Journal of Modern Applied Statistical Methods

Volume 3 | Issue 2 Article 13

11-1-2004

Type I Error Rates For A One Factor Within-Subjects Design With Missing Values

Miguel A. Padilla
University of Florida, mpadilla@ufl.edu

James Algina
University of Florida, algina@ufl.edu

Journal of Modern Applied Statistical Methods
November, 2004, Vol. 3, No. 2,406-416

1538 – 9472/04/$95.00

Early Scholars

Type I Error Rates For A One Factor Within-Subjects Design With Missing Values

Miguel A. Padilla James Algina
Educational Psychology
University of Florida

Missing data are a common problem in educational research. A promising technique, that can be implemented in SAS PROC MIXED and is therefore widely available, is to use maximum likelihood to estimate model parameters and base hypothesis tests on these estimates. However, it is not clear which test statistic in PROC MIXED performs better with missing data. The performance of the Hotelling-Lawley-McKeon and Kenward-Roger omnibus test statistics on the means for a single factor within-subject ANOVA are compared. The results indicate that the Kenward-Roger statistic performed better in terms of keeping the Type I error close to the nominal alpha level.

Key words: Type I error, Within-subjects, missing values, Kenward-Roger omnibus test, robustness

Introduction

A common problem in multivariate analysis is the missing data problem. Data values may be missing for a variety of reasons. For example, a subject may drop out of a longitudinal study because of death or illness, or refuse to respond to sensitive questions on a survey, or neglect to finish the survey because of its length, etc. These, of course, are just a few examples of processes that might cause the missing data.

There are several methods available for use when data are missing. The statistical properties of these procedures depend on the mechanism for the missing data. Rubin (1976, 1987) and Little and Rubin (1987) defined three types of missing data mechanisms. Two of these are missing completely at random (MCAR) and missing at random (MAR).

Miguel A. Padilla is a NIH Graduate Student Fellow in Educational Psychology. His research interests are in applied statistics. Email: mpadilla@ufl.edu. James Algina is Professor of Educational Psychology. His interests are in psychometric theory and applied statistics. Email: algina@ufl.edu.

The third type consists of all other missing data mechanisms. Verbeke and Molenberghs (2000) advocate calling this third type missing not at random (MNAR). These types of missing data mechanisms will be described in the context of the design and analysis considered in this study. The design includes p repeated measurements made on a single group of participants. The purpose of the data analysis is to estimate parameters (i.e., the means, variances, and covariances of the repeated measurements) and to test the omnibus hypothesis that the p means are equal. To simplify the presentation the case of two repeated measurements (the simplest repeated measures design) will be used in the description.

Let X_1 and X_2 be two distinct variables. The missing data mechanism is MCAR when the pattern of missing data on X_1 and X_2 is completely independent of X_1 and X_2. The missing data mechanism is MAR if the pattern of missing data on X_2 is dependent on observed values on X_1 but not on X_2 when X_1 is held constant and the pattern of missing data on X_1 is dependent on observed values on X_2 but not on X_1 when X_2 is held constant.

So what is a researcher to do if missing data are present in his or her study? A large number of methods have been proposed for analyzing incomplete data, but the most common solutions are probably listwise deletion

and maximum likelihood ignoring the missing data mechanism. In listwise deletion all subjects with any missing data are excluded from the analysis. This is the procedure used in popular software packages (e.g., SAS and SPSS) for repeated measures ANOVA and MANOVA. Listwise deletion works reasonably well if the researcher has a large sample, a small percentage of missing data, and a MCAR missing data mechanism.

For example, if the researcher has a sample of 500 and 5% have missing data, the researcher will do the analysis with a sample of 475 and, if the data are MCAR, obtain unbiased estimates while still retaining power. However, if the researcher has a sample of 100 and 35% have missing data, doing the analysis with a sample of 65 could severely compromise power. Regardless of the sample size and amount of missing data, estimates will be biased and sampling distribution based inferences, such as hypothesis tests and confidence intervals, will be invalid if the missing data mechanism is MAR or MNAR.

As noted previously maximum likelihood ignoring the missing data mechanism is another procedure that can be used when data are missing. To understand the concept of ignoring the missing data mechanism, we must recognize that there are two types of data that can be taken into account in the analysis when there are missing data.

