Testing Treatment Effects in Repeated Measures Designs: Trimmed Means and Bootstrapping

by

H. J. Keselman[1] Rhonda K. Kowalchuk
University of Manitoba University of Manitoba

and

James Algina Lisa M. Lix Rand R. Wilcox
University of Florida Saskatchewan Health University of Southern California

Abstract

Nonnormality and covariance heterogeneity between groups affects the validity of the traditional repeated measures methods of analysis, particularly when group sizes are unequal. A nonpooled Welch (1947, 1951)-type statistic (WJ) and the Huynh (1978) Improved General Approximation (IGA) test generally have been found to be effective in controlling rates of Type I error in unbalanced nonspherical repeated measures designs even though data are nonnormal in form and covariance matrices are heterogeneous. However, under some conditions of departure from multisample sphericity and multivariate normality their rates of Type I error have been found to be elevated. Westfall and Young's (1993) results suggest that Type I error control could be improved by combining bootstrap methods with methods based on trimmed means. Accordingly, in our investigation we examined four methods for testing for main and interaction effects in a between- by within-subjects repeated measures design: (a) the IGA and WJ tests with least squares estimators based on theoretically determined critical values, (b) the IGA and WJ tests with least squares estimators based on empirically determined critical values, (c) the IGA and WJ tests with robust estimators based on theoretically determined critical values, and (d) the IGA and WJ tests with robust estimators based on empirically determined critical values. We found that the IGA tests were always robust to assumption violations whether based on least squares or robust estimators or whether critical values were obtained through theoretical or empirical methods. The WJ procedure, however, occasionally resulted in liberal rates of error when based on least squares estimators but always proved robust when applied with robust estimators. Neither approach particularly benefited from adopting bootstrapped critical values. Recommendations are provided to researchers regarding when each approach is best.

Testing Treatment Effects in Repeated Measures Designs: Trimmed Means and Bootstrapping

Traditional tests for mean equality typically are invalid when data are nonnormal in form and heterogeneity exists between groups of subjects, particularly when group sizes are unequal (the design is unbalanced). In particular, rates of Type I error usually are inflated or depressed and the power to detect treatment effects can be substantially reduced from theoretical values. This finding holds in independent and correlated groups designs; furthermore, it applies to univariate and multivariate designs (see Coombs, Algina, & Oltman, 1996; Lix & Keselman, 1998; Wilcox, 1998).

A number of researchers have shown that the deleterious effects of variance heterogeneity generally can be overcome by adopting Welch (1947, 1951)-type statistics (see Coombs et al., 1996; Lix & Keselman, 1995), that is, statistics that do not pool across heterogeneous sources of variability and where error degrees of freedom (df) are estimated from the sample data. The deleterious effects of nonnormality can also generally be overcome by adopting robust measures of central tendency and variability, that is, by using trimmed means and Winsorized variances rather than the usual least squares estimators (see Lix & Keselman, 1998; Wilcox, 1997b, 1998). Within the context of independent groups designs, a number of papers have demonstrated that one can indeed generally achieve robustness to nonnormality and variance heterogeneity in unbalanced designs by using robust estimators with nonpooled statistics (see Keselman, Kowalchuk & Lix, 1998; Lix & Keselman, 1998)

Within the context of correlated groups designs, Keselman, Carriere and Lix (1993) have shown how Johansen's (1980) nonpooled multivariate statistic can be used to test for treatment effects in between- by within-subjects repeated measures designs. Furthermore, they have demonstrated through Monte Carlo methods that this Welch-James (1951, 1954)-type statistic (WJ) is generally robust to nonnormality and covariance heterogeneity in nonspherical unbalanced repeated measures designs. Another

generally robust approach to analyzing treatment effects in repeated measures designs is the Huynh (1978) Improved General Approximation (IGA) test. The IGA procedure uses the traditional univariate F tests for assessing treatment effects, however, df are adjusted to take into account possible violations of multisample sphericity. Algina and Keselman (1998) and Keselman, Algina, Kowalchuk and Wolfinger (1999, in press) have shown that the IGA approach, as well as the WJ approach, are generally robust to the combined effects of nonnormality and covariance heterogeneity in nonspherical unbalanced repeated measures designs.

Even though it has been demonstrated that the IGA and WJ procedures are generally robust to the combined effects of nonnormality and covariance heterogeneity, under some conditions of departure from multisample sphericity and multivariate normality, their rates of Type I error have been found to be inflated (see Algina & Keselman, 1997; Keselman et al., 1993; Keselman, Kowalchuk & Boik, in press). For example, Keselman, Kowalchuk and Boik reported values of 9.68% and 8.46% for the IGA and WJ procedures, respectively, when they were used to test the repeated measures interaction effect. Further improvement in Type I error control should be possible by applying these procedures with robust estimators, that is, with trimmed means and Winsorized variances and covariances and by obtaining critical values through bootstrap methods. Such improvement has been demonstrated with statistics for independent group designs (see Wilcox, Keselman, & Kowalchuk, 1998). Lix, Keselman and Algina (1997) provide limited verification of the utility of using trimmed means in the analysis of repeated measures designs, however, in their study statistical significance was assessed with theoretically determined rather than empirically determined critical values.

Accordingly, the purpose of our paper is to determine whether the IGA and WJ procedures rates' of Type I error can be better controlled when data are nonnormal and covariance matrices across groups are unequal in unbalanced nonspherical repeated measures designs when they are used with trimmed means and Winsorized variances and

covariances and when critical values are obtained through a bootstrapping method. Determining whether the IGA and WJ procedures provide valid tests of repeated measures hypotheses when used in conjunction with robust estimators is important because when data are nonnormal it can be argued that testing hypotheses about robust parameters (e.g., trimmed population means) with robust estimators is a more justifiable approach for comparing the typical performance of treatment groups than is the use of traditional statistics based on least squares estimators (see e.g., Wilcox, 1998).

Test Statistics

The simplest of the higher-order repeated measures designs involves a single between-subjects factor and a single within-subjects factor, in which subjects ($i = 1, \ldots, n_j$, $\Sigma n_j = N$) are selected randomly for each level of the between-subjects factor ($j = 1, \ldots, J$) and observed and measured under all levels of the within-subjects factor ($k = 1, \ldots, K$). In this design, the repeated measures data are modeled by assuming that the random vectors $\mathbf{Y}_{ij} = (Y_{ij1}\ Y_{ij2}\ \ldots\ Y_{ijk})'$ are normal, independent and identically distributed within each level j, with common mean vector $\boldsymbol{\mu}_j$ and where we allow $\boldsymbol{\Sigma}_j \neq \boldsymbol{\Sigma}_{j'}$, $j \neq j'$.

IGA. Huynh (1978) developed tests of the within-subjects main and interaction effects that are designed to be used when multisample sphericity is violated. The test statistic for the within-subjects main effect is $F_K = MS_K/MS_{K\times S/J}$ and the critical value is $bF[\alpha;\ h',\ h]$; the test statistic for the within-subjects interaction effect is $F_{JK} = MS_{JK}/MS_{K\times S/J}$ and the critical value is $cF[\alpha;\ h'',\ h]$. The parameters of the critical values are defined in terms of the $\boldsymbol{\Sigma}_j$ and the n_j. These parameters adjust the critical values to take into account the effect of violating multisample sphericity on F_K and F_{JK}. If multisample sphericity holds,

$$bF[\alpha; h', h] = F[\alpha; (K-1), (N-J)(K-1)] \text{ and}$$
$$cF[\alpha; h'', h] = F[\alpha; (J-1)(K-1), (N-J)(K-1)]. \quad (1)$$

Estimates of the parameters (c, b, h, h′and h″), and the correction due to Lecoutre (1991), are presented in Algina (1994) and Keselman and Algina (1996). A SAS/IML (SAS Institute, 1989) program is also available for computing this test in any repeated measures design (see Algina, 1997).

