DOCUMENT RESUME

ED 387 499 TM 023 614

AUTHOR Lix, Lisa M.; And Others
TITLE Analyzing Multivariate Repeated Measures Designs When Covariance Matrices Are Heterogeneous.
PUB DATE Apr 95
NOTE 32p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 18-22, 1995).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Hypothesis Testing; *Multivariate Analysis; Robustness (Statistics); *Sample Size
IDENTIFIERS Covariance Matrices; Heterogeneity of Variance; *Heteroscedasticity (Statistics); *Repeated Measures Design

ABSTRACT

Methods for the analysis of within-subjects effects in multivariate groups by trials repeated measures designs are considered in the presence of heteroscedasticity of the group variance-covariance matrices and multivariate nonnormality. Under a doubly multivariate model approach to hypothesis testing, within-subjects main and interaction effect procedures are largely robust to the effects of heteroscedasticity when group sizes are equal, even when the data are nonnormal. However, these tests are highly sensitive to the effects of covariance heterogeneity when the design is unbalanced. An approximate degrees of freedom multivariate statistic given by Johansen (1980) is shown to be largely robust to the combined effects of these assumption violations for unbalanced designs, provided that the smallest of the group sizes is sufficiently larger than the product of the number of dependent variables times the number of repeated measurements minus one. (Contains 30 references and 7 tables.) (Author)

Analyzing Multivariate Repeated Measures Designs When Covariance Matrices are Heterogeneous

Lisa M. Lix

Y. H. Lin

H. J. Keselman

Department of Psychology
University of Manitoba

Paper presented at the annual meeting of the
American Educational Research Association

San Francisco
April 18 - 22, 1995

Abstract

Methods for the analysis of within-subjects effects in multivariate groups by trials repeated measures designs are considered in the presence of heteroscedasticity of the group variance-covariance matrices and multivariate nonnormality. Under a doubly multivariate model approach to hypothesis testing, within-subjects main and interaction effect procedures are largely robust to the effects of heteroscedasticity when group sizes are equal, even when the data are nonnormal. However, these tests are highly sensitive to the effects of covariance heterogeneity when the design is unbalanced. An approximate degrees of freedom multivariate statistic given by Johansen (1980) is shown to be largely robust to the combined effects of these assumption violations for unbalanced designs, provided that the smallest of the group sizes is sufficiently larger than the product of the number of dependent variables times the number of repeated measurements minus one.

Analyzing Multivariate Repeated Measures Designs When Covariance Matrices are Heterogeneous

In many experimental situations encountered by educational and psychological researchers, individual response data are repeatedly collected on multiple dependent variables over several experimental conditions. In the simplest of these multivariate repeated measures designs containing a single between-subjects grouping factor, n_j study participants in each of J independent groups ($\Sigma_{j=1}^{J} n_j = N$) are measured on each of L dependent variables over K occasions or trials.

Multivariate repeated measures data may be analyzed from either a multivariate mixed model (MMM) or doubly multivariate model (DMM) perspective; each approach rests on its own set of derivational assumptions. Specifically, the former assumes that the multivariate multisample sphericity assumption is satisfied (Boik, 1988, 1991; Thomas, 1983), and that the observations are independently distributed as multivariate normal variables. For multivariate multisample sphericity to exist, a set of orthonormalized contrast variables on the repeated measurements must exhibit a constant variance across the dependent variables and the covariance matrices of these orthonormalized variables are assumed to be homogeneous across the levels of the between-subjects grouping factor. Under a doubly multivariate approach, no restrictions are placed on the structure of the common covariance matrix, that is, the data need not be multivariate spherical. However, the assumptions of homogeneity of the group covariance matrices, multivariate normality, and independence of observations must be satisfied.

Tests of within-subjects effects are known to be sensitive to departures from the multivariate sphericity assumption under a MMM analysis (Robey & Barcikowski, 1986). However, Boik (1991) has shown that when the data are not multivariate spherical, an adjusted degrees of freedom (df) MMM test can control the Type I error rate for both multivariate within-subjects main and interaction effect tests in repeated measures designs containing a grouping factor (i.e., groups by trials designs). At the same time, Boik observed very few instances in which an adjusted-df MMM analysis was preferable to a DMM analysis, and found that the DMM analysis was almost always more powerful. Furthermore, a DMM procedure can be used in almost all data-analytic situations, the exception being when sample sizes are extremely small.

However, Boik (1991) did not investigate the effects of heteroscedasticity of the group covariance matrices on the Type I error control offered by various multivariate criteria in a DMM analysis. Keselman and Keselman (1990) have shown that in repeated measures designs containing a single dependent variable, multivariate tests of within-subjects main and interaction effects can not provide Type I error control when covariance matrices are heterogeneous and group sizes are unequal. Consequently, tests of within-subjects effects in multivariate repeated measures designs are also likely to be sensitive to violations of this assumption, particularly when the design is unbalanced, an observation made by Thomas (1983).

Keselman, Carriere, and Lix (1993) identified that an approximate df multivariate Welch-James (James, 1951, 1954; Welch, 1938, 1951) statistic given by Johansen (1980) is robust to the effects of covariance heterogeneity in repeated measures designs containing a single dependent variable, even when the data are skewed, provided that sample size is sufficiently large. Keselman et al. observed that the critical factor in determining the robustness of this statistic to the effects of covariance heteroscedasticity and nonnormality in unbalanced designs is the ratio of the smallest group size (i.e., n_{min}) to (K -1). The authors suggest that this ratio should be at least 3 or 4 to 1 for Johansen's procedure to provide effective Type I error control when the data are normally distributed and slightly higher (i.e., 5 to 1) for skewed data.

