ROBUST NONORTHOGONAL ANALYSES REVISITED: AN UPDATE BASED ON TRIMMED MEANS

by

H. J. Keselman, Rhonda K. Kowalchuk and Lisa M. Lix
University of Manitoba

Abstract

Three approaches to the analysis of main and interaction effect hypotheses in nonorthogonal designs were compared in a 2×2 design for data that was neither normal in form nor equal in variance. The approaches involved either least squares or robust estimators of central tendency and variability and/or a test statistic that either pools or does not pool sources of variance. Specifically, we compared the ANOVA F test which used trimmed means and Winsorized variances, the Welch-James test with the usual least squares estimators for central tendency and variability and the Welch-James test using trimmed means and Winsorized variances. As hypothesized, we found that the latter approach provided excellent Type I error control, whereas the former two did not.

ROBUST NONORTHOGONAL ANALYSES REVISITED: AN UPDATE BASED ON TRIMMED MEANS

Introduction

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). Numerous solutions to what is now defined as the "Behrens-Fisher problem" have appeared in the literature (e.g., Alexander & Govern, 1994; Box, 1954; Brown & Forsythe, 1974; James, 1951; Welch, 1938). Extensive empirical evidence suggests that all of these methods can generally control the rate of Type I errors when the data are normally distributed, even under extreme degrees of variance heterogeneity in the underlying population distributions (e.g., Alexander & Govern, 1994; Wilcox, 1995a, 1995b). 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 solutions to the "Behrens-Fisher problem" have limitations, namely their sensitivity to the nature of the population distributions.

The deleterious effects of nonnormality are predictable on theoretical grounds. Cressie and Whitford (1986) have shown that, unless population variances or group sizes are equal, Student's two-sample *t* test is not asymptotically correct when the group distributions have unequal third cumulants; therefore, Type I error inflation is expected.

It is well known that the usual population mean and variance, which are the basis for all of the previously enumerated Behrens-Fisher type solutions, are greatly influenced by the presence of extreme observations in a distribution of scores. Moreover, as Wilcox (1994a) notes, the appropriateness of the population mean as a measure of location is questionable when the underlying distribution is skewed. Adopting a nonrobust measure, such as the usual mean, “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). However, by substituting robust measures of location, and a corresponding robust measure of scale, 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). In particular, while the standard error of the usual mean can become seriously inflated when the underlying distribution has heavy tails (Tukey, 1960), the standard error of the trimmed mean is less affected by departures from normality because extreme observations, that is, observations in the tails of a distribution, are censored or removed. Furthermore, as Gross (1976) notes, "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. While the trimmed mean has been shown to be highly effective, we caution the reader that this measure should only be adopted if one 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. As an illustration of how a trimmed mean may provide a better estimate of the typical score than the usual mean, consider the example given by Wilcox (1995a, p. 57) in which a single score in a chi-square distribution with four df (hence $\mu = 4$) is multiplied by 10 (with probability .1). Because this single deviant score causes a shift in the distribution, the usual mean now equals 7.6, a value closer to the upper tail of the distribution. A trimmed mean based on censoring 20% of the data in each tail of the

distribution however, equals 4.2, a value that is closer to the bulk of scores, hence closer to the typical score in the distribution. Nonetheless, readers should note that the hypothesis tested when the usual mean is used as an estimate of location is not the same as that tested when the trimmed mean is employed. Consequently, we stress that the researcher needs to be clear on the goals of data analysis prior to choosing a particular method of statistical inference.

Wilcox (1994b) has shown that in the two-sample case, by substituting trimmed means and Winsorized variances for the usual least squares estimators, the degree of bias due to skewness can indeed be reduced, compared to the results that would be expected based on Cressie and Whitford's (1986) theoretical derivation. That is, the results in Wilcox can be used to show that the theoretical results due to Cressie and Whitford can be extended to trimmed means. It is also apparent that these findings can be generalized to multi-group designs. Lix and Keselman (1995, 1997) found that in one-way designs, tests of mean equality based on the usual mean and variance (i.e., Alexander & Govern, 1994; Box, 1954; Brown & Forsythe, 1974; James, 1951; Welch, 1951) were indeed biased by the effects of skewness when group sizes were unequal; the degree of bias was reduced when trimmed means and variances based on Winsorized data were used in the computation of these solutions (see also Wilcox, 1994a).

The negative consequences of conducting tests of mean equality in the presence of variance heterogeneity and nonnormality have also been examined within the context of nonorthogonal factorial designs. That is, Keselman, Carriere and Lix (1995, 1996) demonstrated how a Welch (1947, 1951)-James (1951, 1954) (WJ) type statistic due to Johansen (1980) generally could provide a robust test of various hypotheses and/or model comparisons in unbalanced factorial designs when cell variances were heterogeneous, while the usual analysis of variance (ANOVA) F test could not. Though these authors recommended the WJ test, they also noted that it occasionally resulted in a liberal test of

significance when variance heterogeneity occurred in combination with nonnormality. Consequently, the WJ test had its limitations in combatting assumption violations in unbalanced fixed-effects factorial designs. Specifically, its ability to cope with the effects of variance heterogeneity was diminished when the data were also nonnormal in shape.

Based on Wilcox's (1994b) results, one may, however, be able to improve the performance of the WJ test by incorporating robust measures of location and variability instead of relying on the usual least squares estimators of these statistics. That is, using trimmed means should result in a more accurate solution, when distributions have heavy tails, because with this type of nonnormality they have smaller standard errors compared to least squares means. In the present paper, we were primarily concerned with extending the WJ procedure 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 variances based on Winsorized sums of squares be used in conjunction with Welch's (1938) statistic. For heavy-tailed symmetric distributions, Yuen showed that the statistic based on these robust estimators 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 applied her approach to the analysis of nonorthogonal designs. Additionally, to date, no investigator has quantified the reduction in the degree of bias that can be expected by replacing least squares estimators with their robust counterparts, as Wilcox (1994b) has done in the one-way two-sample case.

Background

To reaquaint the reader with the issues related to the analysis of nonorthogonal designs, as illustrated in a two-way J $\times$ K design with disproportionate cell frequencies (i.e., n_{jk}s; $j = 1, \ldots, J$, $k = 1, \ldots, K$), we first enumerate the hypotheses and models that frequently are examined. We adopt the notational scheme used by Keselman et al.

(1995, 1996) where these sample sizes are associated with correspondingly notated population cell means (i.e., μ_{jk}s). The dependent scores (i.e., Y_{ijk}s) are modeled by $Y_{ijk} = \mu_{jk} + \epsilon_{ijk}$, with $i = 1, \ldots, n_{jk}$. For this model, the errors (i.e., ϵ_{ijk}s) are usually assumed to be independent and identically distributed normal random variables with a mean of zero and a variance of σ^2.

