A Generally Robust Approach To Hypothesis Testing in Independent and Correlated Groups Designs

by

H. J. Keselman

University of Manitoba

Rand R. Wilcox

University of Southern California

and

Lisa M. Lix

University of Manitoba

Abstract

Standard least squares analysis of variance methods suffer from poor power under arbitrarily small departures from normality and fail to control the probability of a Type I error when standard assumptions are violated. These problems are vastly reduced when using a robust measure of location; incorporating bootstrap methods can result in additional benefits. This paper illustrates the use of trimmed means with an approximate degrees of freedom heteroscedastic statistic for independent and correlated groups designs in order to achieve robustness to the biasing effects of nonnormality and variance heterogeneity. As well, we indicate when a boostrap methodology can be effectively employed to provide improved Type I error control. We also illustrate, with examples from the psychophysiological literature, the use of a new computer program to obtain numerical results for these solutions.

Descriptors: Heteroscedastic variances, nonnormal distributions, robust estimators, bootstrapping, independent and correlated groups.

Robust Estimation and Testing

On a number of occasions, *Psychophysiology* has published articles that are intended to identify problems with traditional methods of analyzing psychophysiological data and indicate how valid and reliable results could generally be obtained by adopting newer methods (e.g., Keselman, 1998). Our intention in the present article is to extend that body of literature by offering a general framework for statistical tests (omnibus and focused hypothesis tests) that are robust to the biasing effects of variance heterogeneity and nonnormality in both independent and correlated groups designs.

Research has shown that the deleterious effects of (co)variance heterogeneity on the usual omnibus analysis of variance (ANOVA) F and linear contrast tests (Student's t) generally can be overcome by adopting Welch (1938, 1951)-type statistics (see Lix & Keselman, 1998; Keselman, Lix & Kowalchuk, 1998), that is, statistics that do not pool across heterogeneous sources of variability and where error degrees of freedom are estimated from the sample data. The biasing 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 (co)variances rather than the usual least squares estimators (see Lix & Keselman, 1998; Wilcox, 1997). A number of papers have demonstrated that one can indeed generally achieve robustness to nonnormality and (co)variance heterogeneity in unbalanced independent and correlated groups designs by using robust estimators with heteroscedastic test statistics (Keselman, Algina, Wilcox, Kowalchuk, 2000; Keselman, Kowalchuk & Lix, 1998).

Further improvement in Type I error control is often possible by obtaining critical values for test statistics through bootstrap methods. Such improvement has been demonstrated with statistics for independent group designs (Wilcox, Keselman, & Kowalchuk, 1998). Wasserman and Bockenholt (1989) introduced the technique of bootstrapping to psychophysiologists, defining the methodology and indicating how various inferential problems [e.g., correlational and general linear model (GLM) analyses] could be addressed via bootstrapping techniques.

Our paper is a follow-up to Wasserman and Bockenholt (1989) in three important ways: (a) first, we demonstrate how bootstrapping can be applied to tests of significance, rather than just to interval estimates around population parameters; (b) we discuss the use of the bootstrapping methodology with robust estimators (viz. trimmed means) rather than the usual least squares estimates; and (c) we illustrate the use of a new computer program to produce Welch (1938, 1951)-type approximate degrees of freedom (ADF) test statistics in combination with robust estimators and/or bootstrapping. These topics are discussed both for independent and correlated groups designs. [A nontechnical exposition of robust estimation and testing can be found in Wilcox (2001).]

A General ADF Test Statistic

Methods that give improved power and better control over the probability of a Type I error can be formulated using a GLM ADF perspective. Lix and Keselman (1995) showed how the various Welch (1938, 1951) statistics that appear in the literature for testing omnibus main and interaction effects as well as focused hypotheses using contrasts in univariate and multivariate independent and correlated groups designs can be formulated from a GLM ADF perspective, thus allowing researchers to apply one statistical procedure to any testable model effect. We adopt their approach in this paper and begin by presenting, in abbreviated form, its mathematical underpinnings.[1]

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

$$\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\xi} \, , \qquad (1)$$

where $\mathbf{Y}$ is an $\mathrm{N} \times \mathrm{p}$ matrix of scores on p dependent variables or p repeated measurements, N is the total sample size, $\mathbf{X}$ is an $\mathrm{N} \times \mathrm{r}$ design matrix consisting entirely of zeros and ones with rank($\mathbf{X}$) $=$ r, β is an $\mathrm{r} \times \mathrm{p}$ matrix of nonrandom parameters (i.e., population means), and $\boldsymbol{\xi}$ is an $\mathrm{N} \times \mathrm{p}$ matrix of random error components. Let $\mathbf{Y}_{\mathrm{j}}$

($j = 1, \ldots, r$) denote the submatrix of **Y** containing the scores associated with the n subjects in the jth group (cell). It is typically assumed that the rows of **Y** are independently and normally distributed, with mean vector β_j and variance-covariance matrix $\mathbf{\Sigma}_j$ [i.e., $N(\beta_j, \mathbf{\Sigma}_j)$], where the jth row of β, $\beta_j = [\mu_{j1} \ldots \mu_{jp}]$, and $\mathbf{\Sigma}_j \neq \mathbf{\Sigma}_{j'}$ ($j \neq j'$). Specific formulas for estimating β and $\mathbf{\Sigma}_j$, as well as an elaboration of **Y** are given in Lix and Keselman (1995, see their Appendix A).

The general linear hypothesis is

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

where $\mathbf{R} = \mathbf{C} \otimes \mathbf{U}^T$, **C** is a $df_C \times r$ matrix which controls contrasts on the independent groups effect(s), with $\text{rank}(\mathbf{C}) = df_C \leq r$, and **U** is a $p \times df_U$ matrix which controls contrasts on the within-subjects effect(s), with $\text{rank}(\mathbf{U}) = df_U \leq p$, '$\otimes$' is the Kronecker or direct product function, and 'T' is the transpose operator. For multivariate independent groups designs, **U** is an identity matrix of dimension p (i.e., $\mathbf{I}_p$). The **R** contrast matrix has $df_C \times df_U$ rows and $r \times p$ columns. In Equation 2, $\boldsymbol{\mu} = \text{vec}(\beta^T) = [\beta_1 \ldots \beta_r]^T$. In other words, $\boldsymbol{\mu}$ is the column vector with $r \times p$ elements obtained by stacking the columns of β^T. The **0** column vector is of order df_C x df_U [see Lix & Keselman (1995) for illustrative examples].

The generalized test statistic given by Johansen (1980) is

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

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

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

Obtaining Numerical Results Using an ADF Solution with Robust Estimators and/or Bootstrapping

Keselman, Wilcox and Lix (2001) present a SAS/IML (SAS Institute Inc, 1999) program which can be used to obtain numerical results for the general ADF solution. The program can also be obtained from the first author's website at http://www.umanitoba.ca/faculties/arts/psychology/. This program is an extension of the program found in Lix and Keselman (1995). The general ADF solution contained in the current program can be applied with robust estimators, that is, trimmed means and Winsorized variances (covariances) and can also be used in conjunction with a bootstrapping methodology. Tests of omnibus main effects or interaction effects may be performed, in addition to tests of individual contrasts or families of contrasts. The program can be applied in a variety of research designs; several applications of the program will be explored in the following sections of this paper.

The main module, which is called WJGLM, requires as input **Y**, **C**, **NX**, **OPT1**, and **OPT2**. By default, $\mathbf{U} = \mathbf{I}_p$, but for correlated groups designs, the program user must specify the elements of **U**. The vector **NX** is a $1 \times r$ vector containing the number of observations in each group or cell (i.e., the n_js). It is assumed that the order of entry for **Y** and **NX** correspond, so that the first n_1 rows of **Y** correspond to the first element of **NX**, the next n_2 rows of **Y** correspond to the second element of **NX**, and so on. The scalar **OPT1** can assume values of 0 or 1 only; a 0 is specified when the program user does not want to apply robust estimation in conjunction with the ADF solution, while a 1 is specified for robust estimation. The scalar **OPT2** also assumes values of 0 or 1; a 0 is specified when the program user does not want to apply the bootstrapping methodology, while a 1 is used to indicate that bootstrapping should be used. When **OPT1** $= 1$, the program user must also specify a value for **PER**, which represents the proportion of trimming (discussed later in the paper). **PER** can range in value from 0 (no trimming) to a value less than or equal to .5; a common choice might be **PER** $=$ .20, which represents a 20% symmetric trim rule. When **OPT2** $=$ 1, the program user must also specify an integer value for the scalar **NUMSIM**, which represents the number of simulations for the bootstrapping methodology, and for **SEED** which defines the initial argument for the first call to the bootstrap simulation. **SEED** can be any integer up to $2^{31} - 1$. If **SEED** $=$ 0 is specified, the computer's internal clock is used as the argument.

The main module is invoked with a RUN WJGLM statement. The output of the program is determined by the user's choices for **C**, **U**, **OPT1**, and **OPT2**; further details are provided in later sections of the paper.

