Comparing Measures of the "Typical" Score Across Treatment Groups

by

Katherine Fradette

University of Manitoba

Abdul R. Othman

Universiti Sains Malaysia

H. J. Keselman

University of Manitoba

and

Rand R. Wilcox

University of Southern California

Paper presented at The Joint Statistical Meetings in New York City.

Abstract

Researchers can adopt one of many different measures of central tendency and test statistics to examine the effect of a treatment variable across groups. That is, methods for examining the effect of a treatment variable can be based on means, trimmed means, M-estimators, medians, etc. In our paper we compared a number of recently developed statistics with respect to their ability to control Type I errors when data were nonnormal, heterogeneous and the design was unbalanced. We examined: (1) the use of a preliminary test for symmetry which determines whether data should be trimmed symmetrically or asymmetrically, (2) the use of two different transformations to eliminate skewness, (3) the accuracy of assessing statistical significance with a bootstrap methodology and (4) statistics which utilized a robust measure of the typical score that empirically determined whether data should be trimmed, and, if so, in which direction, and by what amount. Though the fifty-six procedures we examined were remarkably robust to extreme forms of heterogeneity and nonnormality, we recommend a number of Welch-James heteroscedastic statistics which are preceded by the Babu Padmanaban and Puri (1999) test for symmetry that either trimmed symmetrically 10% of the data per group or trimmed asymmetrically 20% of the data per group, after which either Johnson's (1978) or Hall's (1992) transformation was applied to the statistic and where significance was assessed through bootstrapping. Close competitors to the best methods were found that did not involve a statistic transformation.

Developing new methods for locating treatment effects in the one-way independent groups design remains a very active area of study. Much of the work centers around comparing measures of the "typical" score when group variances are unequal and/or when data are obtained from nonnormal distributions. This has been and continues to be an important area of work because the classical method of analysis, e.g., the analysis of variance F-test, is known to be adversely affected by heterogeneous group variances and/or nonnormal data. In particular, these conditions usually result in distorted rates of Type I error and/or a loss of statistical power to detect effects. Wilcox and Keselman (2002) discuss why this is so.

Many treatises have appeared on the topic of substituting robust measures of central tendency such as 20% trimmed means or M-estimators for the usual least squares estimator, i.e., the (least squares) means. Indeed, many investigators have demonstrated that one can achieve better control over Type I errors when robust estimators are substituted for least squares estimators in a heteroscedastic statistic such as Johanson's (1980) Welch-James (WJ)-type test (See e.g., Guo & Luh, 2000; Keselman, Kowalchuk, & Lix, 1998; Keselman, Lix, & Kowalchuk, 1998; Keselman, Wilcox, Taylor & Kowalchuk, 2000; Lix & Keselman, 1998; Luh & Guo, 1999; Wilcox (1995, 1997); Wilcox, Keselman & Kowalchuk, 1998).

Another development in this area was to apply a transformation to a heteroscedastic statistic to eliminate the biasing effects of skewness. Indeed, Luh and Guo (1999) and Guo and Luh (2000) demonstrated that better Type I error control was possible when transformations [Hall's (1978) or Johnson's (1992) method] were applied to the WJ statistic with trimmed means.

Despite the advantages of using (20%) trimmed means, a heteroscedastic statistic with 20% trimming suffers from at least two practical concerns. First, situations arise where the proportion of outliers exceeds the percentage of trimming adopted, meaning that more trimming or some other measure of location, that is relatively unaffected by a large proportion of outliers, is needed. Second, if a distribution is highly skewed to the right, say, then at least in some situations it seems more reasonable to trim more observations from the right tail than from both tails. Thus, using a heteroscedastic statistic with robust estimators, with or without transforming the statistic, may still not provide the best Type I error control. Two solutions that we consider in this paper are using a preliminary test for symmetry in order to determine whether data should be trimmed from both tails (symmetric trimming) or just from one tail (asymmetric trimming) and whether an estimator, other than the trimmed mean, that is, one that does not fix the amount of trimming *a priori* but empirically determines the amount and direction, or even the need for trimming, can provide better Type I error control.

The prevalent method of trimming is to remove outliers from each tail of the distribution of scores. In addition, the recommendation is to trim 20% from each tail (See Rosenberger & Gasko, 1983; Wilcox, 1995). However, asymmetric trimming has been theorized to be potentially advantageous when the distributions are known to be skewed, a situation likely to be realized with behavioral science data (See De Wet & van Wyk, 1979; Micceri, 1989; Tiku, 1980, 1982; Wilcox, 1994, Wilcox, 1995). Indeed, if a researcher's goal is to adopt a measure of the "typical" score, that is, a score that is representative of the bulk of the observations, then theory certainly indicates that he/she should trim just from the tail in which outliers are located in order to get a score that represents the bulk of the observations; trimming symmetrically in this circumstance would eliminate representative scores, scores similar to the bulk of observations.

A stumbling block to adopting asymmetric versus symmetric trimming has been the inability of researchers to determine when to adopt one form of trimming over the other. That is, previous work has not identified a procedure which reliably identifies when data are positively or negatively skewed, rather than symmetric; thus researchers have not been able to successfully adopt one method of trimming versus the other. However, work by Hogg, Fisher and Randles (1975), later modified by Babu Padmanaban and Puri (1999), may provide a successful solution to this problem and accordingly enable researchers to successfully adopt asymmetric trimming in cases where it is needed thus providing them with measures of the typical score which more accurately corresponds to the bulk of the observations. The by-product of correctly identifying and eliminating only the outlying values should result in better Type I error control for heteroscedastic statistics that adopt trimmed means.

A concomitant issue that needs to be resolved is knowing how the 20% rule should be applied when trimming just from one tail. That is, should 40% of the longer tail of scores be trimmed since in total that amount is trimmed when trimming 20% in each tail? Or, should just 20% be trimmed from the one tail of the distribution? As well, the 20% rule is not universally recommended; others have had success with values other than 20%. For example, Babu et al. (1999) obtained good Type I error control, for the procedures they investigated, with 15% symmetric trimming. Indeed, as Huber (1993) argues, an estimator should have a breakdown point of at least .1; thus, even 10% trimming might provide effective Type I error control.

A second approach to the problem of direction and amount of trimming would be to adopt another robust estimator that does not *a priori* set the amount of trimming. Wilcox and Keselman (in press) introduced a modified M-estimator which empirically determines whether to trim symmetrically or asymmetrically and by what amount, or whether no trimming at all is appropriate. In the context of correlated groups design, they showed that their estimator does indeed provide effective Type I error control.

A last refinement that we will examine is the use of the bootstrap for hypothesis testing. Bootstrap methods have two practical advantages. First, theory and empirical findings indicate that they can result in better Type I error control than nonbootstrap methods (See Guo & Luh, 2000; Keselman, Kowalchuk, & Lix, 1998; Keselman, Lix, & Kowalchuk, 1998; Keselman, Wilcox, Taylor & Kowalchuk, 2000; Lix & Keselman, 1998; Luh & Guo, 1999; Wilcox (1995, 1997); Wilcox, Keselman & Kowalchuk, 1998). Second, certain variations of the bootstrap method do not require explicit expressions for standard errors of estimators. This makes hypothesis testing in some settings more flexible when other robust estimators (soon to be discussed) are used instead of trimmed means.

Thus, the purpose of our investigation was to compare rates of Type I error for numerous versions of the WJ heteroscedastic statistic versus two test statistics that use the estimator introduced by Wilcox and Keselman (2002). Variations of the WJ statistic will be based on asymmetric versus symmetric trimming, the amount of trimming, transformations of WJ and bootstrap versus nonbootstrap versions.

Methods

The WJ Statistic

Methods that give improved power and better control over the probability of a Type I error can be formulated using a general linear model perspective. Lix and Keselman (1995) showed how the various Welch (1938, 1951) statistics that appear in the literature for testing omnibus main and interaction effects as well as focused hypotheses using contrasts in univariate and multivariate independent and correlated groups designs can be formulated from this perspective, thus allowing researchers to apply one statistical procedure to any testable model effect. We adopt their approach in this paper and begin by presenting, in abbreviated form, its mathematical underpinnings.

A general approach for testing hypotheses of mean equality using an approximate degrees of freedom solution is developed using matrix notation. The multivariate perspective is considered first; the univariate model is a special case of the multivariate. Consider the general linear model:

$$\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\xi} \,, \tag{1}$$

where **Y** is an $N \times p$ matrix of scores on p dependent variables or p repeated measurements, N is the total sample size, **X** is an $N \times r$ design matrix consisting entirely of zeros and ones with rank(**X**) $= r$, β is an $r \times p$ matrix of nonrandom parameters (i.e., population means), and $\boldsymbol{\xi}$ is an $N \times p$ matrix of random error components. Let $\mathbf{Y}_j$ ($j = 1, \ldots, r$) denote the submatrix of **Y** containing the scores associated with the n subjects in the jth group (cell) (For the one-way design considered in this paper $n = n_j$.). It is typically assumed that the rows of **Y** are independently and normally distributed, with mean vector β_j and variance-covariance matrix $\boldsymbol{\Sigma}_j$ [i.e., N(β_j, $\boldsymbol{\Sigma}_j$)], where the jth row of β, $\beta_j = [\mu_{j1} \ldots \mu_{jp}]$, and $\boldsymbol{\Sigma}_j \neq \boldsymbol{\Sigma}_{j'}$ ($j \neq j'$). Specific formulas for estimating β and $\boldsymbol{\Sigma}_j$, as well as an elaboration of **Y** are given in Lix and Keselman (1995, see their Appendix A).

The general linear hypothesis is

$$H_0\text{: } \mathbf{R}\boldsymbol{\mu} = \mathbf{0}, \tag{2}$$

where $\mathbf{R} = \mathbf{C} \otimes \mathbf{U}^T$, **C** is a $df_C \times r$ matrix which controls contrasts on the independent groups effect(s), with rank(**C**) $= df_C \leq r$, and **U** is a $p \times df_U$ matrix which controls contrasts on the within-subjects effect(s), with rank(**U**) $= df_U \leq p$, '$\otimes$' is the Kronecker or direct product function, and 'T' is the transpose operator. For multivariate independent groups designs, **U** is an identity matrix of dimension p (i.e., $\mathbf{I}_p$). The **R** contrast matrix has $df_C \times df_U$ rows and $r \times p$ columns. In Equation 2, $\boldsymbol{\mu} = \text{vec}(\beta^T) = [\beta_1 \ldots \beta_r]^T$. In other words, $\boldsymbol{\mu}$ is the column vector with $r \times p$

elements obtained by stacking the columns of β^T. The **0** column vector is of order df_C x df_U [See Lix & Keselman (1995) for illustrative examples].