First, there are the independent and dependent variables that are the focus of the study and, second, there is a dichotomous indicator variable indicating whether or not a particular data point is missing. The missing data mechanism is a relationship of the indicator variable to the independent variables and the dependent variables and models the probability that data are missing as a function of the independent variables and dependent variables. The relationship might be modeled, for example, as a logistic regression function relating the presence or absence of the data points to the independent and dependent variables. Analyzing only the observed scores on the dependent variables is referred to as ignoring the missing data mechanism.

Rubin (1976) has shown that if the missing data mechanism is MCAR or MAR, ML estimators of the parameters are consistent when the missing data mechanism is ignored. Thus, the MCAR or MAR missing data mechanisms are ignorable for purposes of ML estimation. If the data are MCAR both listwise deletion and ML ignoring the missing data mechanism will produce consistent estimators, but the ML estimators will be more accurate because they use all of the available data. Rubin (1976) has also shown that the MCAR missing data mechanism is ignorable for sampling distribution based inference procedures such as hypothesis tests and confidence intervals. So if the data are MCAR either listwise deletion or ML ignoring the missing data mechanism can be used for inference, but ML will result in more powerful tests and narrower confidence intervals because it does not delete the observed data for participants that have some missing data.

When ML estimation is used, whether the MAR missing data mechanism is ignorable for sampling distribution based inference depends on the how sampling variances and covariances are calculated. The MAR missing data mechanism is ignorable for sampling distribution based inferences on the means if the sampling covariance matrix is estimated from the observed information matrix for the means and the covariance parameter estimates but not if the matrix is estimated from the portion of the observed information matrix that pertains only to the means (Kenward & Molenberghs, 1998).

The MAR mechanism may not be ignorable for sampling distribution based inferences if the sampling covariance matrix is estimated from the expected information matrix. That is, for sampling distribution based inferences to be valid the expected value of the information matrix must be taken under the actual sampling process implied by the MAR mechanism (Kenward & Molenberghs, 1998). Kenward and Molenberghs refer to using this type of expected information matrix as the unconditional sampling framework whereas using the information matrix that ignores this sampling process is called the naïve sampling framework.

Additionally, the sampling covariance matrix for the means must be computed as the inverse of the unconditional information matrix for the means and the covariance parameters.

For a design with one-within subjects factor, as well as for more complicated multivariate designs, maximum likelihood ignoring the missing data mechanism can be implemented, by using PROC MIXED on SAS. However, it should be noted that many of the test statistic options in SAS use the expected information matrix under the naïve sampling framework.

Another method for analyzing incomplete data is multiple imputation (MI) (Little & Rubin, 1987; Rubin, 1976, 1987). In MI, multiple sets of plausible values are used to replace the missing values. This creates *m* data sets with plausible values replacing missing values. Each of the *m* data sets is analyzed to produce parameter estimates. The *m* estimates are then combined to create a single estimate and a standard error of the estimate.

One advantage of MI is that a single set of imputed data sets can be used for a variety of analyses. Second, inferences drawn from multiply imputed data are valid, provided that the missing data mechanism is MAR or MCAR, because MI accounts for missing data uncertainty (Schafer, 1997; Schafer & Olsen, 1998). MI is very efficient in that it only requires a small set of imputed data sets to conduct a valid analysis (Rubin, 1987; Schafer, 1997; Schafer & Olsen, 1998). However, MI can be cumbersome to use because of the need to analyze multiple data sets and combine the results to make one overall inference. This drawback has been overcome for some designs because software is available that combines the estimates automatically.

As noted previously, if the missing data mechanism is MNAR, the missing data mechanism is non-ignorable (NI) for purposes of ML estimation. Thus, if the missing data mechanism is not MAR or MCAR, the pattern of missing data must be taken into account in order to obtain consistent ML estimates. This can be accomplished by using a selection model that incorporates a model for the missing data indicator or by using a pattern mixture model, which stratifies the data on the basis of the pattern of missing data. See Little (1995) for additional details about these two approaches. For examples of these models the reader is referred to Diggle and Keward (1994), Troxel (1998), Kenward (1998), Albert and Follmann (2000), and Fitzmaurice, Laird, and Shneyer (2001).

Sampling based inferences will also be valid under selection modeling that incorporates a model for the missing data and under a pattern mixture model. However, selection modeling incorporating the missing data mechanism and pattern mixture modeling are more difficult to implement than are analyses that ignore the missing data mechanism. For example, for the design considered in this study, the analysis ignoring the missing data mechanism can be implemented using PROC MIXED in SAS, but selection modeling incorporating the missing data mechanism cannot. Thus, it seems very likely that analyses that ignore the missing data mechanism will be widely used in the future. For this reason we focus on ML methods ignoring the missing data mechanism.

