TO TRIM OR NOT TO TRIM: TESTS OF LOCATION EQUALITY UNDER HETEROSCEDASTICITY AND NONNORMALITY

Lisa M. Lix and H.J. Keselman

University of Manitoba

Correspondence concerning this manuscript should be sent to: Lisa M. Lix, Department of Clothing and Textiles, Faculty of Human Ecology, University of Manitoba, Winnipeg, Manitoba R3T 2N2, (204)-474-8064, lix@bldghumec.lan1.umanitoba.ca

Abstract

Tests of mean equality proposed by Alexander and Govern (1994), Box (1954), Brown and Forsythe (1974), James (1951), and Welch (1951), as well as the analysis of variance F test, were compared for their ability to limit the number of Type I errors and to detect true treatment group differences in one-way completely randomized designs where the underlying distributions were nonnormal, variances were nonhomogeneous, and groups sizes were unequal. These tests were compared when the usual method of least squares was applied to estimate group means and variances and when Yuen's (1974) trimmed means and Winsorized variances were adopted. In the former case the procedures can be used to test for population mean equality, while in the latter case they can be used to test for equality of the population trimmed means. Based on the variables examined in this investigation, which included number of treatment groups, degree of population skewness, nature of the pairing of variances and group sizes, and nonnull effects of varying sizes, we recommend that researchers utilize trimmed means and Winsorized variances with either the Alexander and Govern (1994), James (1951) or Welch (1951) tests to test for mean equality.

TO TRIM OR NOT TO TRIM: TESTS OF LOCATION EQUALITY UNDER HETEROSCEDASTICITY AND NONNORMALITY

Testing for mean equality in the presence of unequal variances has a long history in the statistical literature dating back to the time of Behrens (1929) and Fisher (1935). Since this early work, numerous authors have offered potential solutions to the problem. Perhaps the most well-known of these is the approximate degrees of freedom (df) solution for the one-way completely randomized design provided by Welch (1951). Two other solutions that are frequently recommended in the literature are the James (1951) second-order and Brown and Forsythe (1974) approximation methods. Other less well-known solutions have also been proposed. Rubin's (1983) findings regarding the poor asymptotic performance of the Brown and Forsythe (1974) statistic led her to recommend Box's (1954) method which involves modifying the numerator df of the Brown and Forsythe statistic. Alexander and Govern (1994) proposed a solution which is based on a series of one-sample statistics. These statistics are combined, and the final solution, like that of James (1951), is based on large sample theory and utilizes a χ^2 statistic.

All of these procedures, with the exception of the one suggested by Rubin (1983), have been investigated in empirical studies; the evidence suggests that these methods can generally control the rate of Type I error when group variances are heterogeneous and the data are normally distributed (e.g., Alexander & Govern, 1994; Dijkstra & Werter 1981; Oshima & Algina, 1992; Wilcox, 1990). However, the literature also indicates that these tests can become liberal when the data are both heterogeneous and nonnormal, particularly when the design is unbalanced. Thus, these statistics have limitations, namely their sensitivity to the nature of the population distributions.

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 score distributions. In particular, the standard error of the usual

mean can become seriously inflated when the underlying distribution has heavy tails. Accordingly, adopting a nonrobust measure "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). By substituting robust measures of location and scale for the usual mean and variance, it should be possible to obtain test statistics which are 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). 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 censored or 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.

However, it should be noted at the outset that these measures should only be adopted if the researcher is interested in testing for treatment effects across groups using a measure of location that more accurately reflects the typical score within a group when working with heavy-tailed distributions. The hypothesis tested when the usual mean is used as an estimate of location is <u>not</u> the same as that tested when the trimmed mean is employed. Consequently, the researcher needs to be clear on the goals of data analysis prior to choosing a particular method of statistical inference and must clearly communicate these goals to all who will evaluate the results.

In the present paper, we are primarily concerned with extending procedures for comparing treatment groups in the presence of variance heterogeneity in order to also achieve robustness against nonnormality. Yuen (1974) initially suggested that trimmed means and Winsorized variances be used in conjunction with Welch's (1951) statistic. For heavy-tailed symmetric distributions, Yuen found that the statistic based on trimmed means and Winsorized variances could adequately control the rate of Type I errors and resulted in greater power than a statistic based on the usual mean and variance. However, to date, no study has compared all of the previously enumerated tests employing trimmed means and Winsorized variances.

Therefore, the purposes of our investigation were to determine whether the use of trimmed means and Winsorized variances with the Alexander and Govern (1994), Box (1954), Brown and Forsythe (1974), James (1951), and Welch (1951) statistics will result in robust tests for mean equality when the data are both heterogeneous and nonnormal in form and group sizes are unequal, and which of the robust procedures will be most sensitive for detecting treatment effects.

Definition of the Test Statistics

Suppose n_j independent random observations $X_{1j}, X_{2j}, \ldots, X_{n_j j}$ are sampled from population j ($j = 1, \ldots, J$). We assume that the X_{ij}s ($i = 1, \ldots, n_j$) are obtained from a normal population with mean μ_j and unknown variance σ_j^2, with $\sigma_j^2 \neq \sigma_{j'}^2$ ($j \neq j'$). Then, let $\bar{X}_j = \Sigma_i X_{ij}/n_j$ and $s_j^2 = \Sigma_i(X_{ij} - \bar{X}_j)^2/(n_j - 1)$, where $\bar{X}_j$ is the estimate of μ_j and s_j^2 is the usual unbiased estimate of the variance for population j. Further, let the standard error of the mean be denoted as $S_j = (s_j^2/n_j)^{\frac{1}{2}}$ and let $w_j = 1/S_j^2/(\Sigma_j 1/S_j^2)$.

The procedures presented by Alexander and Govern (1994), Brown and Forsythe (1974), James (1951), and Welch (1951) for testing the null hypothesis H_0: $\mu_1 = \mu_2 =$

$... = \mu_J$ in the presence of variance heterogeneity may all be obtained from a single general result. That is, for each group one can compute

$$t_j = \frac{\bar{X}_j - \hat{\mu}}{S_j}, \quad (1)$$

where $\hat{\mu} = \sum_{j=1}^{J} w_j \bar{X}_j$, the variance weighted grand mean.

