e.g., values from the central t-distribution). Diaconis and Efron (1983) provide an accessible introduction to bootstrap concepts. Lunneborg (2000) provides a more comprehensive and technical treatment. Bootstrap sampling allows the data analyst to obtain a critical value that is empirically determined to ascertain statistical significance. For example, the SAS system allows users to obtain both simultaneous and stepwise pairwise MCPs that do not presume normally distributed data. In particular, users can use either bootstrap or permutation methods to compute all possible pairwise comparisons.

If users consider adopting this approach to combat the effects of nonnormality they must consider the cautionary note provided by Westfall et al. (1999, p. 234), namely, the procedure may not control the familywise error (FWE) rate when the data are heterogeneous, particularly when group sizes are unequal. Unfortunately, to date, we do not know what the magnitude of that effect might be, if indeed there is one. Thus, researchers should also consider another approach, that is, pairwise comparisons based on robust estimators and a heteroscedastic statistic, an approach that has been demonstrated to generally control the FWE when data are nonnormal and heterogeneous even when group sizes are unequal.

Specifically, a different type of testing procedure, based on trimmed means, has been discussed by Yuen and Dixon (1973) and Wilcox (1995a, 1995b,1997), and is robust to violations of normality. That is, it is well known that the usual group means and variances, which are the basis for all of the previously described procedures, are greatly influenced by the presence of extreme observations in distributions. In particular, the standard error of the usual mean can become seriously inflated when the underlying distribution has heavy tails and the population mean can lie in the tails of a skewed distribution which "can give a distorted view of how the typical individual in one group compares to the typical individual in another, and about accurate probability coverage, controlling the probability of a Type I error, and achieving relatively high power" (Wilcox, 1995a, p. 66). Theoretical results indicate that substituting robust measures of location and scale for the usual mean and variance, one obtains a test statistic which is relatively insensitive to the combined effects of variance heterogeneity and nonnormality.

While a wide range of robust estimators have been proposed in the literature (see Gross, 1976), the trimmed mean and Winsorized variance are intuitively appealing because of their computational simplicity and good theoretical properties (Wilcox, 1995a, 1995b). The standard error of the trimmed mean is less affected by departures from normality than the usual mean because extreme observations, that is, observations in the tails of a distribution, are removed. Furthermore, as Gross (1976) noted, "the Winsorized variance is a consistent estimator of the variance of the corresponding trimmed mean" (p. 410). In computing the Winsorized variance, the most extreme observations are replaced with less extreme values in the distribution of scores.

Based on the preceding, the purpose of our investigation was to examine the FWE rate of the bootstrap method provided by SAS (1999) (see Westfall et al., 1999, pp. 228-235) under conditions of nonnormality and variance heterogeneity in balanced and unbalanced designs. These findings were then compared to the results reported by Keselman, Lix, and Kowalchuk (1998) who examined MCPs based on robust estimators.

## Design

A mathematical model that can be adopted when examining pairwise mean differences in a one-way completely randomized design is

$$Y_{ij} = \mu_j + \epsilon_{ij},$$

where $Y_{ij}$ is the score of the ith participant ($i = 1, \ldots , n$) in the jth group ($\Sigma_j n = N$), $\mu_j$ is the jth group mean, and $\epsilon_{ij}$ is the random error for the ith participant in the jth group. In the typical application of the model, it is assumed that the $\epsilon_{ij}$s are normally and independently distributed and that the treatment group variances ($\sigma_j^2$s) are equal. Relevant sample estimates include

$$\widehat{\mu}_j = \overline{Y}_j = \sum_{i=1}^{n} Y_{ij}/n \text{ and } \widehat{\sigma}^2 = MSE = \sum_{j=1}^{J}\sum_{i=1}^{n}(Y_{ij}-\overline{Y}_j)^2/J(n-1).$$

A confidence interval for a pairwise difference $\mu_j - \mu_{j'}$ has the form

$$\overline{Y}_j - \overline{Y}_{j'} \pm c_\alpha\, \widehat{\sigma}\sqrt{2/n}\ ,$$

where $c_\alpha$ is selected such that FWE $= \alpha$. In the case of all possible pairwise comparisons, one needs a $c_\alpha$ for the set such that they simultaneously contain the true differences with a specified level of significance. That is, for all $j \neq j'$, $c_\alpha$ must satisfy

$$P(\overline{Y}_j - \overline{Y}_{j'} - c_\alpha\, \widehat{\sigma}\sqrt{2/n} \le \mu_j - \mu_{j'} \le \overline{Y}_j - \overline{Y}_{j'} + c_\alpha\, \widehat{\sigma}\sqrt{2/n}\ ) = 1 - \alpha\ .$$