Tang and Algina (1993) evaluated the performance of a number of procedures which do not depend on the assumption of homogeneity of group covariance matrices in the context of a multivariate independent groups design with more than two groups. The authors considered Johansen's (1980) statistic in addition to multivariate analogs of James' (1951, 1954) first and second order procedures. Johansen's (1980) statistic provided better Type I error control than the other procedures under most situations, but only when the ratio of total sample size (i.e., N) to the number of dependent variables (i.e., L) was at least 15 to 1 when the data were normally distributed. However, the authors did not consider the effects of multivariate nonnormality on the operating characteristics of the approximate df solutions.

Algina, Oshima, and Tang (1991) did investigate the effects of both variance-covariance heteroscedasticity and nonnormality on James' (1951, 1954) and Johansen's (1980) procedures in a multivariate independent groups design, but only for the two-group situation. They found that for symmetric nonnormal distributions, all procedures were able to maintain the Type I error rate close to α. However, for asymmetric distributions, the rate of

Type I errors could become seriously inflated when the data were heteroscedastic and the ratio of N to L was small.

In light of the findings of previous research, the purpose of the present study was two-fold: (1) to examine the operating characteristics of DMM test procedures under departures from the assumptions of homogeneity of the group covariance matrices and multivariate normality and (2) to determine whether the approximate df statistic provided by Johansen (1980) can offer robust tests of within-subjects main and interaction effects in unbalanced multivariate repeated measures designs.

Definition of Test Statistics

The test procedures considered in this investigation of multivariate groups by trials designs were Hotelling's (1931) one-sample T^2 statistic for tests of within-subjects main effects, Hotelling's two-sample statistic for tests of within-subjects interactions in two-group designs, and the Hotelling-Lawley (Hotelling, 1951; Lawley, 1938) trace, Pillai-Bartlett (Bartlett, 1939; Pillai, 1955) trace, and Wilks' (1932) likelihood ratio for tests of interactions in multi-group designs. Roy's (1957) largest root criterion was not considered in this paper since the F approximation to this statistic is not highly accurate (Muller, LaVange, Ramey, & Ramey, 1992). The DMM procedures were compared to the approximate df procedure described by Johansen (1980).

All of these procedures for testing within-subjects effects in groups by trials designs may be described within the context of the general linear model (GLM; See Timm, 1980).

Let

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

where $\mathbf{Y}$ is an N x p matrix of scores, p = KL, K is the number of repeated measurements, L is the number of dependent variables, N is the total sample size, $\mathbf{X}$ is an N x J design matrix with rank($\mathbf{X}$) = J, β is an J x p matrix of nonrandom parameters (i.e., population means), and ξ is an N x p matrix of random error components. Each row of $\mathbf{Y}$ contains the p-dimensional response vector associated with a particular study participant, where the first K columns correspond to the repeated measurements obtained on the first dependent variable, the next K columns correspond to the repeated measurements obtained for the second dependent variable, and so on.

DMM Test Procedures

Under a DMM approach to hypothesis testing, it is assumed that the rows of ξ are independently and identically distributed normal p-vector variates with mean vector $\mathbf{0}$ and variance-covariance matrix Σ_p [i.e., i.i.d. N($\mathbf{0}$, Σ_p)]. To illustrate the DMM approach to the analysis of within-subjects main and interaction effects in a groups by trials designs containing a single between-subjects factor and a single within-subjects factor, let

$$\boldsymbol{\theta} = \mathbf{C}\beta(\mathbf{I}_L \otimes \mathbf{U}) \ , \tag{2}$$

where $\mathbf{C}$, of dimension r x J with rank($\mathbf{C}$) = r, is used to define a set of r contrasts on the between-subjects effect, β is as previously defined and is estimated by

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \ , \tag{3}$$

where superscript T denotes the transpose operator, $\mathbf{I}_L$ is an identity matrix of dimension L, $\otimes$ is the Kronecker product function, and $\mathbf{U}$, of dimension K x q with rank($\mathbf{U}$) = q, is used to define a set of contrasts on the within-subjects effect. Thus, $\boldsymbol{\theta}$ is of dimension r x t, where t = Lq. The statistics that are used to test hypotheses concerning $\boldsymbol{\theta}$ (i.e., H_o: $\boldsymbol{\theta} = \mathbf{0}$) can all be expressed in terms of the matrices $\mathbf{H}$ and $\mathbf{E}$, where

$$\mathbf{H} = \hat{\boldsymbol{\theta}}^T[\mathbf{C}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}]^{-1}\hat{\boldsymbol{\theta}} \ , \tag{4}$$

and

$$\mathbf{E} = (\mathbf{I}_L \otimes \mathbf{U})^T\mathbf{Y}^T[\mathbf{I}_N - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]\mathbf{Y}(\mathbf{I}_L \otimes \mathbf{U}) \ . \tag{5}$$

In Equation 4, $\hat{\boldsymbol{\theta}}$ estimates $\boldsymbol{\theta}$ and in Equation 5, $\mathbf{I}_N$ is an identity matrix of dimension N.

The Hotelling-Lawley (Hotelling, 1951; Lawley, 1938) trace is defined as HT = tr($\mathbf{HE}^{-1}$) where tr denotes the trace operator, the Pillai-Bartlett (Bartlett, 1939, Pillai, 1955) trace is defined as PB = tr($\mathbf{HT}^{-1}$) where $\mathbf{T} = \mathbf{H} + \mathbf{E}$, and Wilks's (1938) likelihood criterion is given by W = det($\mathbf{ET}^{-1}$) where det denotes the determinant of a matrix.

Each of these statistics may be defined in terms of an F variable. For example, for the Pillai-Bartlett trace (Bartlett, 1939; Pillai, 1955), let $\hat{\eta}_{(PB)}$ = PB/s, where s = min(r, t).

Then

$$F_{(PB)} = \frac{\hat{\eta}_{(PB)}/(rt)}{(1 - \eta_{(PB)})/\nu_{(PB)}} , \tag{6}$$

where $\nu_{(PB)} = s[M - t + s]$ and $M = N - J$. The statistic, $F_{(PB)}$, approximately follows an F distribution with $\nu_1 = rt$ and $\nu_{(PB)}$ df. Approximate F statistics for the other multivariate criteria may be found in a number of sources, including Muller et al. (1992). When $s = 1$, all of these statistics are equivalent to Hotelling's (1931) T^2 statistic.