WJ. Since the effects of testing mean equality in repeated measures designs with heterogeneous data is similar to the results reported for independent groups designs, one solution to the problem parallels those found in the context of completely randomized designs. The Johansen (1980) approach, a multivariate extension of the Welch (1951) and James (1951) procedures for completely randomized designs, involves the computation of a statistic that does not pool across heterogeneous sources of variation and estimates error df from sample data. (This is in contrast to the Huynh (1978) approach which, by use of the conventional univariate F statistics, does pool across heterogeneous sources of variance. The Huynh approach adjusts the critical value to take account of the pooling.)

Suppose that we wish to test the hypothesis:

$$H_0: \mathbf{C}\boldsymbol{\mu} = \mathbf{0} \quad (2)$$

where $\boldsymbol{\mu} = (\boldsymbol{\mu}_1', \ldots, \boldsymbol{\mu}_J')'$, $\boldsymbol{\mu}_j = (\mu_{j1}, \ldots, \mu_{jK})'$, j = 1, … , J, and **C** is a full rank contrast matrix of dimension r × JK. Then an approximate df multivariate Welch (Welch, 1947, 1951)-James (James, 1951, 1954)-type statistic according to Johansen (1980) and Keselman et al. (1993) is

$$T_{WJ} = (\mathbf{C}\overline{\mathbf{Y}})'(\mathbf{C}\mathbf{S}\mathbf{C}')^{-1}(\mathbf{C}\overline{\mathbf{Y}}), \quad (3)$$

where $\overline{\mathbf{Y}} = (\overline{\mathbf{Y}}_1', \ldots, \overline{\mathbf{Y}}_J')'$, with $E(\overline{\mathbf{Y}}) = \boldsymbol{\mu}$, and the sample covariance matrix of $\overline{\mathbf{Y}}$ is $\mathbf{S} = \text{diag}(\mathbf{S}_1/n_1, \ldots, \mathbf{S}_J/n_J)$, where $\mathbf{S}_j$ is the sample variance-covariance matrix of the j-th grouping factor. T_{WJ}/c is distributed, approximately, as an F variable with df $f_1 = r$, and $f_2 = r(r+2)/(3A)$, and c is given by $r + 2A - 6A/(r + 2)$, with

$$A = \frac{1}{2}\sum_{j=1}^{J}[\text{tr}\{\mathbf{S}\,\mathbf{C}'(\mathbf{C}\mathbf{S}\,\mathbf{C}')^{-1}\,\mathbf{C}\,\mathbf{Q}_j\}^2 + \{\text{tr}(\mathbf{S}\,\mathbf{C}'(\mathbf{C}\mathbf{S}\,\mathbf{C}')^{-1}\,\mathbf{C}\,\mathbf{Q}_j)\}^2]/(n_j - 1). \quad (4)$$

The matrix $\mathbf{Q}_j$ is a block diagonal matrix of dimension JK × JK, corresponding to the j-th group. The (s,t)-th block of $\mathbf{Q}_j = \mathbf{I}_{K\times K}$ if s = t = j and is $\mathbf{0}$ otherwise. In order to obtain the main and interaction tests with the WJ procedure let $\mathbf{C}'_{K-1}$ be a [(K − 1) × K] contrast matrix and let $\mathbf{C}_{J-1}$ be similarly defined. A test of the main effect can be obtained by letting $\mathbf{C} = \mathbf{1}_J \otimes \mathbf{C}_{K-1}$, where $\mathbf{1}_J$ is the (j × 1) unit vector and $\otimes$ denotes the Kronecker product. The contrast matrix for a test of the interaction effect is $\mathbf{C} = \mathbf{C}_{J-1} \otimes \mathbf{C}_{K-1}$. Lix and Keselman (1995) present a SAS/IML (SAS Institute, 1989) program that can be used to compute the WJ test for any repeated measures design that does not contain quantitative covariates nor has missing values.

Robust Estimation

While a wide range of robust estimators have been proposed in the literature (see Gross, 1976), the trimmed mean and Winsorized (co)variance are intuitively appealing because of their computational simplicity and good theoretical properties (Wilcox, 1995a, 1998). 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 (co)variance, the most extreme observations are replaced with less extreme values in the distribution of scores.

The first step in computing robust estimators within the context of repeated measures designs is to Winsorize the observations. For our J × K design, Winsorization must be performed for every level of the two factors. That is, for fixed j and k, Winsorize the observations Y_{ijk}, $i = 1, \ldots, n_j$ and repeat this process for j = 1, … , J and k = 1, … , K. Let $g_j = [\gamma n_j]$ be the desired amount of trimming where $[\gamma\, n_j]$ is the greatest integer less than or equal to γn_j; we shall set $\gamma = .2$.[2] The Winsorized values are given by

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

Now, for every j there is a K × K Winsorized covariance matrix that must be estimated. The estimated Winsorized covariance between the *m*th and *l*th levels of the within-subjects factor is, for fixed j, estimated with

$$s_{jml} = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (Y_{ijm} - \bar{Y}_{.jm})(Y_{ijl} - \bar{Y}_{.jl}), \tag{5}$$

where $\bar{Y}_{.jm} = \sum_i Y_{ijm}/n_j$, is the Winsorized sample mean for the *j*th level of the between-subjects factor and the *m*th level of the within-subjects factor. For fixed *j*, let $\mathbf{S}_{j_W} = (s_{jml})$. That is, $\mathbf{S}_{j_W}$ estimates the K × K Winsorized covariance matrix for the *j*th level of factor J.

In our study we applied the robust estimators to the IGA and WJ procedures. For example, with the WJ procedure hypotheses about the repeated measures main and interaction effects can now be expressed as

$$H_0: \mathbf{C}\boldsymbol{\mu}_t = \mathbf{0}, \tag{6}$$

where $\boldsymbol{\mu_t}$ is a vector of population trimmed means. Let $\mathbf{S_W} = \text{diag}[(n_1 - 1)\mathbf{S}_{1W}/[h_1(h_1 - 1)] \cdots (n_J - 1)\mathbf{S}_{JW}/[h_J(h_J - 1)]]$ be a block diagonal matrix, where $h_j = n_j - 2g_j$. For each j and k, let $\bar{Y}_{tjk}$ be the trimmed mean based on $Y_{1jk}, \ldots, Y_{n_jjk}$. That is,

$$\bar{Y}_{tjk} = \frac{1}{n_j - 2g_j} \sum_{i=g_j+1}^{n_j-g_j} Y_{(i)jk}, \tag{7}$$

where $Y_{(1)jk} \leq Y_{(2)jk} \leq \ldots \leq Y_{(n_j)jk}$ are the n_j values in the jkth treatment group written in ascending order.