A second module, called BOOTCOM, is also included in the program. It computes the ADF solution for a family of contrasts when the program user wishes to use bootstrapping and control the familywise Type I error rate (FWR). This module, which is invoked with a RUN BOOTCOM statement, requires as input, **Y**, **C**, **NX**, **OPT1**, **NUMSIM**, and **ALPHA**. **ALPHA** sets the FWR. For this module, **C** is used to specify a set of contrasts on the independent groups effect(s). Specific examples will help to illustrate the options that are available when using this program.

Applications of the ADF Solution

One-Way Independent Groups Design

A great deal of evidence indicates that the traditional tests for mean equality are adversely affected by nonnormality, particularly when variances are heterogeneous and group sizes are unequal (see Lix & Keselman, 1998; Wilcox, 1995). That is, Type I error and power rates are substantially affected when these assumptions are jointly violated. In particular, depending on whether there is a positive or negative correlation between group sizes and (within-) group variances, the risk of a Type I error can be inflated or deflated relative to the nominal alpha (e.g., $\alpha = .05$) level and correspondingly, the power to detect a treatment effect may be depressed or enhanced.

Reductions in power occur because the usual population standard deviation (σ) is greatly influenced by the presence of extreme observations (outliers) in a distribution of scores. Consequently, the standard error (SE) of the mean, σ^2/n, can become seriously inflated when the underlying distribution has heavy tails (Wilcox, 1995). Thus, standard errors of t and F are relatively large and power accordingly will be depressed.

One can substitute a robust measures of location, and a corresponding robust measure of scale. Trimmed means and variances based on Winsorized sums of squares enable one to obtain test statistics which achieve minimal losses in power due to nonnormality. Indeed, a considerable amount of evidence has accumulated to date supporting this position (see Wilcox, 1995, 1997, 2001).

With regard to spurious rejections, many investigators have shown that better Type I error results can be obtained by using test statistics designed for heterogeneity combined with robust estimators of central tendency and variability (see Lix & Keselman, 1998; Keselman et al., 1998; Wilcox, 1995; Yuen, 1974). Though rates of Type I error improved when adopting robust estimators with heteroscedastic statistics, these improved methods were nonetheless still occasionally affected when distributions were nonnormal, variances were heterogeneous and group sizes were unequal. That is, Type I error rates did occasionally exceed .075 for $\alpha =$ .05, attaining values close to .10.

Westfall and Young's (1993) results suggest that Type I error control could be improved further by combining a bootstrap method with one based on trimmed means. Wilcox et al. (1998) provide empirical support for the use of robust estimators and test statistics with bootstrap-determined critical values in one-way independent groups designs. This benefit has also been demonstrated in correlated groups designs [see Keselman, Algina, Wilcox & Kowalchuk (2000); Keselman, Kowalchuk, Algina, Lix & Wilcox (2000)].

For an independent groups experiment with n_j subjects ($\Sigma_j n_j = N$) in each of J groups, and using the notation of Equation 1, $\mathbf{Y} = (Y_{ij})$, where Y_{ij} is the score associated with the ith subject in the jth group ($j = 1,...,J$; $i = 1,...,n_j$), $E(Y_j) = \mu_j$, the jth population mean, $\boldsymbol{\beta}^T = [\mu_1 \ldots \mu_J]$ and $\boldsymbol{\xi} = (\epsilon_{ij})$ defines the random error term. The Y_{ij}s are assumed to be $N(\mu_j, \sigma_j^2)$ variates, with $\widehat{\mu}_j$ and $\widehat{\sigma}_j^2$ respectively representing the jth sample mean and unbiased variance.

To test the general linear hypothesis of Equation 2, $\mathbf{R} = \mathbf{C} = \mathbf{C}_j$, because $\mathbf{U} = 1$. That is, $\mathbf{C}_j$ is a $(J - 1) \times J$ matrix for which the rows represent a set of linearly independent contrasts among the levels of the independent groups factor. With respect to Equation 3, $\widehat{\boldsymbol{\mu}} = [\widehat{\mu}_1 \ldots \widehat{\mu}_J]^T$ and $\widehat{\boldsymbol{\Sigma}} = \text{diag}[\widehat{\sigma}_1^2/n_1 \ldots \widehat{\sigma}_J^2/n_J]$.

Pairwise contrasts on the group means are frequently of great interest. Using Equation 2, $\mathbf{R} = \mathbf{C} = \mathbf{c}_{jj'} = (c_1 \ldots c_J)$, the $1 \times J$ vector of coefficients which contrasts the jth and j′th means ($\Sigma_j c_j = 0$). In other words, we test the null hypothesis $H_{jj'}$: $\mu_j = \mu_{j'}$ ($j \neq j'$).

Robust Estimation. In this paper we apply robust estimates of central tendency and variability to the ADF statistic. When researchers feel that they are dealing with populations that are nonnormal in form [Tukey (1960) suggested that outliers are a common occurrence in distributions and others have indicated that skewed distributions frequently depict psychological (reaction time) data] 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 procedures, based on robust estimators, should be adopted. When trimmed means are being compared the null hypothesis pertains to the equality of population trimmed means, i.e., the μ_ts. That is, to test the general linear hypothesis in a one-way independent groups design we specify H_0: $\mathbf{R}\boldsymbol{\mu}_t = \mathbf{0}$.

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

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

Wilcox (1995) suggests that 20% trimming should be used.

The sample Winsorized mean is necessary and is computed as

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

where

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

Thus, one can see that the Winsorized mean is obtained by “recoding” the bottom 20% of the distribution to the 20th percentile and the top 20% of the distribution to the 80th percentile and then computing the mean. The sample Winsorized variance, which is required to get a theoretically valid estimate of the standard error of a trimmed mean, is then given by

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

The SE of the trimmed mean equals $\sqrt{(n_j - 1)\, \widehat{\sigma}^2_{Wj}/[(h_j(h_j - 1)]}$.

Thus, with robust estimation, the trimmed group means ($\widehat{\mu}_{tj}$s) replace the least squares group means ($\widehat{\mu}_j$s), the Winsorized group variances estimators ($\widehat{\sigma}^2_{Wj}$s) replace the least squares variances ($\widehat{\sigma}^2_j$s), and h_j replaces n_j and accordingly one computes the robust version of T_{WJ}, defined as T_{WJt} (see Yuen, 1974).[2]

Bootstrapping. Following Westfall and Young (1993) and as enumerated by Wilcox (1997), let $C_{ij} = Y_{ij} - \widehat{\mu}_{tj}$; thus, the C_{ij} values are the empirical distribution of the jth group, centered so that the observed trimmed mean is zero. That is, the empirical distributions are shifted so that the null hypothesis of equal trimmed means is true in the sample. The strategy behind the bootstrap is to use the shifted empirical distributions to estimate an appropriate critical value. For each j, obtain a bootstrap sample by randomly sampling with replacement n_j observations from the C_{ij} values, yielding $Y^*_1, \ldots, Y^*_{n_j}$. Let F^*_t be the value of a test statistic $[T_{WJt}/c]$ based on the bootstrap sample. The B values of F^*_t, where B represents the number of bootstrap simulations, are put in ascending order, that is, $F^*_{t(1)} \leq \cdots \leq F^*_{t(B)}$, and an estimate of an appropriate critical value is $F^*_{t(a)}$, where $a = (1 - \alpha)B$, rounded to the nearest integer. One will reject H_0: $\mathbf{R}\boldsymbol{\mu}_t = \mathbf{0}$ when $F_t \geq F^*_{t(a)}$, where F_t is the value of the heteroscedastic statistic based on the original nonbootstrapped data. According to Wilcox (1997), B can be set at 599.[3]

Focused contrast tests such as pairwise contrasts are computed by replacing robust estimators for least squares values in the T_{WJ}/c statistic. To control the FWR for a set of contrasts, the following approach is used. Let T^*_{WJt}/c be the value of the statistic based on the bootstrap sample. Set $t^* = \max T^*_{WJt}/c$, the maximum being taken over all $j \neq j'$. Repeat this process B times yielding $t^*_1, \ldots, t^*_B$. Let $t^*_{(1)} \leq \cdots \leq t^*_{(B)}$ be the t^*_b values written in ascending order, and let $q = (1 - \alpha)B$, rounded to the nearest integer. Then a test of a null hypothesis is obtained by comparing T_{WJt}/c to $t^*_{(q)}$ (i.e., whether $T_{WJt}/c \geq t^*_{(q)}$), where q is determined so that the FWR is approximately α.

Numerical Example. To illustrate the procedures just presented for a one-way independent groups design we use the data from a study reported by Karayanidis, Andrews, Ward and Michie (1995).[4] In particular, these authors "compared the performance of normal subjects in three age groups and of medicated Parkinson's disease patients on auditory selective attention processes." (p. 335) The dependent score was reaction time (ms) on an auditory target task. We compared three of their groups: young, middle and old.