The generalized test statistic given by Johansen (1980) is

$$T_{WJ} = (\mathbf{R}\widehat{\boldsymbol{\mu}})^T(\mathbf{R}\widehat{\boldsymbol{\Sigma}}\mathbf{R}^T)^{-1}(\mathbf{R}\widehat{\mu}), \tag{3}$$

where $\widehat{\boldsymbol{\mu}}$ estimates $\boldsymbol{\mu}$, and $\widehat{\boldsymbol{\Sigma}} = \text{diag}[\widehat{\boldsymbol{\Sigma}}_1/n_1 \ldots \widehat{\boldsymbol{\Sigma}}_r/n_r]$, a block matrix with diagonal elements $\widehat{\boldsymbol{\Sigma}}_r/n_r$. This statistic, divided by a constant, c (i.e., T_{WJ}/c), approximately follows an F distribution with degrees of freedom $\nu_1 = df_C \times df_U$, and $\nu_2 = \nu_1(\nu_1 + 2)/(3A)$, where $c = \nu_1 + 2A - (6A)/(\nu_1 + 2)$. The formula for the statistic, A, is provided in Lix and Keselman (1995).

When $p = 1$, that is, for a univariate model, the elements of **Y** are assumed to be independently and normally distributed with mean μ_j and variance σ_j^2 [i.e., $N(\mu_j, \sigma_j^2)$]. To test the general linear hypothesis, **C** has the same form and function as for the multivariate case, but now **U** = 1, $\widehat{\boldsymbol{\mu}} = [\widehat{\mu}_1 \ldots \widehat{\mu}_r]^T$ and $\widehat{\boldsymbol{\Sigma}} = \text{diag}[\sigma_1^2/n_1 \ldots \sigma_r^2/n_r]$. (See Lix & Keselman's 1995 Appendix A for further details of the univariate model.)

Robust Estimation

In this paper we apply robust estimates of central tendency and variability to the T_{WJ} statistic. That is, heteroscedastic ANOVA methods are readily extended to the problem of comparing trimmed means. The goal is to determine whether the effect of a treatment varies across J ($j = 1, \ldots, J$) groups; that is, to determine whether a typical score varies across groups. When trimmed means are being compared the null hypothesis pertains to the equality of population trimmed means, i.e., the μ_ts. That is, to test the omnibus hypothesis in a one-way completely randomized design, the null hypothesis would be H_0: $\mu_{t1} = \mu_{t2} = \cdots = \mu_{tJ}$.

Let $Y_{(1)j} \leq Y_{(2)j} \leq \cdots \leq Y_{(n_j)j}$ represent the ordered observations associated with the jth group. Let $g_j = [\gamma\, n_j]$, where γ represents the proportion of observations that are to be trimmed in each tail of the distribution and $[x]$ is the greatest integer $\leq x$. The effective sample size for the jth group becomes $h_j = n_j - 2g_j$. The jth sample trimmed mean is

$$\widehat{\mu}_{tj} = \frac{1}{h_j} \sum_{i=g_j+1}^{n_j-g_j} Y_{(i)j} \ . \tag{4}$$

Wilcox (1995) suggests that 20% trimming should be used [See Wilcox (1995) and the references he cites for a justification of the 20% rule.].

The sample Winsorized mean is necessary and is computed as

$$\widehat{\mu}_{wj} = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ij}, \tag{5}$$

where

$$\begin{aligned} X_{ij} &= Y_{(g_j+1)j} \text{ if } Y_{ij} \leq Y_{(g_j+1)j} \\ &= Y_{ij} \text{ if } Y_{(g_j+1)j} < Y_{ij} < Y_{(n_j-g_j)j} \\ &= Y_{(n_j-g_j)j} \text{ if } Y_{ij} \geq Y_{(n_j-g_j)j} \ . \end{aligned}$$

The sample Winsorized variance, which is required to get a theoretically valid estimate of the standard error of a trimmed mean, is then given by

$$\widehat{\sigma}^2_{wj} = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (X_{ij} - \widehat{\mu}_{wj})^2. \tag{6}$$

The standard error of the trimmed mean is estimated with

$\sqrt{(n_j - 1)\, \widehat{\sigma}^2_{wj}/[(h_j(h_j - 1)]}$.

Under asymmetric trimming, and assuming, without loss of generality, that the distribution is positively skewed so that trimming takes place in the upper tail, the jth sample trimmed mean is

$$\widehat{\mu}_{tj} = \frac{1}{h_j} \sum_{i=1}^{n_j-g_j} Y_{(i)j},$$

and the jth sample Winsorized mean is

$$\hat{\mu}_{wj} = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ij},$$

where

$$X_{ij} = Y_{ij} \text{ if } Y_{ij} < Y_{(n_j - g_j)j}$$

$$= Y_{(n_j - g_j)j} \text{ if } Y_{ij} \leq Y_{(n_j - g_j)j} .$$

The sample Winsorized variance is again defined as (given the new definition of $\hat{\mu}_{wj}$)

$$\hat{\sigma}^2_{wj} = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (X_{ij} - \hat{\mu}_{wj})^2,$$

and the standard error of the mean again takes its usual form (given the new definition of $\hat{\mu}_{wj}$).

Thus, with robust estimation, the trimmed group means ($\hat{\mu}_{tj}$s) replace the least squares group means ($\hat{\mu}_j$s), the Winsorized group variances estimators ($\hat{\sigma}^2_{wj}$s) replace the least squares variances ($\hat{\sigma}^2_j$s), and h_j replaces n_j and accordingly one computes the robust version of T_{WJ}, T_{WJt} (See Keselman, Wilcox, & Lix, 2001; For another justification of adopting robust estimates see Rocke, Downs & Rocke, 1982).

Bootstrapping

Now we consider how extensions of the ANOVA method just outlined might be improved. In terms of probability coverage and controlling the probability of a Type I error, extant investigations indicate that the most successful method, when using a 20% trimmed mean (or some M-estimator), is some type of bootstrap method.

Following Westfall and Young (1993), and as enumerated by Wilcox (1997), let $C_{ij} = Y_{ij} - \hat{\mu}_{tj}$; thus, the C_{ij} values are the empirical distribution of the jth group, centered so that the sample trimmed mean is zero. That is, <u>the empirical distributions are shifted so that the null hypothesis of equal trimmed means is true in the sample</u>. The strategy behind the bootstrap is to use the shifted empirical distributions to estimate an appropriate critical value. For each j, obtain a bootstrap sample by randomly sampling with replacement n_j observations from the C_{ij} values, yielding $Y^*_1, \ldots, Y^*_{n_j}$. Let T^*_{WJt} be the value of Johansen's (1980) test based on the bootstrap sample. Now we randomly sample (with replacement), B bootstrap samples from the shifted/centered distributions each time calculating the statistic T^*_{WJt}. The B values of T^*_{WJt} are put in ascending order, that is, $T^*_{WJt(1)} \leq \cdots \leq T^*_{WJt(B)}$, and an estimate of an appropriate critical value is $T^*_{WJt(a)}$, where $a = (1 - \alpha)B$, rounded to the nearest integer. One will reject the null hypothesis of location equality (i.e., H_0: $\mu_{t1} = \mu_{t2} = \cdots = \mu_{tJ}$) when $T_{WJt} > T^*_{WJt(a)}$, where T_{WJt} is the value of the heteroscedastic statistic based on the original nonbootstrapped data. Keselman et al. (2001) illustrate the use of this procedure for testing both omnibus and sub-effect (linear contrast) hypotheses in completely randomized and correlated groups designs.

Transformations for the Welch-James Statistic

Guo and Luh (2000) and Luh and Guo (1999) found that Johnson's (1978) and Hall's (1992) transformations improved the performance of several heteroscedastic test statistics when they were used with trimmed means, including the WJ statistic, in the presence of heavy-tailed and skewed distributions.

In our study we, accordingly, compared both approaches for removing skewness when applied to the T_{WJt} statistic. Let $Y_{ij} = \left(Y_{1j}, Y_{2j}, \ldots, Y_{n_j j}\right)$ be a random sample from the jth distribution. Let $\hat{\mu}_{tj}$, $\hat{\mu}_{wj}$ and $\hat{\sigma}^2_{wj}$ be, respectively, the trimmed mean, Winsorized mean and Winsorized variance of group j. Define the Winsorized third central moment of group j as

$$\hat{\mu}_{3j} = \frac{1}{n_j} \sum_{i=1}^{n_j} \left(X_{ij} - \hat{\mu}_{wj}\right)^3.$$

Let

$$\tilde{\sigma}^2_{wj} = \frac{(n_j - 1)}{(h_j - 1)}\hat{\sigma}^2_{wj},$$

$$\tilde{\mu}_{wj} = \frac{n_j}{h_j}\hat{\mu}_{3j},$$

$$q_j = \frac{\tilde{\sigma}^2_{wj}}{h_j},$$

$$w_{tj} = \frac{1}{q_j},$$

$$U_t = \sum_{j=1}^{J} w_{tj},$$

and

$$\hat{\mu}_t = \frac{1}{U_t}\sum_{j=1}^{J} w_{tj}\ \hat{\mu}_{tj}.$$

Luh and Guo (2000) defined a trimmed mean statistic with Johnson's transformation as

$$T_{Johnson_j} = \left(\hat{\mu}_{tj} - \hat{\mu}_t\right) + \frac{\tilde{\mu}_{wj}}{6\ \tilde{\sigma}^2_{wj} h_j} + \frac{\tilde{\mu}_{wj}}{3\ \tilde{\sigma}^4_{wj}}\left(\hat{\mu}_{tj} - \hat{\mu}_t\right)^2. \qquad (7)$$

From Guo and Luh (2000) we can deduce that a trimmed mean statistic with Hall's (1992) transformation would be

$$T_{Hall_j} = \left(\hat{\mu}_{tj} - \hat{\mu}_t\right) + \frac{\tilde{\mu}_{wj}}{6\ \tilde{\sigma}^2_{wj} h_j} + \frac{\tilde{\mu}_{wj}}{3\ \tilde{\sigma}^4_{wj}}\left(\hat{\mu}_{tj} - \hat{\mu}_t\right)^2 + \frac{\tilde{\mu}^2_{wj}}{27\ \tilde{\sigma}^8_{wj}}\left(\hat{\mu}_{tj} - \hat{\mu}_t\right)^3. \qquad (8)$$

Keselman et al. (2001) indicated that sample trimmed means, sample Winsorized variances and trimmed sample sizes can be substituted for the usual sample means, variances and sample sizes in the T_{WJ} statistic. That is,

$$T_{WJ} = \sum_{j=1}^{J} w_{tj}\left(\hat{\mu}_{tj} - \hat{\mu}_{t}\right)^2,$$

which, when divided by c, is distributed as an F variable with df of J − 1 and

$$\nu = \left(J^2 - 1\right)\left[3\sum_{j=1}^{J}\frac{\left(1 - w_{tj}/U_t\right)^2}{h_j - 1}\right]^{-1},$$

where

$$c = (J - 1)\left(1 + \frac{2(J-2)}{J^2 - 1}\sum_{j=1}^{J}\frac{\left(1 - w_{tj}/U_t\right)^2}{h_j - 1}\right).$$

Now we can define

$$T_{WJ_{Johnson}} = \sum_{j=1}^{J} w_{tj}\left(T_{Johnson_j}\right)^2, \qquad (9)$$

and

$$T_{WJ_{Hall}} = \sum_{j=1}^{J} w_{tj}\left(T_{Hall_j}\right)^2, \qquad (10)$$

Then $T_{WJ_{Johnson}}$ and $T_{WJ_{Hall}}$, when divided by c, are also distributed as F variates with no change in degrees of freedom.