Accordingly, the WJ statistic is

$$T_{WJ_t} = (\mathbf{C}\,\bar{\mathbf{Y}}_{\boldsymbol{t}})'(\mathbf{C}\,\mathbf{S_W}\,\mathbf{C}')^{-1}(\mathbf{C}\,\bar{\mathbf{Y}}_{\boldsymbol{t}}), \tag{8}$$

where $\bar{\mathbf{Y}}_{\boldsymbol{t}} = (\bar{\mathbf{Y}}'_{t11}, \ldots, \bar{\mathbf{Y}}'_{tJK})'$ and A is now defined as

$$A = \frac{1}{2}\sum_{j=1}^{J}[\text{tr}\{\mathbf{S_W}\,\mathbf{C}'(\mathbf{C}\mathbf{S_W}\,\mathbf{C}')^{-1}\,\mathbf{C}\,\mathbf{Q}_j\}^2 + \{\text{tr}(\mathbf{S_W}\,\mathbf{C}'(\mathbf{C}\mathbf{S_W}\,\mathbf{C}')^{-1}\,\mathbf{C}\,\mathbf{Q}_j)\}^2]/(h_j - 1). \tag{9}$$

Bootstrapping

Rather than approximate the null distribution of IGA_t and T_{WJt} with an F distribution, a percentile-t bootstrap estimate of the critical value can be used instead. That is, Westfall and Young's (1993) results suggest that Type I error control could be improved by combining bootstrap methods with methods based on trimmed means. The asymptotic results provided by Hall and Padmanabhan (1992) support this conjecture and the results of Wilcox (1997a) provide empirical support. Additional asymptotic results supporting the use of the percentile-*t* bootstrap stem from general conditions where it is second-order accurate, as opposed to only first-order accurate as is obtained with

standard asymptotic methods (see, e.g., Hall, 1986). Roughly, this means that when the goal is to have the probability of a Type I error equal alpha, its error in achieving this goal goes to zero at the rate 1/n, in contrast to standard asymptotic methods where the error goes to zero at the rate of $1/(n)^{\frac{1}{2}}$.

For a fixed value of j randomly sample, with replacement, n_j rows of observations from the matrix

$$\begin{bmatrix} Y_{1j1}, \ldots, Y_{1jK} \\ \vdots \\ Y_{n_j j1}, \ldots, Y_{n_j jK} \end{bmatrix}.$$

Label the results

$$\begin{bmatrix} Y^*_{1j1}, \ldots, Y^*_{1jK} \\ \vdots \\ Y^*_{n_j j1}, \ldots, Y^*_{n_j jK} \end{bmatrix}.$$

Next, set $C_{ijk} = Y^*_{ijk} - \overline{Y}_{tjk}$. That is, shift the bootstrap samples so that, in effect, the bootstrap samples are obtained from a distribution for which the null hypothesis of equal trimmed means is true. Next compute T^*_{WJt} (or IGA^*_t), the value of the statistic T_{WJt} (or IGA_t) (based on the C_{ijk} values). Repeat this process B times yielding T^*_b, $b = 1, \ldots, B$. Let $T^*_{(1)} \leq T^*_{(2)} \leq \ldots \leq T^*_{(B)}$ be the B values written in ascending order and set $[m = (1 - \alpha)B]$. Then an estimate of an appropriate critical value is $T^*_{(m)}$. That is, reject the null hypothesis if T_{WJt} (or IGA_t) $> T^*_{(m)}$.) We set B at 599 (See Hall, 1986; Wilcox, 1997a). Results from Hall (1986) suggest that it may be advantageous to chose B such that $1 - \alpha$ is a multiple of $(B + 1)^{-1}$. (For more details about the percentile-t bootstrap method, see Efron & Tibshirani, 1993.)

Methods of the Simulation

The IGA and WJ approaches for testing repeated measures main and interaction effect hypotheses were examined for balanced and unbalanced designs containing one between-subjects and one within-subjects factor; there were three and four/eight levels of these factors, respectively. Specifically, we computed the IGA and WJ tests of the main and interaction effects in our J $\times$ K repeated measures design with both least squares and robust estimators and obtained critical values from the F distribution or through our percentile-t bootstrap method. Thus, we examined four methods for testing main and interaction effects: (a) the IGA and WJ tests with least squares estimators based on theoretically determined critical values, (b) the IGA and WJ tests with least squares estimators based on empirically determined critical values, (c) the IGA and WJ tests with robust estimators based on theoretically determined critical values, and (d) the IGA and WJ tests with robust estimators based on empirically determined critical values.

Combinations of five factors were investigated which included: (a) equal and unequal covariance structures, (b) equal and unequal group sizes, (c) pairings of covariance matrices and group sizes, (d) the value of the sphericity parameter, and (e) normal and nonnormal data.

Equal as well as unequal between-subjects covariance matrices were investigated. When unequal, the matrices were multiples of one another, namely $\boldsymbol{\Sigma}_1 = \frac{1}{3}\boldsymbol{\Sigma}_2$, and $\boldsymbol{\Sigma}_3 = \frac{5}{3}\boldsymbol{\Sigma}_2$ or $\boldsymbol{\Sigma}_1 = \frac{1}{5}\boldsymbol{\Sigma}_2$, and $\boldsymbol{\Sigma}_3 = \frac{9}{5}\boldsymbol{\Sigma}_2$. These degrees and type of covariance heterogeneity were selected because Keselman and Keselman (1990) found that, of the conditions they investigated, they resulted in the greatest discrepancies between the empirical and nominal rates of Type I error and, therefore, were conditions under which the effects of covariance heterogeneity could readily be examined.

The test statistics were investigated when the number of observations across groups were equal or unequal. Total sample size was based on the recommendations provided by Wilcox (1995b), Keselman et al. (1993), and Algina and Keselman (1997).

First, Wilcox recommends that groups should contain at least 20 observations when data are to be trimmed. Second, according to Keselman et al. and Algina and Keselman, in order to obtain a robust WJ test, the ratio of the smallest group size [$n_{(min)}$] to the number of repeated measurements minus one [$(K - 1)$] should be approximately 2 (4 or 5) to one when testing the main effect, depending on whether data are normally (nonnormally) distributed or 3 or 4 (7 or 8) to one, for the test of the interaction. Based on these recommendations we initially chose to investigate the following cases: (a) (20, 20, 20), (16, 20, 24) and (12, 20, 28) ($N = 60$) for $K = 4$ and (b) (35, 35, 35), (28, 35, 42), and (21, 35, 49) ($N = 105$) for $K = 8$. Note that for each value of N, both a moderate and substantial degree of group size inequality were investigated. The moderately unbalanced group sizes had a coefficient of sample size variation (C) equal to $\simeq .16$, while for the more disparate cases $C \simeq .33$, where C is defined as $(\Sigma_j(n_j - \overline{n})^2/J)^{\frac{1}{2}}/\overline{n}$, and $\overline{n}$ is the average group size. For these initial sample sizes, it is important to note, the above recommendations were not quite satisfied. However, we decided to start at this point and increase sample size if trimming and/or bootstrapping did not improve the Type I error rates for the WJ test, which according to recommendations, could be liberal for these sample sizes when data are nonnormal.

Six pairings of covariance matrices and group sizes were investigated: (a) equal n_j; equal $\boldsymbol{\Sigma}_j$, (b) equal n_j; unequal $\boldsymbol{\Sigma}_j$, (c/c′) unequal n_j; unequal $\boldsymbol{\Sigma}_j$ (positively paired), and (d/d′) unequal n_j; unequal $\boldsymbol{\Sigma}_j$ (negatively paired). The c′/d′ condition refers to the more disparate unequal group sizes case while the c/d condition designates the less disparate unequal group sizes case. A positive pairing results when the largest group size is associated with the covariance matrix containing the largest element values whereas a negative pairing results when the largest group size is associated with the covariance matrix with the smallest element values.

Another issue considered in the current investigation was nonsphericity. In our investigation the sphericity index was set at $\epsilon = 0.75$ or 0.57. When $\epsilon = 1.0$, sphericity is satisfied and for the J $\times$ K design the lower bound of $\epsilon = 1/(K - 1)$. The covariance matrices for each value of ϵ investigated are contained in Table 1.