Three hypotheses that researchers would typically be interested in testing are:

H_{Row}: No row main effect,

H_{Column}: No column main effect, and

$H_{R\times C}$: No interaction effect.

As Keselman et al. (1995, 1996) and others indicate there is no ambiguity concerning how to test $H_{R\times C}$ in a two-way design. However, the proper way to test H_{Row} and H_{Column} has received a great deal of discussion. Most treatises on this topic indicate that one can test hypotheses concerning unweighted or weighted marginal main effect means and these tests correspond to different model comparisons.

Specifically, in a two-way design, the hypotheses that are typically associated with tests of the marginal means are:

i. Unweighted:

H_R: $\sum_k \mu_{jk}/K - \sum_k \mu_{j'k}/K = 0$ or $\mu_{j.} - \mu_{j'.} = 0$ for all j and j', and

H_C: $\sum_j \mu_{jk}/J - \sum_j \mu_{jk'}/J = 0$ or $\mu_{.k} - \mu_{.k'} = 0$ for all k and k'.

ii. Weighted:

H_R^*: $\sum_k n_{jk}\mu_{jk}/n_{j.} - \sum_k n_{j'k}\mu_{j'k}/n_{j'.} = 0$ or $\overline{\mu}_{j.} - \overline{\mu}_{j'.} = 0$ for all j and j', and

H_C^*: $\sum_j n_{jk}\mu_{jk}/n_{.k} - \sum_j n_{jk'}\mu_{jk'}/n_{.k'} = 0$ or $\overline{\mu}_{.k} - \overline{\mu}_{.k'} = 0$ for all k and k',

where $n_{j.} = \Sigma_k n_{jk}$ and $n_{.k} = \Sigma_j n_{jk}$ and

iii. Weighted:

$$H_R^{**}: \sum_k \left(n_{jk} - \frac{n_{jk}^2}{n_{.k}} \right) \mu_{jk} - \sum_{j \neq j'} \sum_k \left(\frac{n_{jk} n_{j'k}}{n_{.k}} \right) \mu_{j'k} = 0$$

for $j = 1, \ldots, \mathrm{J} - 1$, and all j and j' and

$$\mathrm{H}_{\mathrm{C}}^{**}\colon \sum_{j}\left(\mathrm{n}_{jk} - \frac{\mathrm{n}_{jk}^{2}}{\mathrm{n}_{j.}}\right)\mu_{jk} \; - \sum_{k \neq k'}\sum_{j}\left(\frac{\mathrm{n}_{jk}\mathrm{n}_{jk'}}{\mathrm{n}_{j.}}\right)\mu_{jk'} = 0$$

for $k = 1, \ldots, \mathrm{K} - 1$, and all k and k'.

From a model comparison perspective, tests of these respective sets of hypotheses correspond to assessing each main effect by comparing: (i.) the residual sum of squares (SS) for the full nonadditive model ($\mathrm{Y}_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \epsilon_{ijk}$) to the residual SS for a model that does not contain the effect specified under the null hypothesis, (ii.) a model that contains both the specified main effect and constant terms, to a model that only contains the constant term (e.g., for $\mathrm{H}_{\mathrm{R}}^{*}$ the models that are compared are $\mathrm{Y}_{ijk} = \mu + \alpha_j + \epsilon_{ijk}$ vs. $\mathrm{Y}_{ijk} = \mu + \epsilon_{ijk}$) and, (iii.) a model that contains only main effects to a model that excludes the main effect under investigation (e.g., for $\mathrm{H}_{\mathrm{R}}^{**}$ the models that are compared are $\mathrm{Y}_{ijk} = \mu + \alpha_j + \beta_k + \epsilon_{ijk}$ vs. $\mathrm{Y}_{ijk} = \mu + \beta_k + \epsilon_{ijk}$). In terms of the eliminating/ignoring terminology, each main effect is assessed, respectively: (i.) eliminating all other effects specified in the full nonadditive model, (ii.) ignoring all other effects, and (iii.) eliminating the other main effect and ignoring the interaction effect.

Robust Data Analysis Strategies

The customary approach to testing H_{R}, H_{C}, $\mathrm{H}_{\mathrm{R}}^{*}$, $\mathrm{H}_{\mathrm{C}}^{*}$, $\mathrm{H}_{\mathrm{R}}^{**}$, $\mathrm{H}_{\mathrm{C}}^{**}$ and $\mathrm{H}_{\mathrm{R}\times\mathrm{C}}$ is to assume homogeneity of variances and, therefore, use a pooled estimate of within cell variability as the denominator of an ANOVA F statistic. Milligan, Wong and Thompson (1987) and Keselman et al. (1995, 1996) have shown, however, that when variances are heterogeneous, the ANOVA approach leads to seriously biased results.

On the other hand, Keselman et al. (1995, 1996) demonstrated how one can obtain generally robust tests of these hypotheses by employing a WJ type test. For completeness we present the WJ test again. As these authors note, each of the previously delineated hypotheses can, from a full model perspective, be expressed as H_0: $\mathbf{C}\boldsymbol{\mu} = \mathbf{0}$, where $\mathbf{C}$ is the contrast matrix associated with the specific hypothesis, $\boldsymbol{\mu}$ is a vector of means and $\mathbf{0}$ is the null vector. Therefore, one can test any of the previously enumerated hypotheses or models by specifying the appropriate contrast matrices and applying them in a full rank model analysis of the data with a procedure (i.e., the WJ test) that can handle unequal variances.