Approximate DF Test Procedure

To define Johansen's (1980) statistic, denote $\mathbf{Y}_j = \mathbf{Y} \cdot (\mathbf{X}_j \mathbf{1}_p^T)$ as a Hadamard product (Searle, 1987, p. 49), where $\mathbf{X}_j$ is the *j*th column of $\mathbf{X}$ ($j = 1, \ldots, J$) and consists entirely of zeros and ones, $\mathbf{1}_p$ is a p x 1 vector of ones, and $\cdot$ is the dot product function, such that $\mathbf{Y}_j$ is an element-by-element product matrix. It is assumed that the observations in $\mathbf{Y}_j$ are independently distributed normal variates with mean vector β_j and variance-covariance matrix Σ_j [i.e., i.d. $N(\beta_j, \Sigma_j)$], where β_j is the *j*th row of β and $\Sigma_j \neq \Sigma_{j'}$ ($j \neq j'$). Let

$$\hat{\Sigma}_j = \frac{(\mathbf{Y}_j - \mathbf{X}_j\hat{\beta}_j)^T(\mathbf{Y}_j - \mathbf{X}_j\hat{\beta}_j)}{n_j - 1} , \tag{7}$$

estimate $\boldsymbol{\Sigma}_j$, where $n_j = \mathbf{X}_j^T\mathbf{X}_j$, and $\hat{\beta}_j$ estimates β_j.

The general linear hypothesis for Johansen's (1980) solution is

$$H_o: \mathbf{R}\boldsymbol{\mu} = \mathbf{0} , \tag{8}$$

where $\mathbf{R} = \mathbf{C} \otimes (\mathbf{I}_L \otimes \mathbf{U})^T$ and $\mathbf{C}$, $\mathbf{I}_L$, and $\mathbf{U}$ are as previously defined. Furthermore,

$\mu = \text{vec}(\beta^T) = [\beta_1 \ldots \beta_J]^T$, where $\beta_j = [\mu_{j1} \ldots \mu_{jp}]$, such that μ is the column vector with Jp elements obtained by stacking the columns of β^T. The $\mathbf{0}$ column vector is of order rt.

The generalized test statistic given by Johansen (1980) is

$$T_{WJ} = (\mathbf{R}\hat{\mu})^T(\mathbf{R}\hat{\Sigma}\mathbf{R}^T)^{-1}(\mathbf{R}\hat{\mu}) \,, \tag{9}$$

where $\hat{\mu}$ estimates μ, and $\hat{\Sigma} = \text{diag}[\hat{\Sigma}_1/n_1 \ldots \hat{\Sigma}_J/n_J]$, a block matrix with diagonal elements $\hat{\Sigma}_j/n_j$. This test statistic divided by a constant, c, approximately follows an F distribution with $\nu_1 = rt$, 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

$$A = \frac{1}{2}\sum_{j=1}^{J}\left[\text{tr}\left\{\hat{\Sigma}\mathbf{R}^T(\mathbf{R}\hat{\Sigma}\mathbf{R}^T)^{-1}\mathbf{R}\mathbf{Q}_j\right\}^2 + \left\{\text{tr}(\hat{\Sigma}\mathbf{R}^T(\mathbf{R}\hat{\Sigma}\mathbf{R}^T)^{-1}\mathbf{R}\mathbf{Q}_j)\right\}^2\right]/(n_j - 1) \,. \tag{10}$$

The matrix $\mathbf{Q}_j$ is a symmetric block matrix of dimension Jp associated with $\mathbf{X}_j$, such that the (g,h)-th diagonal block of $\mathbf{Q}_j = \mathbf{I}_p$ if $g = h = j$ and is $\mathbf{0}$ otherwise.

In order to test the within-subjects main effect in a multivariate groups by trials design, $\mathbf{C} = \mathbf{1}_J^T$ and $\mathbf{U} = \mathbf{U}_k$, so that for Johansen's (1980) solution, $\mathbf{R} = \mathbf{1}_J^T \otimes (\mathbf{I}_L \otimes \mathbf{U}_k)^T$, where $\mathbf{1}_J$ is a J x 1 vector of ones and $\mathbf{U}_k$ is a K x (K - 1) matrix which defines a set of (K - 1) linearly independent contrasts for the within-subjects factor. To test the within-subjects interaction, $\mathbf{C} = \mathbf{C}_j$ and $\mathbf{U} = \mathbf{U}_k$, so that $\mathbf{R} = \mathbf{C}_j \otimes (\mathbf{I}_L \otimes \mathbf{U}_k)$, where $\mathbf{C}_j$ is a (J - 1) x J matrix which defines a set of (J - 1) linearly independent contrasts for the between-subjects factor.

Methodology

A Monte Carlo simulation study was undertaken to empirically evaluate the Type I error performance of the approximate df solution given by Johansen (1980) to that of DMM procedures for testing within-subjects main and interaction effects. These tests were investigated for a multivariate design containing a single between-subjects factor and a single within-subjects factor.

Nine variables were manipulated in the simulation study. These were: (a) number of levels of the between-subjects factor, (b) number of dependent variables, (c) total sample size, (d) degree of group size equality/inequality, (e) ratio of the smallest group size (i.e., n_{min}) to t, where $t = L(K - 1)$, (f) equality/inequality of the group variance-covariance matrices, (g) nature of the pairing of group sizes and group covariance matrices, (h) multivariate normality/nonnormality, and (i) degree of correlation among the dependent variables.

The one constant in the study was the number of levels of the within-subjects factor, which was set at four across all of the investigated conditions. As well, the multivariate sphericity assumption was not violated, since none of the previously described test procedures are dependent on this assumption.