Rates of Type I error were collected when the simulated data were obtained from multivariate normal or multivariate nonnormal distributions. The algorithm for generating the multivariate normal data can be found in Keselman et al (1993). The nonnormal distribution was a multivariate lognormal distribution with marginal distributions based on $Y_{ijk} = \exp(X_{ijk})$ $(i = 1, \ldots, n_j)$ where X_{ijk} is distributed as N(0, .25); this distribution has skewness (γ_1) and kurtosis (γ_2) values of 1.75 and 5.90, respectively. The procedure for generating the multivariate lognormal data is based on Johnson, Ramberg, and Wang (1982) and is presented in Algina and Oshima (1994). This particular type of nonnormal distribution was selected since applied data, particularly in the behavioral sciences, typically have skewed distributions (Micceri, 1989; Wilcox, 1994b). Furthermore, Sawilowsky and Blair (1992) found in their Monte Carlo investigation of the two independent sample t test that only distributions with extreme degrees of skewness (e.g., $\gamma_1 = 1.64$) affected Type I error control. In addition, Algina and Oshima (1995) found that tests for mean equality are affected when distributions are lognormal and homogeneity assumptions are not satisfied. Thus, we felt that our approach to modeling skewed data would adequately reflect conditions in which the tests might not perform optimally.

Type I error rates were estimated with 3,000 replications per investigated condition.

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 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, there is no one universal standard by which tests are judged to be robust, so different interpretations of the results are possible.

Our initial analysis of the data indicated that rates (percentages) of Type I error were generally well controlled when the observational vectors were obtained from normal distributions. That is, all main effect IGA and WJ rates of error, based on least squares or robust estimators, with either nonbootstrapped or bootstrapped critical values, were close to theoretical expectation regardless of type of pairing of group sizes and covariance matrices (conditions a-d'), value of epsilon ($\epsilon = .75$ and .57), or ratio of unequal covariance matrices (1:3:5 or 1:5:9) investigated. The interaction rates, with the exception of three liberal WJ values (7.57%, 7.77%, 7.80%), based on least squares estimates and nonbootstrapped critical values, were also well controlled. The liberal WJ values occurred in condition d', that is, the case involving the most disparate of the unequal group sizes negatively paired with unequal covariance matrices; it is important to remember that the smallest of the group sizes in condition d' does not conform to the size recommendations previously stipulated. Based on these initial analyses we decided to table only the results when observational vectors were obtained from lognormal distributions. (Nontabled values can be obtained upon request.)

Lognormal Data

K = 4. Rates of Type I error for the test of the repeated measures main and interaction effect for nonnormal data when there were four levels of the repeated measures variable are presented in Tables 2 and 3, respectively. One can see that there were only five liberal values in total from both tables, all associated with the WJ test based on least squares estimators. These liberal values occurred when $\epsilon = .57$. Thus, when there are four levels of the repeated measures variable, one can generally obtain a robust test of the repeated measures effects with either of the four investigated procedures. Specifically, the IGA procedure based on least squares or robust estimators always provided a valid test of the repeated measures main and interaction effect hypotheses. On the other hand, the WJ test based on least squares estimators was occasionally liberal, though well behaved when based on robust estimators. It is also important to note that, for the test of the repeated measures interaction effect, rates of Type I error were frequently (24 cases out of 96) conservative when critical values were obtained via bootstrapping.

K = 8. Rates of Type I error for the test of the repeated measures main and interaction effect when there were eight levels of the repeated measures variable for nonnormal data are presented in Tables 4 and 5, respectively. Once again the rates for the IGA tests were well controlled regardless of whether the tests were based on least squares or robust estimators or whether critical values were obtained via the bootstrap or not. The rates of Type I error for the WJ procedure were well controlled when the procedure was based on robust estimators and often not well controlled when based on least squares estimators. Interestingly, for the test of the main effect, bootstrapping was effective in providing a robust WJ test, while for the test of the interaction effect, bootstrapping resulted in conservative WJ rates of Type I error in every case but two.

Discussion

The Wech-James multivariate test, due to Johansen (1980) and presented by Keselman et al. (1993), was compared to Huynh's (1978) IGA test. Both procedures have been found to be generally robust to violations of multisample sphericity and covariance heterogeneity in unbalanced designs when data are nonnormal in form (see Keselman et al., 1993; Keselman et. al., in press, 1999; Keselman, Kowalchuk & Boik, in press). However, particularly for the WJ test, conditions do arise where rates of Type I error can be liberal, particularly if sample sizes are not as large as those prescribed by Keselman et al. (1993) and Algina and Keselman (1997). Algina and Keselman (1998) nonetheless recommended the WJ procedure over the IGA test, when sample sizes conform to the recommended guidelines, since they found that the WJ test can be substantially more powerful to detect nonnull effects.

The performance of these tests, and WJ in particular, may be improved if they are based on robust rather than least squares estimators and/or if critical values used for assessing statistical significance are obtained through a bootstrap method. Thus, we computed empirical rates of Type I error for the WJ and IGA procedures, when the procedures were based on either least squares or robust estimators (i.e., trimmed means and Winsorized variances and covariances) and when critical values used for assessing statistical significance were obtained through bootstrap or usual methods. The empirical rates of error were compiled when data were either normal/lognormal, covariance matrices were either equal/unequal, group sizes were either equal/unequal, sphericity was either moderately/severely violated, covariance matrices were either moderately/severely unequal, and when these conditions occurred in various combinations.

We found that when data were obtained from normally distributed populations both procedures were generally able to provide very effective Type I error control when they were based on least squares estimators of central tendency and variability. Utilizing

robust estimators or obtaining critical values through a bootstrap method did not generally result in substantially different rates of Type I error.

When data were nonnormal in shape (i.e., lognormal), the IGA procedure based on least squares estimators and its usual critical value continued to effectively control its rates of Type I error while the rates for the WJ test, also based on least squares estimators and its usual critical value, often were liberal (i.e., $> 7.50\%$). On the other hand, both procedures when based on robust estimators and their usual critical values resulted in well behaved rates of Type I error over the conditions examined in our investigation. Obtaining critical values through a bootstrap method did not offer any additional improvement in Type I error control. In fact, rates of Type I error were frequently very conservative (i.e., $< 2.5\%$) when the bootstrap was employed.

Based on our findings and those reported elsewhere we offer the following recommendations. When one is interested in testing main and interaction effect hypotheses pertaining to the usual population means we then recommend that researchers adopt the Welch-James procedure as long as sample sizes meet the prescriptions set forth by Keselman et al. (1993) and Algina and Keselman (1998). When sample sizes meet these prescriptions the WJ procedure will typically provide a robust test of the null hypothesis under most conditions of nonsphericity, covariance heterogeneity, nonnormality, and, as well, will typically be more powerful to detect treatment effects than the IGA test due to Huynh (1978). We make this recommendation even though in our study, rates of Type I error for WJ were often liberal. However, the reader should remember that our sample sizes did not meet the prescribed recommended sizes; we used smaller than recommended sizes because we wanted to see if these smaller sizes would nonetheless provide robust tests when robust estimators were adopted. When sample sizes are smaller than those prescribed, the IGA test involving least squares estimators should be adopted because it is very robust to assumption violations. For completeness we note that Wilcox, Keselman, Muska and Cribbie (in press) have found that the Huynh

and Feldt (1976) univariate corrected df statistic as well as the usual multivariate test statistic based on least squares estimators do not provide adequate Type I error protection under conditions similar to those investigated in our study.