To introduce the WJ test, consider the full rank model $Y_{ijk} = \mu_{jk} + \epsilon_{ijk}$ where the ϵ_{ijk}s are independently and normally distributed within each combination of j and k and where cell variances are <u>not</u> required to be equal. Suppose under these model assumptions that we wish to test the hypothesis:

$$H_0: \mathbf{C}_1\boldsymbol{\mu} = \mathbf{0}\,, \tag{1}$$

where $\boldsymbol{\mu} = (\boldsymbol{\mu}'_1, \ldots, \boldsymbol{\mu}'_J)'$, $\boldsymbol{\mu}_j = (\mu_{j1}, \ldots, \mu_{jK})$, and $\mathbf{C}_1$ is a contrast matrix of dimension r $\times$ JK with r = rank ($\mathbf{C}_1$). Then from Johansen (1980) and Keselman et al. (1995, 1996), the test statistic is

$$T_{WJ} = (\mathbf{C}_1\bar{\mathbf{Y}})'(\mathbf{C}_1\,\mathbf{S}\,\mathbf{C}'_1)^{-1}(\mathbf{C}_1\,\bar{\mathbf{Y}})\,, \tag{2}$$

where $\bar{\mathbf{Y}} = (\bar{\mathbf{Y}}'_1, \ldots, \bar{\mathbf{Y}}'_J)'$, $\bar{\mathbf{Y}}_j = (\bar{Y}_{j1}, \ldots, \bar{Y}_{jK})'$, $\bar{Y}_{jk} = \Sigma_i Y_{ijk}/n_{jk}$, with $\boldsymbol{\mathcal{E}}(\bar{\mathbf{Y}}) = \boldsymbol{\mu}$, and the sample variance matrix of $\bar{\mathbf{Y}}$ is $\mathbf{S} = \text{diag}(s^2_{11}/n_{11}, \ldots, s^2_{JK}/n_{JK})$, where $s^2_{jk} = \sum_{i=1}^{n_{jk}}(Y_{ijk} - \bar{Y}_{jk})^2/(n_{jk} - 1)$. This statistic, divided by a constant, c, has an

approximate F distribution with degrees of freedom $f_1 = r$, and $f_2 = r(r + 2)/(3A)$. The constant $c = r + 2A - 6A/(r + 2)$, with

$$A = \sum_{jk} (1 - P_{jk,jk})^2/(n_{jk} - 1),$$

where $P_{jk,jk}$ is the (jk, jk)-th element of the matrix $\mathbf{I} - \mathbf{S}\mathbf{C}_1'(\mathbf{C}_1\mathbf{S}\mathbf{C}_1')^{-1}\mathbf{C}_1$.

Testing Hypotheses Under Unrestricted Models. As previously indicated, H_R and H_C (and $H_{R\times C}$) are assessed by eliminating all other effects in the full nonadditive model. That is, the model that is postulated as being true is $Y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \epsilon_{ijk}$, and H_R, H_C and $H_{R\times C}$ are evaluated against this model. As there are no restrictions placed on α_j, β_k, or $\alpha\beta_{jk}$ (except as side conditions for a full rank model), one can simply use the WJ statistic as given in Equation 2, with $\mathbf{C}_1$ equal to either $\mathbf{C}_R$, $\mathbf{C}_C$ or $\mathbf{C}_{R\times C}$, to test the respective null hypotheses. One can also evaluate H_R^*, H_C^*, H_R^{**}, and H_C^{**} assuming that the full nonadditive model is true in this case substituting $\mathbf{C}_R^*$, $\mathbf{C}_C^*$, $\mathbf{C}_R^{**}$, or $\mathbf{C}_C^{**}$ for $\mathbf{C}_1$.

Testing Hypotheses Under Restricted (R) Models. **A**s Keselman et al. (1995, 1996) indicate some authors would not recommend substituting $\mathbf{C}_R^*$, $\mathbf{C}_C^*$, $\mathbf{C}_R^{**}$, or $\mathbf{C}_C^{**}$ for $\mathbf{C}_1$ in Equations 1 and 2 (e.g., Keren, 1982; Lewis & Keren, 1977; Scheffe, 1959). Their view is that it is inappropriate to use the error term from the full nonadditive model to test H_R^*, H_C^*, H_R^{**}, or H_C^{**} because under the null hypothesis certain effects are assumed to be zero. For example, when testing H_R^{**}, one compares two models, neither of which contains $\alpha\beta_{jk}$. In other words, one assesses H_R^{**} assuming that all $\alpha\beta_{jk}$s $= 0$. Accordingly, to test $H(R)_R^*$, $H(R)_C^*$, $H(R)_R^{**}$, or $H(R)_C^{**}$, these authors would favor the use of a procedure that restricts ignored effects to zero.

To test $H(R)_R^*$, $H(R)_C^*$, $H(R)_R^{**}$, or $H(R)_C^{**}$ under a restricted model, one tests each hypothesis as in Equation 1 with the restrictions specified by $\mathbf{C}_0\boldsymbol{\mu} = \mathbf{0}$, where the contrast

matrix $\mathbf{C}_0$ denotes the ignored effects that are restricted to zero. As Keselman et al. (1995, 1996) indicate, the contrast matrix $\mathbf{C}_1$ can be constructed by forming a super matrix which contains rows for the effect that is being tested, as well as rows for each effect that is being ignored under H_0.

For restricted models, the test statistic in Equation 2 has the more general but explicit form

$$T_{WJ} = \bar{\mathbf{Y}}'[(\mathbf{S}\,\mathbf{C}_1'(\mathbf{C}_1\,\mathbf{S}\,\mathbf{C}_1')^{-1}\mathbf{C}_1\,) - (\mathbf{S}\,\mathbf{C}_0'(\mathbf{C}_0\,\mathbf{S}\,\mathbf{C}_0')^{-1}\mathbf{C}_0\,)]'\ \mathbf{S}^{-1}\ \times$$
$$[(\mathbf{S}\,\mathbf{C}_1'(\mathbf{C}_1\,\mathbf{S}\,\mathbf{C}_1')^{-1}\,\mathbf{C}_1\,) - (\mathbf{S}\,\mathbf{C}_0'(\mathbf{C}_0\,\mathbf{S}\,\mathbf{C}_0')^{-1}\,\mathbf{C}_0\,)]\,\bar{\mathbf{Y}}\,. \qquad (3)$$

The test statistic in Equation 3 divided by c has an approximate F distribution with $f_1 = r = \text{rank}\,(\mathbf{C}_1) - \text{rank}\,(\mathbf{C}_0)$ and $f_2 = r(r + 2)/\{3(A - B)\}$, where $c = r + 2(A + B) - 6(A - B)/(r + 2)$,

$$A = \sum_{jk}(P_{0jk,jk} - P_{1jk,jk})[(1 - P_{1jk,jk})/(n_{jk} - 1)],$$

and

$$B = \sum_{jk}(P_{0jk,jk} - P_{1jk,jk})[(1 - P_{0jk,jk})/(n_{jk} - 1)],$$

where $P_{i,jk,jk}$ is the (jk, jk)-th element of the matrix $\mathbf{I} - \mathbf{S}\mathbf{C}_i'(\mathbf{C}_i\mathbf{S}\mathbf{C}_i')^{-1}\mathbf{C}_i$, for $i = 0$ or 1.