In order to test the null hypothesis of mean equality, Welch (1951), James (1951), and Brown and Forsythe (1974) derived statistics which relate to $\Sigma_j t_j^2$ (see Alexander & Govern, 1994 for the definition of these approximate statistics). These test statistics reference either the χ^2 or F distributions.

In Alexander and Govern's (1994) solution, a normalizing transformation is first applied to each t_j. These normalized values (say, n-scores) are then used to derive a statistic ($\Sigma_j n_j^2$) that is distributed as a χ^2variable.

As previously noted, Rubin (1983) demonstrated that the Brown and Forsythe (1974) procedure is not asymptotically correct. Furthermore, she found that a better test for mean equality could be obtained by incorporating Box's (1954) procedure of adopting a corrected numerator df, as well as the usual denominator df correction. This statistic (F') is defined as

$$F' = \frac{\sum_{j=1}^{J} n_j(\bar{X}_j - \bar{X})^2}{\sum_{j=1}^{J} [1-(n_j/N)]s_j^2}, \quad (2)$$

where $\bar{X} = \sum_{j=1}^{J} n_j \bar{X}_j / N$ and $N = \sum_{j=1}^{J} n_j$.

According to Box (1954), F' is approximately distributed as an F variable with ν_1' and ν_2' df, where

$$\nu_1' = \frac{\left[\sum_{j=1}^{J}(1-f_j)s_j^2\right]^2}{\left(\sum_{j=1}^{J}s_j^2 f_j\right)^2 + \sum_{j=1}^{J}s_j^4(1-2f_j)}\text{ , and} \tag{3}$$

$$\nu_2' = \frac{\left(\sum_{j=1}^{J}(1-f_j)s_j^2\right)^2}{\sum_{j=1}^{J}\frac{s_j^4(1-f_j)^2}{(n_j-1)}}\text{ , and} \tag{4}$$

$f_j = n_j/N$.

Another consideration in the present paper was the application of robust estimates of the group means and variances to these various test procedures. When trimmed means are being compared the null hypothesis pertains to the equality of population trimmed means, i.e., the μ_ts. That is, H_0: $\mu_{t1} = \mu_{t2} = \ldots = \mu_{tJ}$ and H_{0j}: $\mu_{tj} = \mu_t$, $[H_{A_j} : \mu_{tj} \neq \mu_t]$. Let $X_{(1)j} \leq X_{(2)j} \leq \ldots \leq X_{(n_j)j}$ represent the ordered observations associated with the *j*th group. When trimming let $g_j = [\gamma_s\, n_j]$, where γ_s represents the proportion of observations that are to be trimmed in each tail of the distribution. The effective sample size for the *j*th group becomes $h_j = n_j - 2g_j$, and thus the *j*th sample trimmed mean is

$$\overline{X}_{tj} = \frac{1}{h_j}\sum_{i=g_j+1}^{n_j-g_j} X_{(i)j}\ . \tag{5}$$

In order to compute the sample Winsorized variance, the sample Winsorized mean is necessary and is computed as

$$\overline{X}_{wj} = \frac{1}{n_j}\sum_{i=1}^{n_j} Y_{ij}, \tag{6}$$

where

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

The sample Winsorized variance is then given by

$$s^2_{wj} = \frac{1}{h_j - 1} \sum_{i=1}^{n_j} (Y_{ij} - \overline{X}_{wj})^2, \tag{7}$$

and the standard error of the mean is

$$S_{tj} = \sqrt{\frac{s^2_{wj}}{h_j}} . \tag{8}$$

Under robust estimation, the trimmed group means, Winsorized group variances, and Winsorized group standard errors of the means were substituted in the appropriate equation for a particular test statistic. For example, under trimming, Equation 1 becomes

$$t_{tj} = \frac{\overline{X}_{tj} - \widehat{\mu}_t}{S_{tj}}, \tag{10}$$

where $\widehat{\mu}_t = \sum_{j=1}^{J} w_{tj} \overline{X}_{tj}$,

and

$$w_{tj} = \frac{1/S^2_{tj}}{\sum_{j=1}^{J} 1/S^2_{tj}} .$$

This statistic, t_{tj}, can be approximated as a t-variable with $h_j - 1$ df.

Method

Twelve tests for mean equality were compared for their rates of Type I error under conditions of nonnormality and variance heterogeneity in one-way independent groups designs. These tests resulted from crossing the Alexander and Govern (1994), Box (1954), Brown and Forsythe (1974), James (1951) second-order, Welch (1951), and usual ANOVA F statistics with two methods for estimating group means and variances, Yuen's robust estimation method, which uses a trimmed mean and Winsorized variance (see Wilcox, 1993; Yuen & Dixon, 1973), and the usual least squares estimators for the mean and variance. The ANOVA F test was included only to serve as a baseline measure for comparison purposes.

Six variables were manipulated in the study: (a) number of groups (2, 4, 6 and 10), (b) sample size (two cases), (c) degree/pattern of variance heterogeneity (two cases), (d) pairing of unequal variances and group sizes (positive and negative), (e) population distribution (normal and nonnormal), and (f) magnitude of the nonnull treatment means (two cases).

We chose to investigate completely randomized designs containing two, four, six and ten groups since previous research has looked at these designs (e.g., Wilcox, 1988). In fact, most of the investigated conditions were selected because they were either similar to or employed in previous studies (e.g., Dijkstra & Werter, 1981; Oshima & Algina, 1992; Wilcox, Charlin & Thompson, 1986) and thus allowed us to compare the procedures under conditions which are known to highlight the strength and weaknesses of tests for location equality. For this reason, only unbalanced designs were considered. Table 1 contains the numerical values of the sample sizes and variances investigated in the study, and also the nature of the pairings of the sample sizes and variances. 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.