It is important to note, that although the results from Monte Carlo investigations are, as always, limited to the conditions examined, our recommendations follow, and are generalizable, not only from the conditions we examined, but as well, from findings previously reported. With regard to the conditions we varied, we believe they sufficiently probed the effects of the examined variables and as well permit generalizations across a broad range of conditions likely to be encountered by behavioural science researchers. Specifically, our cases of covariance heterogeneity, nonsphericity and sample size equality/inequality cover a range of values that we believe are sufficiently broad that they should include most data sets that conceivably could be obtained in behavioural science research. That is, covariance matrices whose elemental values differ by a factor of 3:1 and 5:1 or 5:1 and 9:1 were disparate enough to sufficiently represent the effects of covariance heterogeneity for any likely real data set. Likewise, our cases of nonsphericity ($\epsilon = .75$ and .57) were sufficiently broad over the range of values that sphericity can assume. With regard to sample size, we chose our cases according to the results reported by Keselman et al. (1998). According to their survey of statistical practices of behavioural science researchers, unbalanced designs are more prevelant than balanced designs and typical sample size is 60 subjects for between by within repeated measures designs. Another point to consider, with regard to sample size, is that it was not necessary to compare the tests based on robust estimators (i.e., WJ with robust estimators) to their least squares counterparts (WJ-LS) for larger sample size cases because published findings indicate that the WJ-LS procedure will be prone to inflated rates of Type I error in large designs (i.e., $K = 8$) unless sample sizes are very large (e.g., > 300) (Algina & Keselman, 1997). Accordingly, because these sizes are typically not available to researchers (see Keselman et al., 1998), we sought a solution that would be viable with

typical sizes. Finally, with respect to the possible effects of nonnormaility on rates of Type I error, our choice of distribution was based on the results reported by Sawilowsky and Blair (1992) who indicated that it is the degree of skewness that affects rates of Type I error for tests of mean equality and that in their investigation when skewness equalled 1.64 the tests were adversely affected. This conclusion generalizes to repeated measures designs (see e.g., Keselman & Lix, 1997).

When researchers feel that they are dealing with populations that are nonnormal in form [Tukey (1960) suggests that most populations are skewed and/or contain outliers] and thus subscribe to the position that inferences pertaining to robust parameters are more valid than inferences pertaining to the usual least squares parameters, then either the IGA or WJ procedures, based on robust estimators, can be adopted. Our results certainly suggest that these procedures will provide valid tests of the repeated measures main and interaction effect hypotheses (of trimmed population means) when data are non-normal, nonspherical, and heterogeneous.

Finally, it should be noted 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, 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 contain outliers and/or are 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.

Footnotes

1. Other than the first two authors, the order of authorship was determined alphabetically. The research reported in this paper was supported by the National Science and Engineering Research Council of Canada and the Social Sciences and Humanities Research Council of Canada.

2. A choice for the amount of trimming, γ, must be made. Efficiency (achieving a relatively small standard error) is one approach to this problem. If γ is too small, efficiency can be poor when sampling from a heavy-tailed distribution. If γ is too large, efficiency is poor when sampling from a normal distribution. A good compromise is $\gamma = .2$ because efficiency is good when sampling from a normal distribution and little power is lost as compared with using means ($\gamma = 0$) (e.g., Rosenberger & Gasko, 1983; Wilcox, 1997b). In terms of computing confidence intervals and controlling Type I error probabilities, theory tells us that problems associated with means decrease as the amount of trimming increases (Wilcox, 1994a, 1994b). The improvement can be substantial as γ increases from 0 to .2, but for $\gamma > .2$ the benefits of trimming are less dramatic versus using $\gamma = .2$. Huber (1993) argues that in practice, using $\gamma < .1$ is "dangerous," meaning we run the risk of relatively high standard errors, and thus low power. Of course, situations arise where $\gamma < .2$ yields a smaller standard error versus $\gamma = .2$, but the improvement is typically small. In contrast, using $\gamma = .2$ offers a substantial improvement over .1 or 0 in many cases. For these reasons, $\gamma = .2$ is assumed henceforth when referring to the trimmed mean.

References

Algina, J. (1994). Some alternative approximate tests for a split plot design. Multivariate Behavioral Research, 29, 365-384.

Algina, J. (1997). Generalization of Improved General Approximation tests to split-plot designs with multiple between-subjects factors and/or multiple within-subjects factors. British Journal of Mathematical and Statistical Psychology, 50, 243-252.

Algina, J., & Keselman, H. J. (1997). Testing repeated measures hypotheses when covariances are heterogeneous: Revisiting the robustness of the Welch-James test. Multivariate Behavioral Research, 32, 255-274..

Algina, J., & Keselman, H. J. (1998). A power comparison of the Welch-James and Improved General Approximation tests in the split-plot design. Journal of Educational and Behavioral Statistics, 23, 152-169.

Algina, J., & Oshima, T. C. (1994). Type I error rates for Huynh's general approximation and improved general approximation tests. British Journal of Mathematical and Statistical Psychology, 47, 151-165.

Algina, J., & Oshima, T. C. (1995). An Improved General Approximation test for the main effect in a split plot design. British Journal of Mathematical and Statistical Psychology, 48, 149-160.

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

Coombs, W. T., Algina, J., & Oltman, D. O. (1996). Univariate and multivariate omnibus hypothesis tests selected to control Type I error rates when population variances anr not necessarily equal. Journal of Educational and Behavioral Statistics, 66, 137-179.

Efron, B., & Tibshirani, R. J. (1993). An introduction to the bootstrap. New York: Chapman & Hall.

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

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

Hall, P., & Padmanabhan, A. R. (1992). On the bootstrap and the trimmed mean. Journal of Multivariate Analysis, 41, 132-153.

Huber, P.J. (1981). Robust statistics. New York: Wiley.

Huber, P. J. (1993). Projection pursuit and robustness. In S. Morgenthaler, E. Ronchetti, & W. Stahel (Eds.) New directions in statistical data analysis and robustness. Boston: Birkhauser Verlag.

Huynh, H. (1978). Some approximate tests for repeated measurement designs. Psychometrika, 43, 161-175.

Huynh, H. & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. Journal of Educational Statistics, 1, 69-82.

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 of the distribution of the residual sum of squares in weighted linear regression. Biometrika, 67, 85-92.

Johnson, M. F., Ramberg, J. S., & Wang, C. (1982). The Johnson translation system in Monte Carlo studies. Communications in Statistics-Simulation and Computation, 11, 521-525.

Keselman, H. J., & Algina, J. (1996). The analysis of higher-order repeated measures designs. In Advances in Social Science Methodology, Volume 4, ed. B. Thompson, Greenwich, Conneticuit: JAI Press, (pp. 45-70).

Keselman, H. J., Algina, J., Kowalchuk, R. K., & Wolfinger, R. D. (1999). A comparison of recent approaches to the analysis of repeated measurements. British Journal of Mathematical and Statistical Psychology, 52, 63-78.

Keselman, H. J., Algina, J., Kowalchuk, R. K., & Wolfinger, R. D. (in press). The analysis of repeated measurements: A comparison of mixed-model Satterthwaite F tests and a nonpooled adjusted degrees of freedom multivariate test. Communications in Statistics-Simulation and Computation.

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., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., & Levin, J. R. (1998). Statistical practices of Educational Researchers: An analysis of their ANOVA, MANOVA and ANCOVA analyses. Review of Educational Research, 68(3), 350-386.

Keselman, H. J., Kowalchuk, R. K., & Boik, R. J. (in press). An investigation of the Empirical Bayes approach to the analysis of repeated measurements. British Journal of Mathematical and Statistical Psychology.

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

Keselman, H. J. & Lix, L. M. (1997). Analyzing multivariate repeated measures designs when covariance matrices are heterogeneous. British Journal of Mathematical and Statistical Psychology, 50, 319-338.

Keselman, J.C., & Keselman, H.J. (1990). Analysing unbalanced repeated measures designs. British Journal of Mathematical and Statistical Psychology, 43, 265-282.

Lecoutre, B. (1991). A correction for the $\tilde{\epsilon}$ approximate test in repeated measures designs with two or more independent groups. Journal of Educational Statistics, 16, 371-372.

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

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

Lix, L. M., Keselman, H. J., & Algina, J. (1997, April). Trimmed means in split-plot repeated measures designs. Paper presented at the Annual Meeting of The American Educational Research Association (Chicaogo, Illinois)

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

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

SAS Institute. (1989). SAS/IML sowtware: Usage and reference, Version 6. Cary, NC: Author.