The interval is equivalent to

$$P(\max_{j,j'} \frac{|(\overline{Y}_j - \mu_j) - (\overline{Y}_{j'} - \mu_{j'})|}{\widehat{\sigma}\sqrt{2/n}} \le c_\alpha) = 1 - \alpha\ ,$$

where max stands for maximum. Evident from this last expression is that $c_\alpha$ is related to the Studentized range distribution (see Scheffe, 1959, p. 28). Specifically, if $Z_1, Z_2, \ldots, Z_n$ are standard normal independent random variates and V is a random variable, independent of the Zs, and is chi-square distributed with df degrees of freedom, then

$$\mathrm{q}_{(\mathrm{J},\ \mathrm{df})} = \mathrm{max}_{\mathrm{j,j'}} \frac{|\mathrm{Z_j} - \mathrm{Z_{j'}}|}{\sqrt{\mathrm{V/df}}}$$

has a Studentized range distribution with parameters J and df. Another relation that should be noted, is that it can be shown that $c_\alpha$ satisfies

$$\mathrm{P}(\mathrm{q}_{\mathrm{J},\ \mathrm{J(n-1)}})/\sqrt{2} \le \mathrm{c}_\alpha) = 1 - \alpha.$$

The hypothesis $H_c$: $\mu_j - \mu_{j'} = 0$ can be tested with the statistic:

$$\mathrm{t_c} = (\overline{\mathrm{Y}}_\mathrm{j} - \overline{\mathrm{Y}}_{\mathrm{j'}})\ /\ (2\ \mathrm{MSE/n})^{1/2}.$$

The preceding can also be specified from a general linear model perspective (see Westfall et al., 1999, Chapter 5). That is, the data can be conceived as coming from the model

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon,$$

where **Y** is an $N \times 1$ observational vector, **X** is the $N \times p$ design matrix, $\beta$ is the $p \times 1$ vector of unknown parameters and $\epsilon$ is the $N \times 1$ vector of random errors.

The usual assumptions to the model relate to the characteristics of the random errors. Specifically, it is assumed that the $\epsilon_1, \epsilon_2, \ldots, \epsilon_N$ all (a) have a mean of zero, (b) have common variance, $\sigma^2$, (c) are independent random variables, and (d) are normally distributed. Important estimates of the model are obtained in the following manner:

$$\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}.$$

$$\widehat{\sigma}^2 = (\mathbf{Y} - \mathbf{X}\widehat{\beta})'(\mathbf{Y} - \mathbf{X}\widehat{\beta})/\mathrm{df},$$

where $(\bullet)^{-}$ denotes a generalized inverse and df $=$ (N $-$ rank $\mathbf{X}$) (see Westfall et al., 1999, p. 87).

One can specify estimable (see Scheffe, 1959, p. 13) functions of the parameters, $\mathbf{c}'\beta$, where for this chapter, the functions would be the pairwise comparisons, such as say $\mathbf{c}'\beta = \mu_1 - \mu_2$, where $\mathbf{c}' = (0\ 1\ \ -1\ 0\ \cdots 0)$, which would be estimated by $\mathbf{c}'\widehat{\beta}$.

To form simultaneous intervals or obtain simultaneous tests of the estimable functions (pairwise comparisons), one needs to know the dependence structures of the estimable functions. As Westfall et al. (1999) pointed out, simultaneous inferences rely on the joint distribution of the quantities

$$\mathrm{T_i} = \frac{\mathbf{c}_i'\widehat{\beta} - \mathbf{c}_i'\beta}{\widehat{\sigma}\sqrt{\mathbf{c}_i'(\mathbf{X}'\mathbf{X})^{-}\mathbf{c}_i}},$$

where $\widehat{\sigma}\sqrt{\mathbf{c}_i'(\mathbf{X}'\mathbf{X})^{-}\mathbf{c}_i}$ is the standard error (SE) of $\mathbf{c}'\widehat{\beta}$. The joint distribution of the $\mathrm{T_i}$ is a multivariate t distribution, with df $=$ (N $-$ rank$\mathbf{X}$) and dispersion matrix $\mathbf{R} = \mathbf{D}^{-\frac{1}{2}}\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}\mathbf{D}^{-\frac{1}{2}}$, where $\mathbf{C} = (\mathbf{c}_1, \ldots, \mathbf{c}_k)$ and $\mathbf{D}$ is a diagonal matrix where the ith element equals $\mathbf{c}_i'(\mathbf{X}'\mathbf{X})^{-}\mathbf{c}_i$.