Insert Table 1 About Here

With respect to the effects of distributional shape on Type I error, we chose to investigate the normal distributions as well as conditions in which the data were obtained from a wide variety of skewed distributions. In addition to generating data from χ^2_3 and χ^2_6 distributions, 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, 1994a). 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 χ^2_6 distribution was included in our investigation in order to examine the effects of sampling from a distribution with moderate skewness. For this distribution, $\gamma_1 = 1.16$ and $\gamma_2 = 2.00$. The other types of nonnormal distributions were generated from the g- and h-distribution (Hoaglin, 1985). Specifically, we chose to investigate two g- and h- distributions: (a) $g = 1/h = 0$ and (b) $g = 1/h = .5$. 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) $\gamma_1 = 6.2$ and $\gamma_2 = 114$, respectively, and (b) $\gamma_1 = \gamma_2 =$ undefined. Finally, it should be noted that though the selected combinations of g and h result in extremely skewed distributions, these values, according to Wilcox, are representative of psychometric measures.

The last variable manipulated was the magnitude of nonnull treatment effects. Empirical power rates were collected for only two of the designs that were examined when Type I error rates were collected: J = 4 and J = 10. Only these two designs were examined since we felt they would suffice to provide a comparison between the procedures for small and large designs. Mean values were selected such that ceiling and floor effects would be minimized for the conditions investigated.

It should be noted that for all of the investigated distributions, we always applied symmetric trimming, removing 20% of the observations from each tail of a groups' set of scores, since this rule is well established (see Rosenberger & Gasko, 1983; Wilcox, 1994b, 1996b). This rule is based in part on optimizing power for nonnormal as well as normal distributions (see Wilcox, 1994a).

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 $X_{ij} = \mu_j + \sigma_j \times Z_{ij}$ 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 (six) degrees of freedom, three (six) standard normal variates were squared and summed. The variates were standardized, and then transformed to χ_3^2 or χ_6^2 variates having mean μ_j (when comparing the tests based on the least squares estimates) or μ_{tj} (when comparing the tests based on trimed means) and variance σ_j^2 [see Hastings & Peacock (1975), pp. 46-51, for further details on the generation of data from these distributions].

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

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

according to the values of g and h selected for investigation. To obtain a distribution with standard deviation σ_j, each X_{ij} ($j = 1, \ldots, J$) was multiplied by a value of σ_j obtainable from Table 1. 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_{gh} was first subtracted from X_{ij} before multiplying by σ_j. When working with trimmed means, μ_{tj} was first subtracted from each observation.

Lastly, it should be noted that the standard deviation of a g- and h-distribution is not equal to one, and thus the values enumerated in Table 1 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) in Table 1 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.

Results

Type I Error Rates

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 ($\widehat{\alpha}$) 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 was 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. In the tables, bolded entries are used to denote these latter values. 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. Nonetheless, the reader should be aware that there is no one universal standard by which tests are judged to be robust or not and thus with other standards different interpretations of the results are possible.

Preliminary analysis of the data indicated that there was a high degree of similarity in the results obtained for the two investigated cases of sample size. Therefore, the tabled values have been averaged over these two cases. In discussing the tabled values, the ANOVA F, Alexander and Govern (1994), Box (1954), Brown and Forsythe (1974), James (1951), and Welch (1951) tests are denoted by the abbreviations F, AG, BOX, BF, J, and W, respectively.

J = 2

Least Squares Estimation. The J = 2 results are presented in Table 2. When J = 2, the BF, BOX, and W tests are equivalent; the J and AG tests will also be approximately equal to this common value as well. Therefore, in this table, they are presented as a single test, notated as F^*. When the data were obtained from normal

distributions the tests behaved as expected, based on the findings of previous research. That is, when variances and sample sizes were positively (negatively) paired the F test resulted in conservative (liberal) values, while the approximate tests were robust to variance heterogeneity. However, when variance heterogeneity was combined with nonnormality and the design was unbalanced the approximate tests were affected as well. Not surprisingly, the degree of Type I error control was directly related to the degree of skewness of the data. When sampling from the χ^2_6 distribution, which exhibits very mild skewness, the approximate tests were robust. With increasing departures from symmetry the empirical rates (%) of Type I error for the approximate tests became progressively larger such that when sampling from the distribution with the greatest amount of skewness ($g = 1/h = .5$), the rate of Type I error exceeded 40%. These results are consistent with those presented by Wilcox (1994; 1995b).

Insert Table 2 About Here

Robust Estimation. When employing trimmed means and Winsorized variances, the results were very different, at least for the approximate tests. That is, except in one instance, the approximate tests were robust to nonnormality. Indeed, even under the most severe departure from normality, the empirical rates were well controlled and were less than the 5% statistical significance level. Only when sampling from the distribution where $\gamma_1 = 6.2$ and $\gamma_2 = 114$ ($g = 1/h = 0$) did an empirical rate exceed the upper bound of Bradley's (1978) criterion (i.e., 8.06%) and this occurred when the variances were in the ratio of 1:36 and negatively paired with sample sizes. Thus, the approximate procedures were very effective in controlling the rate of Type I error when employing trimmed means and Winsorized variances.

J = 4

Least Squares Estimation. The J = 4 rates of error are contained in Table 3. Once again, when the data were normally distributed or only slightly skewed (i.e., χ^2_6), the ANOVA F test tended to display its characteristic conservative or liberal tendencies when group sizes and variances were either positively or negatively paired; conservative rates were however, above Bradley's (1978) lower bound of 2.50% while the liberal rates were well above the 7.50% upper bound specified in the Bradley interval. The normal and χ^2_6 distributions were associated with well-controlled rates for all of the approximate tests, with the exception of the BF procedure, which was always liberal under conditions F and H. This finding is consistent with that of previous research, which indicates that patterns of variance heterogeneity in which there is a single aberrant value lead to nonrobust tendencies for the BF test (e.g., Tomarkin & Serlin, 1986). When sampling from distributions with more extreme degrees of nonnormality, all of the approximate procedures displayed nonrobust tendencies. Specifically, as the degree of skewness increased, so did the rates of Type I error. Indeed, when sampling from the g- and h-distribution where g = 1/h = .5, the rates of error could exceed 50 %. As well, the approximate tests' rates always resulted in liberal values for the positive pairings of group sizes and variances and frequently were larger than the corresponding F test rates.