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

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 Students' problems 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.

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

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

Wilcox, R. R. (1997a). Pairwise comparisons using trimmed means or M-estimators when working with dependent groups. Biometrical Journal, 39, 677-688.

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

Wilcox, R.R. (1998). The goals and strategies of robust methods. British Journal of Mathematical and Statistical Psychology, 51, 1-39.

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

Wilcox, R. R., Keselman, H. J., Muska, J., & Cribbie, R. (in press). Repeated measures ANOVA: Some new results on comparing trimmed means and means. British Journal of Mathematical and Statistical Psychology.

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

Table 1. Empirical Main Effect Rates of Type I Error (NormalData; K=4; N=60)

Cond	Test	Σ_j=1:3:5				Σ_j=1:5:9			
		LS		RE		LS		RE	
		~B	B	~B	B	~B	B	~B	B
Epsilon=.75									
a	IGA	5.47	5.23	4.33	4.77	5.10	5.20	4.90	5.50
	WJ	5.10	4.97	3.10	4.23	5.10	4.83	3.97	5.23
b	IGA	5.03	4.90	4.67	5.20	4.90	4.90	4.57	4.90
	WJ	4.80	4.67	3.63	4.60	4.93	4.57	3.33	3.87
c	IGA	4.37	4.40	4.10	4.67	5.43	5.60	5.13	5.40
	WJ	4.27	4.37	3.97	4.73	4.97	4.90	3.60	4.47
d	IGA	5.13	4.77	4.63	4.97	5.03	4.60	5.07	5.53
	WJ	4.83	4.53	4.27	4.60	5.47	4.57	3.90	4.40
c’	IGA	5.33	5.30	4.23	4.70	4.70	4.73	4.00	4.47
	WJ	4.97	5.00	3.77	4.73	4.63	4.33	3.47	4.20
d’	IGA	5.20	4.80	5.40	5.70	5.30	4.77	5.47	5.27
	WJ	5.40	4.47	4.07	3.70	5.30	4.07	3.97	3.77
Epsilon=.57									
a	IGA	4.77	4.93	4.93	5.00	4.67	4.87	5.03	5.47
	WJ	4.43	4.30	4.07	4.67	4.33	4.30	4.03	4.50
b	IGA	5.33	5.17	5.03	5.50	5.60	5.67	4.93	5.47
	WJ	4.77	4.73	4.80	5.50	5.57	5.17	3.77	4.20
c	IGA	4.97	5.03	4.60	5.03	4.33	4.43	4.87	5.10
	WJ	3.93	4.03	3.93	4.47	5.03	4.87	3.80	4.63
d	IGA	5.40	5.30	4.50	4.67	5.50	5.33	5.67	5.73
	WJ	4.47	4.23	4.10	4.00	4.83	4.33	4.17	4.37
c’	IGA	5.50	5.60	5.40	5.43	5.10	5.10	5.23	5.87
	WJ	4.40	4.43	3.87	4.43	4.63	4.27	4.03	4.60
d’	IGA	5.13	4.93	5.33	5.20	5.27	5.03	5.20	5.03
	WJ	5.20	4.27	3.57	3.70	5.07	3.77	3.67	3.07

Note:L(Least)S(Squares)/R(Robust)E(estimation); ~B-No bootstrapping/B-bootstrapping. See the Methods Section for a description of conditions a, b, c, c’, d, and d’.

Table 2. Empirical Interaction Effect Rates of Type I Error (Normal Data; K=4; N=60)

Cond	Test	Σ_j=1:3:5				Σ_j=1:5:9			
		LS		RE		LS		RE	
		~B	B	~B	B	~B	B	~B	B
Epsilon=.75									
a	IGA	5.33	5.40	5.13	5.43	4.60	4.37	4.00	4.47
	WJ	5.33	4.57	3.87	3.83	5.80	4.40	3.97	3.60
b	IGA	4.27	4.20	4.53	4.90	5.07	5.10	4.10	4.50
	WJ	5.97	4.77	4.63	4.17	5.40	3.70	3.37	2.70
c	IGA	4.80	4.77	4.07	4.53	4.93	4.90	5.20	5.47
	WJ	4.40	3.23	3.40	3.00	4.73	3.50	3.57	3.20
d	IGA	5.07	4.80	4.57	5.00	5.13	4.97	4.90	5.47
	WJ	5.73	3.10	4.10	2.63	5.57	2.83	4.33	2.57
c’	IGA	4.17	4.30	4.00	4.47	4.67	4.97	3.80	4.47
	WJ	4.67	3.63	3.27	3.17	4.80	3.33	4.10	3.73
d’	IGA	5.27	4.90	5.00	5.07	4.90	4.37	4.87	4.83
	WJ	6.10	3.03	4.53	**2.17**	5.70	**2.33**	4.97	**1.40**
Epsilon=.57									
a	IGA	4.83	4.73	4.93	5.00	4.87	5.03	4.50	5.13
	WJ	5.10	4.00	3.97	3.43	4.57	3.53	3.73	3.10
b	IGA	5.30	4.93	5.20	5.47	5.37	5.10	5.57	5.83
	WJ	5.27	3.73	4.33	3.33	5.17	3.23	4.37	2.90
c	IGA	4.23	4.23	3.90	4.33	4.53	4.63	4.83	5.20
	WJ	5.30	4.10	4.00	3.17	5.43	3.93	4.73	3.77
d	IGA	4.67	4.43	4.80	4.77	5.03	4.63	4.77	4.93
	WJ	5.13	2.90	5.00	2.77	5.57	3.10	5.00	2.60
c’	IGA	4.40	4.43	4.93	5.13	5.10	5.53	4.97	5.90
	WJ	5.23	3.83	3.87	2.90	5.40	4.20	4.67	3.53
d’	IGA	5.23	4.77	5.13	4.93	5.00	4.20	5.50	5.33
	WJ	5.47	**2.10**	4.80	**1.87**	6.43	**2.20**	5.30	**2.03**

Note:L(Least)S(Squares)/R(Robust)E(estimation); ~B-No bootstrapping/B-bootstrapping. See the Methods Section for a description of conditions a, b, c, c’, d, and d’.

Table 2. Empirical Main Effect Rates of Type I Error (Lognormal Data; K=4; N=60)

Cond	Test	Σ_j=1:3:5				Σ_j=1:5:9			
		LS		RE		LS		RE	
		~B	B	~B	B	~B	B	~B	B
Epsilon=.75									
a	IGA	5.17	5.23	4.37	4.90	5.27	5.23	4.80	5.67
	WJ	5.43	4.37	3.43	4.53	5.20	4.33	3.97	5.00
b	IGA	5.20	5.03	4.40	5.17	4.77	4.67	4.60	5.10
	WJ	4.77	3.90	3.57	4.37	4.97	3.80	3.27	3.90
c	IGA	4.70	4.17	3.80	4.73	4.97	4.80	4.70	5.63
	WJ	4.73	3.83	3.63	4.47	5.03	4.40	3.63	4.60
d	IGA	5.00	4.67	3.97	4.57	5.33	5.20	4.97	5.63
	WJ	5.37	4.00	3.70	4.37	5.73	3.60	3.40	4.10
c’	IGA	4.57	4.33	3.73	4.37	4.30	4.07	3.50	4.17
	WJ	5.13	4.40	3.10	4.47	4.87	3.83	3.20	4.20
d’	IGA	5.53	5.10	4.90	5.37	6.00	5.37	5.23	5.43
	WJ	6.13	4.40	3.40	3.53	5.93	3.83	3.37	3.40
Epsilon=.57									
a	IGA	5.63	5.10	4.43	5.07	5.00	4.60	4.40	5.00
	WJ	6.63	5.20	4.40	4.67	6.20	4.53	3.87	4.20
b	IGA	5.67	5.33	5.10	5.37	6.50	6.20	4.87	5.60
	WJ	6.67	4.83	4.87	4.90	**7.77**	5.40	3.83	4.37
c	IGA	5.97	5.57	4.67	5.10	5.53	5.23	4.77	5.03
	WJ	6.87	4.73	3.80	4.27	6.73	4.80	3.80	4.07
d	IGA	6.43	5.83	4.63	4.90	6.37	5.80	5.67	5.60
	WJ	7.10	5.07	3.57	3.53	**7.73**	5.37	3.80	3.67
c’	IGA	5.77	5.63	5.23	5.43	5.00	4.70	5.23	5.77
	WJ	6.60	5.13	4.13	4.57	5.67	4.03	3.93	4.47
d’	IGA	6.87	6.03	5.23	5.17	5.97	5.37	5.03	4.67
	WJ	7.23	4.70	3.87	3.57	7.47	4.70	3.13	**2.37**