A Preliminary Test for Symmetry

A stumbling block to adopting asymmetric versus symmetric trimming has been the inability of researchers to determine when to adopt one form of trimming over the other. Work by Hogg et al. (1975) and Babu et al. (1999), however, may provide a successful solution to this problem. The details of this method are presented in Appendix A.

The Modified One-Step (MOM) estimator

For J independent groups (this estimator can also be applied to dependent groups) consider the MOM estimator introduced by Wilcox and Keselman (in press). In particular, these authors suggested modifying the well-known one-step M-estimator

$$\frac{1.28\,(\mathrm{MADN}_j)(i_2 - i_1) + \sum_{i=i_1+1}^{n_j - i_2} Y_{(i)j}}{n_j - i_1 - i_2}, \tag{11}$$

by removing 1.28 $(\mathrm{MADN}_j)(i_2 - i_1)$, where $\mathrm{MADNj} = \mathrm{MAD}_j/.6745$, $\mathrm{MAD}_j =$ the median of the values $|Y_{ij} - \widehat{M}_j|, \ldots, |Y_{nj,j} - \widehat{M}_j|$, $\widehat{M}_j$ is the median of the jth group, $i_1 =$ the number of observations where $Y_{ij} - \widehat{M}_j < 2.24(\mathrm{MADN}_j)$ and $i_2 =$ the number of observations where $Y_{ij} - \widehat{M}_j > 2.24(\mathrm{MADN}_j)$. Thus, the modified M-estimator suggested by Wilcox and Keselman is

$$\widehat{\theta}_j = \sum_{i=i_1+1}^{n_j - i_2} \frac{Y_{(i)j}}{n_j - i_1 - i_2}. \tag{12}$$

The MOM estimate of location is just the average of the values left after all outliers (if any) are discarded. The constant 2.24 is motivated in part by the goal of having a reasonably small standard error when sampling from a normal distribution. Moreover, detecting outliers with Equation 12 is a special case of a more general outlier detection method derived by Rousseeuw and van Zomeren (1990).

MOM estimators, like trimmed means, can be applied to test statistics to investigate the equality of this measure (θ) of the typical score across treatment groups. The null hypothesis is

$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_J, \tag{13}$$

where θ_j is the population value of MOM associated with the jth group. Two statistics can be used. The first was a statistic mentioned by Schrader and

Hettsmansperger (1980), examined by He, Simpson and Portnoy (1990) and discussed by Wilcox (1997, p. 164). The test is defined as

$$H = \frac{1}{N} \sum_{j=1}^{J} n_j(\widehat{\theta}_j - \widehat{\theta}_.)^2, \tag{14}$$

where $N = \Sigma_j n_j$ and $\widehat{\theta}_. = \Sigma_j \widehat{\theta}_j / J$. To assess statistical significance a (percentile) bootstrap method can be adopted. That is, to determine the critical value one centers or shifts the empirical distribution of each group; that is, each of the sample MOM_js is substracted from the scores in their respective groups (i.e., $C_{ij} = Y_{ij} - MOM_j$). As was the case with trimmed means, the strategy is to shift the empirical distributions with the goal of estimating the null distributioin of H which yields an estimate of an appropriate critical value. Now one randomly samples (with replacement), B bootstrap samples from the shifted/centered distributions each time calculating the statistic H, which when based on a bootstrap sample, is denoted as H^*. The B values of H^* are put in ascending order, that is, $H^*_{(1)} \leq \cdots \leq H^*_{(B)}$, and an estimate of an appropriate critical value is $H^*_{(a)}$, where $a = (1 - \alpha)B$, rounded to the nearest integer. One will reject the null hypothesis of location equality when $H > H^*_{(a)}$.

The second method of analysis presented can be obtained in the following manner (See Liu & Singh, 1997). Let

$$\delta_{jj\prime} = \theta_j - \theta_{j\prime} \ (j < j\prime). \tag{15}$$

Thus, the $\delta_{jj\prime}$s are the all possible pairwise comparisons among the J treatment groups. Now, if all groups have a common measure of location (i.e., $\theta_1 = \theta_2 = \cdots = \theta_J$), then H_0: $\delta_{12} = \delta_{13} = \cdots = \delta_{J-1,\,J} = 0$. A bootstrap method can be used to assess statistical significance, but for this procedure the data does not need to be centered. In contrast to the first method, the goal is not to estimate the null distribution of some appropriate test statistic. Rather, bootstrap samples are obtained for the Y_{ij} values and one rejects if the zero vector is sufficiently far from the center of the bootstrap estimates of the delta values. Thus, bootstrap samples are obtained from the Y_{ij} values rather than the C_{ij}s. For each bootstrap replication ($B = 599$ is again recommended) one computes the robust estimators (i.e., MOM) of location (i.e., $\hat{\theta}^*_{jb}$, $j = 1, \ldots, J$; $b = 1, \ldots, B$) and the corresponding estimates of $\delta_{jj'b}$ ($\hat{\delta}^*_{jj'b} = \hat{\theta}^*_{jb} - \hat{\theta}^*_{j'b}$). The strategy is to determine how deeply $\mathbf{0} = (0\ 0\ \ldots\ 0)$ is nested within the bootstrap values $\hat{\delta}^*_{jj'b}$, where $\mathbf{0}$ is a vector having length $K = J(J-1)/2$. This assessment is made by adopting a modification of Mahalanobis's distance statistic.

For notational convenience, we can rewrite the K differences $\hat{\delta}_{jj'}$ as $\hat{\Delta}_1, \ldots, \hat{\Delta}_K$ and their corresponding bootstrap values as $\hat{\Delta}^*_{kb}$ ($k = 1, \ldots, K$; $b = 1, \ldots, B$). Thus, let

$$\bar{\Delta}^*_k = \frac{1}{B} \sum_{b=1}^{B} \hat{\Delta}^*_{kb},$$

and

$$Z_{kb} = \hat{\Delta}^*_{kb} - \bar{\Delta}^*_k + \hat{\Delta}_k.$$

(Note the Z_{kb}s are shifted bootstrap values having mean $\hat{\Delta}_k$.) Now define

$$S_{kk'} = \frac{1}{B-1}(Z_{kb} - \bar{Z}_k)(Z_{k'b} - \bar{Z}_{k'}), \tag{16}$$

where

$$\bar{Z}_k = \frac{1}{B} \sum_{b=1}^{B} Z_{kb}\ .$$

(Note: The bootstrap population mean of $\bar{\Delta}^*_k$ is known and is equal to $\hat{\Delta}_k$.)

With this procedure, one next computes

$$D_b = (\hat{\mathbf{\Delta}}^*_b - \hat{\mathbf{\Delta}})\mathbf{S}^{-1}(\hat{\mathbf{\Delta}}^*_b - \hat{\mathbf{\Delta}})', \tag{17}$$

where $\hat{\mathbf{\Delta}}^*_b = (\hat{\Delta}^*_{1b}, \ldots, \hat{\Delta}^*_{Kb})$ and $\hat{\mathbf{\Delta}} = (\hat{\Delta}_1, \ldots, \hat{\Delta}_K)$. Accordingly, D_b measures how closely $\hat{\mathbf{\Delta}}_b$ is located to $\hat{\mathbf{\Delta}}$. If the null vector ($\mathbf{0}$) is relatively far from $\hat{\mathbf{\Delta}}$ one rejects H_0. Therefore, to assess statistical significance, put the D_b values in ascending order ($D_{(1)} \leq \cdots \leq D_{(B)}$) and let $a = (1 - \alpha)B$ (rounded to the nearest integer). Reject H_0 if

$$T \geq D_{(a)}, \tag{18}$$

where

$$T = (\mathbf{0} - \widehat{\boldsymbol{\Delta}})\mathbf{S}^{-1}(\mathbf{0} - \widehat{\boldsymbol{\Delta}})'. \tag{19}$$

It is important to note that $\theta_1 = \theta_2 = \cdots = \theta_J$ can be true iff H_0: $\theta_1 - \theta_2 = \cdots = \theta_{J-1} - \theta_J = 0$ (Therefore, it suffices to test that a set of K pairwise differences equal zero). However, to avoid the problem of arriving at different conclusions (i.e., sensitivity to detect effects) based on how groups are arranged (if all MOMs are unequal), we recommend that one test the hypothesis that all pairwise differences equal zero.

Empirical Investigation

Fifty-six tests for treatment group equality were compared for their rates of Type I error under conditions of nonnormality and variance heterogeneity in an independent groups designs with four treatments. The procedures we investigated were:

<u>Trimmed Means with Symmetric Trimming (No preliminary test for symmetry)</u>:

1.-3. WJ10(15)(20)-WJ with 10% (15%) (20%) trimming

4.-6. WJB10(15)(20)-10% (15%) (20%) trimming and bootstrapping

7.-9. WJJ10(15)(20)-10% (15%) (20%) trimming and Johnson's transformation

10.-12. WJJB10(15)(20)-10% (15%) (20%)trimming with Johnson's transformation and bootstrapping

13.-15 WJH10(15)(20)-10% (15%) (20%) trimming and Hall's transformation

16.-18 WJHB10(15)(20)-10% (15%) (20%) trimming and Hall's transformation and bootstrapping

<u>WJ with Q Statistics</u>: <u>Symmetric and Asymmetric Trimming</u>:

19.-21. WJ1010(1515)(2020)-WJ. If data is symmetric use 10% (15%) (20%) symmetric trimming, otherwise use 10% (15%) (20%) one sided trimming.

22.-24. WJB1010(1515)(2020)-WJ with bootstrapping. If data is symmetric use 10% (15%) (20%) symmetric trimming, otherwise use 10% (15%) (20%) one sided trimming.

25.-27. WJJ1010(1515)(2020)-WJ with Johnson's transformation. If data is symmetric use 10% (15%) (20%) symmetric trimming, otherwise use 10% (15%) (20%) one sided trimming.

28.-30. WJJB1010(1515)(2020)-WJ with Johnson's transformation and bootstrapping. If data is symmetric use 10% (15%) (20%) symmetric trimming, otherwise use 10% (15%) (20%) one sided trimming.