Insert Table 3 About Here

Robust Estimation. The empirical rates of Type I error for the approximate tests employing trimmed means and Winsorized variances were not too disimilar from the rates of the tests using least squares estimators when sampling from the normal or χ^2_6 distributions. That is, the rates were generally well controlled for all but the BF test under Conditions F and H. However, for all other investigated distributions the tests with trimmed means and Winsorized variances resulted in much better Type I error control

than their least squares counterparts. When data were obtained from the χ^2_3 distribution, the BF test resulted in liberal values under only the most extreme degree of variance heterogeneity, while the BOX and W procedures each produced only a single value which was slightly liberal (Conditions H and G, respectively). When $g = 1/h = 0$ a slightly higher number of aberrant results were produced; the W, J, and AG tests had liberal rates of 8.61%, 8.36%, and 8.42%, respectively, when all variances were unequal (Condition G; 1:4:9:16). However, as the degree of nonnormality increased, the rates for the approximate tests became substantially smaller, although none of the procedures were associated with conservative results.

J = 6

Least Squares Estimation. Table 4 contains empirical rates of Type I error when sampling from the $J = 6$ normal and nonnormal distributions. The tabled values are quite similar in pattern, though larger in numeric value, to those enumerated in Table 3. That is, the F test resulted in very liberal values when group sizes and variances were negatively paired and sampling was from the normal, χ^2_6, and χ^2_3 distributions, and very liberal rates for both positive and negative pairings of groups sizes and variances when sampling was from the g- and h-distributions. The empirical F values approached 40% for the positive pairing cases and 50% for the negative pairing cases.

Insert Table 4 About Here

The approximate tests, excluding BF, were robust to variance heterogeneity when the data were normally distributed. For the χ^2_6 data, only the BOX and AG tests remained robust; the W and J tests resulted in liberal rates of 7.72% and 7.55%, respectively, in condition K, while the BF test resulted in rates of 11.81% and 11.33%, for conditions J and L, respectively. When the degree of skewness increased (χ^2_3), all of

the approximate tests resulted in liberal rates in at least one of the four conditions examined. Of the five procedures, the BOX test performed best, as only one liberal value was obtained (8.60%). It occurred when all but one of the variances were equal (Condition L; 1:1:1:1:1:36) and were negatively paired with unequal group sizes.

For data that were obtained from the g- and h-distributions, the approximate tests, with rare exception, did not control the rates of Type I error. As was the case when there were four groups, the rates of error were generally very large, even for positive pairings of group sizes and variances. The most extreme values occurred when the data was obtained from the $g = 1/h = .5$ distribution. For positive pairings of group sizes and variances the empirical values approached 40%, while for the negative pairings cases the values approached 70%.

Robust Estimation. When data were obtained from the χ^2_6 distribution all approximate tests with the exception of BF, effectively controlled the rate of Type I error; the W test only very slightly exceeded the upper bound of Bradley's (1978) criterion for condition K (7.53%). As has been reported previously, the BF test can not effectively control the rate of Type I error when there is one very deviant variance (conditions J and L). The values for the J and L conditions when sampling from the χ^2_6 distribution were 10.73% and 10.42%, respectively. When skewness increased in value (χ^2_3), the BF test again resulted in liberal rates for conditions J and L (11.02% and 10.87%, respectively), while the W and J tests each resulted in a single liberal value (i.e., 8.51% and 7.95%, respectively) for condition K, and the BOX test produced a value of 7.99% for condition L.

When the simulated data were obtained from the $g = 1/h = 0$ distribution, all approximate procedures resulted in at least one liberal value; the BF test was again liberal for conditions J and L (11.19% and 11.84%, respectively), while the BOX test was liberal (9.15%) when there was one very deviant variance (condition L). On the other

hand, the W, J, and AG tests resulted in liberal rates (9.80%, 9.19%, and 9.07%, respectively) when most of the variances were unequal (conditon K).

The empirical values of the approximate tests decreased in size however, when sampling from the g = 1/h = .5 distribution. In this case, only the BF test resulted in liberal values. However, the BOX test resulted in two conservative values (i.e., 2.17% and 2.48% respectively for conditions I and K).

J = 10

Least Squares Estimation. The empirical percentages of Type I error which were obtained when J = 10 are contained in Table 5. When the data were from the normal distribution only the BOX, W, J, and AG tests remained robust to variance heterogeneity when least squares estimates were employed. For the χ^2_6 distribution, only the BOX test was not liberal over the four investigated conditions. The BF test resulted in liberal rates in the N and P conditions (i.e., 13.20% and 12.82%, respectively), while the W, J, and AG tests were liberal for condition O (8.51%, 8.01%, and 7.60%, respectively). When skewness was increased in the chi-square distribution (χ^2_3), the W, J, and AG tests were severely affected. All three tests had similar rates of error ranging from approximately 8% in condition M to 11% in condition O. The BF test again was liberal in conditions N and P (14.75% and 14.33%, respectively), while the BOX test only resulted in one liberal value (9.07%).

Insert Table 5 About Here

The rates when sampling from the g- and h-distributions were again generally very inflated, particularly when sampling from the g = 1/h = .5 distribution. Indeed, for this distribution, the rates of Type I error approached 75% for the W, J, and AG procedures under condition M.

Robust Estimation. Consistent with the findings for smaller numbers of groups, the use of trimmed means and Winsorized variances resulted in much better Type I error control. Under the normal distribution, the BF procedure did produce liberal results for three of the four conditions. The W test was also liberal for Conditions N and O (7.58% and 8.20%, respectively). When the data were obtained from the χ^2_6 distribution only the J and AG procedures were not liberal across any of the investigated conditions, although rates for the remaining approximate procedures, excluding BF, were not seriously inflated. For data that was χ^2_3 distributed, all tests were affected by skewness. Specifically, the BF test resulted in rates of 12.72% and 12.80% for conditions N and P, respectively. The BOX test however, was liberal (8.77%) in condition P. On the other hand, the J and AG tests were liberal in conditions O (9.59%, 8.60%, respectively) and P (8.18% and 7.94%, respectively). The W test was found to be liberal for all conditions.

When the data were obtained from the g- and h-distributions, the J test exhibited the best overall Type I error control; only for the g=1/h=0 distribution were liberal values obtained for the O and P conditions. Similar results were obtained for the AG procedure, although it was also liberal for the N condition in that same distribution. Furthermore, while the W procedure was liberal across all conditions for the g=1/h=0 distribution, error rates were controlled within the bounds of Bradley's (1978) criteria for g=1/h=.5.