For example, one would test $H(R)_R^*$, $H(R)_C^*$, $H(R)_R^{**}$ and $H(R)_C^{**}$ in two-way factorial designs by letting

$\mathbf{C}_1 = [\mathbf{C}_R'\ \mathbf{C}_C'\ \mathbf{C}_{R\times C}']'$ and $\mathbf{C}_0 = [\mathbf{C}_C'\ \mathbf{C}_{R\times C}']'$ for $H(R)_R^*$,

$\mathbf{C}_1 = [\mathbf{C}_C'\ \mathbf{C}_R'\ \mathbf{C}_{R\times C}']'$ and $\mathbf{C}_0 = [\mathbf{C}_R'\ \mathbf{C}_{R\times C}']'$ for $H(R)_C^*$,

$\mathbf{C}_1 = [\,\mathbf{C}_R'\ \mathbf{C}_{R\times C}']'$ and $\mathbf{C}_0 = \mathbf{C}_{R\times C}$ for $H(R)_R^{**}$ and

$\mathbf{C}_1 = [\mathbf{C}_C'\ \mathbf{C}_{R\times C}']'$ and $\mathbf{C}_0 = \mathbf{C}_{R\times C}$ for $H(R)_C^{**}$, respectively.

Both $\mathbf{C}_1$ and $\mathbf{C}_0$ can be obtained in the manner described in Keselman et al. (1995, 1996).

Applying Trimmed means and Winsorized variances to the WJ test. As previously indicated we hypothesized that by using trimmed means and Winsorized variances with the WJ test one may be able to obtain a statistic that is robust to the combined effects of nonnormality and variance heterogeneity in nonorthogonal designs. In addition, as a base of comparison we also investigated the use of robust estimators with the ANOVA F test as well as applying the WJ test with the usual least squares estimators.

Trimmed means and Winsorized sums of squares (variances) are based on order statistics. Specifically, let $Y_{(1)jk} \leq Y_{(2)jk} \leq \ldots \leq Y_{(n_{jk})jk}$ represent the ordered observations associated with the *jk*th cell. When trimming symmetrically, let $g_{jk} = [\gamma n_{jk}]$, 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 *jk*th cell becomes $h_{jk} = n_{jk} - 2g_{jk}$. The prevalent method of trimming is to remove outliers from each tail of the distribution of scores; furthermore, it is recommended that 20 percent of the observations be removed from each tail. (see Wilcox, 1995 and the references he cites for a justification of the 20 percent rule.) Under symmetric trimming, the *jk*th sample trimmed mean is

$$\overline{Y}_{tjk} = \frac{1}{h_{jk}} \sum_{i=g_{jk}+1}^{n_j - g_{jk}} Y_{(i)jk} , \tag{4}$$

and the jkth sample Winsorized mean is

$$\overline{Y}_{Wjk} = \frac{1}{n_{jk}} \sum_{i=1}^{n_{jk}} X_{ijk} , \tag{5}$$

where

$$\begin{aligned} X_{ijk} &= Y_{(g_{jk}+1)jk} \text{ if } Y_{ijk} \leq Y_{(g_{jk}+1)jk} \\ &= Y_{ijk} \text{ if } Y_{(g_{jk}+1)jk} < Y_{ijk} < Y_{(n_{jk}-g_{jk})jk} \\ &= Y_{(n_{jk}-g_{jk})jk} \text{ if } Y_{ijk} \geq Y_{(n_{jk}-g_{jk})jk} . \end{aligned}$$

Based on Yuen's (1974) work, a variance based on Winsorized data may be defined as

$$s^2_{Wjk} = \frac{1}{h_{jk}-1} \sum_{i=1}^{n_{jk}} (X_{ijk} - \overline{Y}_{Wjk})^2. \tag{6}$$

Thus, with robust estimation, the trimmed cell means and Winsorized cell variances were substituted for their least squares counterparts into the WJ statistic. In addition, f_1 and f_2 were based on the effective sample sizes (i.e., the h_{jk}s).

Method

In our study we compared three approaches to testing H_R, H_C, H^*_R, H^*_C, H^{**}_R, H^{**}_C, $H(R)^*_R$, $H(R)^*_C$, $H(R)^{**}_R$, $H(R)^{**}_C$, and $H_{R\times C}$: (a) the ANOVA F test with trimmed means and Winsorized variances [F(TM)], (b) the WJ test with least squares estimators [WJ], and (c) the WJ test with robust estimators [WJ(TM)]. In particular, we compared these three approaches by varying three factors: (a) degree of variance heterogeneity, (b) pattern of directional pairing of variances and sample sizes, and (c) distributional shape of the data.

To build on the results provided by Keselman et al. (1995, 1996), we investigated many of their conditions. In particular, we studied a 2×2 design in which cell variances were in the ratio of 1:1:1:9. This degree of heterogeneity has been shown to produce liberal tests in the one-way design (Wilcox, 1987, pp. 30-32). However, we also chose to investigate a more disparate case, specifically, one in which cell variances were in a ratio

of 1:1:1:16. We included this second case of variance heterogeneity because, according to Wilcox, Charlin, and Thompson (1986) and Fenstad (1983), it is not uncommon to have populations with standard deviations that are in a 4:1 ratio. Moreover, Lix, Cribbie and Keselman (1996) found instances in the education literature in which cell standard deviations exceeded a ratio of 20:1.

We chose two patterns of these variances, both of which were also examined by Keselman et al. (1995, 1996), in which the smallest variance was associated with the cell having either the smallest size (Pattern 1 [P^+]) or largest size (Pattern 2 [P^-]). These patterns are known to produce conservative and liberal results, respectively, with many test procedures.

As the WJ procedure has also been shown to be affected by distributional shape in other contexts (Algina, Oshima & Tang, 1991; Keselman, Carriere & Lix, 1993; Keselman et al., 1995), we collected Type I error rates when sampling from both normal and nonnormal distributions. In addition to generating data from a χ^2_3 distribution, we also used the method described in Hoaglin (1985) to generate a distribution with more extreme degrees of skewness and kurtosis. These particular types of nonnormal distributions were selected because 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 identified by Micceri on the robustness of Student's *t* test and found that only distributions with the most extreme degree of skewness which were investigated (e.g., $\gamma_1 = 1.64$) were found to affect the Type I error control of the independent sample t statistic. For the χ^2_3 distribution, skewness and kurtosis values are $\gamma_1 = 1.63$ and $\gamma_2 = 4.00$, respectively. The other nonnormal distribution was generated from the g- and h-distribution (Hoaglin, 1985). Specifically, we chose to investigate the g = 1 and h = 0 (notated throughout the remainder of the paper as g = 1/h = 0) 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 $\gamma_1 = 6.2$ and $\gamma_2 = 114$. Finally, it should be noted that though the selected combinations of g and h result in an extremely skewed distribution, these values, according to Wilcox (1990, 1994a, p. 296), are representative of psychometric measures. Thus, we included the g = 1/h = 0 distribution as an exemplar of a "worst case" distribution that, according to Wilcox (1994a), researchers may encounter.