The following statements are required to establish the data set to be used in the analysis:

```
        Y = {518.29, 548.42, 524.10, 666.63, 488.84, 676.40, 482.43, 531.18, 504.62, 609.53,
584.68, 609.09, 495.15, 502.69, 484.36, 519.10, 572.10, 524.12, 495.24, 335.59, 353.54,
493.08, 469.01, 338.43, 499.10, 404.27, 494.31, 487.30, 485.85, 886.41, 437.50, 558.95,
538.56, 586.39, 530.23, 629.22, 691.84, 557.24, 528.50, 565.43, 536.03, 594.69, 645.69,
558.61, 519.01, 538.83};
```

```
NX = {19 12 15};
```

In the first line, the commas used to separate the individual data points serve to delineate the rows of **Y**, so that **Y** is a column vector with 46 elements.

The **C** matrix to test the omnibus effect is defined as a set of 2 linearly independent contrasts:

```
C = {1 – 1 0, 1 0  – 1};
```

To produce results for the ADF solution without either trimming or bootstrapping, the following additional statements are used:

```
OPT1 = 0;
OPT2 = 0;
PRINT ‘TEST FOR OVERALL GROUP EFFECT- ADF SOLUTION';
RUN WJGLM;
```

To produce results for the ADF solution with 20% symmetric trimming, the following additional statements are used:

```
OPT1 = 1;
PER = .20;
OPT2 = 0;
PRINT ‘TEST FOR OVERALL GROUP EFFECT – ADF SOLUTION WITH
TRIMMING';
RUN WJGLM;
```

To produce results for the ADF solution with 20% trimming and bootstrapping, the following additional statements are used:

```
OPT1 = 1;
PER = .20;
OPT2 = 1;
NUMSIM = 599;
SEED = 40389;
PRINT ‘TEST FOR OVERALL GROUP EFFECT – ADF SOLUTION WITH
TRIMMING AND BOOTSTRAPPING';
RUN WJGLM;
```

Here, we have specified a particular value for the seed for illustration purposes. The program user is free to select any value, given the constraints noted previously.

Means and SEs for each method of estimation are presented in Table 1.[5] The program output gives $T_{WJ}/c = 3.01$, with $\nu_1 = 2$ and $\nu_2 = 22.82$ and $p = .07$. With trimming the corresponding values are $T_{WJt}/c = 6.60$ with $\nu_1 = 2$ and $\nu_2 = 15.11$ and $p = .01$. When the bootstrap is applied to the trimmed data with this seed, the p-value for the omnibus test is $p = .04$. Hence, the least squares solution is not statistical significant while trimming alone or trimming with the bootstrap produce statistically significant results.

Following rejection of an omnibus hypothesis most researchers would be interested in examining focused comparisons among the means, and in particular, pairwise comparisons. To perform all possible pairwise comparisons when using either the usual ADF solution, or the ADF solution with trimmed means and Winsorized variances, the RUN WJGLM statement is issued $J(J - 1)/2$ times, and **C** is redefined before each invocation. The additional programming lines necessary to produce results for the ADF solution with trimmed means and Winsorized variances are:

```
C = {1 –1 0};
PRINT ‘CONTRAST OF YOUNG AND MIDDLE GROUPS – ADF SOLUTION
WITH TRIMMING';
OPT1 = 1;
PER = .20;
OPT2 = 0;
RUN WJGLM;
C = {1 0 –1};
PRINT ‘CONTRAST OF YOUNG AND OLD GROUPS GROUPS – ADF SOLUTION
WITH TRIMMING';
RUN WJGLM;
C = {0 1 –1};
PRINT ‘CONTRAST OF MIDDLE AND OLD GROUPS GROUPS – ADF
SOLUTION WITH TRIMMING';
RUN WJGLM;
```

For the ADF solution with trimmed means, the program gives the following T_{WJt}/c, ν_2 and p-values for these three comparisons, respectively: 6.68, 11.55, .02 (young-middle); 1.97, 19.72, .18 (young-old); 13.41, 9.31, $< .0001$ (middle-old).

The researcher conducting pairwise comparisons would typically wish to adopt a procedure for controlling the FWR, that is, the probability of committing at least one Type I error when conducting multiple significance tests. Keselman, Cribbie and Holland (in press) present a number of methods that researchers can use. In this case, we apply Fisher's two-stage procedure which provides exact FWR control because there are only three levels of the grouping variable. Since the stage one omnibus test was rejected (based on robust estimators), each of the individual pairwise tests can be assessed using $\alpha = .05$. Accordingly, the first and third comparisons can be declared statistically significant.

To produce pairwise comparisons when the ADF solution with trimming is used in combination with the bootstrap, the BOOTCOM module must be invoked if the researcher wishes to control the FWR for the set of contrasts. This module is invoked only a single time, and the **C** matrix contains the entire set of contrasts for which FWR control is required. The following programming statements are used:

```
C = {1 –1 0, 1 0 –1, 0 1 –1};
PRINT ‘PAIRWISE CONTRASTS – ADF SOLUTION WITH
TRIMMING/BOOTSTRAPPING';
OPT1 = 1;
PER = .20;
OPT2 = 1;
NUMSIM = 599;
SEED = 781;
ALPHA = .05;
RUN BOOTCOM;
```

The program gives as output the same T_{WJt}/c and ν_2 results as for the ADF solution with trimming, but specifies a critical value from the bootstrap data. For this example, with the seed specified above, the critical value produced was 12.56, which means that only the third comparison is significant.

Factorial Independent Groups Design

Application of the general ADF solution for hypothesis testing in factorial independent groups designs will only be discussed from the perspective of a two-way design. However, the same concepts may be readily extended to higher-order designs.

Let $\mathbf{Y} = (Y_{ijk})$, where Y_{ijk} represents the score associated with the ith subject in the (j,k)th treatment combination cell ($j = 1, ..., J$; $k = 1, ..., K$; $i = 1, ..., n_{jk}$; $\Sigma_j \Sigma_k n_{jk} = N$). Then $E(Y_{ijk}) = \mu_{jk}$ is the (j,k)th population mean, $\beta^T = [\mu_{11} \ \ \mu_{12} \ldots \mu_{JK}]$, and $\boldsymbol{\xi} = (\epsilon_{ijk})$ defines the random error. The Y_{ijk}s are assumed to be $N(\mu_{jk}, \sigma^2_{jk})$ variates, with $\widehat{\mu}_{jk}$ and $\widehat{\sigma}^2_{jk}$ respectively representing the (j,k)th sample mean and unbiased variance estimate.

The sensitivity of the ANOVA F test to violations of its derivational assumptions for tests of main and interaction hypotheses in factorial designs has been studied in less detail than for one-way designs (Harwell, Rubenstein, Hayes & Olds, 1992). Nevertheless, the evidence available supports the conclusion that the test may become seriously biased when equality of the σ^2_{jk}s is not a tenable assumption, particularly when the $\mathrm{n_{jk}}$s are unequal, that is, for nonorthogonal designs, when hypotheses involving unweighted means are tested (Milligan, Wong & Thompson, 1987). The deleterious effects of nonnormality are described by Wilcox (1997). Keselman, Carriere and Lix (1995, 1996) identified that an ADF solution is largely robust in such situations.

To test the general linear hypothesis of Equation 2, $\mathbf{R} = \mathbf{C} = \mathbf{C}_{\mathrm{JK}}$, $\mathbf{C}_{\mathrm{J}}$, and $\mathbf{C}_{\mathrm{K}}$, respectively for tests of the interaction, row, and column hypotheses, since $\mathbf{U} = 1$ in all cases. Here, $\mathbf{C}_{\mathrm{JK}} = \mathbf{C}_{\mathrm{j}} \otimes \mathbf{C}_{\mathrm{k}}$, where $\mathbf{C}_{\mathrm{j}}$ and $\mathbf{C}_{\mathrm{k}}$ are matrices of order $(\mathrm{J} - 1) \times \mathrm{J}$ and $(\mathrm{K} - 1) \times \mathrm{K}$ respectively, for which the rows represent sets of linearly independent contrasts among the levels of the independent groups factors. Thus, $\mathbf{C}$ is a contrast matrix of order $(\mathrm{J} - 1)(\mathrm{K} - 1) \times \mathrm{JK}$. For the main effect tests, $\mathbf{C}_{\mathrm{J}} = \mathbf{C}_{\mathrm{j}} \otimes \mathbf{1}^{\mathrm{T}}_{\mathrm{K}}$ and $\mathbf{C}_{\mathrm{K}} = \mathbf{1}^{\mathrm{T}}_{\mathrm{J}} \otimes \mathbf{C}_{\mathrm{k}}$, where $\mathbf{1}_{\mathrm{K}}$ and $\mathbf{1}_{\mathrm{J}}$ are column vectors of ones, of order K and J respectively, which serve to sum the means over the appropriate factor. Consequently, $\mathbf{C}_{\mathrm{J}}$, a matrix of order $(\mathrm{J} - 1) \times \mathrm{JK}$, has $(\mathrm{J} - 1)$ contrast rows which sum across the levels of factor K, and $\mathbf{C}_{\mathrm{K}}$, a matrix of order $(\mathrm{K} - 1) \times \mathrm{JK}$ has $(\mathrm{K} - 1)$ contrast rows which sum across the levels of factor J.[6,7,8]

For pairwise comparisons on the row marginal means using the general ADF solution, $\mathbf{R} = \mathbf{C} = \mathbf{c}_{\mathrm{jj'}} \otimes \mathbf{1}^{\mathrm{T}}_{\mathrm{K}}$, a $1 \times \mathrm{JK}$ vector, where $\mathbf{c}_{\mathrm{jj'}}$ contains the coefficients which contrast the jth and j′th row means. Similarly, when $\mathbf{R} = \mathbf{C} = \mathbf{1}^{\mathrm{T}}_{\mathrm{J}} \otimes \mathbf{c}_{\mathrm{kk'}}$, also a $1 \times \mathrm{JK}$ vector, where $\mathbf{c}_{\mathrm{kk'}}$ contains the coefficients which contrast the kth and k′th column means, a pairwise contrast on the column marginal means is formed.