Power Rates

The preceding results led us to compare the AG, BOX, J, and W tests for their sensitivity to detect treatment effects since these procedures exhibited a similar degree of Type I error control. Thus, we examined the sensitivity of the tests to detect true differences among the population trimmed means, that is the μ_{tj}s. In addition, for normally distributed data, we compared these tests to their counterparts that used least

squares estimators to test equality of the μ_js. As previously indicated, we compared these tests under two non null effect sizes for the J = 4 and J = 10 designs.

The J = 4 and J = 10 power values (%) are presented in Table 6. The values in Table 6 were obtained by averaging the power values across the equal/unequal group sizes/variances conditions and the different effect sizes. Apparent from an examination of the tabled values are the following general conclusions. The BOX test was always less powerful than the other procedures, while the remaining tests had very similar power values. Furthermore, the magnitude of the difference between AG, J, W, and BOX was substantial. For both designs, the W, J and AG tests had average power values which were approximately 25 percentage points larger than the BOX values, a difference that can not be attributed to their somewhat differential rates of Type I error control. The values tabled when sampling from normally distributed populations indicate that the test statistics based on trimmed means and Winsorized variances were not substantially less powerful than the test statistics that used the usual least squares estimators. The test statistics using least squares estimators were approximately 4-5 percentage points and 2-5 percentage points higher than the tests based on trimmed means for the J = 4 and J = 10 designs, respectively.

Insert Table 6 About Here

Discussion

This investigation compared six procedures that can be used to test for location equality among two or more groups when population variances are heterogeneous.

Specifically, we compared the procedures due to Alexander and Govern (1994), Box (1954), Brown and Forsythe (1974), James (1951), and Welch (1951), as well as the ANOVA F test. When utilizing group means and variances (i.e., least squares estimators), these procedures test for the equality of population means, while the use of trimmed means and Winsorized variances (i.e., robust estimators) results in tests of equality of population trimmed means.

Results from our study indicate that when the variance homogeneity and normality assumptions are not satisfied and the design is unbalanced, the use of *any* of these test statistics with the usual least squares estimators can not generally be recommended. Indeed, for the skewed distributions that Micceri (1989) and Wilcox (1995b) maintain characterize most psychological data, the rates of Type I error for these test statistics can become very liberal when the variances and group sizes are jointly unequal.

On the other hand, our results also indicate that the approximate tests due to Alexander and Govern (1994), Box (1954), James (1951) and, to a lesser extent, Welch (1951), generally exhibit very good Type I error control when computed with trimmed means and Winsorized variances. However, it is important to remind the reader that none of the procedures were able to control the rate of Type I error in all of the investigated conditions.

It is also important to note that, based on other simulations we conducted, our reported Type I error findings are representative of what happens when group sizes are equal as well (i.e., $n_j = 10$). That is, even when groups were of equal size, the combined effects of nonnormality and variance heterogeneity were consistent with the pattern of results reported in the tables for unbalanced designs. For example, the F and BF tests were most affected by the combined assumption violations and the test statistics based on least squares estimators were most prone to Type I errors as data became progressively

more nonnormal. Power findings were also similar to those reported for unbalanced designs. That is, the AG, J, and W tests of trimmed means all had similar rates, and these were substantially larger than the BOX power values.

In conclusion, we recommend that researchers use either the Alexander and Govern (1994), James (1951) or Welch (1951) statistics with trimmed means and Winsorized variances to test omnibus hypotheses regarding treatment group equality. This recommendation is based on the superior power one will achieve by using one of these tests in comparison to the BOX test. That is, though the BOX procedure occasionally displayed better Type I error control than the other tests, we feel it is reasonable to sacrifice some Type I error control for the substantially increased power one obtains with either AG, J, or W.

Though these statistics test a null hypothesis which stipulates that the population trimmed means are equal we believe this is a reasonable hypothesis to examine since trimmed means, as oppposed to the usual least squares means, provide better estimates of the typical individual in distributions that either contain outliers or are skewed in shape. Since a number of surveys suggest that the data obtained in applied settings, including psychology, are characterized by heavy tailed distributions, then as our data indicate, robust statistics utilizing trimmed means and Winsorized variances will, in addition, provide Type I error control, by in large, where test statistics based on least squares estimators will not.

Also, as Wilcox (in press, b) notes, a single outlier in just one group of a multigroup design can adversely affect the power to reject the omnibus null hypothesis. Thus, researchers should consider adopting robust methods even when data is skewed in just one of their treatment groups. In addition, as Wilcox (in press, a, Sections 8.8 and 8.9) indicates, inferential and descriptive procedures based on these robust estimators

will also provide better probability coverage for interval estimation and better estimates of effect size.

Furthermore, as is the case with omnibus test statistics which compare the usual treatment group means, researchers can choose to follow-up statistically significant omnibus tests of trimmed means with multiple comparison procedures which also employ trimmed means and Winsorized variances (see Keselman, Lix & Kowalchuk, in press; Wilcox, in press, b). And lastly, test statistics utilizing trimmed means and Winsorized variances are available for other research paradigms as well; specifically, the procedures have been extended to factorial designs as well as repeated measures designs (see Wilcox, 1995a).

Finally, we want to acknowledge that our conclusions and recommendations are based on the factors manipulated in this investigation and thus we do not believe we have provided the final word on this topic. If anything, we hope our paper will stimulate others to explore yet other factors that we could not examine given the already extensive number of conditions that we did vary. For example, will our results be qualified if nonnormality differs across the treatment groups? Research into this area is relatively new and further refinements therefore are surely forthcoming.

Thus, researchers must apply our recommendations judiciously always remembering that the best decisions regarding the alternative ways in which one's data may be examined will always only be forthcoming after one has completely emersed oneself in the data, i.e., by knowing the shape of the treatment groups (e.g., boxplots), the variance of the treatment groups, the degree of skewness and kurtosis, whether outliers are present, etc., etc. That is, the final caveat that we want to leave the reader with is that nonnormality of one's data should not automatically signal the adoption of trimmed means and robust test statistics. Under such circumstances researchers should seriously

consider the reasons why data are nonnormal, examining the methods of data collection, measurement instruments, data generating process.