The sample cell sizes used in this study were $n_{11} = 8$, $n_{12} = 20$, $n_{21} = 20$ and $n_{22} = 32$. The unequal cell sizes follow from Wilcox's (in press) recommendation that, in the equal cell size case, sizes should not be less than 20 when statistics are based on trimmed means and Winsorized variances. The degree of cell size inequality, as indexed by a coefficient of variation, equalled .424, and corresponds with a case investigated by Keselman et al. (1995, 1996), though our N is larger than those investigated by these authors; sample size was increased since observations would be deleted due to trimming. It is also important to note, that according to Lix et al. (1996) unbalanced designs are the norm, rather than the exception, in educational research and that the degree of sample size disparity that we investigated does occur in this research field as well.

To generate pseudo-random normal variates, we used the SAS generator RANNOR (SAS Institute, 1989). If Z_{ijk} is a standard normal variate, then $Y_{ijk} = \mu_{jk} + \sigma_{jk} \times Z_{ijk}$ is a normal variate with mean equal to μ_{jk} and variance equal to σ^2_{jk}.

To generate pseudo-random variates having a χ^2 distribution with three degrees of freedom, three standard normal variates were squared and summed. The variates were standardized, and then transformed to χ^2_3 variates having mean μ_{jk} (when investigating tests based on least squares estimates) or μ_{tjk} (when investigating tests based on trimmed

means) and variance σ_{jk}^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 (Z_{ijk}s) were converted to the random variable

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

according to the values of g and h selected for investigation. To obtain a distribution with standard deviation σ_{jk}, each Y_{ijk} ($jk = 1, \ldots, JK$) was multiplied by a value of σ_{jk}. 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

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

(see Hoaglin, 1985, p. 503). Thus, because $g > 0$ in our investigation, μ_{gh} was first subtracted from Y_{ijk} before multiplying by σ_{jk}. When working with trimmed means, μ_{tjk} 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 previously reflect only the amount that each random variable is multiplied by and not the actual values of the standard deviations (see Wilcox, 1994a, p. 298). As Wilcox notes, 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 significance level.

Results

To evaluate the particular conditions under which a test was insensitive to assumption violations, Bradley's (1978) liberal criterion of robustness was employed. According to this criterion, in order for a test to be considered robust, its empirical rate of Type I error ($\widehat{\alpha}$) must be contained in the interval $0.5\alpha \leq \widehat{\alpha} \leq 1.5\alpha$. Therefore, for the five percent level of 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 because 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 universal standard by which tests are judged to be robust; with other standards, different interpretations of the results are possible.

Normally distributed data. Table 1 contains empirical rates of Type I error (%) for the three approaches to the analyses of H_R, H_C, H_R^*, H_C^*, H_R^{**}, H_C^{**}, $H(R)_R^*$, $H(R)_C^*$, $H(R)_R^{**}$, $H(R)_C^{**}$, and $H_{R\times C}$ in the 2×2 nonorthogonal design for the two cases of variance heterogeneity (1:1:1:9 and 1:1:1:16) and two patterns of variances and sample sizes (P^+: positive pairing and P^-: negative pairing). Most evident from this table is that WJ and WJ(TM) effectively controlled Type I error rates for all hypotheses across all conditions while F(TM) never did. WJ and WJ(TM) had rates which occasionally exceeded five percent, and thus could be described as slightly liberal, however, averaged

over all tabled values their respective rates were only 5.15% and 5.54%. For F(TM), the empirical rates ranged from a low of 0.56% to a high of 26.48%. Thus, for normally distributed data, the two approaches involving a statistic that does not pool heterogeneous sources of variation (i.e., WJ) provided effective Type I error control regardless of the type of estimators for location and scale, while the use of robust (trimmed) estimators with a nonrobust statistic (F), did not.

Insert Table 1 About Here

χ^2_3 distributed data. The combined effects of variance heterogeneity and nonnormality can be gleaned from Table 2. The empirical rates reported in Table 2 were obtained when data were χ^2_3 distributed. Once again F(TM) resulted in very conservative and liberal rates when sample sizes and variances were positively and negatively paired, respectively. On the other hand, rates for WJ were only occasionally liberal in condition P^-. For the less disparate case of variance heterogeneity (1:1:1:9) only three of the 11 values exceeded 7.50% and occurred for tests of H_R (8.50%), H_C (8.04%) and $H_{R\times C}$ (8.24%). However, when the variances were in a 1:1:1:16 ratio, five of the 11 values exceeded 7.50%, ranging in value from 7.68% to 8.30%, and were associated with the tests of unrestricted hypotheses. Using robust estimators with WJ [WJ(TM)] was effective in combating the combined effects of nonnormality and variance heterogeneity. That is, for χ^2_3 distributed data, the rates of error for WJ(TM) were always within Bradley's (1978) interval. The average rates of error for the P^+ condition for the less and more disparate cases of heterogeneity were 5.21% and 5.44%, respectively, while the corresponding P^- values were 6.09%, and 6.22%.

Insert Table 2 About Here

---

g = 1/h = 0 distributed data. The effect of combining heterogeneous cell variances with nonnormal data which was also very skewed and kurtotic is evident from the values presented in Table 3. Once again F(TM) was substantially affected by variance heterogeneity though the values for the P^+ condition were not as conservative as those reported in Tables 1 and 2. The WJ rates, however, were substantially larger than the values reported in Tables 1 and 2. Thus, though the values reported by Keselman et al. (1995, 1996) suggest that WJ with the usual least squares estimators for central tendency and variability was somewhat effected by data that is neither normal in form nor equal in variability, the data reported in Table 3 clearly indicate the limitations of WJ to the combined violations. Indeed, for $g = 1/h = 0$ data, rates of error were always in excess of Bradley's (1978) upper bound for the P^- condition, typically attaining values between 10 and 15 percent. Furthermore, the empirical rates were even liberal under the P^+ condition. On the other hand, *all* of the 44 WJ(TM) rates were contained within Bradley's interval. The average rates of error for the P^+ condition for the less and more disparate cases of heterogeneity were 4.96% and 5.04%, respectively, while the corresponding P^- values were 6.23%, and 6.55%. Thus, WJ(TM) proved very effective in controlling the rate of Type I error.