Confidence intervals of the estimable functions have the form

$$\mathbf{c}_i'\widehat{\beta} \pm \mathrm{c}_\alpha\ \mathrm{SE}(\mathbf{c}_i'\widehat{\beta}),$$

where $\mathrm{c}_\alpha$ is chosen such that the FWE $= \alpha$. Bonferroni-type methods can be used to set the simultaneous intervals such the confidence coefficient will not exceed $1 - \alpha$. However, because the Bonferroni procedure is overly conservative, we know that these intervals will simultaneously contain the true values more than $100(1 - \alpha)$ percent of the time. This approach however can be improved by taking the correlational structure among the estimable functions into account, that is, by setting a simultaneous critical value via the multivariate t distribution. That is,

$$P\left(\left|\frac{\mathbf{c}_i'\widehat{\beta}-\mathbf{c}_i'\beta}{\widehat{\sigma}\sqrt{\mathbf{c}_i'(\mathbf{X}'\mathbf{X})^{-}\mathbf{c}_i}}\right| \leq c_\alpha, \text{ for all } i\right) = 1-\alpha.$$

As Westfall et al. (1999, p. 89) noted, “The value of $c_\alpha$ is the $1-\alpha$ quantile of the distribution of $\max_i |T_i|$, where the vector $\mathbf{T}' = (T_1, \ldots, T_k)$ has the multivariate t distribution.”

FWE control is currently favored by social science researchers. In its typical application, researchers compare a test statistic to a FWE critical value. Another approach for assessing statistical significance is with adjusted p-values, $\tilde{p}_c$, $c = 1, \ldots, C$ (Westfall et al., 1999; Westfall & Wolfinger, 1997; Westfall & Young, 1993). As Westfall and Young noted "$\tilde{p}_c$ is the smallest significance level for which one still rejects a given hypothesis in a family, given a particular (familywise) controlling procedure (p. 11)." Thus, authors do not need to look up (or determine) FWE critical values and moreover consumers of these findings can apply their own assessment of statistical significance from the adjusted p-value rather than from the standard (i.e., FWE) significance level of the experimenter. The latter point is consistent with the current practice of reporting a p-value for a single test statistic rather than stating that the ‘result was significant' at the say .05 value; that is, current practice allows the consumer to take a p-value and apply his/her own personal standard of significance in judging the importance of the finding. For example, if $\tilde{p}_c = 0.09$, the researcher/reader can conclude that the test is statistically significant at the FWE $= 0.10$ level, but not at the FWE $= 0.05$ level.

To illustrate the calculation of an adjusted p-value consider the usual Bonferroni procedure. In its usual application, $H_{0c}$ is rejected if the p-value is less than or equal to $\alpha$/C, where C denotes the total number of statistical tests ($c = 1, \ldots, C$). Note that this is equivalent to rejecting any $H_{0c}$ for which $C \cdot p_c$ is less than or equal to $\alpha$. Therefore, Bonferroni adjusted p-values are:

$$\tilde{p}_c = \begin{cases} C \cdot p_c & \text{if} \quad C \cdot p_c \leq 1 \\ 1 & \text{if} \quad C \cdot p_c > 1. \end{cases}$$

Adjusted p-values are provided by the SAS (1999) system for many popular MCPs (See Westfall et al., 1999).

MCPs

<u>Bootstrap and Permutation Tests</u>. The SAS (1999) system allows users to obtain both simultaneous and stepwise pairwise comparisons of means with methods that do not presume normally distributed data. In particular, users can use either bootstrap or permutation methods to compute all possible pairwise comparisons. The availability of the SAS programs (e.g., PROC MULTTEST, see Westfall et al., 1999) is a particularly attractive inducement for researchers to employ bootstrap sampling to overcome the deleterious effects of nonnormality because it alleviates the need to write bootstrap programs.

Bootstrap sampling allows users to create their own empirical distribution of the data and hence adjusted p-values are based on the empirically obtained distribution, not a theoretically presumed distribution. For example, the empirical distribution, say $\widehat{F}$, is obtained by sampling, <u>with replacement</u>, the pooled sample residuals $\widehat{\epsilon}_{ij} = Y_{ij} - \widehat{\mu}_j = Y_{ij} - \overline{Y}_j$. That is, rather than assume that residuals are normally distributed, one uses empirically generated residuals to estimate the true shape of the distribution. From the pooled sample residuals one generates bootstrap data.

Adjusted p-values are calculated as $\tilde{p}_c = P(\max_c |T_c| \geq |t_c|)$. That is, adjusted p-values are based on the multivariate t distribution. As Westfall et al. (1999, p. 229) noted, in many cases, this is equivalent to $\tilde{p}_c = P(\min_c P_c \leq p_c)$. Their PROC MULTTEST computes adjusted p-values in this fashion [i.e., $\tilde{p}_c = P(\min_c P_c \leq p_c|\widehat{F})$. With this in mind, bootstrapping of adjusted p-values with their MULTTEST program is performed in the following manner:

- Bootstrap data, $Y^*_{ij}$, is generated by sampling with replacement from the pooled sample of residuals.
- Based on the bootstrapped data, $p^*_1$, $p^*_2$, $\ldots$, $p^*_C$ values are obtained from the pairwise tests.
- The above process is repeated many times (PROC MULTTEST allows the user to set the number of replications.).
- For stepwise testing, PROC MULTTEST uses minima over appropriate restricted subsets to obtain the adjusted p-values (Further details about step-down bootstrap methodology can be found in Westfall & Young, 1993, pp. 62-68).