References

Alexander, R.A., & Govern, D.M. (1994). A new and simpler approximation for ANOVA under variance heterogeneity. Journal of Educational Statistics, 19, 91-101.

Behrens, W.V. (1929). Ein beitrag zur fehlerberechnung bei wenigen beobachtungen. Landwirtsch Jahrbucher, 68, 807-837.

Box, G.E.P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. Annals of Mathematical Statistics, 25, 290-302.

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

Brown, M.B., & Forsythe, A.B. (1974). The small sample behavior of some statistics which test the equality of several means. Technometrics, 16, 129-132.

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.

Dijkstra, J.B., & Werter, P.S.P.J. (1981). Testing the equality of several means when the population variances are unequal. Communications in Statistics, Simulation and Computation, B10(6), 557-569.

Fisher, R.A. (1935). The fiducial argument in statistical inference. Annals of Eugenics, 6, 391-398.

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.

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.

James, G.S. (1951). The comparison of several groups of observations when the ratios of the population variances are unknown. Biometrika, 38, 324-329.

Keselman, H.J., Lix, L.M., & Kowalchuk, R. K. (in press). Multiple comparison procedures for trimmed means. Psychological Methods.

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

Oshima, T.C., & Algina, J. (1992). Type I error rates for James's second-order test and Wilcox's H_m test under heteroscedasticity and non-normality. British Journal of Mathematical and Statistical Psychology, 45, 255-263.

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.

Rubin, A.S. (1983). The use of weighted contrasts in analysis of models with heterogeneity of variance. Proceedings of the Business and Economic Statistics Section, American Statistical Association, 347-352.

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

Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. Journal of the American Statistical Association, 62, 626-633.

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.

Tomarkin, A.J., & Serlin, R.C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. Psychological Bulletin, 99, 90-99.

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

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. (1989). Adjusting for unequal variances when comparing means in one-way and two-way fixed effects ANOVA models. Journal of Educational Statistics, 14, 269-278.

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

Wilcox, R.R. (1992). An improved method for comparing variances when distributions have non-identical shapes. Computational Statistics and Data Analysis, 13, 163-172.

Wilcox, R.R. (1993). Robustness in ANOVA. In L.K. Edwards (Ed.), Applied analysis of variance in behavioral science (pp. 345-374). New York: Marcel Dekker.

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(1), 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. (in press-b). Three multiple comparison procedures for trimmed means. Biometrical Journal.

Wilcox, R.R. (1996a). Statistics for the social sciences. New York: Academic Press.

Wilcox, R.R. (1996b). Simulation results on performing pairwise comparisons of trimmed means. Unpublished manuscript.

Wilcox, R.R., Charlin, V.L., & Thompson, K.L. (1986). New Monte Carlo results on the robustness of the ANOVA F, W and F^* statistics. Communications in Statistics, Simulation and Computation, 15(4), 933-943.

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

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

Acknowledgements

This research was supported by a National Sciences and Engineering Research Council of Canada grant (#OGP0015855) to the second author. The authors would like to express their gratitude to Rand Wilcox for his many helpful comments on the topic of robust estimation and testing.

Table 1. Investigated Sample Size and Variance Conditions

Condition	Sample Sizes (Two Cases)	Population Variances
A	10, 20; 15, 25	1, 16
B	10, 20; 15, 25	1, 36
C	10, 20; 15, 25	16, 1
D	10, 20; 15, 25	36, 1
E	10, 15, 20, 25; 15, 20, 25, 30	1, 4, 9, 16
F	10, 15, 20, 25; 15, 20, 25, 30	1, 1, 1, 36
G	10, 15, 20, 25; 15, 20, 25, 30	16, 9, 4, 1
H	10, 15, 20, 25; 15, 20, 25, 30	36, 1, 1, 1
I	10, 15(2), 20(2), 25; 15, 20(2), 25(2), 30	1(2), 4, 9(2), 16
J	10, 15(2), 20(2), 25; 15, 20(2), 25(2), 30	1(5), 36
K	10, 15(2), 20(2), 25; 15, 20(2), 25(2), 30	16, 9(2), 4, 1(2)
L	10, 15(2), 20(2), 25; 15, 20(2), 25(2), 30	36, 1(5)
M	10(2), 15(3), 20(3), 25(2); 15(2), 20(3), 25(3), 30(2)	1(3), 4(3), 9(3), 16
N	10(2), 15(3), 20(3), 25(2); 15(2), 20(3), 25(3), 30(2)	1(9), 36
O	10(2), 15(3), 20(3), 25(2); 15(2), 20(3), 25(3), 30(2)	16, 9(3), 4(3), 1(3)
P	10(2), 15(3), 20(3), 25(2); 15(2), 20(3), 25(3), 30(2)	36, 1(9)

Table 2. Percentages of Type I Error (J=2)

Cond	Normal		χ^2_6		χ^2_3		g=1/h=0		g=1/h=.5	
	F	F^*	F	F^*	F	F^*	F	F^*	F	F^*
	Least Squares Estimation									
A	**1.53**	5.00	**2.04**	5.55	2.67	6.13	7.29	**10.75**	**31.13**	**35.63**
B	**1.61**	5.43	**1.98**	5.54	2.76	6.84	6.87	**11.24**	**34.04**	**39.80**
C	**14.19**	4.93	**16.20**	6.95	**16.83**	**8.31**	**20.98**	**14.27**	**45.61**	**41.03**
D	**14.85**	5.08	**16.19**	6.53	**17.80**	**8.02**	**23.07**	**14.70**	**50.47**	**44.18**
	Robust Estimation									
A	**1.68**	5.24	**2.04**	5.24	2.67	5.24	7.29	5.89	**31.13**	5.14
B	**2.02**	5.55	**1.98**	5.45	2.76	6.19	6.87	6.88	**34.04**	6.09
C	**15.49**	5.42	**16.20**	6.42	**16.83**	7.22	**20.98**	7.44	**45.61**	6.25
D	**17.35**	5.71	**16.19**	6.19	**17.80**	7.09	**23.07**	**8.06**	**50.47**	6.78

Note: Normal= Normal distribution; χ^2_6, χ^2_3 =Chi Square distribution with six (three) df; g=1/h=0, g=1/h=.5 Hoaglin's (1985) g- and h- distributions; F= ANOVA F; F^*= Alternative test procedures; COND= sample size/variance condition; See Table 1 for definitions of the investigated conditions.