A significant interaction effect could be probed using a variety of procedures, including tetrad contrasts. These contrasts are used to test for the presence of an interaction in a 2 × 2 submatrix of the J × K data matrix. Tetrad contrasts are defined as $\mathbf{R} = \mathbf{c}_{\mathrm{jj'}} \otimes \mathbf{c}_{\mathrm{kk'}}$. For such contrasts, $\mathbf{R}$ is of order $1 \times \mathrm{JK}$.

Robust Estimation. With robust estimation, the trimmed cell means and Winsorized cell variances are substituted for their least squares counterparts into the ADF statistic. The definitions of trimmed means and Winsorized variances can be found in Keselman et al. (1995, 1996). In addition, ν_2s are based on the effective sample sizes.

The results reported by Keselman et al. (1995, 1996) and Keselman et al. (1998) indicate that for moderate degrees of skewness (e.g., χ^2_3-type data) and variance heterogeneity (σ^2_{jk} ratio of 1:1:1:9), the T_{WJ}/c test with the usual least squares estimators for central tendency and variability typically is robust in nonorthogonal designs. However, for more disparate assumption violations, the ADF test using trimmed means and Winsorized variances provides better Type I error control. Thus, the T_{WJt}/c statistic 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.

Bootstrapping. Bootstrap methods can be generalized to factorial designs from the one-way methodology. That is, empirical sampling distributions can be created for each effect by using the resampled bootstrapped data. At this time it is uncertain whether researchers have much to gain by determining statistical significance through bootstrap methods. That is, the results presented by Keselman et al. (1998) indicate that robust and powerful tests can be obtained by using trimmed means and Winsorized variances in nonorthogonal heterogeneous designs when data are nonnormal. And, though their findings are limited to 2 × 2 designs, we see no reason why they would not generalize to higher-order designs.

Nonetheless, if researchers feel bootstrapping methods are necessary for their factorial designs (based on the specific conditions of nonnormality and variance heterogeneity present in their data), the methodology we presented for one-way independent groups designs is applicable to higher-order factorial designs when estimating an appropriate critical value. The crucial issue is to make certain that the values of C_{ijk} are constructed such that the empirical distribution is shifted so that any null hypothesis one might want to examine is true.[9]

For each cell of the design, obtain a bootstrap sample by randomly sampling with replacement n_{jk} observations from the C_{ijk} values, yielding $Y^*_1, \ldots, Y^*_{n_{jk}}$. For omnibus effects (J, K, J $\times$ K) let F^*_t be the value of the test statistic [T_{WJt}/c] based on the bootstrap sample. As previously indicated, one will reject the appropriate null hypothesis when $F_t \geq F^*_{t(a)}$, where F_t is the value of the heteroscedastic statistic (for J, K and/or J $\times$ K) based on the original nonbootstrapped data.

Marginal mean or interaction contrasts can be obtained in a manner that is analogous to contrast testing in the one-way design. That is, let T^*_{WJt}/c be the value of the statistic based on the bootstrap sample. Set $t^* = \max T^*_{WJt}/c$, the maximum being taken over all contrasts in the set {e.g., J(J − 1)/2 tests for all possible pairwise tests on the J marginal means or [J(J − 1)/2][K(K − 1)/2] tests for all possible tetrad contrasts on the cell means}. Again, after repeating the process B times, T_{WJt} is compared to $t^*_{(q)}$.

Numerical example. To illustrate the use of the program for a factorial design, a data set was generated from summary data presented by Phillips, Jones, Rieger and Snell (1999). According to these authors "Research has indicated that performance on heartbeat counting tasks may be influenced by beliefs about heart rate. Subjects were administered the Schandry heartbeat counting task after viewing fast, slow, or no heart rate feedback." (p. 504) Subjects were presented with two tests [Schandry and Whitehead (Signal-detection type of task)]. The dependent variable was a perception score based on Schandry's original error score.[10] The design contained two independent groups factors; feedback (factor J = 3: No, Fast, Slow) and order (factor K = 2: Order1, Order2). Order refers to whether the Schandry or Whitehead was presented first or second in the order of testing. In order to illustrate how the program is applied to an unbalanced design, unequal numbers of observations were generated for the cells of the design.

The following lines of code are used to specify the data set and cell sizes for this example:

```
        Y = {0.59, 0.60, 0.32, 0.67, 0.61, 1.76, 0.38, 0.63, 0.88, 0.08, 1.75, 0.81, 0.64, 0.67,
0.67, 0.08, 1.13, 0.56, 0.84, 0.74, 0.92, 0.89, 1.17, 0.75, 0.90, 0.66, 0.81, 0.67, 0.92, 1.42,
0.80, 1.09, 0.79, 1.28, 0.76, 0.84, 0.88, 0.79, 0.87, 0.86, 0.54, 0.37, 0.47, 0.63, 0.58, 0.43,
0.39, 0.57, 0.83, 0.74, 1.67, 1.02, 0.80, 0.87, 0.94, 0.72, 0.81, 0.67, 0.70, 0.69};
        NX = {12 8 8 12 8 12};
```

The data are entered for each successive cell of the design, with the data for the order factor being entered within each level of the feedback factor. (e.g., order1 data then order2 data within feedback1, etc., etc.) The means and variances for these data are found in Table 2, for both the least squares and trimmed solutions.

To test the omnibus main and interaction effects, one would specify the following program lines:

```
        CJ = {1 –1 0, 1 0 –1};
        CK = {1 –1};
        C = CJ@CK;
        OPT1 = 0;
        OPT2 = 0;
        PRINT ‘TEST OF ORDER X FEEDBACK INTERACTION EFFECT – ADF SOLUTION';
        RUN WJGLM;
        CK = {1 1};
        C = CJ@CK;
        PRINT ‘TEST OF FEEDBACK MAIN EFFECT – ADF SOLUTION';
        RUN WJGLM;
        CJ = {1 1 1};
        CK = {1 –1};
        C = CJ@CK;
        PRINT ‘TEST OF ORDER MAIN EFFECT – ADF SOLUTION';
        RUN WJGLM;
```

To test these omnibus effects using the ADF solution with 20% trimming, the same program lines would be used, with the following modifications:

```
        OPT1 = 1;
        PER = .20;
        OPT2 = 0;
```

To test these omnibus effects using the ADF solution with trimming and bootstrapping, the program user would instead specify (any starting seed could be selected):

```
        OPT1 = 1;
        PER = .20;
        OPT2 = 1;
        NUMSIM = 599;
        SEED = 651332;
        RUN WJGLM;
```

With this code, the following results are obtained for the interaction effect: $T_{WJ}/c = 4.03$ ($\nu_1 = 2$; $\nu_2 = 31.84$; p = .03); $T_{WJt}/c = 4.38$ ($\nu_1 = 2$; $\nu_2 = 22$; p = .02).

When the bootstrap is used, the p-value for this test is .01. All three procedures produce a significant result.