Much of the previous research which has investigated methods for testing within-subjects main and interaction effects in groups by trials designs has focussed on the situation in which the number of levels of the between-subjects grouping factor is held constant (i.e., Keselman & Keselman, 1990; Keselman et al., 1993). In their meta-analysis of the repeated measures robustness literature, Keselman, Lix, and Keselman (1994) recommended that researchers consider the effect of variation in this variable on Type I error performance when the effects of violation of the assumption of covariance heteroscedasticity is under investigation. Accordingly, in this study a groups by trials design containing either two or three levels of the between-subjects factor was considered.

Keselman et al. (1993) found that a critical determinant of the performance of Johansen's (1980) approximate df solution in univariate groups by trials designs was the ratio $n_{min}/(K - 1)$. While the value of K was held constant in this study, the value of L was set at either two or four and consequently $t = L(K - 1)$ assumed values of either six or 12.

The third variable in this study was total sample size. Based on the findings of Algina et al. (1991) and Tang and Algina (1993), N was selected such that the ratio of N/t ranged from five to 20. Thus, for $t = 6$, $N = 30$, 60, and 90 for $J = 2$ and $N = 60$, 120, and 180 for $J = 3$. For $t = 12$, $N = 60$, 90, and 120 for $J = 2$ and $N = 120$, 180, and 240 for $J = 3$. Thus, for both values of J, small, medium, and large sample sizes were considered.

The operating characteristics of the various test procedures were investigated for both balanced and unbalanced designs, given that DMM test procedures are likely to perform less optimally when group sizes are unequal. Table 1 provides the $J = 2$ and $J = 3$ group sizes that were investigated for each value of total sample size when the design was unbalanced. Table 1 also provides the n_{min}/t ratios for each of these conditions, which ranged in value from 2 to 5. For equal group sizes, this ratio equalled 2.5, 5.0 and 7.5 for $J = 2$, for the small, medium, and large sample size conditions, respectively, while for $J = 3$, the corresponding values were 3.33, 5.00, and 6.67. Finally, Table 1 contains the values associated with a coefficient of variation of group size inequality, Δn_j, where

$$\Delta n_j = \frac{\sqrt{\sum_{j=1}^{J} (n_j - \bar{n})^2 / J}}{\bar{n}}, \tag{11}$$

and $\bar{n}$ is the average group size. This coefficient of variation has a value of zero when group sizes are equal and increases in value as the group sizes become more disparate.

Insert Table 1 about here

The DMM and Johansen (1980) procedures were investigated when the group variance-covariance matrices were equal and unequal. In the latter case, elements of the matrices were in the ratio of 1:5 for $J = 2$ and 1:3:5 for $J = 3$. The degree and type of covariance heterogeneity selected for $J = 3$ corresponds to that investigated by Keselman et al. (1993) for a univariate groups by trials design. The ratio selected for $J = 2$ was chosen for purposes of consistency in the relationship between the elements of the largest and smallest group variance-covariance matrices.

Both positive and negative pairings of group sizes and covariance matrices were investigated. A positive pairing refers to the case in which the largest n_j is associated with the covariance matrix containing the largest element values; a negative pairing refers to the case in which the largest n_j is associated with the covariance matrix with the smallest element values. These pairings are known to produce liberal and conservative Type I error rates respectively for tests of within-subjects main and interaction effects in univariate groups by trials designs for mixed model test procedures (Keselman & Keselman, 1990).

Error rates were obtained when the data were both normal and nonnormal in form. With respect to the latter, the data were sampled from a χ^2_3 distribution, which is skewed to the right, which was also investigated by Keselman et al. (1993).

Pseudorandom observation vectors $\mathbf{Y}_{ij} = [Y_{ij11}, Y_{ij12}, \ldots, Y_{ijKL}]$, ($i = 1, \ldots, n_j$) from a multivariate normal distribution with mean vector β_j and covariance matrix $\mathbf{\Sigma}_j$ were obtained using the SAS (SAS Institute, 1989) generator RANNOR. A row vector of p deviates in which each element had a standard normal distribution (i.e., $\mathbf{Z}_{ij}$), was transformed to a vector of multivariate observations via a triangular (Cholesky) decomposition,

$$\mathbf{Y}_{ij} = \beta_j + \mathbf{L}\mathbf{Z}_{ij}^{T} , \tag{12}$$

where $\mathbf{L}$ is an upper triangular matrix of dimension p satisfying the equality $\mathbf{L}'\mathbf{L} = \mathbf{\Sigma}_j$.

The RANNOR generator was also used in generating the χ^2_3 data. Each element of the p-vector $\mathbf{Z}_{ij}$ was obtained by squaring and summing three standard normal deviates. These chi-square deviates were then standardized; the multivariate observations were obtained via the transformation of Equation 12.

This particular type of distribution was selected for two reasons. First, skewed distributions are representative of educational and psychological research data (see Micceri, 1989). Second, this type of distribution has been reported to affect the Type I error rates of statistics that are related to the approximate df solution of Johansen (1980). Specifically, Sawilowsky and Blair (1992) investigated the effects of eight nonnormal distributions that were identified by Micceri (1989) as representative of educational and psychological research data on the robustness of Student's t test. Only distributions with the most extreme degree of skewness considered (e.g., $\gamma_1 = 1.64$) were found to affect the Type I error rates of the t statistics. For the χ_3^2 distribution, skewness and kurtosis are, respectively, $\gamma_1 = 1.63$ and $\gamma_2 = 4.00$.

The final variable investigated was the degree of correlation, ρ, among the dependent variables at each level of the within-subjects factor. While Robey and Barcikowski (1986) found that the value of ρ had little effect on error rates, this observation was made within the context of testing within-subjects effects in designs containing only a single group of subjects. Robey and Barcikowski set $\rho = .2, .5$, and .8; only the two extreme values, that is, $\rho = .2$ and .8 were considered in this investigation.

The simulation program was written in the SAS/IML (SAS Institute, 1989) programming language. Five thousand replications of each condition were performed using a .05 significance level. For each replication, the various DMM statistics for testing hypotheses concerning main and interaction effects were converted to F statistics and compared to an appropriate critical value from an F distribution. Johansen's (1980) approximate df F statistic was also computed.