The adjusted p-values are obtained through a shortcut closure testing procedure similar to Holm's (1979) step-down Bonferroni procedure, except that the method used by Westfall et al. (1999, pp. 149-151;157-158; 229) takes the correlational structure of the tests into account. An example program for all possible pairwise comparisons is given by Westfall et al. (1999, p. 229).

As well, pairwise comparisons of means (or ranks) can be obtained through permutation of the data with the program provided by Westfall et al. (1999, pp. 233-234). Permutation tests also do not require that the data be normally distributed. Instead of resampling with replacement from a pooled sample of residuals, permutation tests take the observed data ($Y_{11}, \ldots, Y_{n_1 1}, \ldots, Y_{1J}, \ldots, Y_{n_J J}$) and randomly redistributes them to the treatment groups, and summary statistics (i.e., means or ranks) are then computed on the randomly redistributed data. The original outcomes (all possible pairwise differences from the original sample means) are then compared to the randomly generated values (e.g., all possible pairwise differences in the permutation samples). That is, if, $\bar{Y}_1^* - \bar{Y}_2^*$ is the difference between the first two treatment group means based on a permutation of the data, then a permutational p-value can be computed as $p = P(\bar{Y}_1^* - \bar{Y}_2^* \geq \bar{Y}_1 - \bar{Y}_2)$. Accordingly, for pairwise comparisons, the adjusted p-values are calculated as $\tilde{p}_c = P(\min_c P_c^* \leq p_c)$, where the $P_c^*$ are computed from the permutated data. As Westfall et al. (1999, p. 234) note, the major difference between these two approaches "concerns inferential philosophy rather than actual results." Accordingly, in our study, we just examined bootstrap resampling.

Trimmed Means MCP. Trimmed means are computed by removing a percentage of observations from each of the tails of a distribution (set of observations). Let $Y_{(1)} \leq Y_{(2)} \leq \ldots \leq Y_{(n)}$ represent the ordered observations associated with a group. Let $g = [\gamma\, n]$, where $\gamma$ represents the proportion of observations that are to be trimmed in each tail of the distribution and $[x]$ is notation for the largest integer not exceeding $x$. Wilcox (1995a, 1995b) suggested that 20% trimming should be used. The effective sample size becomes $h = n - 2g$. Then the sample trimmed mean is

$$\bar{Y}_t = \frac{1}{h} \sum_{i=g+1}^{n-g} Y_{(i)} \; .$$

An estimate of the standard error of the trimmed mean is based on the Winsorized mean and Winsorized sum of squares. The sample Winsorized mean is

$$\bar{Y}_w = \tfrac{1}{n}[(g+1)Y_{(g+1)} + \; Y_{(g+2)} + \ldots + Y_{(n-g-1)} + (g+1)Y_{(n-g)}],$$

and the sample Winsorized sum of squared deviations is

$$\begin{aligned} SSD_w = {} & (g+1)(Y_{(g+1)} - \bar{Y}_w)^2 + (Y_{(g+2)} - \bar{Y}_w)^2 + \ldots + (Y_{(n-g-1)} - \bar{Y}_w)^2 \\ & + (g+1)(Y_{(n-g)} - \bar{Y}_w)^2. \end{aligned}$$

Accordingly, the squared standard error of the mean is estimated as (Staudte & Sheather, 1990)

$$d = \frac{SSD_w}{h\,(h-1)} \, .$$

To test a pairwise comparison null hypothesis compute $\bar{Y}_t$ and d for the jth group, label the results $\bar{Y}_{tj}$ and $d_j$. The robust pairwise test (see Keselman, Lix & Kowalchuk, 1998) becomes

$$t_W = \frac{\bar{Y}_{tj} - \bar{Y}_{tj'}}{\sqrt{d_j + d_{j'}}} \, ,$$

with estimated df

$$\nu_W = \frac{(d_j + d_{j'})^2}{d_j^2/(h_j - 1) + d_{j'}^2/(h_{j'} - 1)}.$$

When trimmed means are being compared the null hypothesis relates to the equality of population trimmed means, instead of population means. Therefore, instead of testing $H_0$: $\mu_j = \mu_{j'}$, a researcher would test the null hypothesis, $H_0$: $\mu_{tj} = \mu_{tj'}$, where $\mu_t$ represents the population trimmed mean (Many researchers subscribe to the position that inferences pertaining to robust parameters are more valid than inferences pertaining to the usual least squares parameters when they are dealing with populations that are nonnormal in form.).

