INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.
Using multiple imputation in Dunnett's multiple comparison procedure for the one-way repeated measures model

Yamashita, Darryl Tadao, Ph.D.

The Ohio State University, 1993
USING MULTIPLE IMPUTATION IN DUNNETT’S MULTIPLE COMPARISON PROCEDURE FOR THE ONE-WAY REPEATED MEASURES MODEL

DISSERTATION

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Darryl Tadao Yamashita, B.A., M.S.

* * * * *

The Ohio State University

1993

Dissertation Committee:
Professor Elizabeth A. Stasny
Professor Saul Blumenthal
Professor Jason C. Hsu
Professor Steven N. MacEachern

Approved by

[Signature]
Adviser
Department of Statistics
To my wife,

Carrie.
ACKNOWLEDGMENTS

I would like to thank Professor Jason Hsu for introducing the problem to me and his help in the computer programming. Professors Saul Blumenthal and Steven MacEachern merit thanks for their advice throughout this research. I wish to express my deepest gratitude to Professor Elizabeth Stasny for her guidance and dedication in serving as my adviser. Her never-ending support and expertise made this research possible. Finally, I wish to thank my family and friends for their support throughout my graduate career, especially my wife, Carrie, for her love and support (and her grammatical skills).
VITA

September 11, 1965 ............................... Born: Greenbrae, California

1987 ............................................................. B.S., University of California at Berkeley
Berkeley, California

1990 ............................................................. M.S., The Ohio State University
Columbus, Ohio

1990–Present ................................. Research Assistant
The Ohio State University
Columbus, Ohio

FIELDS OF STUDY

Major Field: Statistics

Studies in survey methodology and nonresponse.
TABLE OF CONTENTS

DEDICATION ............................................................................................................................... ii

ACKNOWLEDGMENTS ............................................................................................................. iii

VITA ................................................................................................................................................ iv

LIST OF TABLES ........................................................................................................................ viii

LIST OF FIGURES .................................................................................................................... x

CHAPTER PAGE

I Introduction ............................................................................................................................ 1

1.1 One-way analysis of variance ........................................................................ 4
1.1.1 Frequentist perspective of the one-way layout ............................. 5
1.1.2 Bayesian perspective of the one-way layout .............................. 6

II Background on the method of multiple imputation ........................................ 8

2.1 Introduction ............................................................................................................ 8
2.2 Missing data mechanism ..................................................................................... 10
2.3 The method of multiple imputation ............................................................. 12
2.4 Distributions for a small number of repetitions .................................... 15
2.4.1 Q scalar .......................................................................................................... 15
2.4.2 Q Multidimensional .................................................................................... 18
2.4.3 Multiple imputation procedure for a multivariate normal distribution. 22
2.5 Multiple imputation in linear models ......................................................... 23
2.6 Multiple imputation for the one-way layout ............................................. 26
III  The one-way repeated measures model ................................................................. 36
   3.1  Introduction .......................................................................................................... 36
   3.2  Frequentist-based inferences for the one-way repeated measures model 37
   3.3  Review of Bayesian-based one-way repeated measures model .......................... 39
       3.3.1  The likelihood function ............................................................................. 39
       3.3.2  The prior and posterior distributions ...................................................... 41

IV  Multiple comparison procedures .............................................................................. 44
   4.1  Introduction .......................................................................................................... 44
   4.2  Fixed-effects linear models ..................................................................................... 45
       4.2.1  Scheffe’s multiple comparison procedure ................................................. 46
       4.2.2  Tukey’s multiple comparison procedure ................................................... 47
       4.2.3  Dunnett’s multiple comparison procedure ................................................. 48
   4.3  Multiple comparison procedures for the one-way repeated measures model ................................................................. 49

V  An imputation procedure for the repeated measures model .................................. 54
   5.1  One-way repeated measures multiple imputation procedure ........................... 54
   5.2  The repeated measures normal imputation procedure ...................................... 56
   5.3  Simulation of a repeated measures data set with missing data .......................... 61

VI  Modified Dunnett’s multiple comparison procedure using multiple imputation .... 69
   6.1  Special case: one missing value per treatment .................................................... 71
   6.2  Simulations .......................................................................................................... 75
       6.2.1  Simulations for the special case: one missing value per treatment ............ 79
       6.2.2  Simulations for the random pattern of missing data ................................. 83
       6.2.3  Simulations for a specific pattern of missing data ....................................... 85
       6.2.4  The fourth set of simulations ..................................................................... 88

VII  Conclusions and further research .......................................................................... 92
    7.1  Conclusions ......................................................................................................... 92
    7.2  Further research .................................................................................................. 97
APPENDICES

A  Uncommon distributions ................................................................. 102
   A.1 The Studentized range distribution ................................. 102
   A.2 The multivariate T-distribution ................................. 103