Note:L(Least)S(Squares)/R(Robust)E(estimation); ~B-No bootstrapping/B-bootstrapping. See the Methods Section for a description of conditions a, b, c, c’, d, and d’.

Table 3. Empirical Interaction Effect Rates of Type I Error (Lognormal Data; K=4; N=60)

Cond	Test	Σ_j=1:3:5				Σ_j=1:5:9			
		LS		RE		LS		RE	
		~B	B	~B	B	~B	B	~B	B
Epsilon=.75									
a	IGA	4.50	4.37	4.20	4.97	3.80	3.47	3.13	3.80
	WJ	4.77	2.53	3.23	3.23	4.83	**2.40**	3.07	3.10
b	IGA	4.70	4.50	4.00	4.90	4.10	4.00	3.57	4.03
	WJ	5.50	2.73	4.00	3.80	5.47	**2.37**	2.93	**2.43**
c	IGA	4.37	4.20	3.73	4.43	4.17	4.07	4.37	5.10
	WJ	4.17	**2.27**	2.73	**2.37**	5.00	2.50	3.07	2.60
d	IGA	4.07	3.73	4.13	4.80	4.73	4.33	4.07	4.80
	WJ	5.30	**2.00**	3.33	**2.00**	6.10	**1.93**	3.47	**2.40**
c’	IGA	3.73	4.00	3.67	4.37	3.73	3.87	3.50	4.20
	WJ	4.37	**2.23**	3.00	2.73	4.40	**1.97**	3.20	2.73
d’	IGA	4.53	4.13	4.30	5.03	4.50	3.97	4.00	4.27
	WJ	6.20	**1.83**	4.30	**1.70**	6.50	**1.73**	4.40	**1.53**
Epsilon=.57									
a	IGA	3.93	3.70	4.00	4.57	3.97	3.67	3.87	4.83
	WJ	4.37	2.17	3.17	2.63	4.00	**1.93**	3.33	2.57
b	IGA	5.17	5.10	4.73	5.60	5.03	4.80	5.20	5.40
	WJ	6.60	3.43	4.40	2.97	7.17	2.93	4.27	2.67
c	IGA	4.37	4.03	3.50	4.53	4.37	4.27	4.50	4.87
	WJ	5.90	2.77	3.40	2.70	6.60	2.80	4.43	3.17
d	IGA	4.50	4.07	4.17	4.60	5.30	4.97	4.53	5.07
	WJ	6.60	**2.37**	4.17	**2.27**	**8.00**	3.20	5.33	2.57
c’	IGA	4.30	4.13	4.13	4.93	4.57	4.60	4.73	5.83
	WJ	5.10	**2.17**	3.37	**2.40**	5.57	**2.23**	4.00	3.10
d’	IGA	5.47	4.77	4.40	4.77	5.23	4.63	4.70	5.07
	WJ	**7.60**	2.53	4.50	**1.87**	**9.23**	**2.37**	5.33	**1.63**

Note:L(Least)S(Squares)/R(Robust)E(estimation); ~B-No bootstrapping/B-bootstrapping. See the Methods Section for a description of conditions a, b, c, c’, d, and d’.

Table 3. Empirical Main Effect Rates of Type I Error (Normal Data; K=8; N=105)

Cond	Test	Σ_j=1:3:5				Σ_j=1:5:9			
		LS		RE		LS		RE	
		~B	B	~B	B	~B	B	~B	B
Epsilon=.75									
a	IGA	4.90	4.50	5.07	5.37	5.27	5.07	5.10	5.27
	WJ	5.23	5.33	3.93	4.90	5.13	5.27	4.10	5.40
b	IGA	5.50	5.47	5.63	5.63	5.33	4.90	4.87	5.10
	WJ	5.83	5.27	3.73	4.47	5.13	4.50	3.83	4.43
c	IGA	4.53	4.50	4.83	4.73	5.37	5.23	5.07	5.23
	WJ	4.80	4.47	3.47	4.40	4.50	4.20	3.87	4.73
d	IGA	5.40	5.23	5.20	5.00	4.73	4.57	5.17	5.30
	WJ	5.60	4.47	3.63	4.20	4.53	3.70	3.77	3.83
c’	IGA	4.87	4.57	4.27	4.40	5.60	5.33	4.67	4.97
	WJ	4.70	4.47	3.37	4.47	5.43	5.33	3.93	4.83
d’	IGA	4.60	4.20	3.87	4.73	5.13	4.20	4.23	4.47
	WJ	5.23	3.13	4.70	3.63	4.43	3.03	3.37	2.73
Epsilon=.57									
a	IGA	5.20	5.13	5.57	5.47	4.93	4.87	5.20	5.27
	WJ	4.90	4.67	3.30	4.47	4.63	4.57	3.43	4.23
b	IGA	4.77	4.47	4.93	5.17	4.80	4.53	4.00	4.20
	WJ	5.00	4.50	3.77	4.53	4.90	4.07	3.53	4.17
c	IGA	5.17	5.13	4.60	4.80	5.60	5.47	5.07	4.77
	WJ	5.43	5.07	4.30	5.23	5.47	5.03	3.90	4.73
d	IGA	4.67	4.63	4.27	4.60	5.10	4.97	4.83	4.80
	WJ	5.13	4.43	3.63	4.10	5.30	4.17	3.60	3.97
c’	IGA	4.37	4.20	4.20	4.60	4.83	4.60	5.07	5.10
	WJ	4.53	4.60	3.90	4.60	4.80	4.53	3.57	4.60
d’	IGA	4.57	4.27	4.63	4.60	5.77	5.37	4.43	4.50
	WJ	5.57	3.53	4.10	3.87	5.67	3.53	4.30	3.77

Note:L(Least)S(Squares)/R(Robust)E(estimation); ~B-No bootstrapping/B-bootstrapping. See the Methods Section for a description of conditions a, b, c, c’, d, and d’.