Yuen and Dixon (1973) and Wilcox (1995a, 1995b) reported that for long tailed distributions, tests based on trimmed means and Winsorized variances can be much more powerful than tests based on the usual mean and variance. Accordingly, when researchers feel they are dealing with nonnormal data they can replace the usual least squares estimators of central tendency and variability with robust estimators and apply these estimators in MCPs (see Keselman, Lix & Kowalchuk, 1998).

## Methods

In the simulation study six variables were manipulated: (a) the total sample size, (b) the degree of sample size imbalance, (c) the magnitude of the ratio between the largest and smallest variance, (d) the pairing of group sizes and variances, (e) the configuration of population means and (f) the form of the generated data.

For $J = 4$ groups and equal sample sizes in each group, the total sample size was $N = 40$, $N = 60$, or $N = 100$. According to a survey of the educational and psychological literature, the median sample size in one-way completely randomized designs is 64; however, in a third of the studies reviewed sample size ranged between 20 and 40 (see Lix, Cribbie & Keselman, 1996). Therefore, the $N = 40$ and $N = 100$ cases were intended to cover the range of values identified by Lix et al. The $N = 100$ case however, was intended to assess whether the accuracy of the bootstrap methodology (i.e., estimating the true distribution through resampling) improves with increases in sample size as suggested by Westfall et al. (1999, p. 228).

We also varied sample size balance/imbalance. According to a recent survey of the educational and psychological literature's for papers published in 1995-6, unbalanced designs are the norm, not the exception (Keselman, et al., 1998). Furthermore, since the effects of variance heterogeneity are exacerbated by sample size imbalance, we included three cases of balance/imbalance for each sample size investigated. In particular, sample sizes were either equal, moderately unequal, or very unequal, where the degree of balance/imbalance was quantified with a coefficient of sample size variation (SCV); SCV is defined as $(\Sigma_j(n_j - \overline{n})^2/J)^{\frac{1}{2}}/\overline{n}$, where $\overline{n}$ is the average group size. When sample sizes were equal $SCV = 0$; the moderately unequal cases had values of $SCV \simeq .10$, while $SCV \simeq .40$ for the largest case of imbalance investigated. Keselman et al. report that $SCV \simeq .40$ values, or greater, are common. Sample sizes are enumerated in Table 1 for each case of N.

We also considered two cases of variance heterogeneity, where in one case the ratio of the largest to smallest variance was 4:1 while in the second case the ratio was 8:1. Keselman et al. (1998) also reported that an 8:1 ratio for unequal variances is not uncommon. Variances are enumerated in Table 1.

When variances were unequal, they were both positively and negatively paired with the group sizes. For positive (negative) pairings, the group having the fewest (greatest) number of observations was associated with the population having the smallest variance, while the group having the greatest (fewest) number of observations was associated with the population having the largest variance. These conditions were chosen since they typically produce conservative and liberal results, respectively.

Both complete and partial null hypotheses were investigated. In particular, we investigated the following numerical value mean configurations for the four population means: (a) 0.0, 0.0, 0.0, 0.0 (b) 0.0, 0.0, 0.0, 0.917, (c) 0.0, 0.0, 0.477, 0.954 and (d) 0.0, 0.0, 0.791, 0.791. Case (a) is a complete null hypothesis configuration while cases (b) through (d) are partial null hypothesis configurations.

With respect to the effects of distributional shape on Type I error, we chose to investigate conditions in which the statistics were likely to be prone to an excessive number of Type I errors as well as a normally distributed case. Thus, we generated data from a skewed distribution. Specifically, we sampled from a $\chi_3^2$ distribution. This particular type of nonnormal distribution was selected since data obtained in applied settings (e.g., behavioral science data) typically have skewed distributions (Micceri, 1989; Wilcox, 1994a, 1994b, 1995a,b). Furthermore, Sawilowsky and Blair (1992) investigated the effects of eight non-normal distributions identified by Micceri on the robustness of Student's t test and found that only distributions with the most extreme degree of skewness which were investigated (e.g., $\gamma_1 = 1.64$) were found to affect the Type I error control of the independent sample t statistic. Thus, since the statistics we investigated have operating characteristics similar to those reported for the t statistic, we felt that our approach to modeling skewed data would adequately reflect conditions in which those statistics might not perform optimally. For the $\chi_3^2$ distribution, skewness and kurtosis values are $\gamma_1 = 1.63$ and $\gamma_2 = 4.00$, respectively.