Insert Table 3 About Here

Discussion

Three approaches to the analysis of nonorthogonal designs were compared in a 2×2 design when data were nonnormal and variances were nonhomogeneous. Specifically, we compared the usual ANOVA F test which used trimmed means and Winsorized variances, the WJ test using the usual least squares estimators of central

tendency and variability, and the WJ test based on trimmed means and Winsorized variances. Thus, the first approach involved robust estimators with a nonrobust statistic, the second approach a robust statistic with nonrobust estimators, while the third approach was robust with respect to estimators as well as statistic. Not surprisingly, based on the theoretical results presented by Cressie and Whitford (1986) and Wilcox (1994b), we hypothesized that our third approach would provide the best opportunity for controlling the rate of Type I errors when the data were both nonnormal in form and nonhomogeneous in variability.

The results reported in this investigation support the conclusions of Keselman et al. (1995, 1996). That is, for moderate degrees of skewness (e.g., χ^2_3) and variance heterogeneity (σ^2_{jk} ratio of 1:1:1:9), the WJ test with the usual least squares estimators for central tendency and variability typically is robust in nonorthogonal designs. However, the data reported in this investigation support our hypothesis regarding the best overall method of analysis. That is, the WJ test using trimmed means and Winsorized variances provided excellent Type I error control over *all* the investigated conditions, while the other two approaches did not. This finding is particularly impressive in light of the fact that the conditions that were varied in this investigation were extreme. That is, the degrees of skewness, kurtosis, variance heterogeneity, and sample size disparity were large and thus probably represent the most extreme conditons that applied researchers are ever likely to encounter with actual data. However, as we have indicated previously, the conditions we investigated, including the extreme assumption violation cases, according to other researchers working in the area, have occurred, or are likely to occur, in behavioral science experiments (Fenstad, 1983; Lix et al., 1996; Micceri, 1989; Wilcox, 1994a, 1995a, 1995b; Wilcox et al., 1986). Thus, WJ with trimmed means and Winsorized variances appears to us to be the more versatile procedure in that it controls rates of Type I error when conditions are moderately as well as substantially unfavorable.

Moreover, as Wilcox (1995a) notes, procedures that perform well over a wide range of simulation conditions, including extreme conditions, encouragingly suggests that the positive operating characteristics of the procedure might hold over conditons not considered in the simulation and thus positively reflect on the procedures versatility. Consequently, we are most comfortable in recommending this approach for the analysis of effects in nonorthogonal designs.

Furthermore, though our investigation was limited to tests of effects with one df, results presented elsewhere suggest one can expect comparable findings with tests of effects involving more than one df (see Lix & Keselman, 1995, 1997). At the same time, we must emphasize that the hypotheses tested when trimmed means are used are not equivalent to those tested when the usual means are adopted. Nonetheless, we feel it is most reasonable to use these estimators when one is interested in determining if there has been an effect due to a treatment variable and one wants to compare the groups on a measure of location that is insensitive to nonnormality.

As a post script to our recommendation we want to acknowledge that the results presented in this paper and the paper by Keselman et al. (1995) provide a great deal of assurance that the WJ test with least squares estimators works reasonably well under many conditions, that is, when assumptions are minimally violated. Accordingly, researchers may be tempted to use least squares estimators with the robust WJ solution. The consequence of adopting this approach is that spurious findings may result, as applied researchers will, in all likelihood, encounter data that are markedly skewed and/or contains outliers. Unfortunately, sample estimates of the third and fourth moments of a distribution are prone to large sampling variability and thus are not likely to be useful to researchers attempting to select one approach over the other. This two-stage approach therefore can not be recommended without further investigation. Another strategy that might be considered by some researchers is to detect and remove outliers and then proceed with the a WJ solution; Wilcox (in press) indicates the pitfalls of this

strategy. Accordingly, we recommend uniform adoption of a single testing strategy, in this case, the WJ solution with trimmed means. This approach, according to the present study, is most likely to result in control of Type I errors.

Also worth noting is that inferential and descriptive procedures based on these robust estimators will also provide better probability coverage for interval estimation and better estimates of effect size (see Wilcox, 1996, Sections 8.8 and 8.9). Furthermore, as is the case with omnibus test statistics which compare the usual treatment group means, researchers can choose to follow-up 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, 1994b, 1996). 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 repeated measures designs (see Wilcox, 1995a).

Finally, it is important that we note that although we have not compared the WJ test with trimmed means and Winsorized variances with the WJ test based on least squares estimators with regard to power, theory and prior work indicates that this was not necessary. That is, as previously indicated, theory tells us that procedures based on sample means result in poor power because the standard error of the mean is inflated when distributions have heavy tails; however, this is less of a problem when working with trimmed means (see Tukey, 1960; Wilcox, 1995b). This phenomenon is illustrated in a number of sources. For example, Wilcox (1994b, 1995b) has presented results indicating that in the two sample and one-way problem, tests (i.e., t and F) based on the usual least squares estimators lose power when data contains outliers and/or is heavy tailed. Specifically, in the two sample problem, Wilcox (1994b) compared the Welch (1938) and Yuen (1974) procedures and found that when data were obtained from contaminated normal distributions (distributions that have thicker tails compared to the

normal) the power of Welch's test was considerably diminished compared to its sensitivity to detect nonnull effects when data were normally distributed and, as well, was less sensitive than Yuen's test. Indeed, the power of Welch's test to detect nonnull effects went from .931 when distributions were normally distributed to .278 and .162 for the two contaminated normal distributions that were investigated; the corresponding power values for Yuen's test were .890, .784, and .602, respectively. Wilcox (1995b) presented similar results for four independent groups. To complement these results we illustrate how power can be affected in nonorthogonal designs when the data contains outliers and/or is heavy-tailed, by modifying Wilcox's (1994a; Wilcox, in press) one way hypothetical data set (see his Table 7), letting his four groups correspond to the four cells {[Group $1 \equiv$ (1,1)-cell], [Group $2 \equiv$ (1,2)-cell], [Group $3 \equiv$ (2,1)-cell], and [Group $4 \equiv$ (2,2)-cell]} of a
2(R) $\times$ 2(C) design (see Table 4). For simplicity of presentation we use this balanced data set and thus the various nonorthogonal solutions provide identical results. Additionally, we remind the reader that the example data set contains 20 observations per cell because this is the minimum size recommended in order to obtain good Type I error control (see Wilcox, in press).

Insert Table 4 About Here

Also contained in Table 4 are summary statistics for the four cells based on least squares and robust estimators. The striking feature about this data set is the one extreme observation, outlier, ($n_{20,2,2} = 40$) in cell (2,2); boxplots of the data clearly flag this outlier. To better understand the nature of the data, measures of skewness ($\sqrt{b_1}$) and kurtosis (b_2) were computed for each cell (see D'Agostino, Belanger & D'Agostino, 1990). The reader should note that $\sqrt{b_1} = 0$ indicates a symmetric distribution with $\sqrt{b_1} > 0$ indicating skewness to the right and $\sqrt{b_1} < 0$ indicating skewness to the left.