Let p denote the number of levels of the within-subjects factor, $\mathbf{\Sigma}$ the $p \times p$ population covariance matrix, $\mathbf{S}$ the $p \times p$ estimated covariance matrix, and $\mathbf{\Sigma}_i$ and $\mathbf{S}_i$ the $p_i \times p_i$ section of the population and sample covariance matrices, respectively, that pertain to the dependent variables on which subject i has observed scores. In addition let $\mathbf{A}_i$ denote a $p_i \times p$ indicator matrix obtained by eliminating the j^{th} row from the $p \times p$ identity matrix if the data for subject i is missing on x_j. Ignoring the missing data mechanism, the generalized least squares estimate of the mean vector is

$$\hat{\boldsymbol{\mu}} = \left(\sum_i \mathbf{A}_i' \mathbf{\Sigma}_i^{-1} \mathbf{A}_i \right)^{-} \left(\sum_i \mathbf{A}_i' \mathbf{\Sigma}_i^{-1} \mathbf{x}_i \right) \qquad (1)$$

In practice $\mathbf{\Sigma}_i$ must be estimated and the estimated sample mean vector is

$$\bar{\mathbf{x}} = \left(\sum_i \mathbf{A}_i' \mathbf{S}_i^{-1} \mathbf{A}_i \right)^{-} \left(\sum_i \mathbf{A}_i' \mathbf{S}_i^{-1} \mathbf{x}_i \right).$$

If $\mathbf{S}$ is obtained by maximum likelihood or restricted maximum likelihood, $\bar{\mathbf{x}}$ is the maximum likelihood estimate.

Let $\mathbf{C}$ be a $(p-1) \times p$ matrix of full row rank. Each row of $\mathbf{C}$ is a contrast vector. The

hypothesis that all p population means are equal is

$$H_0 : \mathbf{C}\boldsymbol{\mu} = \mathbf{0}$$

where the bold zero is a vector of length $(p-1)$ with all elements equal to zero. The default test statistic in PROC MIXED for testing the null hypothesis is

$$\frac{1}{(p-1)}\bar{\mathbf{x}}'\mathbf{C}'\left[\mathbf{C}\left(\sum_i \mathbf{A}_i'\mathbf{S}_i^{-1}\mathbf{A}_i\right)^{-}\mathbf{C}'\right]^{-1}\mathbf{C}\bar{\mathbf{x}} \quad (2)$$

with critical value $F_{\alpha,\ p\text{-}1,\ n\text{-}1}$. An alternative is to use the test statistic

$$\frac{n-p+1}{(p-1)(n-1)}\bar{\mathbf{x}}'\mathbf{C}'\left[\mathbf{C}\left(\sum_i \mathbf{A}_i'\mathbf{S}_i^{-1}\mathbf{A}_i\right)^{-}\mathbf{C}'\right]^{-1}\mathbf{C}\bar{\mathbf{x}} \quad (3)$$

with critical value $F_{\alpha,\ p\text{-}1,\ n\text{-}p+1}$. In SAS this is referred to as the Hotelling-Lawley-McKeon (HLM) test. If there are no missing data the test statistic simplifies to the usual F transformation of Hotellings T^2 for a repeated measures design with no between-subjects factors. According to Wolfinger and Chang (1995), when data are complete and the unstructured option for the covariance matrix is selected, the default test statistic tends to be liberal with small samples sizes and the HLM performs more satisfactorily.

In equations (1) and (2), the expression $\left(\sum_i \mathbf{A}_i'\mathbf{S}_i^{-1}\mathbf{A}_i\right)^{-}$ is the estimated sampling covariance matrix of the mean vector $\bar{\mathbf{x}}$ and is based on the expected information matrix calculated under the naïve sampling framework. Even when data are MCAR or there are no missing data, using $\left(\sum_i \mathbf{A}_i'\mathbf{S}_i^{-1}\mathbf{A}_i\right)^{-}$ has two drawbacks

1. $\left(\sum_i \mathbf{A}_i'\mathbf{S}_i^{-1}\mathbf{A}_i\right)^{-}$ is an estimate of $\left(\sum_i \mathbf{A}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\right)^{-}$, the sampling covariance matrix of $\hat{\boldsymbol{\mu}}$ in equation (1). Results by Kackar and Harville (1984) show that, as a sampling covariance matrix for $\bar{\mathbf{x}}$, $\left(\sum_i \mathbf{A}_i'\mathbf{S}_i^{-1}\mathbf{A}_i\right)^{-}$ tends to be too small because it fails to take into account the uncertainty in $\bar{\mathbf{x}}$ introduced by substituting $\mathbf{S}_i$ for $\boldsymbol{\Sigma}_i$.