Table 3. Percentages of Type I Error (J=4)

	Least Squares Estimation						Robust Estimation					
	F	BF	BOX	W	J	AG	F	BF	BOX	W	J	AG
Cond	Normal											
E	3.24	6.76	4.92	5.33	5.32	5.14	3.44	6.54	4.75	5.58	5.52	5.46
F	4.48	**8.85**	5.07	4.80	4.77	4.60	4.69	**8.76**	5.39	5.35	5.27	5.00
G	**13.85**	6.22	4.88	5.30	5.20	5.10	**14.53**	6.22	5.24	6.25	5.93	5.91
H	**22.73**	**8.21**	5.18	5.18	5.07	4.94	**23.97**	**8.58**	6.43	6.14	5.89	5.49
	χ^2_6											
E	3.62	6.91	4.98	6.13	6.11	5.98	3.29	6.64	4.61	5.73	5.61	5.73
F	5.15	**9.73**	6.00	5.97	5.94	6.01	5.14	**9.21**	5.83	6.31	6.22	6.11
G	**14.33**	6.71	5.27	7.16	7.02	7.04	**14.28**	6.19	5.17	7.32	6.96	6.89
H	**23.91**	**10.15**	7.25	6.58	6.46	6.31	**23.98**	**9.16**	7.06	6.50	6.22	5.93
	χ^2_3											
E	3.49	6.40	4.73	5.98	5.98	5.94	3.23	5.95	4.31	5.96	5.90	5.91
F	5.99	**10.08**	6.57	6.15	6.09	6.17	5.27	**9.05**	5.84	5.85	5.75	5.79
G	**14.55**	7.27	5.75	**9.22**	**9.09**	**9.00**	**13.68**	6.03	4.88	**7.70**	7.36	7.21
H	**24.89**	**11.53**	**8.49**	**7.77**	**7.68**	7.25	**25.18**	**9.71**	**7.72**	7.05	6.71	6.59
	g=1/h=0											
E	4.37	6.37	4.24	**11.07**	**11.04**	**11.06**	3.02	5.15	3.57	5.78	5.69	5.78
F	**9.96**	**14.08**	**10.65**	**8.98**	**8.91**	**9.16**	5.79	**9.24**	6.36	5.41	5.27	5.47
G	**15.00**	**8.56**	6.52	**18.44**	**18.27**	**18.15**	**12.60**	5.45	4.06	**8.61**	**8.36**	**8.42**
H	**28.36**	**17.57**	**14.88**	**12.73**	**12.59**	**11.90**	**24.56**	**10.39**	**8.59**	7.10	6.93	6.45
	g=1/h=.5											
E	**15.76**	**16.30**	**12.95**	**48.70**	**48.70**	**48.03**	2.54	4.34	2.63	4.16	4.09	4.15
F	**38.38**	**41.22**	**37.20**	**33.79**	**33.69**	**34.15**	5.31	**8.28**	5.73	3.96	3.87	4.01
G	**27.29**	**22.84**	**19.67**	**56.98**	**56.76**	**57.31**	**10.96**	4.08	3.02	6.31	6.10	6.18
H	**49.88**	**43.21**	**39.44**	**39.58**	**39.39**	**37.85**	**22.68**	**8.90**	7.20	5.27	5.05	4.48

Note: F=ANOVA; BF=Brown and Forsythe (1974); BOX=Box (1954); W=Welch (1951); J=James (1951); AG=Alexander and Govern (1994). See Table 2 note.

Table 4. Percentages of Type I Error (J=6)

	Least Squares Estimation						Robust Estimation (Symmetric Trimming)					
	F	BF	BOX	W	J	AG	F	BF	BOX	W	J	AG
Cond	Normal											
I	3.99	7.21	4.65	5.35	5.23	5.17	4.20	7.02	4.71	6.31	6.02	5.97
J	6.81	**11.27**	5.35	5.19	5.09	5.10	**7.60**	**10.86**	5.89	6.07	5.72	5.79
K	**15.36**	7.06	5.10	5.54	5.37	5.19	**15.52**	7.44	5.36	7.44	6.90	6.63
L	**24.81**	**10.38**	5.35	5.34	5.09	5.00	**26.30**	**10.05**	7.03	6.87	6.28	5.92
	χ^2_6											
I	3.96	7.44	4.52	6.29	6.15	6.20	4.14	6.96	4.53	6.52	6.22	6.17
J	7.12	**11.81**	5.71	5.89	5.70	5.86	7.19	**10.73**	5.58	6.17	5.88	5.98
K	**14.47**	6.83	4.54	**7.72**	**7.55**	7.35	**14.64**	6.58	4.65	**7.53**	6.95	6.72
L	**25.34**	**11.33**	6.98	6.84	6.72	6.59	**26.25**	**10.42**	7.36	6.71	6.30	6.14
	χ^2_3											
I	3.95	7.12	4.28	7.41	7.25	7.25	3.98	6.84	4.20	6.55	6.26	6.31
J	**8.01**	**12.17**	6.66	**7.57**	7.43	**7.60**	**7.62**	**11.02**	6.15	7.18	6.85	7.02
K	**14.37**	7.04	4.78	**9.73**	**9.40**	**9.05**	**14.11**	6.47	4.26	**8.51**	**7.95**	7.50
L	**26.17**	**12.63**	**8.60**	**8.19**	**7.96**	**7.80**	**26.24**	**10.87**	**7.99**	7.49	6.90	6.74
	g=1/h=0											
I	4.76	6.90	3.49	**14.64**	**14.45**	**14.75**	3.67	5.49	3.21	6.95	6.69	6.70
J	**12.50**	**15.61**	**11.24**	**10.74**	**10.54**	**11.30**	**8.22**	**11.19**	6.94	6.17	5.87	6.36
K	**13.98**	7.27	4.73	**20.24**	**19.96**	**19.33**	**12.94**	5.55	3.45	**9.80**	**9.19**	**9.07**
L	**29.95**	**19.25**	**15.69**	**14.03**	**13.81**	**13.27**	**25.82**	**11.84**	**9.15**	7.47	7.00	6.73
	g=1/h=.5											
I	**15.13**	**15.47**	**10.66**	**63.25**	**63.01**	**63.17**	2.95	4.22	**2.17**	4.92	4.64	4.81
J	**38.36**	**39.85**	**33.88**	**32.55**	**32.36**	**33.14**	7.48	**10.10**	6.19	4.39	4.08	4.54
K	**24.60**	**19.31**	**14.38**	**69.30**	**68.91**	**68.64**	**11.01**	4.11	**2.48**	7.04	6.56	6.45
L	**46.52**	**40.51**	**34.95**	**38.62**	**38.27**	**36.31**	**23.57**	**10.28**	7.38	4.96	4.63	4.30

Note: See the notes from Tables 2-3.