Table 4. Empirical Interaction Effect Rates of Type I Error (Normal Data; K=8; N=105)

Cond	Test	Σ_j=1:3:5				Σ_j=1:5:9			
		LS		RE		LS		RE	
		~B	B	~B	B	~B	B	~B	B
Epsilon=.75									
a	IGA	5.30	5.17	4.77	5.17	4.17	3.83	4.23	4.30
	WJ	5.07	3.13	3.20	3.00	5.07	3.53	3.17	2.70
b	IGA	4.83	4.43	4.17	4.23	5.43	5.10	4.43	4.80
	WJ	5.43	3.33	3.33	2.50	6.30	3.23	3.70	**2.43**
c	IGA	4.67	4.37	3.87	4.10	5.40	4.90	4.23	4.57
	WJ	5.10	3.33	3.07	2.73	5.07	3.23	3.20	**2.47**
d	IGA	5.17	4.87	4.63	4.83	4.83	4.77	4.13	4.33
	WJ	6.70	2.87	3.97	**2.10**	5.93	2.33	3.57	**1.67**
c’	IGA	4.90	4.67	4.47	4.77	5.50	5.27	4.97	5.20
	WJ	5.40	3.50	3.43	2.73	5.47	3.93	3.43	2.77
d’	IGA	4.30	4.00	4.00	4.23	4.60	4.07	3.67	3.87
	WJ	**7.57**	**1.47**	5.53	**1.43**	7.17	**1.23**	5.03	**0.53**
Epsilon=.57									
a	IGA	4.73	4.43	4.73	4.93	5.20	4.97	5.00	5.40
	WJ	4.80	2.93	3.53	2.80	5.50	3.40	3.20	3.10
b	IGA	4.27	4.00	4.27	4.57	5.00	5.13	4.40	4.97
	WJ	5.70	3.40	3.73	2.50	5.60	2.83	3.40	**1.97**
c	IGA	4.77	4.53	4.73	5.23	4.37	4.20	4.73	5.13
	WJ	5.27	3.40	3.20	2.53	5.30	3.17	3.60	2.67
d	IGA	4.83	4.60	4.60	4.97	5.40	5.03	5.27	5.27
	WJ	6.67	2.70	3.73	**1.97**	6.57	2.60	4.10	**1.87**
c’	IGA	4.80	4.77	4.23	4.50	4.87	4.97	4.03	4.23
	WJ	5.83	3.50	4.00	3.10	4.73	3.10	3.40	2.57
d’	IGA	4.80	4.57	4.97	5.33	4.67	4.47	4.20	4.47
	WJ	**7.77**	**1.60**	5.20	**0.97**	**7.80**	**1.57**	5.57	**1.10**

Note:L(Least)S(Squares)/R(Robust)E(estimation); ~B-No bootstrapping/B-bootstrapping. See the Methods Section for a description of conditions a, b, c, c’, d, and d’.

Table 4. Empirical Main Effect Rates of Type I Error (Lognormal Data; K=8; N=105)

Cond	Test	Σ_j=1:3:5				Σ_j=1:5:9			
		LS		RE		LS		RE	
		~B	B	~B	B	~B	B	~B	B
Epsilon=.75									
a	IGA	5.17	4.83	5.13	5.40	4.43	4.10	4.97	5.37
	WJ	7.23	5.10	3.90	5.03	5.77	4.23	3.70	4.83
b	IGA	5.37	5.07	5.30	5.70	4.70	4.57	4.77	5.30
	WJ	7.47	4.67	3.90	4.57	6.03	4.10	3.73	4.20
c	IGA	4.67	4.27	4.37	4.73	5.17	5.33	5.00	5.50
	WJ	6.07	4.40	3.30	4.13	6.67	4.37	3.60	4.77
d	IGA	4.60	4.27	4.40	4.73	3.93	3.60	4.70	5.37
	WJ	6.20	3.13	3.27	3.57	6.03	3.20	3.27	3.53
c’	IGA	4.37	4.07	3.93	4.07	5.03	4.90	4.83	5.10
	WJ	5.83	4.10	3.23	4.27	5.97	4.47	3.87	4.70
d’	IGA	3.73	3.23	4.37	4.43	4.00	3.63	3.40	3.87
	WJ	6.30	3.03	3.57	3.23	7.20	**2.43**	3.23	2.67
Epsilon=.57									
a	IGA	4.63	4.37	5.03	5.37	4.47	4.27	4.53	4.93
	WJ	7.00	4.47	3.50	4.23	**7.90**	5.00	3.47	4.07
b	IGA	4.70	4.33	4.70	4.90	4.80	4.57	3.77	4.33
	WJ	7.20	4.27	3.73	4.50	**7.97**	4.07	3.90	4.47
c	IGA	4.80	4.77	4.47	4.87	5.23	5.07	4.53	5.10
	WJ	**7.93**	4.90	4.03	4.60	**7.63**	4.77	3.87	4.67
d	IGA	4.53	4.20	4.23	4.77	5.80	5.40	4.90	5.17
	WJ	**9.03**	4.90	3.87	3.93	**9.00**	4.50	4.23	4.10
c’	IGA	4.60	4.60	4.23	4.67	4.40	4.50	5.03	5.47
	WJ	**7.67**	4.83	3.83	4.70	6.93	4.53	4.10	4.87
d’	IGA	4.63	4.53	4.13	4.53	4.67	4.37	4.07	4.43
	WJ	**10.00**	3.77	3.97	3.17	**9.20**	3.90	4.17	3.10

Note:L(Least)S(Squares)/R(Robust)E(estimation); ~B-No bootstrapping/B-bootstrapping.

Table 5. Empirical Interaction Effect Rates of Type I Error (Lognormal Data; K=8; N=105)

Cond	Test	Σ_j=1:3:5				Σ_j=1:5:9			
		LS		RE		LS		RE	
		~B	B	~B	B	~B	B	~B	B
Epsilon=.75									
a	IGA	4.63	4.40	4.47	5.03	3.13	3.10	3.43	4.10
	WJ	4.93	**1.63**	2.53	2.50	4.77	**1.40**	**2.43**	**2.27**
b	IGA	3.80	3.77	4.17	4.77	4.03	4.07	3.73	4.53
	WJ	6.83	**1.50**	2.77	**2.20**	7.00	**1.93**	2.97	**1.90**
c	IGA	3.63	3.47	3.80	4.50	4.67	4.40	4.33	4.83
	WJ	5.63	**1.47**	**2.33**	**2.17**	6.47	**1.87**	2.83	**2.13**
d	IGA	3.87	3.47	4.27	4.90	4.17	3.97	3.53	4.10
	WJ	7.17	**1.90**	3.30	**1.67**	**8.33**	**1.70**	3.03	**1.13**
c’	IGA	4.00	3.87	3.93	4.60	4.67	4.43	4.27	4.90
	WJ	5.27	**1.50**	2.53	**2.17**	5.67	**1.97**	2.70	**2.03**
d’	IGA	3.70	3.30	3.17	4.03	3.53	3.27	3.03	3.50
	WJ	**9.13**	**1.07**	4.83	**1.40**	**9.33**	**0.90**	4.73	**0.73**
Epsilon=.57									
a	IGA	4.27	4.13	4.07	4.57	4.10	4.13	4.47	5.17
	WJ	4.47	**1.20**	**2.47**	**2.33**	5.10	**1.33**	2.93	2.50
b	IGA	3.83	3.73	4.03	4.50	4.37	4.10	4.53	5.07
	WJ	7.07	**1.93**	2.90	**2.03**	**8.13**	**1.67**	2.93	**1.67**
c	IGA	4.43	4.47	4.47	5.13	4.00	3.90	4.43	4.93
	WJ	5.93	**1.67**	2.97	**2.47**	7.40	**1.93**	3.17	**2.20**
d	IGA	3.27	3.10	4.07	4.77	5.10	4.87	4.63	5.13
	WJ	**8.37**	**2.33**	3.23	**1.70**	**11.60**	**2.47**	3.97	**1.77**
c’	IGA	4.43	4.43	3.97	4.73	4.50	4.50	3.50	3.90
	WJ	5.50	**1.30**	2.97	**2.33**	5.57	**1.73**	2.63	**1.90**
d’	IGA	3.93	3.60	4.27	4.87	3.80	3.43	3.90	4.53
	WJ	**12.23**	**1.37**	4.90	**1.10**	**13.77**	**1.43**	5.77	**1.10**

Note:L(Least)S(Squares)/R(Robust)E(estimation); ~B-No bootstrapping/B-bootstrapping.