B  Programs used in the simulations .................................................. 105
   B.1 The dunnett.f program ........................................... 106
   B.2 The conf.f program ........................................... 116

BIBLIOGRAPHY ................................................................. 119
## List of Tables

<table>
<thead>
<tr>
<th>TABLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>64</td>
</tr>
<tr>
<td>2</td>
<td>80</td>
</tr>
<tr>
<td>3</td>
<td>82</td>
</tr>
<tr>
<td>4</td>
<td>82</td>
</tr>
<tr>
<td>5</td>
<td>84</td>
</tr>
<tr>
<td>6</td>
<td>84</td>
</tr>
<tr>
<td>7</td>
<td>86</td>
</tr>
<tr>
<td>8</td>
<td>87</td>
</tr>
<tr>
<td>9</td>
<td>88</td>
</tr>
<tr>
<td>10</td>
<td>89</td>
</tr>
</tbody>
</table>
11 Ratio of $S^2$ to $2MSE/n$ when $n = 30$, $t = 5$, and $(\sigma_r^2, \sigma_l^2) = (0.8, 0.2)$, (9,1), (1,9). ......................................................... 90

12 The largest ratio of half-widths and the ratio of the number of observations for the four sets of simulations. .................................................... 95
# List of Figures

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>66</td>
</tr>
<tr>
<td>2</td>
<td>67</td>
</tr>
<tr>
<td>3</td>
<td>68</td>
</tr>
</tbody>
</table>

1. Plots of the estimators for $\theta_i, i = 1, \ldots, 5$, using the actual data (horizontal) versus the imputed data (vertical) over the 5000 replications.

2. Plot of the estimator for $\sigma^2$ using the actual data (horizontal) versus the imputed data (vertical) over the 5000 replications.

3. Plot of the estimator for $\sigma^2$ using the actual data (horizontal) versus the imputed data (vertical) over the 5000 replications.
Dunnett's multiple comparison procedure is a well-known method for constructing simultaneous confidence intervals to compare the various treatments with a control for fixed-effects linear models. The procedure depends on the calculation of a critical point of a multidimensional $T$-distribution which amounts to a numerical approximation of a multidimensional integral. As few as five or six treatments can make this integration computationally prohibitive. For certain models, this high dimensional integration can be reduced to a two dimensional integral. This occurs when the design is balanced with respect to treatments; balanced designs will be discussed in more detail in Chapter 4. Many designs are balanced with respect to treatments if they have the same number of observations for each treatment. Even in this situation, the computation of the critical value takes a few seconds of computer time.

The balanced one-way repeated measures model is a two-way mixed model with one fixed effect and one random effect. The design requires that each subject is given each treatment exactly once. Here we call the design balanced because the generalized least squares estimates are equivalent to the ordinary least squares estimates. If the generalized least squares estimates are not equal to the ordinary least squares estimates, then the design is called unbalanced. Dunnett's multiple comparison pro-
procedure can be applied to the balanced one-way repeated measures model because (1) the design is balanced with respect to treatments, and (2) the variance of the random effect algebraically drops out of the confidence interval. This is shown in Chapter 4.

If the one-way repeated measures model is unbalanced, then both (1) and (2) from the above paragraph may be invalid. In this instance, Dunnett's multiple comparison procedure is also inappropriate. A unbalanced design can occur when an experimenter begins with a balanced one-way repeated measures design, but some of the observations are unavailable or missing.

We propose using a technique introduced by Rubin (1986) called multiple imputation to create confidence intervals based on Dunnett's multiple comparison procedure for the balanced one-way repeated measures model with missing data. Multiple imputation is a technique used in survey sampling to find estimates from a survey with missing data by replacing the missing data with plausible values to create a number of completed-data sets. These completed-data sets, also known as multiply-imputed data sets, can be used to make inferences similar to inferences possible from the complete-data set (inferences if there were no missing data). The method of multiple imputation was extended to the fixed-effects linear models by Schenker and Welsh (1988) and to multivariate normal distributions by Rubin and Schafer (1990).

We will develop a multiple imputation procedure, calling it the repeated measures normal procedure (RMNI), to generate the multiply imputed values for the missing data and calculate the multiple imputation estimates. We will then use the estimates from the RMNI procedure to derive simultaneous confidence intervals based on the
Dunnett's multiple comparison procedure.

We will start the discussion with some background material about multiple imputation procedures in Chapter 2. We will discuss why multiple imputation is a viable technique, the parameter estimates for multiple imputation procedures, and the properties of the estimates. We will also give examples of various multiple imputation procedures and show the expectation of estimates from a multiple imputation procedure for the one-way analysis of variance model.