Results

A quantitative measure of robustness suggested by Bradley (1978) was used to evaluate the Type I error performance of the DMM and Johansen (1980) procedures. According to Bradley's liberal criterion, in order for a test to be considered robust, its empirical rate of Type I error (i.e., $\hat{\alpha}$) must be contained in the interval $.5\alpha \leq \hat{\alpha} \leq 1.5\alpha$. For the 5% level of significance used in this study, therefore, a test was declared robust for a particular condition if its empirical rate of Type I error fell within the interval

$.025 \le \hat{\alpha} \le .075$. 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 of values reported in this section, the latter values are bolded. Also, in these tables, the error rates associated with the DMM procedures are denoted by the abbreviation DM, while the results associated with Johansen's procedure are denoted by the abbreviation WJ. Both main and interaction effect test results are given; for the latter when J = 3, only the results associated with the Pillai-Bartlett (Bartlett, 1939; Pillai, 1955) trace are reported, since the the Hotelling-Lawley (Hotelling, 1951; Lawley, 1938) trace and Wilks' (1932) likelihood ratio procedures proved to be more sensitive to the combined effects of nonnormality and covariance heterogeneity than the Pillai-Bartlett procedure, which is consistent with the findings of other researchers, including Olson (1974). Finally, the tabled values have been averaged across the two values of ρ due to similarities in results. However, it is worth noting that the error rates obtained when the degree of correlation was strong tended to be slightly larger that those obtained when the degree of correlation was weak.

Tables 2 and 3 contain the empirical percentages of Type I error associated with balanced designs for both J = 2 and J = 3, when t = 6 and 12, respectively (i.e., L = 2 and 4). All of the procedures had error rates which were contained within the bounds of Bradley's (1978) criterion when group covariance matrices were equal, even when the data were nonnormal in form. The maximum value obtained for equal Σ_js was 7.02% and was associated with the WJ interaction test procedure.

Insert Tables 2 and 3 about here

The Table 2 values (t = 6) indicate that when group sizes were equal, the DM and WJ statistics were largely robust to the effects of covariance heterosecdasticity even under violations of the multivariate normality assumption, as only a small number of values exceeded the upper bound of Bradley's (1978) criterion. When J = 2, the DM main effect procedure was liberal when covariance matrices were unequal and $n_{min}/t = 2.5$ for both normal and nonnormal data, while this was true for the WJ main effect procedure only when the data were nonnormal. On the other hand, the DM interaction effect test was liberal for this same ratio of n_{min}/t when the data were normally distributed. When J = 3, only a single liberal value was obtained (7.93%) and was associated with the WJ interaction test procedure when the data were sampled from a χ^2_3 distribution and n_{min}/t was at its minimum value.

The results obtained for balanced designs when t = 12 (see Table 3) reveal a larger number of liberal values. When the covariance homogeneity assumption was violated and J = 2, the DM tests were liberal when $n_{min}/t = 2.5$, for both the normal and χ^2_3 data; when this ratio equalled five they were only liberal for the χ^2_3 data. The maximum value obtained for the DM procedures was 11.09%. At this same value of J, the WJ main and interaction test procedures were also liberal when the data were both heteroscedastistic and nonnormal for the smallest sample size condition; the maximum value was 9.11%. However, when the number of groups was increased to three, only the WJ interaction test procedure was liberal when sample size was at a minimum.

A comparison of corresponding normal and nonnormal values in Tables 2 and 3 reveal that the latter were almost always higher than the former for the main effect test procedures. For the interaction effect test procedures, however, this was not always the case.

Tables 4 and 5 contain the empirical percentages of Type I error associated with unbalanced designs for J = 2, when t = 6 and 12, respectively. The DM procedures were generally conservative for positive pairings of group sizes and group covariance matrices, except when $n_{min}/t = 2.0$ and N = 30, for t = 6. In fact, for a fixed value of n_{min}/t, these procedures became increasing conservative as total sample size increased in value, due to a corresponding increase in the magnitude of Δn_j. The minimum value obtained for the positive pairing conditions was less than .001%. For negative pairings, the DM procedures were always liberal and error rates were, in many cases, extremely inflated. The maximum value obtained for t = 6 was 51.63% while for t = 12 it was 71.55%. As with the positive pairings, the DM procedures became more liberal for a fixed value of n_{min}/t as the degree of group size inequality increased. However, for a fixed value of total sample size, error rates became less extreme for both positive and negative pairings as n_{min}/t increased in magnitude.

Insert Tables 4 and 5 about here

As Tables 4 and 5 reveal, error rates for the WJ procedures were contained within the bounds of Bradley's (1978) criterion when group sizes and covariance matrices were positively paired. This was true even when the data were nonnormal in form.

For negative pairings, consistent with the findings of Keselman et al. (1993), the WJ test procedures performed best when the ratio of n_{min}/t was not too small. When t = 6, the WJ main and interaction test procedures were liberal for the smallest value of n_{min}/t, even when the data were normal in form. For the χ^2_3 data associated with t = 6, this finding also held when $n_{min}/t = 3$ and $N = 90$, although error rates were only marginally greater than the upper bound of Bradley's criterion (i.e., 7.97%). For $t = 12$, liberal values were obtained when n_{min}/t was 2 and 3, for both the normal and χ^2_3 data.

The results associated with the three-group multivariate design are contained in Tables 6 and 7, for $t = 6$ and 12, respectively. Consistent with the $J = 2$ results, for positive pairings of group sizes and covariance matrices, the DMM procedures were generally conservative; error rates only slightly exceeded the lower bound of Bradley's (1978) criterion when $n_{min}/t = 3$ and total sample size was small. For negative pairings, these test procedures were always liberal. As well, error rates became more inflated as t increased in value.