31.-33. WJH1010(1515)(2020)-WJ with Hall's transformation. If data is symmetric use 10% (15%) (20%) symmetric trimming, otherwise use 10% (15%) (20%) one sided trimming.

34.-36. WJHB1010(1515)(2020)-WJ with Hall's transformation and bootstrapping. If data is symmetric use 10% (15%) (20%) symmetric trimming, otherwise use 10% (15%) (20%) one sided trimming.

37.-39. WJ1020(1530)(2040)-WJ. If data is symmetric use 10% (15%) (20%) symmetric trimming, otherwise use 20% (30%) (40%) one sided trimming.

40.-42. WJB1020(1530)(2040)-WJ with bootstrapping. If data is symmetric use 10% (15%) (20%) symmetric trimming, otherwise use 20% (30%) (40%) one sided trimming.

43.-45. WJJ1020(1530)(2040)-WJ with Johnson's transformation. If data is symmetric use 10% (15%) (20%) symmetric trimming, otherwise use 20% (30%) (40%) one sided trimming.

46.-48. WJJB1020(1530)(2040)-WJ with Johnson's transformation and bootstrapping. If data is symmetric use 10% (15%) (20%) symmetric trimming, otherwise use 20% (30%) (40%) one sided trimming.

49.-51. WJH1020(1530)(2040)-WJ with Hall's transformation. If data is symmetric use 10% (15%) (20%) symmetric trimming, otherwise use 20% (30%) (40%) one sided trimming.

52.-54. WJHB1020(1530)(2040)-WJ with Hall's transformation and bootstrapping. If data is symmetric use 10% (15%) (20%) symmetric trimming, otherwise use 20% (30%) (40%) one sided trimming.

Modified M-Estimators:

55. MOMH

56. MOMT

We examined: (a) the effect of using a preliminary test to determine whether data are symmetric or not in order to determine whether symmetric or asymmetric trimming should be adopted (We present in Appendix A a SAS/IML program that can be used to obtain the Q-statistics.), (b) the percentage of symmetric (10%, 15% or 20%) and asymmetric (10%, 15%, 20%, 30% or 40%) trimming used, (c) the utility of transforming the WJ statistic with either Johnson's (1978) or Hall's (1992) transformation, (d) the utility of bootstrapping the data, and (e) the use of two statistics with an estimator (MOM) that empirically determines whether data should be symmetrically or asymmetrically trimmed and by what amount, allowing also for the option of no trimming.

Additionally, four other variables were manipulated in the study: (a) sample size, (b) pairing of unequal variances and group sizes, and (c) population distribution.

We chose to investigate an unbalanced completely randomized design containing four groups since previous research has looked at this design (e.g., Lix & Keselman, 1998; Wilcox, 1988). The two cases of total sample size and the group sizes were $N = 70$ (10, 15, 20, 25) and $N = 90$ (15, 20, 25, 30). We selected our values of n_j from those used by Lix and Keselman (1998) in their study comparing omnibus tests for treatment group equality; their choice of values was, in part, based on having group sizes that others have found to be generally sufficient to provide reasonably effective Type I error control (e.g., see Wilcox, 1994). The unequal variances were in a 1:1:1:36 ratio. Unequal variances and unequal group sizes were both positively and negatively paired. For positive (negative) pairings, the group having the fewest (greatest) number of observations was associated with the population having the smallest (largest) variance, while the group having the greatest (fewest) number of observations was associated with the population having the largest (smallest) variance. These conditions were chosen since they typically produce conservative (liberal) results.

With respect to the effects of distributional shape on Type I error, we chose to investigate nonnormal distributions in which the data were obtained from a variety of skewed distributions. In addition to generating data from a χ^2_3 distribution, we also used the method described in Hoaglin (1985) to generate distributions with more extreme degrees of skewness and kurtosis. These particular types of nonnormal distributions were selected since educational and psychological research data typically have skewed distributions (Micceri, 1989; Wilcox, 1994). Furthermore, Sawilowsky and Blair (1992) investigated the effects of eight nonnormal distributions, which were identified by Micceri on the robustness of Student's t test, and they found that only distributions with the most extreme degree of skewness (e.g., $\gamma_1 = 1.64$) affected 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 χ^2_3 distribution, skewness and kurtosis values are $\gamma_1 = 1.63$ and $\gamma_2 = 4.00$, respectively. The other nonnormal distributions were generated from the g and h distribution (Hoaglin, 1985). Specifically, we chose to investigate two g and h distributions: (a) g = .5 and h = 0 and (b) g = .5 and h = .5, where g and h are parameters that determine the third moments of a distribution. To give meaning to these values it should be noted that for the standard normal distribution g = h = 0. Thus, when g = 0 a distribution is symmetric and the tails of a distribution will become heavier as h increases in value. Values of skewness and kurtosis corresponding to the investigated values of g and h are (a) $\gamma_1 = 1.75$ and $\gamma_2 = 8.9$, respectively, and (b) $\gamma_1 = \gamma_2 =$ undefined. These values of skewness and kurtosis for the g and h distributions are theoretical values; Wilcox (1997, p. 73) reports computer generated values, based on 100,000 observations, for these values--namely $\hat{\gamma}_1 = 1.81$ and $\hat{\gamma}_2 = 9.7$ for $g = .5$ and $h = 0$ and $\hat{\gamma}_1 = 120.10$ and $\hat{\gamma}_2 = 18{,}393.6$ for $g = .5$ and $h = .5$. Thus, the conditions we chose to investigate could be described as extreme. That is, they are intended to indicate the operating characteristics of the procedures under substantial departures from homogeneity and normality, with the premise being that, if a procedure works under the most extreme of conditions, it is likely to work under most conditions likely to be encountered by researchers.

In terms of the data generation procedure, to obtain pseudo-random normal variates, we used the SAS generator RANNOR (SAS Institute, 1989). If Z_{ij} is a standard unit normal variate, then $Y_{ij} = \mu_j + \sigma_j \times Z_{ij}$ is a normal variate with mean equal to μ_j and variance equal to σ_j^2. To generate pseudo-random variates having a χ^2 distribution with three degrees of freedom, three standard normal variates were squared and summed.

To generate data from a g- and h-distribution, standard unit normal variables were converted to random variables via

$$Y_{ij} = \frac{\exp(g Z_{ij}) - 1}{g} \exp\left(\frac{h Z_{ij}^2}{2}\right),$$

according to the values of g and h selected for investigation. To obtain a distribution with standard deviation σ_j, each Y_{ij} was multiplied by a value of σ_j. It is important to note that this does not affect the value of the null hypothesis when $g = 0$ (See Wilcox, 1994, p. 297). However, when $g > 0$, the population mean for a g- and h-distributed variable is

$$\mu_{gh} = \frac{1}{g(1-h)^{\frac{1}{2}}} (e^{g^2/2(1-h)} - 1)$$

(See Hoaglin, 1985, p. 503). Thus, for those conditions where $g > 0$, μ_{tj} was first subtracted from Y_{ij} before multiplying by σ_j. When working with MOMs, θ_j was first subtracted from each observation (The value of θ_j was obtained from generated data from the respective distributions based on one million observations.). Specifically, for procedures using trimmed means, we substrated μ_{tj} from the generated variates under every generated distribution. Correspondingly, for procedures based on MOMs, we substracted out θ_j for all distributions investigated.

Lastly, it should be noted that the standard deviation of a g- and h-distribution is not equal to one, and thus the values reflect only the amount that each random variable is multiplied by and not the actual values of the standard deviations (See Wilcox, 1994, p. 298). As Wilcox noted, the values for the variances (standard deviations) more aptly reflect the ratio of the variances (standard deviations) between the groups. Five thousand replications of each condition were performed using a .05 statistical significance level. According to Wilcox (1997) and Hall (1986), B was set at 599; that is, their results suggest that it may be advantageous to chose B such that $1 - \alpha$ is a multiple of $(B + 1)^{-1}$.

Results

For previous investigations, when we have evaluated Type I error rates, we adopted Bradley's (1978) liberal criterion of robustness. According to this criterion, in order for a test to be considered robust, its empirical rate of Type I error ($\widehat{\alpha}$) must be contained in the interval $0.5\alpha \leq \widehat{\alpha} \leq 1.5\alpha$. Therefore, for the five percent level of statistical significance used in this study, a test would be considered robust in a particular condition if its empirical rate of Type I error fell within the interval $.025 \leq \widehat{\alpha} \leq .075$. Correspondingly, a test was considered to be nonrobust if, for a particular condition, its Type I error rate was not contained in this interval. We have adopted this standard because we felt that it provided a reasonable standard by which to judge robustness. That is, it has been 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.

Type I error rates can be obtained from the first author's website at http://www.umanitoba.ca/faculties/arts/psychology. Based on this criterion of robustness, the procedures we investigated were remarkably robust to the cases of heterogeneity and nonnormality. That is, out of the 672 empirical values tabled (Tables 1-10) only 24, or approximately 3.5 percent of the values, did not fall within the .025-.075 interval (Values not falling in this interval are in boldface in the tables.).

Even though, in general, the procedures exhibited good Type I error control from the Bradley (1978) liberal criterion perspective, in the interest of making discriminations between the procedures, we went on to a second examination of the data adopting Bradley's stringent criterion of robustness. For this criterion, a statistic is considered robust, under a .05 significance level, if the empirical value falls in the interval .045-.055 (Nonbolded values not falling in this interval are underlined in the tables.). The tables as well contain information regarding the average Type I error rate and the number of empirical values not falling in the stringent interval for each procedure investigated; these values (excluding MOMH and MOMT values), along with the range of values over the 12 investigated conditions, are reproduced in summary form in Table 1.

Tests Based on MOMs. Of the 12 conditions examined, MOMH values ranged from .027 to .073, with an average value of .049; nine values fell outside of Bradley's (1978) stringent interval. MOMT values ranged from .014 to .060, with an average value of .038; six values fell outside the interval and most occurred when data were obtained from the $g = .5$ and $h = .5$ distribution. We describe our results predominately from Table 1; however, we, occasionally, also rely on the detailed information contained in the ten tables not contained in the paper.

<u>20% Symmetric and 20% (40%) Asymmetric Trimming</u>. Empirical results for 20% symmetric trimming conform to those reported in the literature. That is, the WJ test is generally robust with the liberal criterion of robustness, occasionally, however, resulting in a liberal rate of error (See Wilcox et al., 1998). Adopting a transformation for skewness improves rates of Type I error and further improvement is obtained when adopting bootstrap methods (See Luh & Guo, 1999). However, most of the values reported in the tables did not fall within the bounds of the stringent criterion. In particular, the number of these deviant values ranged from a low of 9 (WJJ20, WJH20, WJJB20) to a high of 12 (WJ20).