Given the Feedback × Order effect is statistically significant, psychophysiological researchers would most likely be interested in probing this interaction. As indicated, one can follow-up a significant interaction with tetrad contrasts. The code to produce the entire set of interaction contrasts for two of the solutions are:

```
CK = {1 − 1};
CJ12 = {1 − 1 0};
C = CJ12@CK;
OPT1 = 0;
OPT2 = 0;
PRINT 'INTERACTION CONTRAST WITH J1 AND J2 - ADF SOLUTION';
RUN WJGLM;
CJ13 = {1 0 − 1};
C = CJ13@CK;
PRINT 'INTERACTION CONTRAST WITH J1 AND J3 - ADF SOLUTION';
RUN WJGLM;
CJ23 = {0 1 − 1};
C = CJ23@CK;
PRINT 'INTERACTION CONTRAST WITH J2 AND J3 - ADF SOLUTION';
RUN WJGLM;
CJ = {1 − 1 0, 1 0 − 1, 0 1 − 1};
CK = {1 − 1};
C = CJ@CK;
OPT1 = 1;
PER = .20;
OPT2 = 1;
NUMSIM = 599;
ALPHA = .05;
SEED = 19744;
PRINT 'INTERACTION CONTRASTS - ADF SOLUTION WITH TRIMMING &
BOOTSTRAP';
RUN BOOTCOM;
```

For the three tetrad contrasts [(K1 vs K2 by J1 vs J2), (K1 vs K2 by J1 vs J3), (K1 vs K2 by J2 vs J3)] the values of T_{WJ}/c, ν_2 and p are 0.86, 25.09, .36; 5.38, 25.07, .0286; 5.36, 31.76, .0272 ($\nu_1 = 1$ in all cases). If one would have used the module with trimming, the corresponding are .02, 20.79, .89; 5.12, 18.08, .0363; 6.70, 22.41, .0166. Applying Hochberg's step-up Bonferroni procedure (see Keselman, 1998; Keselman et al., in press), none of the tetrad contrasts based on least squares means are significant, while the third contrast is significant when trimmed means are used. For the ADF solution with trimming and bootstrapping, the critical value is 5.90 (the test statistics and df are the same as for trimming), accordingly, only the third tetrad contrast is again significant.

For the marginal main effect of feedback, $T_{WJ}/c = 6.27$, with $\nu_1 = 2$ and $\nu_2 = 31.84$ ($p = .01$). As well, $T_{WJt}/c = 9.42$ with $\nu_1 = 2$, $\nu_2 = 22$ ($p = .001$). The p-value based on the bootstrap for the trimmed data is $< .0001$. Again, all three approaches produce a significant result. For the marginal main effect of order, $T_{WJ}/c = 3.04$, with $\nu_1 = 1$; $\nu_2 = 33.26$ ($p = .09$). For the trimmed data, $T_{WJt}/c = 8.44$ with $\nu_1 = 1$ and $\nu_2 = 28.57$ ($p = .01$). When the trimmed data are bootstrapped, the p-value is $< .0001$. Hence for the order effect, trimming and trimming with bootstrapping produce significant results, while results based on least squares estimators do not.

Correlated Groups Design

Keselman et al. (1993) have shown how T_{WJ}/c can be used to test for treatment effects in between- by within-subjects correlated groups designs (see also Keselman, 1998). Furthermore, they have demonstrated through Monte Carlo methods that this statistic is generally robust to nonnormality and covariance heterogeneity in nonspherical unbalanced repeated measures designs.[11]

Even though it has been demonstrated that the ADF procedure is generally robust to the combined effects of nonnormality and covariance heterogeneity, under some conditions of departure from multisample sphericity and multivariate normality, its rate of Type I error has been found to be inflated (see Algina & Keselman, 1997; Keselman et al., 1993). Further improvement in Type I error is possible by applying the procedure with robust estimators, that is, with trimmed means and Winsorized variances and covariances and/or by obtaining critical values through bootstrap methods (see Keselman, Algina, Wilcox & Kowalchuk, 2000; Keselman, Kowalchuk, Algina, Lix & Wilcox, 2000). Furthermore, the sample sizes necessary to achieve robustness with these estimators and/or bootstrapping can be substantially smaller than the sizes required to achieve robustness with the ADF procedure based on least squares estimators. Thus, though we subscribe to the analysis procedures advocated by Keselman (1998) for the analysis of repeated measures effects, procedures based on the usual least squares estimators, his analyses should, when appropriate, be adopted with robust estimators.[12]

Consider the design in which n_j subjects ($\Sigma_j n_j = N$) in each of J groups are measured on a single dependent variable at K points in time, or under each of K treatments. Using the notation of Equation 1, the observations $\mathbf{Y} = (\mathbf{Y}_{ij})$, where $\mathbf{Y}_{ij} = [Y_{ij1} \;\; ... \;\; Y_{ijK}]$ $(j = 1 \;\; ,...,\;\; J; \;\; i = 1 \;\; ,...,\;\; n_j)$. The $\beta^T = (\boldsymbol{\mu}_j) = (\mu_{j1} \ldots \mu_{jK})$ and $\boldsymbol{\xi} = (\boldsymbol{\epsilon}_{ij}) = (\epsilon_{ij1} \ldots \epsilon_{ijK})$. The $\mathbf{Y}_{ij}$s are assumed to be $N(\boldsymbol{\mu}_j, \, \boldsymbol{\Sigma}_j)$ K-vector variates, with $\widehat{\boldsymbol{\mu}}_j$ and $\widehat{\boldsymbol{\Sigma}}_j$ denoting the jth sample mean vector and variance-covariance matrix, respectively.

To test the general linear hypothesis of Equation 2, both **C** and **U** are defined in terms of the effect to be tested, to create the appropriate **R** contrast matrix. For the within-subjects interaction effect, $\mathbf{R} = \mathbf{C}_j \otimes \mathbf{U}_j^T$, because $\mathbf{C} = \mathbf{C}_j$, where $\mathbf{C}_j$ has the same form and function as for the one-way univariate independent groups design, and $\mathbf{U} = \mathbf{U}_k$, where $\mathbf{U}_k$ is a $K \times (K - 1)$ matrix whose columns form a set of linearly independent contrasts among the levels of the within-subjects factor. Thus, **R** is of order $(J - 1)(K - 1) \times JK$. For tests of the within-subjects main effect, $\mathbf{C} = \mathbf{1}_J^T$ and $\mathbf{U} = \mathbf{U}_k$; for the independent groups main effect, $\mathbf{C} = \mathbf{C}_j$ and $\mathbf{U} = \mathbf{1}_K$. Consequently, these **R** matrices are of order $(K - 1) \times JK$ and $(J - 1) \times JK$, respectively.

As we have indicated, significant interaction effects in correlated groups designs can be probed using a variety of procedures, including tetrad contrasts. FWR control can be obtained with a procedure described in Lix and Keselman (1996). Main effects may be probed using pairwise comparisons of the marginal means. To test within-subjects pairwise comparison hypotheses using an ADF approach, $\mathbf{R} = \mathbf{1}_J^T \otimes \mathbf{u}_{kk'}^T$, of order $1 \times JK$, where $\mathbf{C} = \mathbf{1}_J^T$ and $\mathbf{U} = \; \mathbf{u}_{kk'}$.

Robust Estimation. Keselman, Kowalchuk, Algina, Lix and Wilcox (2000) indicate how one Winsorizes the observations in order to compute the Winsorized covariance matrices and, as well, indicate how to compute trimmed means in this J × K design. Robust estimators can then be applied to T_{WJ}/c.

Bootstrapping. Keselman, Algina, Wilcox, and Kowalchuk (2000) and Keselman, Kowalchuk, Algina, Lix and Wilcox (2000) found that in the context of J × K repeated measures designs, bootstrapping did not result in better control of Type I errors than test statistics that just adopted trimmed means and Winsorized variances and covariances. However, their findings are applicable to a limited number of designs, and therefore may not generalize to other repeated measures designs. Accordingly, we present bootstrap methodology for those who believe it could be beneficial for the designs they utilize.

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 $C_{ijk} = Y^*_{ijk} - \widehat{\mu}_{tjk}$. That is, shift the bootstrap samples. Next compute T^*_{WJt}/c, the value of the statistic which is based on the C_{ijk} values. Repeat this process B times yielding T^*_b, $b = 1\ , \ldots\ , B$. Once again, effects are significant if $T_{WJt}/c \geq T^*_{(q)}$. Keselman, Algina, Wilcox, and Kowalchuk (2000) recommend that B be set at 599. Focused hypothesis tests using contrasts are accomplished in the same manner previously enumerated.

Numerical Example. A data set was generated from the summary measures provided by Jonkman, Kemner, Verbaten, Van Engeland, Kenemans, Camfferman, Buitelaar and Koelega (1999). In their study "Children with attention-deficit hyperactivity disorder (ADHD) and normal (control) children were compared with respect to stimulus- and response-related processes. Performance and electrophysiological measures such as P2, N2, and P3 components of event-related potential and electromyogram (EMG) activity were measured during an Eriksen flanker task." (p. 419) Reaction times (ms) to a target alone, or an arrow stimuli incongruent, congruent or neutral to the target were obtained for each subject. To create an unbalanced design group sizes of 20 and ten were created.