2. Booth and Hobert (1998) and Prasad and Rao (1990) show that $\left(\sum_i \mathbf{A}_i'\mathbf{S}_i^{-1}\mathbf{A}_i\right)^{-}$ is biased for $\left(\sum_i \mathbf{A}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\right)^{-}$.

Harville and Jeske (1992) developed a better approximation, denoted by $\widehat{m}^{@}$, that can be used to estimate the sampling covariance matrix of $\bar{\mathbf{x}}$. Subsequently, Kenward and Roger (1997) developed an alternative estimator, denoted by $\widehat{\Phi}_A$, that can also be used to estimate the sampling covariance matrix for $\bar{\mathbf{x}}$. Kenward and Roger also proposed a test statistic, which in the context of comparing p means is

$$\frac{\lambda}{p-1}\bar{\mathbf{x}}'\mathbf{C}'\left(\mathbf{C}\widehat{\Phi}_A\mathbf{C}'\right)^{-1}\mathbf{C}\bar{\mathbf{x}}$$

with critical value $F_{\alpha,\ p\text{-}1,\ df}$ where λ and df are estimated from the data. The Kenward-Roger (KR) procedure is implemented in PROC MIXED. However, $\widehat{m}^{@}$ is used in place of $\widehat{\Phi}_A$.

The Current Study

The purpose of this article is to compare Type I error rates for two procedures available in SAS: the HLM procedure and the Kenward-Roger (KR) procedure. Simulation methods were used to make the comparison. Data were generated under the MAR and MCAR missing data mechanisms because of the properties enjoyed by ML estimation under these mechanisms if the missing data mechanism is ignored. For comparison purposes data were also generated under a MNAR missing data mechanism. None of the procedures were expected to work well under this missing data mechanism.

Related literature

Fai and Cornelius (1996) developed and compared four alternative test procedures that can be used to test linear hypotheses on means in multivariate studies. The four test statistics, specialized to the context of this paper are shown in Table 1. For each of the four statistics Fai and Cornelius showed how to use the data to estimate the second degrees of freedom. The F_2 and F_4 statistics have a scale factor estimated from the data. The F_1 and F_2 statistics use $\left(\sum_i \mathbf{A}_i'\mathbf{S}_i^{-1}\mathbf{A}_i\right)^-$ to estimate the covariance matrix of the mean vector whereas F_3 and F_4 use $\widehat{m}^{@}$. The F_4 statistic is similar to the statistic obtained by using the Kenward-Rogers option in PROC MIXED, but the formula for the scale factors and the degrees of freedom are not identical to those used when the Kenward-Rogers option is employed in PROC MIXED. The test using F_1 is available in SAS when the Satterthwaite option is used in PROC MIXED.

Fai and Cornelius (1996) applied their tests to split-plot designs with a between-subjects factor with three levels and a within-subjects factor with four levels. The covariance structure was compound symmetric. The design was unbalanced in that the number of subjects varied across levels of the between-subjects factor and data were not generated for some combinations of subjects and the within-subjects treatment. Because the missing data were never generated, the missing data mechanism was effectively MCAR. Estimated Type I error rates and power were reported for the main effect of the between-subjects factor. All four tests provided reasonable control of the Type I error rate. The performance of F_1 and F_3, which do not include a scale factor was very similar. Type I error rates and power for F_4 were always larger than for F_3.

Schaalje, McBride, and Fellingham (2001), reporting on a study conducted by McBride (2002), reported Type I error rates for F_1 and the test obtained using the Kenward-Roger option in PROC MIXED. McBride investigated performance of these tests in a split-plot design.