Insert Tables 6 and 7 about here

For both values of t and $J = 3$, the WJ main effect procedure always had rates of Type I error which were contained within the bounds of Bradley's (1978) criterion, even when there was a negative relationship between group sizes and covariance matrices. When $t = 6$, the WJ interaction effect procedure was liberal for normal data in only a single instance, when the ratio of $n_{min}/t = 3.0$ and $N = 120$, when t was increased in value to 12 and the data were normal, liberal values resulted for $N = 180$ and 240. For the χ^2_3 data, n_{min}/t had to be at least 4 to 1 for the interaction test to control the error rate to α, for both values of t.

Conclusions

The performance of tests of within-subjects main and interaction effects in multivariate groups by trials designs which are based on a doubly multivariate analytic approach was consistent with findings obtained for univariate designs and was therefore not unexpected. In most instances, these procedures can effectively control the rate of Type I errors rate for balanced designs when group variance-covariance matrices are heterogeneous, even when the data are nonnormal in form. However, they are extremely sensitive to departures from the covariance homogeneity assumption when the design is unbalanced. Furthermore, this sensitivity may increase when multivariate normality is not a tenable assumption. Consequently, researchers who adopt the doubly multivariate analysis strategy when group sizes are unequal may be drawing inaccurate and misleading conclusions about their data.

Researchers do have an alternative strategy available to them. The results from this study indicate that, under certain condition, the approximate df solution given by Johansen (1980) can be used to test repeated measures hypotheses when covariance homogeneity is not a tenable assumption and the design is unbalanced. However, the issue of sample size must be attended to carefully. When it can be assumed that the data are normal in form, the number of observations in the smallest of the groups should be at least three times the product of the number of dependent variables times the number of repeated measurements minus one. To obtain a robust test in the presence of multivariate nonnormality, this ratio may need to be increased to at least 4 or 5 to 1, particularly if tests of the within-subjects interaction effect are to be valid. It should be noted that these results concur with those obtained by Keselman et al. (1993) for univariate groups by trials designs.

Implementation of Johansen's (1980) solution to test for mean equality is easily accomplished using a SAS/IML (SAS Institute, 1989) program developed by Lix and

Keselman (in press) which is based on the general linear model. This program only requires that the researcher enter the data, the group sizes, and one or more contrast matrices which specify the hypothesis to be tested. As a final note, this progam may also be used to test specific contrasts on multivariate data that may be useful in probing the nature of a significant within-subjects main or interaction effect.

References

Algina, J., Oshima, T. C., & Tang, K. L. (1991). Robustness of Yao's, James', and Johansen's tests under variance-covariance heteroscedasticity and nonnormality. Journal of Educational Statistics, 16, 125-139.

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

Bartlett, M. S. (1939). A note on tests of significance in multivariate analysis. Proceedings of the Cambridge Philosophical Society, 35, 180-185.

Boik, R. J. (1988). The mixed model for multivariate repeated measures: Validity conditions and an approximate test. Psychometrika, 53, 469-486.

Boik, R. J. (1991). Scheffe's mixed model for multivariate repeated measures: A relative efficiency evaluation. Communications in Statistics, Theory and Methods, 20, 1233-1255.

Hotelling, H. (1931). The generalization of Student's ratio. Annals of Mathematical Statistics, 2, 360-378.

Hotelling, H. (1951). A generalization t-test and measure of multivariate dispersion. In Proceedings of the Second Berkeley Symposium on Mathematics and Statistics (pp. 23-41). University of California Press, Berkley.

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.

Keselman, H. J., Carriere, K. C., & Lix, L. M. (1993). Testing repeated measures hypotheses when covariance matrices are heterogeneous. Journal of Educational Statistics, 18, 305-309.

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

Keselman, J. C., Lix, L. M., & Keselman, H. J. (1994). The analysis of repeated measurements: A quantitative research synthesis. Manuscript submitted for publication.

Lawley, D. N. (1938). A generalization of Fisher's z-test. Biometrika, 30, 180-187.

Lix, L. M., & Keselman, H. J. (in press). Approximate degrees of freedom tests: A unified perspective on testing for mean equality. Psychological Bulletin.

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

Muller, K. E., LaVange, L. M., Ramey, S. L., & Ramey, C. T. (1992). Power calculations for general linear multivariate models including repeated measures applications. Journal of the American Statistical Association, 87, 1209-1226.

Olson, C. L. (1974). Comparative robustness of six tests in multivariate analysis of variance. Journal of the American Statistical Association, 69, 894-908.

Pillai, K. C. S. (1955). Some new test criteria in multivariate analysis. Annals of Mathematical Statistics, 26, 117-121.

Robey, R. R., & Barcikowski, R. S. (1986, April). Type I error in the single group repeated measures design with multiple measures per occasion. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Roy, S. N. (1957). Some aspects of multivariate analysis. New York: Wiley.

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

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

Searle, S. S. (1987). Linear models for unbalanced data. New York: Wiley.

Tang, K. L., & Algina, J. (1993). Performance of four multivariate tests under variance-covariance heteroscedasticity. Multivariate Behavioral Research, 28, 391-405.

Thomas, D. R. (1983). Univariate repeated measures techniques applied to multivariate data. Psychometrika, 48, 451-464.

Timm, N. H. (1980). Multivariate analysis of variance of repeated measurements. In P. R. Krishnaiah (Ed.), Handbook of statistics, Vol. 1 (pp. 41-87). Amsterdam: North-Holland.

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

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

Wilks, S. S. (1932). Certain generalizations in the analysis of variance. Biometrika, 24, 471-494.