The following statements are required to define the data set to be used in the analysis:

```
        Y = {568.52  433.80  658.51  711.33,  1034.82  864.79  639.42  815.18,  817.92  680.11
499.49  1364.28,  1729.87  1707.13  1272.20  1110.98,  410.26  485.44  367.90  329.94,  514.95
669.29 430.10 438.18, 294.32 1452.33 266.79 754.27,545.39 749.46 1047.34 830.82,390.28
1483.94  217.67  1393.37,  376.48  547.35  441.32  1390.05,  397.44  2206.82  693.42  1178.00,
297.75 423.55 333.06 536.85, 892.58 871.68 639.42 617.32, 341.38 288.38 617.80 1662.21,
477.07 703.05 569.79 788.49, 706.26 610.19 481.23 589.75, 385.93 479.29 1163.69 1166.98,
496.68 492.88 545.05 664.45, 346.00 782.20 232.52 523.59, 386.75   486.97 340.63 359.87,
538.60  480.81  475.36  693.31,  745.53  637.51  994.77  927.96,  477.63  391.84  483.56  699.92,
679.19  2009.67  483.18  815.24,  514.84  402.62  835.36  680.97,452.15  358.31  417.43  342.63,
587.09 461.73 670.07 1026.69, 514.69 430.82 953.16 1328.55, 457.65 691.47 417.86 755.62,
707.15 872.39 645.83 677.84};
        NX = {20 10};
```

Note that a comma is used to separate scores for different subjects, so that **Y** is a 30 × 4 matrix.

The next set of programming lines are used to test the omnibus main and interaction effects using the ADF solution:

```
        C = {1 – 1};
        U = {1 – 1 0 0, 1 0 – 1 0, 1 0 0 – 1}`;
        OPT1 = 0;
        OPT2 = 0;
        PRINT 'TEST FOR INTERACTION EFFECT - ADF SOLUTION';
        RUN WJGLM;
        C = {1 1};
        PRINT 'TEST FOR STIMULUS MAIN EFFECT - ADF SOLUTION';
        RUN WJGLM;
        C = {1 – 1};
        U={1 1 1 1}`;
        PRINT 'TEST FOR GROUP MAIN EFFECT - ADF SOLUTION';
        RUN WJGLM;
```

Because the same **U** matrix is used to test both the interaction effect and the within-subjects main effect, it need not be respecified before the second invocation of the program.

To test these omnibus effects using the ADF solution with trimming, the same program lines would be used, with the following modifications:

```
OPT1 = 1;
PER = .20;
OPT2 = 0;
```

To test these omnibus effects using the ADF solution with trimming and bootstrapping, the program user would instead specify:

```
OPT1 = 1;
PER = .20;
OPT2 = 1;
NUMSIM = 599;
SEED = 61112;
```

Table 3 gives the means (least-squares and trimmed) for each group, at each level of the within-subjects factor, stimulus. For the interaction effect, $T_{WJ}/c = .57$, $\nu_1 = 3$ and $\nu_2 = 21.02$ ($p = .64$); for the within-subjects main effect, $T_{WJ}/c = 5.66$ ($\nu_1 = 3$; $\nu_2 = 21.02$; $p < .01$). The ADF test statistic for the group main effect produces a value $T_{WJ}/c = .22$ ($\nu_1 = 1$; $\nu_2 = 24.84$; $p = .64$) for the group main effect.

When the data are trimmed, the value of $T_{WJt}/c = 2.12$ ($\nu_1 = 3$; $\nu_2 = 11.22$; $p = .15$) for the interaction effect. For the within-subjects main effect the corresponding results are 5.74 with $\nu_1 = 3$, $\nu_2 = 11.22$, and $p < .01$, and for the independent groups main effect, $T_{WJt}/c = .02$ with $\nu_1 = 1$ and $\nu_2 = 13.48$ ($p = .89$). When bootstrapping is applied to the data, the p-value for the interaction effect is .18, for the within-subjects main effect $p = .02$, and for the independent groups main effect, $p = .92$.

Notice that the only effect that is significant is the repeated measures stimulus effect. Effects significant within either of the three ADF solutions can be probed in a similar manner as was demonstrated for the factorial independent groups design, remembering that **U** must be appropriately defined.

For example, consider testing all possible pairs of within-subjects marginal means using the ADF solution. The following SAS/IML program lines are used:

```
C = {1 1};
U = {1 –1 0 0}`;
OPT1 = 0;
OPT2 = 0;
PRINT ‘CONTRAST K1 AND K2 – ADF SOLUTION';
```

```
RUN WJGLM;
U = {1 0 –1 0}`;
PRINT ‘CONTRAST K1 AND K3 – ADF SOLUTION';
RUN WJGLM;
U = {1 0 0 – 1}`;
PRINT ‘CONTRAST K1 AND K4 – ADF SOLUTION';
RUN WJGLM;
U = {0 1 –1 0}`;
PRINT ‘CONTRAST K2 AND K3 – ADF SOLUTION';
RUN WJGLM;
U = {0 1 0 – 1}`;
PRINT ‘CONTRAST K2 AND K4 – ADF SOLUTION';
RUN WJGLM;
U = {0 0 1 –1}`;
PRINT ‘CONTRAST K3 AND K4 – ADF SOLUTION';
RUN WJGLM;
```

The code for the ADF with trimming should be self-explanatory. For the bootstrap, the code would be:

```
C = {1 1};
U = {1 –1 0 0, 1 0 –1 0, 1 0 0 –1, 0 1 – 1 0, 0 1 0  – 1, 0 0 1 – 1}`;
OPT1 = 1;
PER = .20;
OPT2 = 1;
NUMSIM = 599;
SEED = 19216;
ALPHA = .05;
PRINT ‘PAIRWISE COMPARISONS – ADF SOLUTION WITH TRIMMING &
BOOTSTRAP';
RUN BOOTCOM;
```

The critical value would be produced with this invocation of the program; all pairwise comparisons having a test statistic greater than or equal to the critical value would be declared statistically significant.

The results from these analyses are presented in Table 4. We again use Hochberg's step-up Bonferroni procedure to assess statistical significance. The ADF results based on least squares result in two significant pairwise differences: K1 vs K4 and K3 vs K4. With trimming there is one additional significant pairwise difference--K2 vs K4. On the other hand, only K1 vs K4 and K3 vs K4 are significant when bootstrapping is applied with robust estimators.

Summary

When psychophysiological researchers feel that they are dealing with populations that are nonnormal in form 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 procedures based on robust estimators, should be adopted. In our paper we presented an approximate degrees of freedom test statistic for one-way and factorial completely randomized and correlated groups designs based on robust estimators (trimmed means and Winsorized variances and covariances) in order to circumvent the biasing effects of variance heterogeneity and nonnormality. As well, we indicated when testing could be improved by determining statistical significance through a bootstrap method.

We presented this methodology in order to encourage psychophysiological researchers to adopt a procedure that has been shown to be generally robust to variance heterogeneity and nonnormality. That is, the empirical literature has indicated that distortion in rates of Type I error can generally be eliminated by applying robust estimators with heteroscedastic test statistics. Moreover, the power to detect treatment effects is also improved through the use of robust estimators in the presence of nonnormal data.

Within the context of independent groups designs, we indicated that for one-way designs, the use of a bootstrap methodology does indeed result in better Type I error control. For factorial designs, the current literature suggests that the adoption of robust estimators should be sufficient to eliminate the biasing effects of variance heterogeneity and nonnormality, though researchers, if they choose, can apply bootstrap methodology to determine statistical significance, if they feel this would improve the validity of their results.

With respect to the analysis of effects in correlated groups designs, we strongly support the recommendations presented by Keselman (1998). Keselman recommended the use of the ADF statistic in repeated measures designs based on least squares estimators. However, as was pointed out, sample sizes must meet the prescriptions enumerated by Keselman et al. (1993) and Algina and Keselman (1998), in order to obtain a robust test with the ADF solution. When sample sizes do not meet these presciptions, researchers can still obtain a robust test of treatment effects by applying trimmed means and Winsorized variances and covariances with the ADF statistic. Indeed, the results reported by Keselman, Algina, Wilcox and Kowalchuk (2000) indicate that robustness can be achieved with very modest sample sizes (e.g., $n_j = 22$).

The ADF solution can be applied to a wide range of designs by using a GLM framework to define the hypothesis of interest and the computer program demonstrated in this paper. Lix and Keselman (1995) show how to specify multivariate designs and the associated tests of model effects. The ADF solution has been explored in a limited manner in multivariate repeated measures designs (Keselman & Lix, 1997), but not with robust estimators and/or bootstrapping.

References

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., & Olejnik, S. F. (1984). Implementing the Welch-James procedure with factorial designs. Educational and Psychological Measurement, 44, 39-48.

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

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing monte carlo results in methodological research: The one- and two-factor fixed effects ANOVA cases. Journal of Educational Statistics, 17, 315-339.

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.

Jonkman, L. M., Kemner, C., Verbaten, M. N., Van Engeland, H., Kenemans, J. L., Camfferman, G., Buitelaar, J. K., & Koelega, H. S. (1999). Perceptual and response interference in children with attention-deficit hyperactivity disorder, and the effects of methylphenidate. Psychophysiology, 36, 419-429.

Karayanidis, F., Andrews, S., Ward, P. B. & Michie, P. T. (1995). ERP indicies of auditory selective attention in aging and Parkinson's disease. Psychophysiology, 32, 335-350.