The following provides a social science example of the design investigated by McBride. Suppose three methods for structuring interactions among students in a mathematics classroom are to be compared; n schools are randomly assigned to each method, where n was three in half of the conditions studied by McBride and five in the other half. The methods will be implemented for three, six, or nine weeks. Each school contributes K classes. Each class is assigned a single interaction quality score. In half of the conditions studied by McBride, $K = 3$ and the design was balanced. In the other half, $K = 5$ so that within each school two classes would be assigned to two of the implementation periods and one class would be assigned to the remaining implementation period. In these conditions the design is unbalanced, but no data are missing.

McBride also investigated the effect of the covariance structure, including five structures: compound symmetric (equal correlations and equal variance for the repeated measures), heterogeneous compound symmetric (equal correlations, but unequal variances for the repeated measures), Toeplitz, heterogeneous first-order autoregressive (correlations conform to a first-order autoregressive pattern, but the variances for the repeated measures are unequal), and first-order ante-dependence (see Wolfinger, 1995, for examples of these covariance structures). The results indicated that employing the Kenward-Roger option provided better control than did employing the Satterthwaite option in PROC MIXED. Type I error rates were closer to the nominal level for balanced designs than for unbalanced designs.

For unbalanced designs, Type I error rates improved as *n* increased.

Kenward and Roger (1997) investigated how well the original Kenward-Roger procedure controlled Type I error rates in four situations: (a) a four-treatment, two-period cross-over design, (b) a row-column-α design, (c) a random coefficients regression model for repeated measures data, and (d) a split-plot design. In (c) and (d) there were missing data. In (c) the missing data mechanism was MCAR. The missing data mechanism in (d) was not specified. In all situations, the Kenward-Roger test controlled the Type I error rate well.

Table 1. Test Statistics from Fai and Cornelius (1996).

Test Statistics	Critical values
$F_1 = \frac{1}{(p-1)}\bar{\mathbf{x}}'\mathbf{C}'\left[\mathbf{C}\left(\sum_i \mathbf{A}_i'\mathbf{S}_i^{-1}\mathbf{A}_i\right)^{-}\mathbf{C}'\right]^{-1}\mathbf{C}\bar{\mathbf{x}}$	$F_{\alpha,(p-1),df_1}$
$F_2 = \frac{\lambda_2}{(p-1)}\bar{\mathbf{x}}'\mathbf{C}'\left[\mathbf{C}\left(\sum_i \mathbf{A}_i'\mathbf{S}_i^{-1}\mathbf{A}_i\right)^{-}\mathbf{C}'\right]^{-1}\mathbf{C}\bar{\mathbf{x}}$	$F_{\alpha,(p-1),df_2}$
$F_3 = \frac{1}{(p-1)}\bar{\mathbf{x}}'\mathbf{C}'\left[\mathbf{C}\left(\widehat{m}^{@}\right)^{-}\mathbf{C}'\right]^{-1}\mathbf{C}\bar{\mathbf{x}}$	$F_{\alpha,(p-1),df_3}$
$F_4 = \frac{\lambda_4}{(p-1)}\bar{\mathbf{x}}'\mathbf{C}'\left[\mathbf{C}\left(\widehat{m}^{@}\right)^{-}\mathbf{C}'\right]^{-1}\mathbf{C}\bar{\mathbf{x}}$	$F_{\alpha,(p-1),df_4}$

Methodology

The design of the simulation had three between-subject factors and three within-subjects factors. The between subjects-factors were number of variables (p), ratio of the number of subjects to number of variables (n/p), and correlation (ρ) for each pair of variables. The number of variables factor had three levels, p = 2, 4 and 6. The ratio factor had two levels, n/p = 5 and 10. The actual sample sizes are presented in Table 2.

The correlation factor had three levels, ρ = .25, .50, and .75 with all pairs of variables equally correlated (compound symmetric). The within-subjects factors were type of missing data mechanism (type), percent of missing data (percent), and test statistic (test). The type of missing data mechanism factor had three levels: MAR, MCAR, and MNAR. The percent of missing data factor had two levels: 10% and 20%. Finally, the test factor has two levels: HLM and KR. All factors in the design were crossed.

Table 2. Sample Size (n) According to Number of Variables and Sample Size Ratio (n/p).