On the other hand, values of $b_2 > 3(n-1)/(n+1)$ indicate heavy tails while $b_2 < 3(n-1)/(n+1)$ is indicative of light tailed distributions. Thus, cell (2,2) is extreme with regard to these two measures (see Table 4).

Results based on least squares estimators are: (a) $WJ_R = 2.23$ ($\nu = 21.6583$), $p = .150$; (b) $WJ_C = 2.56$ ($\nu = 21.6583$), $p = .124$; and (c) $WJ_{RC} = 1.12$ ($\nu = 21.6583$), $p = .302$. Accordingly, treatment effects are not detected for any of the investigated effects. On the other hand, the results based on robust estimators are: (a) $WJ(TM)_R = 4.07$ ($\nu = 39.3104$), $p = .050$; (b) $WJ(TM)_C = 6.73$ ($\nu = 39.3104$), $p = .013$; and (c) $WJ(TM)_{RC} = 0.0$ ($\nu = 39.3104$), $p = 1.0$. Thus, based on trimmed means and Winsorized variances, one would conclude that there are main effect differences, though effects were not detected for the interaction.

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.

Algina, J., Oshima, T.C., & Tang, K.L. (1991). Robustness of Yao's, James', and Johansen's tests under variance-covariance heteroscedasticity and nonnormality. Journal of Educational Statistics, 16, 125-140.

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.

Cressie, N.A.C., & Whitford, H.J. (1986). How to use the two sample *t*-test. Biometrical Journal, 28, 131-148.

D'Agostino, R.B., Belanger, A., D'Agostino, R.B., Jr. (1990). A suggestion for using powerful and informative tests of normality. The American Statistician, 44, 316-321.

Fenstad, G.U. (1983). A comparison between U and V tests in the Behrens-Fisher problem. Biometrika, 70, 300-302.

Fisher, R.A. (1935). The design of experiments. Edinburgh and London: Oliver & Boyd.

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.

James, G.S. (1954). Tests of linear hypotheses in univariate and multivariate analysis when the ratios of the population variances are unknown. Biometrika, 41, 19-43.

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.

Keren, G. (1982). A balanced approach to unbalanced designs. In G. Keren (Ed.), Statistical and methodological issues in psychology and social science research (pp. 155 – 186). Hillsdale, NJ: Lawrence Erlbaum.

Keselman, H.J., Carriere, K.C., & Lix, L.M. (1993). Testing repeated measures hypotheses when covariance matrices are heterogeneous. Journal of Educational Statistics, 18, 305-319.

Keselman, H.J., Carriere, K.C., & Lix, L.M. (1995). Robust and powerful nonorthogonal analyses. Psychometrika, 60, 395-418.

Keselman, H.J., Carriere, K.C., & Lix, L.M. (1996). Errata to "Robust and powerful nonorthogonal analyses". Psychometrika, 61, 191.

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

Lewis, C., & Keren, G. (1977). You can't have your cake and eat it too: Some considerations of the error term. Psychological Bulletin, 84, 1150-1154.

Lix, L.M., Cribbie, R., & Keselman, H.J. (1996). The analysis of completely randomized univariate designs. Paper presented at the annual meeting of the Psychometric Society, Banff, AB.

Lix, L.M., & Keselman, H.J. (1995). To trim or not to trim: Tests of mean equality under heteroscedasticity and nonnormality. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Lix, L.M., & Keselman, H.J. (1997). To trim or not to trim: Tests of mean equality under heteroscedasticity and nonnormality, Manuscript submitted for publication.

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

Milligan, G.W., Wong, D.S., & Thompson, P.A. (1987). Robustness properties of nonorthogonal analysis of variance. Psychological Bulletin, 101, 464-470.

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.

Scheffe, H. (1959). The analysis of variance. New York: Wiley.

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

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. (1947). The generalization of Student's problem when several different population variances are involved. Biometrika, 34, 28-35.

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

Wilcox, R.R. (1987). New designs in the analysis of variance. Annual Review of Psychology, 38, 29-60.

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

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

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

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

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

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

Wilcox, R.R. (in press). Three multiple comparison procedures for trimmed means. Biometrical Journal.

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.

Acknowledgements

Financial support for this research was provided by grants to the first author from the National Sciences and Engineering Research Council of Canada (#OGP0015855) and the Social Sciences and Humanities Research Council (#410-95-0006). The authors would like to express their appreciation to the Associate Editor as well as the reviewers who provided valuable comments on an earlier version of this paper.

Table 1. Percentages of Type I Error (Normal/n_{jk}=8, 20, 20, 32)

	F (TM)		WJ		WJ (TM)	
	P^+	P^-	P^+	P^-	P^+	P^-
H_0	Variance Ratio=1:1:1:9					
Row						
H	**0.94**	**22.98**	4.78	5.22	5.34	5.82
H^*	**1.98**	**11.42**	4.68	5.08	4.66	5.54
H^{**}	**2.26**	**12.42**	4.70	5.32	4.68	5.58
$H(R)^*$	**1.66**	**8.46**	4.76	4.80	6.26	5.56
$H(R)^{**}$	**2.28**	**11.24**	5.08	5.16	5.96	5.70
Column						
H	**1.06**	**22.84**	4.96	5.38	4.74	5.68
H^*	**2.30**	**11.94**	5.00	5.20	5.20	5.42
H^{**}	**2.46**	**12.88**	5.08	5.32	5.06	5.46
$H(R)^*$	**1.86**	**8.92**	5.40	5.32	5.60	5.86
$H(R)^{**}$	**2.46**	**11.50**	5.42	5.60	5.60	5.92
H_{RXC}	**0.96**	**22.38**	5.10	5.36	5.46	5.62

Table 1, continued

	F (TM)		WJ		WJ (TM)	
	P^+	P^-	P^+	P^-	P^+	P^-
H_0	Variance Ratio=1:1:1:16					
Row						
H	**0.70**	**26.32**	5.48	5.30	5.38	5.92
H^*	**2.08**	**14.72**	5.10	5.64	5.06	6.08
H^{**}	**2.38**	**15.78**	5.20	5.40	5.14	6.08
$H(R)^*$	**1.54**	**10.82**	5.56	4.92	5.74	6.20
$H(R)^{**}$	**2.40**	**14.22**	5.46	5.54	5.78	5.94
Column						
H	**0.56**	**26.46**	4.70	4.88	4.84	5.34
H^*	**1.88**	**14.32**	4.62	4.88	5.38	5.40
H^{**}	**2.20**	**15.52**	4.76	5.02	5.38	5.34
$H(R)^*$	**1.54**	**10.02**	4.82	5.70	5.64	6.08
$H(R)^{**}$	**2.08**	**13.40**	5.30	5.58	5.18	6.34
H_{RXC}	**0.64**	**26.48**	4.90	5.00	5.24	5.78

Note: F(TM)=ANOVA F test with trimmed means and Winsorized variances; WJ=Welch-James test with usual least squares estimators; WJ (TM)=Welch-James test with trimmed means and Winsorized variances; P+/P-=positive/negative pairings of sample sizes and variances.