Keeping the total amount of trimmed values at 40%, regardless of whether data were trimmed symmetrically or asymmetrically, based on the preliminary test for symmetry, resulted in liberal rates of error, except when bootstrapping methods were adopted. Indeed, when bootstrapping was adopted for assessing statistical significance and a transformation was/was not applied to the statistic (WJJB2040, WJHB2040, WJB2040), rates of Type I error were well controlled; the number of values falling outside the stringent interval were two, two and four, respectively, with corresponding average rates of error of .048, .047 and .045.

<u>15% Symmetric and 15% (30%) Asymmetric Trimming</u>. Similar results were found to those previously reported, however, a few differences are noteworthy. First, none of the values fell ouside the liberal criterion, though with the exception of WJJ15 and WJH15, the number of values outside of the stringent criterion was large, obtaining values of 8 and 9. Also noteworthy is that for 15% symmetric trimming bootsrapping did not result in improved rates of Type I error.

On the other hand, bootstrapping was quite effective for contolling errors when trimming was based on the preliminary test for symmetry and either 15% or 30% of the data were trimmed symmetrically or asymmetrically. Without bootstrapping, rates, on occasion, reached values above .075 and the number of values falling outside the stringent criterion ranged from 7 to 12. With bootstrapping, no value exceeded .075, in fact no value exceeded .054, and the number of values outside the stringent criterion was small--3 (WJB1530), 3 (WJJB1530) and 2 (WJHB1530).

When trimming was 15%-symmetric or 15%-asymmetric, based on the preliminary test for symmetry, again, all empirical values were contained in the liberal interval, ranging from a low value of .025 (WJB1515) to a high value of .073 (WJH1515). However, the number of values falling outside the stringent interval varied over the tests examined, ranging from a low of 4 values (WJJB1515) to a high value of 9 values (WJB1515). The best two procedures were WJJB1515 (4 values outside the stringent criterion) and WJHB1515 (5 values outside the stringent criterion).

<u>10% Symmetric and 10% (20%) Asymmetric Trimming</u>. Results are not generally dissimilar from those reported for the other two trimming rules. That is, when adopting a 10% symmetric rule, all rates were contained in the liberal interval, though with the 10% rule, bootstrapping and transforming the statistic for skewness was effective in limiting the number of deviant values (WJJJB10 and WJHB10), while the remaining methods were not nearly as successful.

For 10% symmetric trimming or 20% asymmetric trimming, based on the preliminary test for symmetry, empirical rates were again best controlled when bootstrapping methods were applied. In particular, the number of deviant values ranged from 2 to 5, with fewer deviant values occurring when a transformation for skewness was applied to WJ (i.e., WJJB1020 and WJHB1020). The nonbootstrapped tests, on the other hand, frequently had rates falling outside the stringent interval; 8 for WJ1020 and 11 for WJJ1020 and WJH1020.

Adopting 10% symmetric or asymmetric trimming resulted in rates that generally also fell within the liberal criterion of Bradley (1978), except for two exceptions--.076 for WJH1010 and .023 for WJB1010. Once again, using a transformation to eliminate skewness and adopting bootstrapping to assess statistical significance resulted in relatively good Type I error control. That is, WJJB1010 and WJHB1010 had, respectively, 6 and 5 values falling outside the stringent interval, with corresponding average rates of error of .048 and .042.

Symmetric Trimming (10% vs 15% vs 20%). Our last examination of the data was a comparison of the rates of Type I error across the various percentages of symmetric trimming. Only two liberal values (.076 and .079), according to the .025-.075 criterion, were found across the three cases of symmetric trimming and they occurred under 20% symmetric trimming. The total number of values outside the .045-.055 criterion for 20%, 15% and 10% symmetric trimming were 58, 41 and 45, respectively; the corresponding average Type I error rates (across the six averages reported in the table) were .049, .047 and .050. The four procedures with the fewest values (i.e., 4) outside the stringent interval were WJJ15, WJH15, WJJB10 and WJHB10.

Discussion

In our investigation we examined various test statistics that can be used to compare treatment effects across groups in a one-way independent groups design. Issues that we examined were whether: (1) a preliminary test for symmetry can be used effectively to determine whether data should be trimmed symmetrically or asymetrically when used in combination with a heteroscedastic statisic that compares trimmed means, (2) the amount of trimming effects error rates of these heteroscedastic statistics, (3) transformations to these heteroscedastic statistics improve results, (4) bootstrapping methodology provides yet additional improvements and (5) an estimator (MOM) that empirically determines whether one should trim, and, if so, by what amount and from which tail(s) of the distribution, can effectively control rates of Type I error, and how those rates compare to the other methods investigated.

We found that the fifty-six procedures examined performed remarkably well. Of the 672 empirical values, only 24, or approximately 3.5 percent of the values, did not fall within the bounds of .025-.075, a criterion that many investigators have used to assess robustness. Based on this criterion, only six procedures did not perform well--namely MOMT, WJ2040, WJJ2040, WJH2040, WJJ1530 and WJH1530; that is, they all had two or more values less than .025 or greater than .075. The vast majority of these nonrobust values occurred under our most extreme case of nonnormality--$g = .5$ and $h = .5$.

On the basis of the more stringent criterion defined by Bradley (1978), five methods demonstrated exceptionally tight Type I error control. They were WJJB2040, WJHB2040, WJHB1530, WJJB1020 and WJHB1020. The number of values not falling in the stringent interval was two for each procedure. In addition, the average rate of error was .048, .047, .048, .049 and .049, respectively. Common to these six procedures is the use of a transformation to eliminate skewness [either Hall's (1978) or Johnson's (1992)] and the use of bootstrapping methodology to assess statistical significance. Two close competitors were the WJB1530 and WJJB1530 tests, each had three values outside .045-.055, with average rates of error of .045 and .049, respectively.

Based on our results we recommend WJJB1020 or WJHB1020; that is, the WJ heteroscedastic statistic which trims, based on a preliminary test for symmetry, 10% in each tail or 20% in one of the two tails and then transforms the test with a transformation to eliminate the effects of skewness [either Hall (1978) or Johnson (1992)] and where statistical significance is determined from bootstapping methodology. We recommend one of these methods, over the other three tests which also limited the number of discrepant values to two, because the other methods can result in greater numbers of data being discarded. It is our impression that applied researchers would prefer a method that compared treatment performance across groups with a measure of the typical score which was based on as much of the original data as possible--a very reasonable view. It is also worth mentioning that relatively good results are also possible by adopting a simpler WJ method--namely the WJ test with just bootstapping. In particular, WJB1530 and WJB2040 resulted in 3 and 4 values outside the stringent interval and each had an average Type I error rate of .045.

Another noteworthy finding was that other percentages of symmetric trimming work better in the one-way design than 20% symmetric trimming. In particular, we found four methods involving less trimming than 20% (WJJ15, WJH15, WJJB10 and WJHB10) that provided good Type I error control, resulting in fewer values outside .045-.055 than identical procedures based on 20% trimming. For two of the methods (WJJ15 and WJH15), bootstrapping methodology is not required.

We want to conclude by reminding the reader that we examined fifty-six test statistics under conditions of <u>extreme</u> heterogeneity and nonnormality. Thus, we believe we have identified procedures that are truly robust to cases of heterogeneity and nonnormality likely to be encountered by applied researchers and therefore we are very comfortable with our recommendation. That is, we believe we have found a very important result--namely, very good Type I error control is possible with relatively modest amounts of trimming.

We demonstrate the computations involved for obtaining the test of symmetry in Appendix A. We include this illustration, even though we provide software in Appendix B to obtain numerical results, because we believe it is instructive to see how Q_2 and Q_1 are obtained.

Appendix A

Test for Symmetry

Consider the problem of comparing distributions $F_1 = F_2 = \cdots = F_J$. One way of approaching this problem is to again consider the one-way ANOVA problem of comparing means $\mu_1 = \mu_2 = \cdots = \mu_J$ from J distributions $F_1(y) = F(y - \mu_1)$, $F_2(y) = F(y - \mu_2)$, $\ldots$, $F_J(y) = F(y - \mu_J)$. When the distributions are unknown and one cannot assume that they are normal with equal variances, Babu et al. (1999) suggested the following procedure:

1. First, one determines if the distributions are symmetric. To do so, they applied a procedure that uses two indices, Q_2 and Q_1, originally proposed by Hogg et al. (1975). The Q_2 index is first used to determine the nature of the tails of the distributions and then the Q_1 index is used to determine if the distributions are symmetric.
2. If the distributions are found to be symmetric then any of the WJ test statistics, based on symmetrically trimmed means is used to test for differences between distributions.
3. Otherwise, if the distributions are skewed, then any of the WJ statistics, based on asymmetrically (from the left or right tail for each treatment group) trimmed means is used to test for differences between distributions.

What we will be enumerating then are the <u>combined</u> Type I error rates for WJ tests where the preliminary test for symmetry determines whether in a particular simulation data are trimmed symmetrically or asymmetrically.

Let $Y_{ij} = (Y_{1j}, Y_{2j}, \ldots, Y_{n_j j})$ be a sample from an unknown distribution F_j. Again, let $Y_{(1)j} \leq Y_{(2)j} \leq \cdots \leq Y_{(n_j)j}$ represent the ordered observations associated with the jth group. Let γ be the proportion of the data in the sample that are of interest as either the proportion of data to be trimmed or the proportion of data to be used in the calculation of several intermediate variables leading to the Q_2 and Q_1 indices. Let $g = [\gamma n_j] + 1$, where $[x]$ represents the greatest integer less than γn_j and $r = g - \gamma n_j$. It is important to note that trimming here, and the amount trimmed, is just for purposes of assessing symmetry. Once the omnibus test (WJ) is used, the amount of trimming is based on a 15% or 20% rule.

<u>Preliminaries</u>

The Babu et al. (1999) procedure is based, in part, on the work of Hogg et al. (1975). Specifically, for these authors, the hypothesis of interest was

H_0: $\theta = 0$ against H_A: $\theta > 0$, where θ is the location parameter of interest. They proposed a test to detect the nature of the underlying distribution before proceeding with (nonparametric) tests of H_0.

In particular, they defined $Y_1, Y_2, \ldots, Y_m$ as a random sample from $F(y)$, and $Y_{m+1}, Y_{m+2}, \ldots, Y_n$ as a random sample from $F(y-\theta)$. Then $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$ are the ordered statistics of the combined random samples and Y_{med} and $\overline{Y}$ are, respectively, the median and mean of the combined samples.

Hogg et al.'s (1975) procedure to detect the nature of the underlying distribution, is composed of two tests, a test of the heaviness of the tail of the distribution using the Q_2 statistic and a test of symmetry using the Q_1 statistic. Their work was based on papers by Uthoff (1970, 1973).