	Variable		
Ratio	2	4	6
5	10	20	30
10	20	40	60

The model used to generate the data was

$$X_{ij} = \mu + e_{ij},$$

i = 1, 2, …, n and j = 1, 2, …, p. In matrix terms

$$\begin{bmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{ip} \end{bmatrix} = \begin{bmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{bmatrix} + \begin{bmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{ip} \end{bmatrix}$$

where $\mathbf{x}$ is a $p \times 1$ vector containing the random variables for the i^{th} subject on the p variables and $\boldsymbol{\mu}$ is a $p \times 1$ vector of means, with all elements equal. All of the means are equal because the study is concerned with Type I error rates. The common element was arbitrarily set to zero. The vector $\mathbf{e}$ is a $p \times 1$vector of random errors with the following assumption, $\mathbf{e} \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$. In all conditions the diagonal elements of $\boldsymbol{\Sigma}$ were equal to one.

All data simulations were conducted using SAS version 9.0. For each combination of levels of the between-subjects factors, the following steps were used to simulate the data.

1. Simulate $\mathbf{Z}$, a $n \times p$ matrix of pseudorandom standard normal variables.
2. Calculate $\mathbf{T}$, the $p \times p$ upper triangular Cholesky factor of the covariance matrix $\boldsymbol{\Sigma}$.
3. Calculate $\mathbf{E} = \mathbf{ZT}$, an $n \times p$ matrix of error scores.
4. Set $\mathbf{X} = \mathbf{E}$
5. Copy $\mathbf{X}$ five times, yielding six copies of the data set. The six copies were used to create data matrices with missing data for the six combinations of type of missing data mechanism and percentage of missing data.
6. Select data points for elimination. In all conditions there were no data missing on x_1.

a. For the MCAR missing data mechanism, x_{ij} was eliminated from the matrix if $U_{ij} < \pi$ where π is the expected proportion of missing data on x_j.

b. For an MAR missing data mechanism, x_{ij} was missing if

$$U_{ij} < \Phi(kx_{i1} + c),\ j = 2, \ldots, p$$

where Φ is the cumulative standard normal distribution function and k and c are parameters that control the dependence of the missing data on the x variables and the expected proportion of missing data.

c. For the MNAR missing data mechanism, x_{ij} was deleted if

$$U_{ij} < \Phi(kx_{ij} + c).$$

That is, the probability that x_{ij} was missing depended on x_i. All conditions were replicated 5,000 times.

Setting *k* and c

The parameter *k* controls how dependent the missing data are on *x* in the MAR and MNAR conditions and was set equal to one. Let R_{ij} = 1 if X_{ij} is missing and zero otherwise. With *k* = 1, in the MAR conditions the biserial correlation between R_j and x_1 was .5 for *j* = 2, …, *p*; in an MNAR condition the biserial correlation between x_j and R_j was .5. Thus the missing data indicators depend fairly heavily on the *x* variables. With *k* = 1, the expected proportion of missing data on X_{ij} is dependent on *c*. In the procedure described in the preceding paragraphs the probability that R_{ij} = 1 is related to x_j or x_1 by a normal ogive (or probit model). Using well-known facts about the normal ogive model (see, for example, Lord & Novick, 1968, equations 16.9.3 and 16.9.4), it can be shown that

$$c = \sqrt{1+k^2}\{\Phi^{-1}(\pi)\}.$$

Thus, when *k* = 1,

$$c = \sqrt{2}\{\Phi^{-1}(\pi)\}.$$

For 10% and 20% missing data the expression becomes $c = -1.28\sqrt{2}$ and $c = -.84\sqrt{2}$, respectively.

Results

For each combination of the between-subjects factors (number of variables, correlation, and sample size) and the within-subjects factors (missing data mechanism, percent of missing data, and type of test) the Type I error rates for the HLM and KR tests were estimated as the proportion of the 5000 replications that resulted in a significant test statistic. This proportion variable was then analyzed by a 3 (number of variables) × 3 (correlation) × 2 (sample size ratio) × 3 (missing data mechanism) × 2 (percent of missing data) × 2 (test) ANOVA with missing data mechanism, percent of missing data, and test type as within-subjects factors. The main effect of test was significant with $F(1, 4) = 1066.70$, $p = .000$. The mean Type I error rates for the two tests were M_{HLM} = .083 and M_{KR} = .065. Inspection of the estimated Type I error rates indicated that, with the exception of four conditions, the estimated Type I error rate for the KR test was closer to the true Type I error rate than was the Type I error rate for the HLM test. Consequently, results for the HLM test statistic were dropped from the model and Type I error rates for the KR test statistic were reanalyzed.