Accordingly, our simulated $\chi^2_3$ distribution mirrors data found in behavioral science experiments with regard to skewness.

To generate pseudo-random normal variates, we used the SAS generator RANNOR (SAS Institute, 1989). If $Z_{ij}$ is a standard normal variate, then $Y_{ij} = \mu_j + (\sigma_j \times Z_{ij})$ is a normal variate with mean equal to $\mu_j$ and variance equal to $\sigma^2_j$. To generate pseudo-random variates having a $\chi^2$ distribution with six (three) degrees of freedom, six (three) standard normal variates were squared and summed. The variates were standardized, and then transformed to $\chi^2_3$ variates having mean $\mu_j$ and variance $\sigma^2_j$ [see Hastings & Peacock (1975), pp. 46-51, for further details on the generation of data from this distribution].

Our simulation program was written in SAS/IML (SAS, 1989). One thousand replications of each condition were performed using a .05 significance level. The step-down bootstrap tests were obtained with the program (PROC MULTTEST) provided by Westfall et al. (1999, see pp. 228-231); the number of bootstrap samples was set at 10,000.

## Results

To evaluate the particular conditions under which a test was insensitive to assumption violations, Bradley's (1978) liberal criterion of robustness was employed. According to this criterion, in order for a test to be considered robust, its empirical rate of Type I error ($\hat{\alpha}$) must be contained in the interval $0.5\alpha \leq \hat{\alpha} \leq 1.5\alpha$. Therefore, for the five percent level of statistical significance used in this study, a test was considered robust in a particular condition if its empirical rate of Type I error fell within the interval $.025 \leq \hat{\alpha} \leq .075$. Correspondingly, a test was considered to be nonrobust if, for a particular condition, its Type I error rate was not contained in this interval. In the tables, bolded entries are used to denote liberal values, that is, values greater than .075. We chose this criterion since we feel that it provides a reasonable standard by which to judge robustness. That is, in our opinion, applied researchers should be comfortable working with a procedure that controls the rate of Type I error within these bounds, if the procedure limits the rate across a wide range of assumption violation conditions.

Empirical FWE rates for $N = 40$, $N = 60$, and $N = 100$ are contained in Tables 1 through 3, respectively (Partial null hypothesis results were obtained by averaging rates of error over the three partial null cases investigated.). Since the rates were similar for normal and nonnormal $\chi^2_3$ data, we only tabled the rates for the nonnormal case. Results were similar across the investigated sample sizes and indicate that the SAS (Westfall et al., 1999) step-down bootstrap procedure for pairwise comparisons was: (a) able to control Type I errors when group sizes were equal and when group sizes and variances were positively paired, (b) not able to control the rate of Type I error when group sizes and variances were negatively paired, with rates approaching 20 percent, and (c) liberal for negative pairings of group sizes and variances under the partial null cases, with rates exceeding 10 percent.

To further investigate the effect of sample size on Westfall et al.'s conjecture that the stability of the bootstrap estimates should improve with increases in sample size we collected FWE rates for the complete null hypothesis for four similar conditions that produced liberal rates in Tables 1-3 when there were 100 observations per group ($N = 400$). In particular, we investigated the rates of error when (a) $n_j = 90, 100, 100, 110$ and $\sigma^2_j = 4, 2, 1, 1$; (b) $n_j = 90, 100, 100, 110$ and $\sigma^2_j = 8, 5, 3, 1$; (c) $n_j = 70, 90, 110, 130$ and $\sigma^2_j = 4, 2, 1, 1$; and (d) $n_j = 70, 90, 110, 130$ and $\sigma^2_j = 8, 5, 3, 1$. The empirical

FWE values were .079, .071, .102, and .098, respectively. Thus, rates of error marginally improve with increases in sample size.

## Discussion

The rates we presented in Tables 1 through 3 indicate that the step-down bootstrap MCP available through the SAS (1999) system of programs cannot control the FWE rate when data are nonnormal and are as well heterogeneous, when the design is unbalanced and variances and group sizes are negatively paired. That is, as Westfall et al. (1999) suspected, this approach to pairwise testing with nonnormal data does not work when variances are heterogeneous in unbalanced designs. However, when group sizes are equal the bootstrap procedure does provide acceptable Type I error control. Furthermore, our data suggest that some improvement in Type I error control can be achieved with increases in sample size, though the required sample size would be much larger than those typically found in educational and psychological research.

The results tabled by Keselman et al. (1998) indicate that when trimmed means and Winsorized variances are substituted into Welch's (1938) heteroscedastic statistic, rates of Type I error can indeed be controlled under these same conditions with many stepwise MCPs [e.g., Shaffer's (1986) sequentially rejective Bonferroni procedure, Hayter's (1986) two-stage modified LSD procedure, range-type procedures, Hochberg's (1988) step-up sequentially acceptive Bonferroni procedure].