Hogg et al. (1975) chose a test statistic enumerated by Uthoff (1973, Equation 2) as a basis to define their Q_2 index. This index determined whether the tail of the underlying distribution is light or heavy. They first approximated it as

$$\frac{Y_{(n)} - Y_{(1)}}{2\sum \left| Y_{(i)} - Y_{med} \right| / n}.$$

They then transformed this ratio into

$$Q_2 = \frac{(U_{0.05} - L_{0.05})}{(U_{0.5} - L_{0.5})},$$

where $U_{0.05}$ and $L_{0.05}$ are, respectively, the means of the upper and lower 5% of the order statistics of the combined samples and $U_{0.5}$ and $L_{0.5}$ are, respectively, the means of the upper and lower 50% of the order statistics of the combined sample.

Again, based on the work of Uthoff (1970, Equation 1), Hogg et al. (1975) derived their Q_1 index:

$$Q_1 = \frac{(U_{0.05} - MID)}{(MID - L_{0.05})},$$

where MID is the mean of the middle 50% of the combined sample. Thus, this index determines the symmetry of the underlying distribution.

Babu et al. (1999) extended the use of these two indices to testing the location parameters of more than two groups. They proposed that both indices be calculated within the groups and weighted means of these indices be the overall estimates of Q_2 and Q_1. They also proposed adjustments to the Q_1 index whereby the amount of data needed to calculate the index depended on the outcome of the calculation of the Q_2 index.

Determination of Symmetry

Q_2 Index. Prior to determining the symmetry of the distributions, the nature of their tails is examined. The Q_2 index determines whether $F_1(y), F_2(y), \ldots, F_J(y)$ are normal-tailed, heavy-tailed or very heavy-tailed. Tail classification is determined in the following manner:

1. Define $U_{\gamma,j}$ and $L_{\gamma j}$ as the means of the upper and lower γn_j order statistics, respectively, of the sample Y_j.

Case 1. If $\gamma n_j \leq 1$, then $U_{\gamma j} = Y_{(n_j)j}$ and $L_{\gamma j} = Y_{(1)j}$.

Case 2. If $\gamma n_j > 1$ then

$$U_{\gamma,\, i} = \frac{1}{\gamma n_j}\left(\sum_{j=n_i-g+2}^{n_j} Y_{(i)j} + (1-r)\, Y_{(n_j-g+1),\, j} \right) \text{ and}$$

$$L_{\gamma,\, j} = \frac{1}{\gamma n_j}\left(\sum_{i=1}^{g-1} Y_{(i)j} + (1-r)\, Y_{(g)j} \right).$$

2. Calculate $U_{0.05,\, j}$ and $L_{0.05,\, j}$ as the mean of the upper and lower $0.05n_j$ order statistics of Y_j, respectively.

3. Calculate $U_{0.5,\, j}$ and $L_{0.5,\, j}$ as the mean of the upper and lower $0.5n_j$ order statistics of Y_j, respectively.

4. For each j, set $Q_{2,\, j} = (U_{0.05,\, j} - L_{0.05,\, j}) / (U_{0.5,\, j} - L_{0.5,\, j})$.

5. Using $Q_{2,\, j}$, $j = 1, 2, \ldots, J$, from # 4 compute

$$Q_2 = \left(\sum_{j=1}^{J} n_j Q_{2,\, j} \right) \Big/ \left(\sum_{j=1}^{J} n_j \right). \tag{A.1}$$

6. If $Q_2 < 3$ then F is classified as normal-tailed. If $3 \leq Q_2 < 5$ then F is classified as heavy-tailed. If $Q_2 \geq 5$ then F is classified as very heavy-tailed.

Q_1 Index. Once the nature of the tails of the distributions is known, the Q_1 index, which determines the symmetry of the distributions, is calculated. To calculate the Q_1 index one should:

1. Based on Q_2, determine the number of sample points in each sample Y_j to be used. Define this as n_j^*. [This is Babu et al.'s (1999) modification of Hogg et al.'s (1975) proposal for computing Q_1.] Specifically, if $Q_2 < 3$ then use all sample points in Y_j. If $3 \leq Q_2 < 5$ then trim the top and bottom 10% of the sample points and use the middle 80% in Y_j. If $Q_2 \geq 5$ then trim the top and bottom 20% of the sample points and use the middle 60% in Y_j.

2. Let MID_j to be the mean of the middle 50% of the order statistics of the sample points in sample Y_j defined in #1. According to A. R. Padmanaban

(personal communication, June 26, 2001), MID_j is calculated in the following manner:

- Discard the top and bottom 25% of the order statistics of Y_j.
- The remainder is the middle 50% of the order statistics of Y_j.

Hence, $g^* = [0.25n_j^*] + 1$ and $r^* = g^* - 0.25n_j^*$. Therefore, MID_j is given by

$$MID_j = \frac{1}{0.5n_j^*}\left[\sum_{i=g^*+1}^{n_j^*-g^*} Y_{(i)j} + r^*\left(Y_{(g^*)\,j} + Y_{(n_j^*-g^*+1)\,j}\right)\right].$$

3. For each j, set $Q_{1,\,j} = (U_{0.05,\,j} - MID_j)/(MID_j - L_{0.05,\,j})$.
4. Using $Q_{1,\,j}$, $j = 1, 2, \ldots, J$, from # 3 compute

$$Q_1 = \left(\sum_{j=1}^{J} n_j^* Q_{1,\,j}\right) \Big/ \left(\sum_{j=1}^{J} n_j^*\right). \qquad \text{(A.2)}$$

5. If $Q_1 < \frac{1}{2}$, F is deemed to be left skewed. If $\frac{1}{2} \le Q_1 \le 2$, then F is considered to be symmetric. If $Q_1 > 2$, then F is designated as right skewed.

Computational Example

Suppose we want to test the null hypothesis, H_o: $F_1(x) = F_2(x) = F_3(x)$ based on the following data set.

Groups	Order Statistics	n_j
1	30 32 32 34 35 35 39 40 40 41 42 48 50 52 99	15
2	35 36 40 40 41 42 43 49 56 64	10
3	48 48 51 51 51 55 55 60 63 83	10

Note: The tabled values were chosen so that the data would be classified as heavy-tailed.

Calculating Q_2 (Tail thickness)

Notice that $0.05n_j < 1$ for $j = 1, 2, 3$, Therefore, $U_{0.05,\,1} = X_{(15,\,1)} = 99$, $U_{0.05,\,2} = X_{(10,\,2)} = 64$, $U_{0.05,\,3} = X_{(10,\,3)} = 83$, and $L_{0.05,\,1} = X_{(1,1)} = 30$, $L_{0.05,\,2} = X_{(1,\,2)} = 35$, $L_{0.05,\,3} = X_{(1,\,3)} = 48$. When $\gamma = 0.5$, the calculations for $U_{0.5,\,j}$, $L_{0.5,\,j}$ and $Q_{2,\,j}$ for each group are as follows.

Group 1. $n_1 = 15$, $0.5n_1 = 7.5$, $g = 8$ and $r = 0.5$.

$$U_{0.5,\,1} = \frac{1}{7.5}\left(\sum_{i=9}^{15} X_{(i1)} + 0.5X_{(8,\,1)}\right)$$
$$= \frac{1}{7.5}((40 + 41 + \cdots + 99) + (0.5)40)$$
$$= 52.2667$$

$$L_{0.5,1} = \frac{1}{7.5}\left(\sum_{i=1}^{7} X_{(i1)} + 0.5X_{(8,\,1)}\right)$$
$$= \frac{1}{7.5}((30 + 32 + \cdots + 39) + (0.5)40)$$
$$= 34.2667$$

$$Q_{2,\,1} = \frac{(99 - 30)}{(52.2667 - 34.2667)} = 3.8333$$

<u>Group 2.</u> $n_2 = 10$, $0.5n_2 = 5$, $g = 6$ and $r = 0$.

$$U_{0.5,\,2} = \frac{1}{5}\left(\sum_{1=6}^{10} X_{(i2)} + (0)X_{(5,\,2)}\right)$$
$$= \frac{1}{5}((42 + 43 + \cdots + 64) + 0)$$
$$= 50.8$$

$$L_{0.5,\,2} = \frac{1}{5}\left(\sum_{i=1}^{5} X_{(i2)} + (0)X_{(6,\,2)}\right)$$
$$= \frac{1}{5}((35 + 36 + \cdots + 41) + 0)$$
$$= 38.4$$

$$Q_{2,\,2} = \frac{(64 - 35)}{(50.8 - 38.4)} = 2.3387$$

<u>Group 3.</u> $n_3 = 10$, $0.5n_3 = 5$, $g = 6$ and $r = 0$.

$$U_{0.5,\,3} = \frac{1}{5}\left(\sum_{i=6}^{10} X_{(i3)} + (0)X_{(5,\,3)}\right)$$
$$= \frac{1}{5}((55 + 55 + \cdots + 83) + 0)$$
$$= 63.2$$

$$
\begin{aligned}
L_{0.5,\,3} &= \frac{1}{5}\left(\sum_{i=1}^{5} X_{(i3)} + (0)X_{(6,\,3)}\right) \\
&= \frac{1}{5}((48+48+\cdots+51)+0) \\
&= 49.8
\end{aligned}
$$

$$
Q_{2,\,3} = \frac{(83-48)}{(63.2-49.8)} = 2.6119
$$

Therefore

$$
Q_2 = \frac{(15(3.8333)+10(2.3387)+10(2.6119))}{(15+10+10)} = 3.0573 \text{ and}
$$

F is classified as heavy-tailed.

<u>Calculating Q_1</u>

Since F if classified as heavy-tailed, we have to symmetrically trim 10% of the data before calculating Q_1. (See the following table.)

Groups	Order Statistics Following 10% Symmetric Trimming	n_j^*
1	32 32 34 35 35 39 40 40 41 42 48 50 52	13
2	36 40 40 41 42 43 49 56	8
3	48 51 51 51 55 55 60 63	8

Notice that $0.05n_j^* < 1$ for $j = 1, 2, 3$, Therefore, $U^*_{0.05,\,1} = X^*_{(13,\,1)} = 52$, $U^*_{0.05,\,2} = X^*_{(8,\,2)} = 56$, $U^*_{0.05,\,3} = X^*_{(8,\,3)} = 63$, and $L^*_{0.05,\,1} = X^*_{(1,\,1)} = 32$, $L^*_{0.05,\,2} = X^*_{(1,\,2)} = 36$, $L^*_{0.05,\,3} = X^*_{(1,\,3)} = 48$. Let us calculate MID_j and $Q_{1,\,j}$, for $j = 1, 2, 3$.