The new analysis showed no significant effects for correlation. The highest-order significant interaction was the interaction of missing data mechanism, percent missing data, and sample size ratio, $F(2, 8) = 15.58$, $p = .002$.

In addition the main effect of number of variables was significant, $F(2, 4) = 23.10$, $p = .006$. Because of this pattern of effects we present, in Table 3, the Type I error rates averaged over levels of the correlation factor. Bradley (1978) presented a conservative and liberal criterion for identifying conditions in which hypothesis testing procedures work adequately. His conservative criterion is $.9\alpha \le \tau \le 1.1\alpha$ ($.045 \le \tau \le .055$) and his liberal criterion is $.5\alpha \le \tau \le 1.5\alpha$ ($.025 \le \tau \le .075$). For this study, the liberal criterion was used to identify conditions in which the average Type I error rate was unacceptable. These are indicated in bold in Table 3.

Inspection of the results indicates that, as expected, Type I error rates for the KR test may be unacceptable when the missing data mechanism is MNAR. It appears that the error rate for the KR test is more likely to be unacceptable as the percent of missing data, sample size ratio, and number of variables increases. In regard to the effect of the number of variables, in our simulation the number of variables on which data were MNAR increased as the number of variables increased. Different results might have emerged if there had been missing data on only one of the variables, regardless of the number of variables.

When the data were MCAR or MAR, average Type I error rates were acceptable in all conditions. Inspection of the Type I error rates for individual cells in the design (i.e., not collapsing over correlation) indicated that when the data were MCAR or MAR, the Type I error rate was acceptable in all conditions. Reanalysis of the data, after dropping the results for MNAR conditions indicated that number of variables did not have a significant main effect and did not enter into any significant interactions.

Table 3. Mean Type I Error Rates for KR by Number of Variables, Sample Size Ratio, Percent of Missing Data, and Missing Data Mechanism

Number	Ratio	Percent	MCAR	MAR	MNAR
2	10	10	0.051	0.050	0.053
		20	0.061	0.052	0.068
	20	10	0.048	0.050	0.066
		20	0.049	0.057	**0.098**
4	10	10	0.053	0.049	0.063
		20	0.055	0.060	0.072
	20	10	0.052	0.054	0.072
		20	0.050	0.061	**0.146**
6	10	10	0.051	0.048	0.060
		20	0.058	0.059	**0.096**
	20	10	0.050	0.054	**0.082**
		20	0.052	0.062	**0.184**

Note. Each mean Type I error rate is an average of Type I error rates for three conditions. Unacceptable mean Type I error rates are in boldface.

The only significant effects were a two-way interaction of type of missing data and sample size ratio, $F(1,4) = 8.25$, $p = .045$, and a main effect of percent of missing data, $F(1,4) = 15.45$, $p = .017$. Average Type I error rates by type of missing data and sample size are presented in Table 4.

The results suggest that increasing the sample size ratio improves control of the Type I error rate when the data are MCAR, but not when the data are MAR. The means when 10% and 20% of the data were missing and the mechanism was MCAR or MAR were .051 and .056, respectively, suggesting that Type I error rate control declines as the percentage of missing data increases.

Table 4 Mean Type I Error Rates for KR by Sample Size Ratio and Missing Data Mechanism.

Ratio	MCAR	MAR
10	0.055	0.053
20	0.050	0.056

Conclusion

The aim of this study was to determine whether, when there are missing data and the sample size is small, using ML estimates of the means for a single factor repeated measures design in testing the omnibus hypothesis results in control of the Type I error rate. The specific methods used to test the hypothesis were the KR test and the HLM test as implemented in SAS. The results clearly showed that KR test provided better control of the Type I error rate than did the HLM test.

The results of this study support the conclusion that, in a single-factor repeated measures design, sampling distribution based inferences on the means using the KR test may not control the Type I error rate for the MNAR missing data mechanisms but do control the Type I error rate for the MCAR and MAR missing data mechanisms. However, sample size and percent of missing data may be key factors that affect ML based inferences for MCAR and MAR missing data conditions using the KR test.

For both MCAR and MAR data, the results suggest that increasing the percent of missing data tends to inflate the Type I error rates. The effect of increasing the sample size depended on the missing data mechanism, with a stabilizing effect when the data were MCAR, but not when the data were MAR.