Accordingly, we recommend that for pairwise comparisons of treatment group means researchers adopt one of the MCPs enumerated by Keselman et al. when data are nonnormal, variances are unequal and the design is unbalanced, conditions that, according to various authors, characterize behavioral science investigations. The reader should note that Wilcox and Keselman (2000) have enumerated a number of bootstrap MCPs that utilize trimmed means and Winsorized variances. However, when group sizes are equal, researchers can confidently rely on the bootstrap (permutation) procedure provided by Westfall et al. (1999) to examine pairwise mean differences under conditions of nonnormality and variance heterogeneity. That is, bootstrapping provides effective Type I error control for comparisons of means; however, the reader should take note that comparisons of means with bootstrapping methods can still fall short with respect to power considerations. Lastly, though likely least attractive, researchers can write their own bootstrap sampling programs for examining pairwise comparisons when data are nonnormal and heterogeneous (see Westfall & Young, 1993, pp. 88-89).

References

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

Diaconis, P., & Efron, B. (1983). Computer-intensive methods in statistics. Scientific American, 248(5), 116-130.

Gross, A. M. (1976). Confidence interval robustness with long-tailed symmetric distributions. Journal of the American Statistical Association, 71, 409-416.

Hastings, N. A. J., & Peacock, J. B. (1975). Statistical distributions: A handbook for students and practitioners. New York: Wiley.

Hayter, A. J. (1986). The maximum familywise error rate of Fisher's least significant difference test. Journal of the American Statistical Association, 81, 1000-1004.

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

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

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C. & Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. Review of Educational Research, 68, 350-386.

Keselman, H.J., Lix, L.M., & Kowalchuk, R. K. (1998). Multiple comparison procedures for trimmed means. Psychological Methods, 3, 123-141.

Lix, L. M., Cribbie, R. A., & Keselman, H. J. (1996, June). The analysis of between-subjects univariate designs. Paper preseneted at the annual meeting of the Psychometric Society, Banff, Canada.

Lunneborg, C. E. (2000). Data analysis by resampling: Concepts and applications. Pacific Grove, CA: Duxbury.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. Psychological Bulletin, 105, 156-166.

Miller, J. (1988). A warning about median reaction time. Journal of Experimental Psychology: Human Perception and Performance, 14, 539-543.

Pearson, E. S. (1931). The analysis of variance in cases of nonnormal variation. Biometrika, 23, 114-133.

Sawilowsky, S.S., & Blair, R.C. (1992). A more realistic look at the robustness and Type II error probabilities of the $t$ test to departures from population normality. Psychological Bulletin, 111, 352-360.

SAS Institute Inc. (1999). SAS/STAT user's guide, Version 7. Cary, NC: SAS Institute Inc.

SAS Institute Inc. (1989). SAS/IML software: Usage and reference, version 6 (1st ed.). Cary, NC: Author.

Scheffé, H. (1959). The analysis of variance. Wiley, 1959.

Shaffer, J.P. (1986). Modified sequentially rejective multiple test procedures. Journal of the American Statistical Association, 81, 826-831.

Staudte, R. G., & Sheather, S. J. (1990). Robust estimation and testing. New York: Wiley.

Tukey, J. W. (1953). The problem of multiple comparisons. Unpublished manuscript, Princeton University, Department of Statistics.

Tukey, J.W. (1960). A survey of sampling from contaminated normal distributions. In I. Olkin et al. (Eds.) Contributions to probability and Statistics (pp. 448-485), Stanford, CA: Stanford University Press.

Welch, B. L. (1938). The significance of the difference between two means when population variances are unequal. Biometrika, 38, 330-336.

Westfall, P. H., Tobias, R. D., Rom, D., Wolfinger, R. D., & Hochberg, Y. (1999). Multiple comparisons and multiple tests. Cary, NC: SAS Institute, Inc.

Westfall, P. H., & Wolfinger, R. D. (1997). Multiple tests with discrete distributions. The American Statistician, 51, 3-8.

Westfall, P. H., & Young, S. S. (1993). Resampling-based multiple testing: Examples and methods for p-value adjustment. New York: Wiley.

Wilcox, R. R. (1990). Comparing the means of two independent groups. Biometrics Journal, 32, 771-780.

Wilcox, R.R. (1994a). A one-way random effects model for trimmed means. Psychometrika, 59, 289-306.

Wilcox, R.R. (1994b). Some results on the Tukey-McLaughlin and Yuen methods for trimmed means when distributions are skewed. Biometrical Journal, 36, 259-273.

Wilcox, R.R. (1995a). ANOVA: A paradigm for low power and misleading measures of effect size? Review of Educational Research, 65, 51-77.