Table 1

Group Sizes For Unbalanced Designs

t	N	n_{min}/t	Group Sizes	Δn_j
J = 2				
6	30	2	12; 18	.20
	60	2	12; 48	.60
		3	18; 42	.40
		4	24; 36	.20
	90	3	18; 72	.60
		4	24; 66	.47
		5	30; 60	.33
12	60	2	24; 36	.20
	120	2	24; 96	.60
		3	36; 84	.40
		4	48; 72	.20
	180	3	36; 144	.60
		4	48; 132	.47
		5	60; 120	.33
J = 3				
6	60	3	18; 20; 22	.08
	90	3	18; 30; 42	.33
		4	24; 30; 36	.16
	120	3	18; 40; 62	.45
		4	24; 40; 56	.33
		5	30; 40; 50	.20
12	120	3	36; 40; 44	.08
	180	3	36; 60; 84	.33
		4	48; 60; 72	.16
	240	3	36; 80; 124	.45
		4	48; 80; 112	.33
		5	60; 80; 100	.20

Table 2

Empirical Percentages of Type I Error (Equal Group Sizes; t = 6)

			Normal				χ^2_3			
n_{min}/t	N	$\overset{=}{\neq} \Sigma_j$	DM Main	WJ Main	DM Int	WJ Int	DM Main	WJ Main	DM Int	WJ Int
						J = 2				
2.5	30	$= \Sigma_j$	4.35	4.80	5.20	5.63	5.94	6.39	4.27	4.69
		$\neq \Sigma_j$	**8.10**	6.51	**8.39**	6.86	**8.94**	**7.62**	7.36	6.24
5.0	60	$= \Sigma_j$	5.54	5.46	5.22	5.17	6.08	5.96	5.30	5.25
		$\neq \Sigma_j$	6.82	5.20	6.73	5.36	**7.60**	6.08	6.71	5.20
7.5	90	$= \Sigma_j$	5.10	5.03	4.25	4.13	4.90	4.79	5.12	4.97
		$\neq \Sigma_j$	6.69	5.60	6.09	5.00	6.98	5.94	6.78	5.82
						J = 3				
3.33	60	$= \Sigma_j$	5.09	4.95	4.78	6.12	5.82	5.66	5.06	6.87
		$\neq \Sigma_j$	6.04	5.05	5.59	6.47	6.22	5.28	6.19	**7.93**
5.0	90	$= \Sigma_j$	5.31	5.12	4.99	5.08	5.43	5.28	4.55	5.16
		$\neq \Sigma_j$	5.75	5.00	6.11	6.15	5.77	5.09	6.24	5.81
6.67	120	$= \Sigma_j$	4.46	4.34	4.47	4.61	4.97	4.82	4.67	5.32
		$\neq \Sigma_j$	4.97	4.29	6.04	5.21	5.77	5.14	5.67	5.41

Note: DM Main = Doubly multivariate main effect test; WJ Main = Welch-James main effect test; DM Int = Doubly multivariate interaction effect test; WJ Int = Welch-James interaction effect test; Bold values are not contained in the interval 2.5 - 7.5.

Table 3

Empirical Percentages of Type I Error (Equal Group Sizes; $t = 12$)

			Normal				χ^2_3			
n_{min}/t	N	$\overset{=}{\neq}\ \Sigma_j$	DM Main	WJ Main	DM Int	WJ Int	DM Main	WJ Main	DM Int	WJ Int
						$J = 2$				
2.5	60	$= \Sigma_j$	5.25	6.04	4.76	5.47	6.07	6.89	4.15	4.76
		$\neq \Sigma_j$	**9.44**	**7.53**	**9.23**	7.21	**11.09**	**9.11**	**9.72**	**7.95**
5.0	120	$= \Sigma_j$	4.44	4.44	5.18	5.18	5.42	5.41	4.73	4.71
		$\neq \Sigma_j$	7.37	5.68	7.15	5.72	**7.74**	5.95	**8.22**	6.48
7.5	180	$= \Sigma_j$	5.08	4.98	4.52	4.49	5.56	5.51	4.92	4.81
		$\neq \Sigma_j$	6.76	5.40	6.13	4.92	6.75	5.30	6.29	5.05
						$J = 3$				
3.33	120	$= \Sigma_j$	5.20	5.19	4.73	6.42	5.57	5.51	4.80	7.02
		$\neq \Sigma_j$	7.05	5.93	6.70	**7.64**	7.09	5.98	6.61	**7.91**
5.00	180	$= \Sigma_j$	5.27	5.17	5.14	5.64	5.21	5.10	4.91	5.49
		$\neq \Sigma_j$	6.14	5.28	6.63	5.91	6.20	5.30	6.78	6.32
6.67	240	$= \Sigma_j$	5.09	4.96	5.34	5.48	5.09	4.96	4.84	5.30
		$\neq \Sigma_j$	5.58	4.91	6.15	5.20	5.78	5.17	6.17	5.59

Note: See the note from Table 2.

Table 4

Empirical Percentages of Type I Error (Unequal Group Sizes; J = 2; t = 6)

			Normal				χ^2_3			
n_{min}/t	N	Pairing	DM Main	WJ Main	DM Int	WJ Int	DM Main	WJ Main	DM Int	WJ Int
2.0	30	+P	2.97	5.83	2.90	5.95	3.28	6.24	3.27	6.07
		−P	**18.75**	**9.11**	**18.29**	**8.05**	**19.38**	**9.92**	**17.33**	**9.18**
	60	+P	**0.03**	4.74	**0.05**	5.19	**0.14**	5.70	**0.04**	4.49
		−P	**51.56**	**11.14**	**51.11**	**11.30**	**50.44**	**11.70**	**50.13**	**12.02**
3.0	60	+P	**0.25**	4.79	**0.28**	4.84	**0.52**	5.19	**0.55**	4.77
		−P	**32.30**	6.72	**31.28**	6.63	**30.98**	7.24	**30.78**	6.60
	90	+P	**0.00**	4.86	**0.00**	5.69	**0.00**	5.69	**0.04**	5.37
		−P	**51.63**	7.24	**50.95**	6.88	**50.85**	**7.73**	**49.18**	**7.97**
4.0	60	+P	**1.83**	4.95	**2.24**	5.27	**2.29**	5.26	**2.02**	5.42
		−P	**16.45**	5.92	**16.56**	6.10	**17.99**	7.05	**17.59**	6.71
	90	+P	**0.10**	4.73	**0.09**	4.93	**0.16**	5.14	**0.18**	4.55
		−P	**37.28**	5.94	**36.29**	6.33	**37.71**	7.38	**38.69**	7.24
5.0	90	+P	**0.68**	5.18	**0.50**	5.12	**0.73**	5.45	**0.58**	5.02
		−P	**25.50**	5.27	**25.44**	5.44	**26.26**	6.68	**25.49**	6.21

Note: +P = Positive pairing of n_j and Σ_j; −P = Negative pairing of n_j and Σ_j; See the note from Table 2.