Keselman, H. J. (1998). Testing treatment effects in repeated measures designs: An update for psychophysiological researchers. Psychophysiology, 35, 470-478.

Keselman, H. J., Algina, J., Wilcox, R. R., & Kowalchuk, R. K. (2000). Testing repeated measures hypotheses when covariance matrices are heterogeneous: Revisiting the robustness of the Welch-James test again. Educational and Psychological Measurement, 60, 925-938.

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., Cribbie, R. A., Holland, B. (in press). Pairwise multiple comparison test procedures. In B. Thompson (Ed.). Advances in social science methodology, Vol. 6, JAI Press.

Keselman, H. J., Kowalchuk, R. K., Algina, J., Lix, L.M., & Wilcox, R. R. (2000). Testing treatment effects in repeated measures designs: Trimmed means and bootstrapping. British Journal of Mathematical and Statistical Psychology, 53, 175 -191.

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, H. J., & Lix, L. M. (1995). Improved repeated measures stepwise multiple comparison procedures. Journal of Educational and Behavioral Statistics, 20, 83-99.

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

Keselman, H. J., & Rogan, J. C. (1980). Repeated measures F tests and psychophysiological research: Controlling the number of false positives. Psychophysiology, 17, 499-503.

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

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. (1996). Interaction contrasts in repeated measures designs. British Journal of Mathematical and Statistical Psychology, 49, 147-162.

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

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

Phillips, G. C., Jones, G. E., Rieger, E. J., & Snell, J. B. (1999). Effects of the presentation of false heart-rate feedback on the performance of two common heartbeat-detection tasks. Psychophysiology, 36, 504-510.

SAS Institute. (1999). SAS/IML: User's guide, Version 8, Cary, NC: Author.

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

Wasserman, S. & Bockenholt, U. (1989). Bootstrapping: Applications to Psychophysiology. Psychophysiology, 26, 208-221.

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

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

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

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

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

Wilcox, R. R. (2001). Fundamentals of modern statistical methods: Substantially improving power and accuracy. New York: Springer.

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. (2000). Repeated measures ANOVA: Some new results on comparing trimmed means and means. British Journal of Mathematical and Statistical Psychology, 53, 69-82.

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

Footnotes

1. A complete understanding of this presentation is not necessary for applied researchers to use the computer program.

2. The strategy of trimming and then applying standard methods for means to the remaining data results in the wrong standard error (see Wilcox, 1997).

3. Results from Hall (1986) suggest that it may be advantageous to choose B such that $1 - \alpha$ is a multiple of $(B + 1)^{-1}$.

4. We obtained data for all of our numerical examples by taking summary statistics from articles published in *Psychophysiology* and then generating skewed data from distributions having the means and variances reported in these articles. Though the data that we generated were obtained from actual psychophysiological investigations, they do not always demonstrate the efficacy of our recommendations. Articles where summary values were presented were limited in number. The literature is very clear, however, regarding the benefits of utilizing robust estimators and applying a bootstrap methodology.

5. The program provides the values of the $\widehat{\boldsymbol{\Sigma}}$ matrices, which are the squared standard errors.

6. See Lix and Keselman (1995) footnote #2 for an example of **C** for a 3×4 design.

7. For higher-order designs, Algina and Olejnik (1984) have developed a set of general rules which can be used to form **C**.

8. In nonorthogonal designs, the researcher may test main effect hypotheses involving either weighted or unweighted means, depending on the values assigned to the elements of the **C** [see Keselman et al. (1995, 1996)]. For the sake of simplicity, we have assumed that the researcher is interested in testing hypotheses of unweighted means.

9. Recall that in the one-way design we let $C_{ij} = Y_{ij} - \widehat{\mu}_{tj}$. This is done to estimate an appropriate critical value. That is, the C_{ij} values were the empirical distribution of the jth group, centered so that the sample trimmed mean was zero. Thus, when testing effects

(main, interaction, contrast) in a two-way design $C_{ijk} = Y_{ijk} - \hat{\mu}_{tjk}$. When the data are centered in this manner (i.e., all cell trimmed means are being equated) any effect null hypothesis would be true.

10. According to the authors, ”The error score is determined by subtracting the total number of counted heartbeats from the actual number and dividing the absolute value of the difference by the actual number of heartbeats.” (p. 506) In their study, ”the error score was subtracted from 1.0 to produce the perception score. Good perception is then associated with a score approaching 1.0, and poor perception with a low score (e.g., 0).” (p. 506)

11. Wilcox et al. (2000) found that applying robust estimators with a multivariate statistic did not result in good Type I error control under conditions of nonnormality.

12. Neither single factor nor multifactor within-subjects designs are considered since covariance heterogeneity is not an issue when the design does not contain an independet groups variable. However, users should always use a procedure and critical value which can either contend with violations of the sphericity assumption, such as an adjusted-df test (see Keselman & Rogan, 1980), or bypass it altogether, such as a multivariate test. As well, users should be cognizant of the normality assumption.