Table 2. Percentages of Type I Error (χ_3^2/n_{jk}=8, 20, 20, 32)

	F (TM)		WJ		WJ (TM)	
	P^+	P^-	P^+	P^-	P^+	P^-
H_0	Variance Ratio=1:1:1:9					
Row						
H	**1.38**	**21.66**	4.98	**8.50**	4.64	6.06
H^*	2.66	**10.94**	5.52	6.48	4.94	5.56
H^{**}	2.96	**12.06**	5.88	6.70	5.02	5.74
$H(R)^*$	**2.16**	**8.34**	5.68	6.34	5.50	6.98
$H(R)^{**}$	2.98	**10.80**	6.22	7.30	5.40	6.86
Column						
H	**1.30**	**21.68**	5.10	**8.04**	4.96	5.84
H^*	2.66	**10.58**	5.62	6.26	5.34	5.06
H^{**}	3.06	**11.76**	5.62	6.76	5.36	5.30
$H(R)^*$	**2.40**	**7.80**	5.88	6.40	5.82	6.48
$H(R)^{**}$	3.08	**10.12**	5.88	7.28	5.60	6.64
H_{RXC}	**0.88**	**22.28**	5.38	**8.24**	4.76	6.50

Table 2, continued

	F (TM)		WJ		WJ (TM)	
	P^+	P^-	P^+	P^-	P^+	P^-
H_0	Variance Ratio=1:1:1:16					
Row						
H	**1.14**	**26.54**	5.74	**8.16**	5.56	6.20
H^*	2.68	**13.22**	6.04	7.06	5.84	5.66
H^{**}	3.22	**14.64**	6.06	7.12	5.92	5.72
$H(R)^*$	**2.10**	**9.02**	6.26	6.26	5.60	6.18
$H(R)^{**}$	3.06	**12.66**	6.82	6.82	5.80	6.30
Column						
H	**1.04**	**27.78**	5.30	**8.30**	5.08	7.00
H^*	**2.10**	**14.50**	5.66	**7.68**	5.30	6.26
H^{**}	2.50	**15.76**	5.76	**7.82**	5.46	6.34
$H(R)^*$	**1.70**	**9.98**	5.86	6.46	5.20	6.42
$H(R)^{**}$	**2.42**	**13.80**	6.64	7.12	5.36	6.00
H_{RXC}	**0.52**	**26.36**	5.52	**8.26**	4.72	6.36

Note: See the note from Table 1.

Table 3. Percentages of Type I Error ($g=1/h=0/n_{jk}=8, 20, 20, 32$)

	F (TM)		WJ		WJ (TM)	
	P^+	P^-	P^+	P^-	P^+	P^-
H_0	Variance Ratio=1:1:1:9					
Row						
H	**2.06**	**22.54**	4.64	**13.96**	4.28	6.80
H^*	3.02	**10.50**	5.68	**9.66**	4.84	5.58
H^{**}	3.38	**11.60**	6.08	**10.22**	4.88	5.72
$H(R)^*$	2.70	**7.70**	7.12	**10.58**	6.28	6.86
$H(R)^{**}$	3.34	**10.24**	**7.98**	**11.94**	5.56	6.52
Column						
H	**1.96**	**22.18**	4.56	**13.84**	4.00	6.60
H^*	2.90	**10.82**	5.90	**9.48**	4.62	5.28
H^{**}	3.20	**11.76**	6.38	**10.14**	4.66	5.50
$H(R)^*$	**2.40**	7.44	6.66	**10.22**	5.98	6.80
$H(R)^{**}$	3.24	**10.44**	7.18	**11.76**	5.02	6.34
H_{RXC}	**1.52**	**22.30**	6.52	**13.62**	4.48	6.56

Table 3, continued

	F (TM)		WJ		WJ (TM)	
	P^+	P^-	P^+	P^-	P^+	P^-
H_0	Variance Ratio=1:1:1:16					
Row						
H	**1.88**	**26.70**	6.28	**14.54**	5.10	7.08
H^*	3.14	**13.56**	**7.80**	**10.68**	5.30	6.28
H^{**}	3.54	**15.30**	**8.52**	**11.70**	5.34	6.32
$H(R)^*$	2.62	**9.10**	6.80	**9.56**	5.28	6.44
$H(R)^{**}$	3.52	**13.04**	**7.90**	**11.32**	5.34	6.86
Column						
H	**1.44**	**26.24**	5.88	**14.48**	4.66	6.90
H^*	2.68	**12.94**	7.46	**11.44**	5.16	5.78
H^{**}	3.00	**14.22**	**8.44**	**11.90**	5.28	5.96
$H(R)^*$	**2.20**	**8.54**	6.16	**9.74**	4.86	6.86
$H(R)^{**}$	2.96	**12.24**	**7.86**	**11.48**	4.50	6.84
H_{RXC}	**0.94**	**26.20**	**8.04**	**14.46**	4.64	6.76

Note: See the note from Table 1.

Table 4. Hypothetical Data Set and Summary Statistics

(1,1)	(1,2)	(2,1)	(2,2)
2	5	3	6
2	4	6	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	2	3	6
4	5	4	4
3	3	4	3
3	4	3	4
4	4	2	5
3	3	6	4
3	4	6	4
3	3	3	5
5	5	5	40

Table 4, continued

Statistics	(1,1)	(1,2)	(2,1)	(2,2)
n_{jk}	20	20	20	20
$\bar{X}_{jk}$	3.50	4.00	3.90	6.35
s^2_{jk}	1.4211	1.2632	1.6737	63.6079
$\sqrt{b_1}$	.8607	0	.1886	4.0354
b_2	2.8601	2.5000	2.0789	17.5633
h_{jk}	12	12	12	12
$\bar{X}_{tjk}$	3.25	4.00	3.83	4.58
s^2_{wjk}	.4136	1.0909	1.2545	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).