Table 5

Empirical Percentages of Type I Error (Unequal Group Sizes; J = 2; t = 12)

			Normal				χ^2_3			
			DM Main	WJ Main	DM Int	WJ Int	DM Main	WJ Main	DM Int	WJ Int
n_{min}/t	N	Pairing								
2.0	60	+P	**2.20**	5.87	**2.34**	6.40	**2.23**	6.56	**2.28**	6.48
		−P	**24.99**	**10.22**	**25.00**	**10.24**	**25.69**	**11.19**	**25.01**	**10.51**
	120	+P	**0.02**	5.27	**0.00**	4.98	**0.00**	5.94	**0.00**	5.32
		−P	**71.38**	**13.00**	**70.47**	**12.71**	**71.33**	**14.47**	**70.78**	**14.42**
3.0	120	+P	**0.12**	4.49	**0.05**	4.83	**0.17**	5.32	**0.11**	5.25
		−P	**45.51**	7.42	**46.29**	6.98	**47.26**	**8.71**	**47.22**	**8.81**
	180	+P	**0.00**	5.02	**0.00**	4.75	**0.00**	5.70	**0.00**	4.76
		−P	**71.55**	**8.08**	**71.74**	**8.17**	**70.88**	**8.84**	**71.16**	**8.65**
4.0	120	+P	**1.30**	4.47	**1.09**	4.83	**1.61**	5.25	**1.21**	4.65
		−P	**22.73**	6.34	**22.57**	6.42	**23.01**	6.81	**22.72**	6.48
	180	+P	**0.03**	4.88	**0.04**	5.20	**0.01**	5.05	**0.02**	4.62
		−P	**55.19**	6.56	**54.40**	6.54	**55.19**	6.96	**54.45**	6.55
5.0	180	+P	**0.13**	4.53	**0.22**	5.00	**0.17**	5.35	**0.20**	5.21
		−P	**36.84**	6.04	**37.28**	5.75	**38.16**	6.53	**37.88**	6.72

Note: See the notes from Tables 2 and 4.

Table 6

Empirical Percentages of Type I Error (Unequal Group Sizes; J = 3; t = 6)

			Normal				χ^2_3			
n_{min}/t	N	Pairing	DM Main	WJ Main	DM Int	WJ Int	DM Main	WJ Main	DM Int	WJ Int
3.0	60	+P	4.33	5.39	4.19	5.89	4.32	5.10	4.26	6.77
		−P	**7.94**	4.88	**8.02**	6.75	**9.81**	6.35	**8.15**	**7.76**
	90	+P	**1.18**	5.40	**1.08**	5.18	**1.12**	5.25	**1.12**	5.47
		−P	**20.29**	5.88	**20.80**	7.10	**19.95**	6.05	**19.09**	**7.91**
	120	+P	**0.30**	5.23	**0.54**	4.71	**0.30**	5.29	**0.74**	5.55
		−P	**28.32**	5.67	**29.64**	**7.92**	**28.68**	6.79	**28.95**	**9.28**
4.0	90	+P	**2.43**	5.05	2.64	5.15	2.84	5.54	2.61	5.41
		−P	**10.84**	4.85	**11.96**	5.92	**11.39**	5.37	**11.17**	6.90
	120	+P	**0.77**	4.82	**1.05**	4.50	**1.00**	5.00	**1.37**	5.42
		−P	**19.76**	5.08	**19.57**	5.12	**19.79**	5.41	**19.56**	**7.51**
5.0	120	+P	**1.84**	5.07	**2.49**	5.62	**2.21**	5.57	2.58	5.95
		−P	**13.04**	4.99	**13.71**	5.73	**12.46**	5.06	**13.01**	5.91

Note: See the notes from Tables 2 and 4.

Table 7

Empirical Percentages of Type I Error (Unequal Group Sizes; J = 3; t = 12)

			Normal				χ^2_3			
n_{min}/t	N	Pairing	DM Main	WJ Main	DM Int	WJ Int	DM Main	WJ Main	DM Int	WJ Int
3.0	120	+P	3.84	5.08	3.74	6.19	4.04	5.42	3.67	7.01
		−P	**10.09**	4.99	**9.87**	5.73	**10.39**	6.00	**9.36**	**8.16**
	180	+P	**0.31**	4.77	**0.74**	5.73	**0.62**	5.52	**0.72**	5.76
		−P	**27.69**	5.46	**28.94**	**7.86**	**27.94**	6.14	**28.96**	**9.08**
	240	+P	**0.09**	4.49	**0.30**	5.65	**0.08**	5.29	**0.14**	5.80
		−P	**42.89**	6.07	**43.18**	**7.90**	**42.56**	7.00	**44.60**	**9.41**
4.0	180	+P	**1.85**	4.92	**2.31**	5.34	**1.75**	4.65	**2.28**	5.80
		−P	**13.10**	4.89	**14.39**	5.82	**14.69**	6.00	**14.60**	6.39
	240	+P	**0.34**	4.87	**0.71**	5.60	**0.35**	5.17	**0.68**	5.08
		−P	**28.53**	5.49	**28.76**	6.28	**27.97**	5.67	**28.56**	7.47
5.0	240	+P	**1.30**	5.25	**1.76**	5.03	**0.91**	5.14	**1.63**	5.22
		−P	**16.81**	4.88	**17.98**	5.86	**16.92**	5.32	**18.26**	6.17

Note: See the notes from Tables 2 and 4.