Table 5. Percentages of Type I Error (J=10)

Cond	Least Squares Estimation						Robust Estimation (Symmetric Trimming)					
	F	BF	BOX	W	J	AG	F	BF	BOX	W	J	AG
	Normal											
M	3.75	**7.98**	4.75	5.49	5.24	5.27	3.87	**7.85**	4.52	7.01	6.21	6.08
N	**9.21**	**13.44**	5.36	5.39	5.10	5.03	**9.75**	**12.60**	5.87	**7.58**	6.56	6.59
O	**16.59**	7.10	4.29	5.54	5.10	4.95	**16.74**	6.88	4.49	**8.20**	7.21	6.72
P	**25.59**	**11.38**	5.34	5.21	4.89	4.94	**27.14**	**11.34**	7.03	**7.52**	6.57	6.32
	χ^2_6											
M	3.95	7.49	4.52	6.99	6.69	6.62	3.80	7.32	4.26	7.13	6.20	6.46
N	**9.83**	**13.20**	6.19	7.12	6.77	6.90	**10.06**	**12.83**	6.10	**7.75**	6.88	6.80
O	**16.49**	7.11	4.25	**8.51**	**8.01**	**7.60**	**16.36**	6.35	3.92	**8.46**	7.24	6.91
P	**26.96**	**12.82**	7.15	**7.72**	7.23	7.23	**27.69**	**11.84**	**7.68**	**8.02**	7.04	6.82
	χ^2_3											
M	3.80	7.36	3.98	**8.49**	**8.09**	**8.17**	3.86	7.02	3.94	**7.69**	6.93	7.08
N	**10.98**	**14.75**	7.20	**8.44**	**8.21**	**8.38**	**10.33**	**12.72**	6.31	**8.01**	7.01	7.49
O	**16.48**	6.89	3.92	**11.51**	**11.03**	**10.51**	**16.02**	6.39	3.61	**10.69**	**9.59**	**8.60**
P	**28.03**	**14.33**	**9.07**	**9.45**	**8.97**	**8.90**	**28.48**	**12.80**	**8.77**	**9.13**	**8.18**	**7.94**
	g=1/h=0											
M	3.84	6.53	**2.44**	**16.01**	**15.44**	**15.99**	3.52	6.04	2.70	**7.68**	6.96	7.25
N	**15.05**	**18.02**	**11.64**	**13.75**	**13.25**	**14.23**	**10.19**	**12.70**	6.85	**7.70**	6.83	**7.64**
O	**14.62**	5.89	2.66	**25.04**	**24.19**	**23.36**	**14.70**	4.83	**2.32**	**11.86**	**10.59**	**9.96**
P	**30.35**	**19.09**	**14.30**	**15.72**	**15.12**	**15.14**	**26.56**	**12.23**	**8.78**	**8.91**	**7.75**	**8.03**
	g=1/h=.5											
M	**10.83**	**10.59**	5.15	**73.96**	**73.39**	**73.65**	2.95	4.45	**1.73**	4.77	4.15	4.45
N	**37.03**	**36.29**	**27.82**	**32.27**	**31.70**	**33.18**	**9.18**	**11.24**	5.94	4.25	3.71	4.28
O	**19.98**	**13.06**	7.33	**8.18**	**8.13**	**8.09**	**12.73**	3.41	**1.43**	7.46	6.54	6.22
P	**41.41**	**35.04**	**26.47**	**36.83**	**36.26**	**34.17**	**24.32**	**10.06**	7.00	5.19	4.40	4.52

Note: See the notes from Tables 2-3.

Table 6. Power Rates (Collapsed Over Effect Sizes and Conditions of Group sizes/Variances Homogeneity/Heterogeneity)

Distribution	Test Statistic			
	BOX	W	J	AG
	J=4			
Normal (Yes)	30	63	63	63
Normal (No)	34	68	68	68
χ^2_6	6	16	15	16
χ^2_3	11	29	27	29
g=1/h=0	24	58	57	58
g=1/h=.5	17	49	48	49
	J=10			
Normal (Yes)	17	50	47	49
Normal (No)	19	53	52	53
χ^2_6	6	14	12	12
χ^2_3	7	22	20	21
g=1/h=0	12	49	47	48
g=1/h=.5	7	39	37	39

Note: Yes/No-indicates that trimmed means were/were not utilized.

Table 7. Hypothetical Data Set and Summary Statistics

Statistics	J_1	J_2	J_3	J_4
	2	5	3	6
	2	4	1	3
	2	4	4	6
	3	4	3	5
	5	6	5	4
	3	2	2	5
	3	5	5	6
	6	4	4	5
	3	4	4	4
	3	3	4	4
	4	6	2	4
	6	3	2	6
	4	5	4	4
	3	3	4	3
	3	4	3	4
	4	4	2	5
	3	3	1	4
	3	4	6	4
	3	3	3	5
	5	5	5	16
n_j	20	20	20	20
$\bar{X}_j$	3.60	4.00	3.40	5.15
s^2_j	.6737	1.2632	1.8316	7.3974
$\sqrt{b_1}$	.2578	0	-.1098	3.3487
b_2	2.3711	2.5000	2.3805	14.0903
h_j	12	12	12	12
$\bar{X}_{tj}$	3.50	4.00	3.42	4.58
s^2_{wj}	.4545	1.0909	1.2500	1.2500

Note: $\sqrt{b_1}$=sample estimate of the third moment (skewness) . b_2= sample estimate of the fourth moment (kurtosis) (See D'Agostino, Belanger & D'Agostino (1990).