Appendix

```
****INVOKE THE IML PROGRAM AND DEFINE THE MODULE WJGLM****;
PROC IML;
RESET NONAME;

****DEFINE MODULES TO PERFORM ALL CALCULATIONS****;
**DEFINE MODULE FOR INITIAL SPECIFICATIONS****;
START INITIAL(C,U,Y,R,X) GLOBAL(NX,NTOT,BOBS,WOBS,WOBS1);
IF NROW(U)=0 THEN U=I(NCOL(Y));
IF NCOL(U)>NROW(U) THEN PRINT
  'ERROR: NUMBER OF COLUMNS OF U EXCEEDS NUMBER OF ROWS';
DO I=1 TO NCOL(NX);
  X1=J(NX[I],1,I);
  IF I=1 THEN X=X1;
  ELSE X=X//X1;
END;
X=DESIGN(X);
NTOT=NROW(Y);
WOBS=NCOL(Y);
BOBS=NCOL(X);
WOBS1=WOBS-1;
R=C@U`;
FINISH;

****DEFINE MODULE TO COMPUTE LEAST SQUARES OR TRIMMED MEANS****;
START MNMOD (Y,OPT1,BHAT,BHATW,MUHAT,YT,DF)
GLOBAL(BOBS,WOBS,WOBS1,NTOT,NX,PER,X);
IF OPT1=0 THEN DO;
  BHAT=INV(X`*X)*X`*Y;
  BHATW=BHAT;
  YT=Y;
  DF=NX-1;
END;
IF OPT1=1 THEN DO;
  BHAT=J(BOBS,WOBS,0);
  BHATW=BHAT;
  YT=J(NTOT,WOBS,0);
  DF=J(1,BOBS,0);
  F=1;
  M=0;
  DO J=1 TO NCOL(NX);
    SAMP=NX[J];
    L=M+SAMP;
    G=INT(PER#SAMP);
    DF[J]=SAMP-2#G-1;
    DO K=1 TO NCOL(Y);
    TEMP=Y[F:L,K];
    NV=TEMP;
    TEMP[RANK(NV),]=NV;
    TRIMY=TEMP[G+1:SAMP-G,];
    TRIMMN=SUM(TRIMY)/(DF[J]+1);
    BHAT[J,K]=TRIMMN;
    MINT=MIN(TRIMY);
    MAXT=MAX(TRIMY);
    DO P=1 TO NROW(NV);
          IF NV[P]<=MINT THEN NV[P]=MINT;
          IF NV[P]>=MAXT THEN NV[P]=MAXT;
        END;
    YT[F:L,K]=NV;
    WINMN=SUM(NV)/SAMP;
    BHATW[J,K]=WINMN;
   END;
   M=L;
   F=F+NX[J];
  END;
END;
MUHAT=SHAPE(BHAT,BOBS#WOBS);
FINISH;

***DEFINE MODULE TO COMPUTE SIGMA MATRIX****;
START SIGMOD (YT,BHATW,DF,SIGMA)
GLOBAL(BOBS,WOBS,WOBS1,NTOT,NX,PER,X);
SIGMA=J(WOBS#BOBS,WOBS#BOBS,0);
DO I=1 TO BOBS;
  SIGB=(YT#X[,I]-X[,I]*BHATW[I,])`*(YT#X[,I]-X[,I]*BHATW[I,])/((DF[I]+1)#DF[I]);
  F=I#WOBS-WOBS1;
  L=I#WOBS;
  SIGMA[F:L,F:L]=SIGB;
END;
FINISH;

****DEFINE MODULE TO COMPUTE TEST STATISTIC****;
START TESTMOD(SIGMA,MUHAT,R,DF,FSTAT,DF1,DF2)
GLOBAL(BOBS,WOBS,WOBS1,NTOT,NX,PER,X);
T=(R*MUHAT)`*INV(R*SIGMA*R`)*(R*MUHAT);
A=0;
IMAT=I(WOBS);
DO I=1 TO BOBS;
  QMAT=J(BOBS#WOBS,BOBS#WOBS,0);
  F=I#WOBS-WOBS1;
  L=I#WOBS;
  QMAT[F:L,F:L]=IMAT;
  PROD=(SIGMA*R`)*INV(R*SIGMA*R`)*R*QMAT;
  A=A+(TRACE(PROD*PROD)+TRACE(PROD)**2)/DF[I];
END;
A=A/2;
```

```
DF1=NROW(R);
DF2=DF1#(DF1+2)/(3#A);
CVAL=DF1+2#A-6#A/(DF1+2);
FSTAT=T/CVAL;
FINISH;

****DEFINE MODULES TO PERFORM BOOTSTRAP****;
***DEFINE MODULE TO GENERATE BOOTSTRAP DATA AND CENTRE DATA****;
START BOOTDAT (Y,BHAT,YB)
GLOBAL(BOBS,WOBS,WOBS1,NTOT,NX,PER,X,SEED);
  F=1;
  M=0;
  DO J=1 TO BOBS;
   L=M+NX[J];
   TEMP=Y[F:L,];
   BVAL=TEMP;
   DO P=1 TO NROW(TEMP);
     RVAL=UNIFORM(SEED);
         BVAL[P,]=TEMP[CEIL(NROW(TEMP)#RVAL),];
   END;
   IF J=1 THEN YB=BVAL;
   ELSE YB=YB//BVAL;
   M=L;
   F=F+NX[J];
 END;
****CENTRE THE BOOTSTRAP DATA****;
M=0;
F=1;
DO I=1 TO BOBS;
  L=M+NX[I];
  MVAL=BHAT[I,];
  DO Q=F TO L BY 1;
    YB[Q,]=YB[Q,]-MVAL;
  END;
  M=L;
  F=F+NX[I];
END;
FINISH;

****DEFINE MODULE TO COMPUTE BOOTSTRAP STATISTIC****;
START BOOTSTAT(YB,OPT1,R,FSTATB)
GLOBAL(BOBS,WOBS,WOBS1,NTOT,NX,PER,X,SEED);
  CALL MNMOD(YB,OPT1,BHATB,BHATBW,MUHATB,YTB,DFB);
  CALL SIGMOD(YTB,BHATBW,DFB,SIGMAB);
  CALL TESTMOD(SIGMAB,MUHATB,R,DFB,FSTATB,DF1B,DF2B);
FINISH;

****COMPUTE WELCH-JAMES STATISTIC****;
START WJGLM;
```

```
CALL INITIAL(C,U,Y,R,X);
CALL MNMOD(Y,OPT1,BHAT,BHATW,MUHAT,YT,DF);
CALL SIGMOD(YT,BHATW,DF,SIGMA);
CALL TESTMOD(SIGMA,MUHAT,R,DF,FSTAT,DF1,DF2);
IF OPT2=1 THEN DO;
   DO SIMLOOP=1 TO NUMSIM;
      CALL BOOTDAT(Y,BHAT,YB);
      CALL BOOTSTAT(YB,OPT1,R,FSTATB);
         IF SIMLOOP=1 THEN FMAT=FSTATB;
       ELSE FMAT=FMAT//FSTATB;
   END;
TEMPB=FMAT;
FMAT[RANK(FMAT)]=TEMPB;
END;
***CALCULATE SIGNIFICANCE LEVEL****;
RESULTS=J(4,1,0);
RESULTS[1]=FSTAT;
RESULTS[2]=DF1;
RESULTS[3]=DF2;
IF OPT2=0 THEN RESULTS[4]=1-PROBF(RESULTS[1],DF1,DF2);
IF OPT2=1 THEN DO;
  AVEC=(FSTAT<=FMAT);
  PVAL=SUM(AVEC)/NUMSIM;
  RESULTS[4]=PVAL;
END;
****PRINT WELCH-JAMES RESULTS****;
PRINT'WELCH-JAMES APPROXIMATE DF SOLUTION';
IF OPT1=0 THEN PRINT'LEAST SQUARES MEANS & VARIANCES';
IF OPT1=1 THEN PRINT'TRIMMED MEANS & WINSORIZED VARIANCES';
IF OPT1=1 THEN DO;
  PRINT 'PERCENTAGE OF TRIMMING:';
  PRINT PER[FORMAT=4.2];
END;
IF OPT2=0 THEN PRINT'F DISTRIBUTION CRITICAL VALUE';
IF OPT2=1 THEN PRINT'BOOTSTRAP CRITICAL VALUE FOR SINGLE TEST
STATISTIC';
IF OPT2=1 THEN DO;
  PRINT 'NUMBER OF BOOTSTRAP SAMPLES:';
  PRINT NUMSIM[FORMAT=4.0],;
  PRINT 'STARTING SEED:';
  PRINT SEED[FORMAT=15.0],;
END;
PRINT'CONTRAST MATRIX:';
PRINT R[FORMAT=4.1],;
MUHAT=MUHAT`;
PRINT 'MEAN VECTOR:';
PRINT MUHAT[FORMAT=10.4],;
PRINT 'SIGMA MATRIX:';
PRINT SIGMA[FORMAT=10.4],;
```

```
RESLAB={"TEST STATISTIC" "NUMERATOR DF" "DENOMINATOR DF" "P-VALUE"};
PRINT 'SIGNIFICANCE TEST RESULTS:';
PRINT RESULTS[ROWNAME=RESLAB FORMAT=10.4]/;
FINISH;

****DEFINE MODULE TO COMPUTE BOOTSTRAP RESULTS WITH FWR
CONTROL****;
START BOOTCOM;
CALL INITIAL(C,U,Y,R,X);
CALL MNMOD(Y,OPT1,BHAT,BHATW,MUHAT,YT,DF);
CALL SIGMOD(YT,BHATW,DF,SIGMA);
DO I=1 TO NROW(R);
CM=R[I,];
CALL TESTMOD(SIGMA,MUHAT,CM,DF,FSTAT,DF1,DF2);
IF I=1 THEN DO;
   CMMAT=FSTAT;
   DF1MAT=DF1;
   DF2MAT=DF2;
END;
IF I>1 THEN DO;
    CMMAT=CMMAT||FSTAT;
    DF1MAT=DF1MAT||DF1;
    DF2MAT=DF2MAT||DF2;
END;
END;
DO SIMLOOP=1 TO NUMSIM;
CALL BOOTDAT(Y,BHAT,YB);
DO Q=1 TO NROW(R);
  CM=R[Q,];
  CALL BOOTSTAT(YB,OPT1,CM,FSTATB);
  IF Q=1 THEN FROW=FSTATB;
    ELSE FROW=FROW||FSTATB;
END;
IF SIMLOOP=1 THEN FMAT=FROW;
  ELSE FMAT=FMAT//FROW;
END;
FMAX=J(NUMSIM,1,0);
DO K=1 TO NUMSIM;
 FMAX[K]=MAX(FMAT[K,]);
END;
TEMPB=FMAX;
FMAX[RANK(FMAX)]=TEMPB;
RESULTS=J(3,NROW(R),0);
RESULTS[1,]=CMMAT;
RESULTS[2,]=DF1MAT;
RESULTS[3,]=DF2MAT;
QCRIT=ROUND((1-ALPHA)#NUMSIM);
CRITV=FMAX[QCRIT];
****PRINT RESULTS****;
```

```
PRINT'WELCH-JAMES APPROXIMATE DF SOLUTION';
IF OPT1=0 THEN PRINT'LEAST SQUARES MEANS & VARIANCES';
IF OPT1=1 THEN PRINT'TRIMMED MEANS & WINSORIZED VARIANCES';
IF OPT1=1 THEN DO;
  PRINT 'PERCENTAGE OF TRIMMING:';
  PRINT PER[FORMAT=4.2];
END;
PRINT'BOOTSTRAP CRITICAL VALUE FOR FWR CONTROL';
PRINT 'NUMBER OF BOOTSTRAP SAMPLES:';
PRINT NUMSIM[FORMAT=4.0];
PRINT 'STARTING SEED:';
PRINT SEED[FORMAT=15.0],;
PRINT 'SIGNIFICANCE LEVEL:';
PRINT ALPHA[FORMAT=3.2],;
PRINT'CONTRAST MATRIX:';
PRINT R[FORMAT=4.1],;
MUHAT=MUHAT`;
PRINT 'MEAN VECTOR:';
PRINT MUHAT[FORMAT=10.4],;
PRINT 'SIGMA MATRIX:';
PRINT SIGMA[FORMAT=10.4],;
RESLAB={"TEST STATISTIC" "NUMERATOR DF" "DENOMINATOR DF"
"SIGNIFICANCE"};
PRINT 'SIGNIFICANCE TEST RESULTS:';
PRINT RESULTS[ROWNAME=RESLAB FORMAT=10.4],;
PRINT 'CRITICAL VALUE:';
PRINT CRITV[FORMAT=5.2]/;
FINISH;
```