<u>Group 1.</u> $n_1^* = 13$, $0.25n_1^* = 3.25$, $g^* = 4$ and $r^* = 0.75$.

$$\begin{aligned} MID_1 &= \frac{1}{6.5}\left(\sum_{i=5}^{9} X^*_{(i1)} + 0.75\left(X^*_{(4,\,1)} + X^*_{(10,\,1)}\right)\right) \\ &= \frac{1}{6.5}((35+39+40+40+\ 41)+(0.75)(35+42)) \\ &= 38.8846 \end{aligned}$$

$$Q_{1,\,1} = \frac{(52-38.8846)}{(38.8846-32)} = 1.905$$

<u>Group 2.</u> $n^*_2 = 8,\ \ 0.25n^*_2 = 2.,\ \ g^* = 3$ and $r^* = 0.$

$$\begin{aligned} MID_2 &= \frac{1}{4}\left(\sum_{i=3}^{6} X^*_{(i2)}\right) \\ &= \frac{1}{4}(40+41+42+43) \\ &= 41.5 \end{aligned}$$

$$Q_{1,\,2} = \frac{(56-41.5)}{(41.5-36)} = 2.6364$$

<u>Group 3.</u> $n^*_3 = 8,\ \ 0.25n^*_3 = 2,\ \ g^* = 3$ and $r^* = 0.$

$$\begin{aligned} MID_3 &= \frac{1}{4}\left(\sum_{i=3}^{6} X^*_{(i3)}\right) \\ &= \frac{1}{4}(51+51+55+55) \\ &= 53 \end{aligned}$$

$$Q_{1,\,3} = \frac{(63-53)}{(53-48)} = 2$$

Therefore

$$Q_1 = \frac{(13(1.905)+8(2.6364)+8(2))}{(13+8+8)} = 2.133 \text{ and}$$

F is classified as right skewed.

Appendix B

SAS/IML Program for Q-Statistics

```
*Checking for symmetry using the Q2 and Q1 indices presented in Babu,
 Padmanabhan and Puri (1999);
*This program details all the steps in obtaining the Q2 and Q1 indices;

OPTIONS NOCENTER;
PROC IML;
RESET NONAME;

*Although the Q2 and Q1 calculations differ, both share common steps;
*Hence, they are incorporated into one module QMOD with the variable
 QCHOICE being the switch that activates Q2 or Q1: 1 activates Q1 and 2
 activates Q2;

START QMOD(QCHOICE,Y,OSY,GINFO,Q) GLOBAL(NY,WOBS,BOBS,PER);
  G = INT(PER#NY);
  NYPRIME = NY - 2#G;
  NPRIME = SUM(NYPRIME);
  *Initialize group information matrix;
  IF QCHOICE = 1 THEN GINFO = J(BOBS,8,0);
  ELSE IF QCHOICE = 2 THEN GINFO = J(BOBS,9,0);
  *Initialize for first pass;
  F = 1;
  M = 0;
  DO J = 1 TO BOBS;
    SAMP = NY[J];
    SAMPPR = NYPRIME[J];
    L = M + SAMP;
    YT = Y[F:L];
    TEMP = YT;
    *Sorting group elements in ascending order;
    YT[RANK(TEMP),] = TEMP;
    FIRST = G[,J] + 1;
    LAST = SAMP - G[,J];
    FPRIME = F + FIRST - 1;
    LPRIME = F + LAST - 1;
    *Get group information;
    GINFO[J,1] = J;        *Group number;
    IF QCHOICE = 1 THEN DO;
      GINFO[J,2] = SAMPPR; *Possibly trimmed group size;
      GINFO[J,3] = FPRIME; *Starting position in possibly trimmed data
                           stream for group j;
      GINFO[J,4] = LPRIME; *Ending position in possibly trimmed data
```

```
                        stream for group j;
END; *if QCHOICE = 1;
ELSE IF QCHOICE = 2 THEN DO;
  GINFO[J,2] = SAMP; *Group size;
  GINFO[J,3] = F;    *Starting position in data stream for group j;
  GINFO[J,4] = L;    *Ending position in data stream for group j;
END; *if QCHOICE = 2;
*Calculating the mean of the upper and lower 5% of data in group j;
*This is common in both Q1 and Q2;
NJP05 = (LAST-FIRST+1)#0.05;
IF NJP05 <= 1 THEN DO;
  UP05J = YT[LAST];
  LP05J = YT[FIRST];
END; *if NJP05 <=1;
ELSE DO;
  A = INT(NJP05);
  FR = NJP05 - A;
  UP05 = YT[LAST-A+1:LAST];
  UP05J = (FR#YT[LAST-A] + SUM(UP05))/NJP05;
  LP05 = YT[FIRST:FIRST+A-1];
  LP05J = (SUM(LP05) + FR#YT[FIRST+A])/NJP05;
END; **if NJP05 > 1;
GINFO[J,5] = UP05J; *Upper 5% mean of group j;
GINFO[J,6] = LP05J; *Lower 5% mean of group j;
IF QCHOICE = 1 THEN DO;
  *Calculating the mean of the middle 50% of data in group j;
  *This calculation is done in Q1 only;
  NJP25 = (LAST-FIRST+1)#0.25;
  A = INT(NJP25);
  FR = NJP25 - A;
  ME = YT[FIRST+A+1:LAST-A-1];
  MIDJ = ((1-FR)#YT[FIRST+A] + SUM(ME) + (1-FR)#YT[LAST-A])/(2#NJP25);
  Q1J = (UP05J - MIDJ)/(MIDJ - LP05J);
  GINFO[J,7] = MIDJ; *Middle 50% mean of possibly trimmed group j;
  GINFO[J,8] = Q1J;  *Q1 index of group j;
END; *if QCHOICE = 1;
IF QCHOICE = 2 THEN DO;
  *Calculating the mean of the upper and lower 50% of data in group j;
  *This calculation is done in Q2 only;
  NJP5 = (LAST-FIRST+1)#0.5;
  A = INT(NJP5);
  FR = NJP5 - A;
  UP5 = YT[LAST-A+1:LAST];
  UP5J = (FR#YT[LAST-A] + SUM(UP5))/NJP5;
  LP5 = YT[FIRST:FIRST+A-1];
  LP5J = (SUM(LP5) + FR#YT[FIRST+A])/NJP5;
```

```
    Q2J = (UP05J - LP05J)/(UP5J - LP5J);
    GINFO[J,7] = UP5J; *Upper 50% mean of group j;
    GINFO[J,8] = LP5J; *Lower 50% mean of group j;
    GINFO[J,9] = Q2J;  *Q2 index of group j;
   END; *if QCHOICE = 2;
   *Update for next pass;
   M = L;
   F = F + NY[J];
   IF J = 1 THEN OSY = YT;
   ELSE OSY = OSY//YT;
  END; *DO J;
  IF QCHOICE = 1 THEN Q = SUM(GINFO[1:3,8]`#NYPRIME)/NPRIME;
  ELSE IF QCHOICE = 2 THEN Q = SUM(GINFO[1:3,9]`#NYPRIME)/NPRIME;
FINISH; *QMOD;

START SHOWGRP(X, GINFO);
  X1 = X[GINFO[1,3]:GINFO[1,4]]`;
  X2 = X[GINFO[2,3]:GINFO[2,4]]`;
  X3 = X[GINFO[3,3]:GINFO[3,4]]`;
  PRINT 'GRP1:' X1[FORMAT=3.0];
  PRINT 'GRP2:' X2[FORMAT=3.0];
  PRINT 'GRP3:' X3[FORMAT=3.0];
FINISH; *SHOWGRP;

START Q2Q1AD;
  PRINT 'DETAILED OUTPUT FOR THE Q-STATISTICS';
  *Calculating Q2;
  PER = 0; *Q2 does not require trimming of data;
  QCHOICE = 2;
  CALL QMOD(QCHOICE,Y,OSY,Q2INFO,Q2);
  PRINT ,;
  PRINT 'Y IN THE VARIOUS GROUPS';
  CALL SHOWGRP(Y,Q2INFO);
  PRINT ,;
  PRINT 'ORDER STATISTICS OF Y';
  CALL SHOWGRP(OSY,Q2INFO);
  OUTQ2 = Q2INFO[,1:2]||Q2INFO[,5:9];
   C1 = {"GRP" "GRP SIZE" "UP5% MEAN" "LO5% MEAN" "UP50% MEAN"
"LO50% MEAN" "Q2J"};
  PRINT ,;
  PRINT 'INTERMEDIATE OUTPUTS FOR Q2';
  PRINT OUTQ2[COLNAME=C1 FORMAT=10.4];
  PRINT 'Q2 =' Q2[FORMAT=10.4];
  IF Q2 < 3 THEN DO;
   PER = 0;
```

```
  PRINT 'DATA DISTRIBUTION IS NORMAL-TAILED. USE ALL DATA TO
DETERMINE Q1.';
  END; *if Q2 < 3;
  ELSE IF Q2 > 5 THEN DO;
    PER = 0.2;
  PRINT 'DATA DISTRIBUTION IS VERY HEAVY-TAILED. DO 20% SYMMETRIC
TRIMMING TO DETERMINE Q1.';
  END; *if Q2 > 5;
  ELSE DO; *if 3 <= Q2 <= 5;
    PER = 0.1;
  PRINT 'DATA DISTRIBUTION IS HEAVY-TAILED. DO 10% SYMMETRIC
TRIMMING TO DETERMINE Q1.';
  END; *if 3 <= Q2 <=5;
  *Calculating Q1;
  QCHOICE = 1;
  CALL QMOD(QCHOICE,Y,OSY,Q1INFO,Q1);
  PRINT /;
  PRINT 'ORDER STATISTICS OF POSSIBLY TRIMMED Y';
  CALL SHOWGRP(OSY,Q1INFO);
  OUTQ1 = Q1INFO[,1:2]||Q1INFO[,5:8];
   C2 = {"GRP" "GRP SIZE" "UP5% MEAN" "LO5% MEAN" "MID50% MEAN"
"Q1J"};
  PRINT ,;
  PRINT 'INTERMEDIATE OUTPUTS FOR Q1';
  PRINT OUTQ1[COLNAME=C2 FORMAT=10.4];
  PRINT 'Q1 =' Q1[FORMAT=10.4];
  IF Q1 < 0.5 THEN PRINT 'DATA DISTRIBUTION IS LEFT-SKEWED.';
  ELSE IF Q1 > 2 THEN PRINT 'DATA DISTRIBUTION IS RIGHT-SKEWED.';
  ELSE PRINT 'DATA DISTRIBUTION IS SYMMETRIC.'; *if 0.5 <= Q1 <= 2;
FINISH; *Q2Q1AD;

***INPUT DATA VECTOR;

*Data is purposely typed in the following manner to show where Groups 1-3
 entries are;

*SAS treats this as a 35x1 column vector;
Y = {42, 40, 32, 48, 32, 52, 41, 35, 30, 99, 40, 35, 34, 39,
50, 49, 35, 43, 36, 40, 56, 41, 40, 64, 42,
48, 51, 63, 51, 60, 51, 83, 55, 55, 48};

*Group sizes are entries in the following 1x3 row vector;
NY = {15 10 10};

*WOBS and BOBS are variable names carried over from past programs;
*WOBS = within subjects groups;
```

```
WOBS = NCOL(Y);
*BOBS = between subject groups;
BOBS = NCOL(NY);
RUN Q2Q1AD;
```

DETAILED OUTPUT FOR THE Q-STATISTICS

Y IN THE VARIOUS GROUPS

GRP1: 42 40 32 48 32 52 41 35 30 99 40 35 34 39 50

GRP2: 49 35 43 36 40 56 41 40 64 42

GRP3: 48 51 63 51 60 51 83 55 55 48

ORDER STATISTICS OF Y

GRP1: 30 32 32 34 35 35 39 40 40 41 42 48 50 52 99

GRP2: 35 36 40 40 41 42 43 49 56 64

GRP3: 48 48 51 51 51 55 55 60 63 83

INTERMEDIATE OUTPUTS FOR Q2

GRP	GRP SIZE	UP5% MEAN	LO5% MEAN	UP50% MEAN	LO50% MEAN	Q2J
1	15	99	30	52.2667	34.2667	3.8333
2	10	64	35	50.8	38.4	2.3387
3	10	83	48	63.2	49.8	2.6119

Q2 = 3.0573

DATA DISTRIBUTION IS HEAVY-TAILED. DO 10% SYMMETRIC TRIMMING TO DETERMINE Q1.