Wilcox, R.R. (1995b). ANOVA: The practical importance of heteroscedastic methods, using trimmed means versus means, and designing simulation studies. British Journal of Mathematical and Statistical Psychology, 48, 99-114.

Wilcox, R. R. (1997). Three multiple comparison procedures for trimmed means. Biometrical Journal, 37, 643-656.

Wilcox, R. R., & Keselman, H. J. (2000). Using trimmed means to compare K measures corresponding to two independent groups. Manuscript submitted for publication.

Yuen, K.K., & Dixon, W.J. (1973). The approximate behaviour and performance of the two-sample trimmed t. Biometrika, 60, 369-374.

Zumbo, B. D., & Coulombe, D. (1997). Investigation of the robust rank-order test for non-normal populations with unequal variances: The case of reaction time. Canadian Journal of Experimental Psychology, 51, 139-150.

Table 1. Empirical Rates of Type I Error (Chi-Squared Data; N=40)

| Sample Sizes | Variances | Complete Null | Partial Null |
|---|---|---|---|
| 10, 10, 10, 10 | 1, 1, 2, 4 | .065 | .015 |
| 10, 10, 10, 10 | 1, 3, 5, 8 | .067 | .024 |
| 9, 10, 10, 11 | 1, 1, 2, 4 | .051 | .017 |
| 9, 10, 10, 11 | 1, 3, 5, 8 | .054 | .020 |
| 9, 10, 10, 11 | 4, 2, 1, 1 | **.099** | .054 |
| 9, 10, 10, 11 | 8, 5, 3, 1 | **.076** | .048 |
| 5, 8, 12, 15 | 1, 1, 2, 4 | .042 | .008 |
| 5, 8, 12, 15 | 1, 3, 5, 8 | .038 | .009 |
| 5, 8, 12, 15 | 4, 2, 1, 1 | **.138** | **.097** |
| 5, 8, 12, 15 | 8, 5, 3, 1 | **.178** | **.104** |

Note: Sample sizes and variances are paired according to the order in which they are enumerated in the table. The numerical values for the population means investigated were: (a) 0.0, 0.0, 0.0, 0.0 (complete null), (b) 0.0, 0.0, 0.0, 0.917 (partial null), (c) 0.0. 0.0, 0.477, 0.954 (partial null), and (d) 0.0, 0.0, 0.791, 0.791 (partial null). The empirical rates tabled under the partial null column are an average value over the three partial null cases. Empirical values not contained in Bradley's (1978) liberal interval (0.25 through .075) are set in bold face type.

Table 2. Empirical Rates of Type I Error (Chi-Squared Data; N=60)

| Sample Sizes | Variances | Complete Null | Partial Null |
|---|---|---|---|
| 15, 15, 15, 15 | 1, 1, 2, 4 | .074 | .015 |
| 15, 15, 15, 15 | 1, 3, 5, 8 | .074 | .018 |
| 13, 15, 15, 17 | 1, 1, 2, 4 | .059 | .016 |
| 13, 15, 15, 17 | 1, 3, 5, 8 | .048 | .014 |
| 13, 15, 15, 17 | 4, 2, 1, 1 | **.083** | .065 |
| 13, 15, 15, 17 | 8, 5, 3, 1 | **.097** | .056 |
| 7, 12, 18, 23 | 1, 1, 2, 4 | .041 | .009 |
| 7, 12, 18, 23 | 1, 3, 5, 8 | .028 | .007 |
| 7, 12, 18, 23 | 4, 2, 1, 1 | **.139** | **.111** |
| 7, 12, 18, 23 | 8, 5, 3, 1 | **.157** | **.118** |

Note: See the note from Table 1.

Table 3. Empirical Rates of Type I Error (Chi-Squared Data; N=100)

| Sample Sizes | Variances | Complete Null | Partial Null |
|---|---|---|---|
| 25, 25, 25, 25 | 1, 1, 2, 4 | .059 | .019 |
| 25, 25, 25, 25 | 1, 3, 5, 8 | .069 | .021 |
| 20, 25, 25, 30 | 1, 1, 2, 4 | .048 | .012 |
| 20, 25, 25, 30 | 1, 3, 5, 8 | .060 | .013 |
| 20, 25, 25, 30 | 4, 2, 1, 1 | **.088** | .066 |
| 20, 25, 25, 30 | 8, 5, 3, 1 | **.090** | .071 |
| 10, 20, 30, 40 | 1, 1, 2, 4 | .026 | .007 |
| 10, 20, 30, 40 | 1, 3, 5, 8 | .031 | .007 |
| 10, 20, 30, 40 | 4, 2, 1, 1 | **.150** | **.107** |
| 10, 20, 30, 40 | 8, 5, 3, 1 | **.182** | **.130** |

Note: See the note from Table 1.