Although the design investigated in this study was a simple one factor repeated measures design, the findings suggest further simulation work on using ML to directly estimate models with missing data with more complicated designs and with additional variation in the factors investigated in this study. One condition that can be introduced is a between-subjects factor. Designs with between-subjects factors and within-subjects factors, also known as split-plot designs, are even more common than the one investigated in this study. Split-plot designs are used in longitudinal studies with two or more treatment groups. In such designs, the number of time point at which observations are made may be larger than six, which is the largest number of measurements investigated in this study. Consequently, a repeated measures factor, with more levels than six, should be investigated in future work.

Although several correlation matrices were used in this study and the correlation matrix had little or no impact on the Type I error rate, in each correlation matrix the off-diagonal elements were the same (i.e., the matrices were compound symmetric). This type of matrix may occur in studies in which the levels of the within-subjects factor are treatments and the order of the treatment has been randomized. Nevertheless, the exclusive use of compound symmetric correlation matrices may have limited the generality of the results. And, in other repeated measures studies (e.g., longitudinal studies) the correlation matrix is not likely to be compound symmetric. Thus, another condition that can be fruitfully investigated in future work is correlation matrices that have varying off-diagonal elements.

The Type I error rates of the KR test were acceptable in both the MCAR and MAR conditions. However, the percent of missing data at which the KR test will begin to breakdown is still not clear, nor is it clear whether sample sizes larger than those studied in this research will improve the Type I error rate for the KR test applied to MAR data. Consequently, future work should increase both the sample size ratio and percent of missing data beyond what was used in this study.

Last, recall that in the MAR missing data mechanism the missing data pattern on one variable is related to or dependent on another variable in the model but not to the variable itself. Therefore, one question that can be asked is how does the KR test statistic perform with different degrees of dependence? So another condition that can be investigated in future work is different degrees of dependence for the MAR condition.

References

Albert, P. S., & Follmann, D. A. (2000). Modeling repeated count data subject to information dropout. *Biometrics, 56*, 667-677.

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

Diggle, P. D., & Keward, M. G. (1994). Informative dropout in longitudinal data analysis. *Applied Statistics, 43*, 49-93.

Efron, B., & Hinkley, D. V. (1978). Assesing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika, 65*, 457-481.

Fai, H. T., & Cornelius, P. L. (1996). Approximate F-tests for multiple degree of freedom hypotheses in generalized least squares analyses of unbalanced split-plot experiments. *Journal of Statistical Computation and Simulation, 54*, 363-378.

Fitzmaurice, G. M., Laird, N. M., & Shneyer, L. (2001). An alternative parameterization of the general linear mixture model for longitudinal data with non-ignorable drop-outs. *Statistics in Medicine, 20*, 1009-1021.

Harville, D. A., & Jeske, D. R. (1992). Mean squared error of estimation or prediction under a general linear model. *Journal of the American Statistical Association, 87*, 724-731.

Heitjan, D. F. (1994). Ignorability in general incomplete-data models. *Biometrika, 81*, 701-708.

Kenward, M. G. (1998). Selection models for repeated measurements with non-random dropout: An illustration of sensitivity. *Statistics in Medicine, 17*, 2723-2732.

Kenward, M. G., & Molenberghs, G. (1998). Likelihood based frequentist inference when data are missing at random. *Statistical Science, 13*, 236-247.

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics, 53*, 983-997.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association, 90*, 1112-1121.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: John Wiley & Sons.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Menlo Park, CA: Addison-Wesley.

McBride, G. B. (2002). Statistical methods helping and hindering environmental science and management. *Journal of Agricultural, Biological, and Environmental Statistics*, 7, 300-305.

Rubin , D. B. (1976). Inference and missing data. *Biometrika, 63*, 581-592.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.

Schaalje, G. B., McBride, J. B., & Fellingham, G. W. (2002). Adequacy of approximations to distributions of test statistics in complex mixed linear models. *Journal of Agricultural, Biological, and Environmental Statistics, 7*, 512-524.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research, 33*, 545-571.

Troxel, A. B. (1998). Analysis of longitudinal data with non-ignorable non-monotone missing values. *Applied Statistics, 47*, 425-438.

Verbeke, G., & Molenbergh, G. (2000). *Linear mixed models for longitudinal data*. New York: Springer-Verlag.

Wolfinger, R. D., & Chang, M. (1995). *Comparing the SAS GLM and Mixed procedures for repeated measures*. Proceedings of the Twentieth Annual SAS Users Groups Conference.