ORDER STATISTICS OF POSSIBLY TRIMMED Y

GRP1: 32 32 34 35 35 39 40 40 41 42 48 50 52

GRP2: 36 40 40 41 42 43 49 56

GRP3: 48 51 51 51 55 55 60 63

INTERMEDIATE OUTPUTS FOR Q1

GRP	GRP SIZE	UP5% MEAN	LO5% MEAN	MID50% MEAN	Q1J
1	13	52	32	38.8846	1.9050
2	8	56	36	41.5	2.6364
3	8	63	48	53	2

Q1 = 2.1330

DATA DISTRIBUTION IS RIGHT-SKEWED.

References

Babu, J. G., Padmanabhan A. R., & Puri, M. P. (1999). Robust one-way ANOVA under possibly non-regular conditions. Biometrical Journal, 41(3), 321-339.

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

De Wet, T., & van Wyk, J. W. J. (1979). Efficiency and robustness of Hogg's adaptive trimmed means. Communications in Statistics, Theory and Methods, A8(2), 117-128.

Guo, J. -H., & Luh, W. -M. (2000). An invertible transformation two-sample trimmed t-statistic under heterogeneity and nonnormality. Statistics & Probability Letters, 49, 1-7.

Hall, P. (1986). On the number of bootstrap simulations required to construct a confidence interval. Annals of Statistics, 14, 1431-1452.

Hall, P. (1992). On the removal of skewness by transformation. Journal of the Royal Statistical Society, Series B, 54, 221-228.

He, X., Simpson, D. G., & Portnoy, S. L. (1990). Breakdown robustness of tests. Journal of the American Statistical Association, 85, 446-452.

Hoaglin, D. C. (1985). Summarizing shape numerically: The g- and h-distributions. In D. Hoaglin, F. Mosteller, & J. Tukey (Eds.), Exploring data tables, trends, and shapes (pp. 461-513). New York: Wiley.

Hogg, R. V., Fisher, D. M., & Randles, R. H. (1975). A two-sample adaptive distribution free test. Journal of the American Statistical Association, 70, 656-661.

Huber, P. J. (1993). Projection pursuit and robustness. In S. Morgenthaler, E. Ronchetti, & W. Stahel (Eds.) New Directions in Statistical Data Analysis and Robustness. Boston: Verlag.

Johansen, S. (1980). The Welch-James approximation to the distribution of the residual sum of squares in a weighted linear regression. Biometrika, 67, 85-92.

Johnson, N. J. (1978). Modified t tests and confidence intervals for asymmetrical populations. Journal of the American Statistical Association, 73, 536-544.

Keselman, H. J., Kowalchuk, R. K., & Lix, L. M. (1998). Robust nonorthogonal analyses revisited: An update based on trimmed means. Psychometrika, 63, 145-163.

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

Keselman, H. J., Wilcox, R. R., & Lix, L. M. (2001, April). A robust approach to hypothesis tesing. Paper presented at the annual meeting of the Western Psychological Association, Maui, HI.

Keselman, H. J., Wilcox, R. R., Taylor, J., & Kowalchuk, R. K. (2000). Tests for mean equality that do not require homogeneity of variances: Do they really work? Communications in Statistics, Simulation and Computation, 29, 875-895.

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

Lix, L. M., & Keselman, H. J. (1998). To trim or not to trim: Tests of location equality under heteroscedasticity and non-normality. Educational and Psychological Measurement, 58, 409-429 (58, 853).

Lix, L. M., & Keselman, H. J. (1995). Approximate degrees of freedom tests: A unified perspective on testing for mean equality. Psychological Bulletin, 117, 547-560.

Luh, W. & Guo, J. (1999). A powerful transformation trimmed mean method for one-way fixed effects ANOVA model under non-normality and inequality of variances. British Journal of Mathematical and Statistical Psychology, 52, 303-320.

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

Rocke, D. M., Downs, G. W., Rocke, A. J. (1982). Are robust estimators really necessary? Technometrics, 24(2), 95-101.

Rousseeuw, P. J. & van Zomeren, B. C. (1990). Unmasking multivariate outliers leverage points. Journal of the American Statistical Association, 85, 633-639.

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

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

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.

Schrader, R. M., & Hettmansperger, T. P. (1980). Robust analysis of variance. Biometrika, 67, 93-101.

Tiku, M.L. (1980). Robustness of MML estimators based on censored samples and robust test statistics. Journal of Statistical Planning and Inference, 4, 123-143.

Tiku, M.L. (1982). Robust statistics for testing equality of means and variances. Communications in Statistics, Theory and Methods, 11(22), 2543-2558.

Uthoff, V. A. (1970). An optimum test property of two well-known statistics. Journal of the American Statistical Association, 65, 1597-1600.

Uthoff, V. A. (1973). The most powerful scale and location invariant test of the normal versus the double exponential. Annals of Statistics, 1, 170-174.

Welch B. L. (1938). The significance of the difference between two means when the population variances are unequal. Biometrika, 29, 350-362.

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. Biometrika, 38, 330-336.

Westfall, P. H., & Young, S. S. (1993). Resampling-based multiple testing. New York: Wiley.

Wilcox, R. R. (1988). A new alternative to the ANOVA F and new results on James's second-order method. British Journal of Mathematical and Statistical Psychology, 41, 109-117.

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

Wilcox, R. R. (1995). ANOVA: A paradigm for low power and misleading measures of effect size? Review of Educational Research, 65(1), 51-77.

Wilcox, R. R. (1997). Introduction to robust estimation and hypothesis testing. San Diego: Academic Press.

Wilcox, R. R., & Keselman, H. J. (in press). Repeated measures one-way ANOVA based on a modified one-step M-estimator. British Journal of Mathematical and Statistical Psychology.

Wilcox, R. R., & Keselman, H. J. (2002). Some Modern data analysis methods: Basics and recent developments. Manuscript submitted for publication.

Wilcox, R. R., Keselman, H. J., & Kowalchuk, R. K. (1998). Can tests for treatment group equality be improved?: The bootstrap and trimmed means conjecture. British Journal of Mathematical and Statistical Psychology, 51, 123-134.

Table 1. WJ Summary Statistics

20% Symmetric Trimming						
	WJ20	WJJ20	WJH20	WJB20	WJJB20	WJHB20
Range	.041-.079	.043-.075	.043-.076	.030-.047	.033-.047	.033-.047
Average	.058	.056	.056	.040	.041	.041
# of Nonrobust Values	12	9	9	10	9	10
20% Symmetric and 40% Asymmetric Trimming						
	WJ2040	WJJ2040	WJH2040	WJB2040	WJJB2040	WJHB2040
Range	.059-.084	.051-.077	.051-.079	.040-.053	.037-.053	.037-.052
Average	.071	.066	.068	.045	.048	.047
# of Nonrobust Values	12	11	11	4	2	2
20% Symmetric and 20% Asymmetric Trimming						
	WJ2020	WJJ2020	WJH2020	WJB2020	WJJB2020	WJHB2020
Range	.048-.075	.054-.071	.054-.072	.030-.051	.033-.055	.034-.054
Average	.059	.060	.060	.043	.047	.046
# of Nonrobust Values	8	9	9	6	4	4
15% Symmetric Trimming						
	WJ15	WJJ15	WJH15	WJB15	WJJB15	WJHB15
Range	.036-.067	.047-.067	.048-.067	.025-.047	.033-.048	.032-.048
Average	.051	.053	.054	.039	.042	.041
# of Nonrobust Values	8	4	4	9	8	8
15% Symmetric and 30% Asymmetric Trimming						
	WJ1530	WJJ1530	WJH1530	WJB1530	WJJB1530	WJHB1530
Range	.057-.078	.050-.079	.050-.082	.035-.049	.041-.054	.039-.054
Average	.064	.063	.064	.045	.049	.048
# of Nonrobust Values	12	7	9	3	3	2

15% Symmetric and 15% Asymmetric Trimming						
	WJ1515	WJJ1515	WJH1515	WJB1515	WJJB1515	WJHB1515
Range	.043-.065	.053-.072	.053-.073	.025-.045	.037-.050	.036-.050
Average	.053	.059	.060	.039	.046	.045
# of Nonrobust Values	7	8	8	9	4	5
10% Symmetric Trimming						
	WJ10	WJJ10	WJH10	WJB10	WJJB10	WJHB10
Range	.038-.075	.053-.072	.055-.073	.025-.048	.033-.053	.033-.053
Average	.053	.059	.060	.039	.045	.043
# of Nonrobust Values	10	9	9	9	4	4
10% Symmetric and 20% Asymmetric Trimming						
	WJ1020	WJJ1020	WJH1020	WJB1020	WJJB1020	WJHB1020
Range	.047-.075	.055-.072	.056-.074	.032-.052	.039-.057	.041-.057
Average	.059	.062	.063	.044	.049	.049
# of Nonrobust Values	8	11	12	5	2	2
10% Symmetric and 10% Asymmetric Trimming						
	WJ1010	WJJ1010	WJH1010	WJB1010	WJJB1010	WJHB1010
Range	.038-.075	.055-.075	.056-.076	.023-.050	.033-.058	.032-.058
Average	.054	.064	.065	.039	.048	.042
# of Nonrobust Values	10	11	12	7	6	5

Note: Nonrobust values are those outside the interval .045-.055.