In Chapter 3, we will present background material for the balanced one-way repeated measures model from both the frequentist and Bayesian perspectives. The frequentist perspective is needed to develop the estimates and their distributions used in Dunnett's multiple comparison procedure. The posterior distributions of the parameters from the Bayesian perspective will be used in the multiple imputation procedure.

In Chapter 4, we will introduce Scheffé's, Tukey's, and Dunnett's multiple comparison procedures for fixed-effects linear models. We will also show how Dunnett's multiple comparison procedure can be extended to the balanced one-way repeated measures model and why it cannot be extended to the unbalanced one-way repeated measures model.

In Chapter 5, we will propose the RMNI procedure for the balanced one-way repeated measures model with missing data. A simulation study is presented to compare the actual estimates with estimates from the multiple imputation procedure.

In Chapter 6, we will show how estimates from the RMNI procedure can be applied
to Dunnett's multiple comparison procedure. Several sets of simulation studies are presented for different proportions of missing data, numbers of imputations, values of the variances, numbers of treatments, and numbers of subjects. Each set of simulations has a different pattern of missing data, starting with the simplest case—one missing observation from each treatment—to a censored data response mechanism which leads to almost 50% of the last treatment missing. We will use the simulations to estimate the $\alpha$-level and the half-widths for the multiply-imputed confidence intervals and to see what effect the two variances have on the widths of the multiply-imputed confidence intervals.

Chapter 7, presents the conclusions of our research and suggests further extensions of this work.

Appendix A lists distributions not commonly found in the literature which are used in the dissertation. The two distributions listed are the Studentized range distribution and the multivariate $T^*$-distribution which are used in Chapter 4.

Appendix B lists the two FORTRAN programs used in the simulation studies in Chapters 5 and 6.

The rest of this chapter will cover basic properties of the estimators from the one-way analysis of variance model. This model will be used as an example throughout the dissertation.

1.1 One-way analysis of variance

The one-way analysis of variance model, also referred to as the one-way layout, is given in many experimental design texts, such as Scheffé (1959). The one-way layout
with \( t \) treatments and \( n_i \) observations in the \( i \)-th treatment has a model of the form

\[
Y_{ij} = \theta_i + \epsilon_{ij}, \quad i = 1, \ldots, t, \quad j = 1, \ldots, n_i,
\]  
(1.1)

where the \( \theta_i \)'s are the fixed (or treatment) effects, and \( \epsilon_{ij} \)'s are the error terms which are assumed to be independent and identically distributed \( N(0, \sigma^2) \) random variables.

For the one-way layout, we will call the design balanced if \( n_i = n_k \), for all \( i \neq k \), or unbalanced if the \( n_i \) are not all the same.

Two different views exist for making inferences about the parameters in (1.1) and both are used in the multiple imputation theory. We will refer to these views as the frequentist perspective and the Bayesian perspective. In the frequentist perspective, the theoretical distributions of the estimators are used to make inferences about the parameters. In the Bayesian perspective, posterior distributions are used to make inferences. Under the frequentist perspective, we will find properties for the estimators; under the Bayesian perspective, we will use a Jeffrey's prior which will lead to estimators from the posterior distributions which are the same as the frequentist estimators.

\subsection{1.1.1 Frequentist perspective of the one-way layout}

In the frequentist perspective in the design of experiments context, the least squares estimates are often used. The least squares estimates for \( \theta_1, \ldots, \theta_t \), and \( \sigma^2 \) are

\[
\hat{Y}_i = \bar{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}, \quad i = 1, \ldots, t
\]
(1.2)

and

\[
\hat{s}^2 = \frac{1}{\sum_i (n_i - 1)} \sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2,
\]
(1.3)
respectively. In this text we will assume that a dot, ",, in the index represents the averaging over that index. For example,

\[ Y_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} \]  

(1.4)

and

\[ Y_{..} = \frac{1}{t \sum_{i} n_i} \sum_{i=1}^{t} \sum_{j=1}^{n_i} Y_{ij}. \]  

(1.5)

The distributions of the estimators for \( \theta_i, i = 1, \ldots, t \), are independent normal distributions

\[ \hat{Y}_i \sim N(\theta_i, \sigma^2/n_i), \quad i = 1, \ldots, t, \]  

(1.6)

and the estimate for \( \sigma^2 \) is unbiased, independent of \((\hat{Y}_1, \ldots, \hat{Y}_t)\), and has the distribution

\[ \hat{s}^2 \sim t(n-1) \sigma^2 / \sum_{i=1}^{t} (n_i - 1). \]  

(1.7)

1.1.2 Bayesian perspective of the one-way layout

In the Bayesian perspective, different priors lead to different estimators. If the design is balanced (which implies that \( n = n_1 = \cdots = n_t \)), Rubin (1987) suggests using the Jeffries prior distribution

\[ [\theta_1, \ldots, \theta_t, \sigma^2] \propto \sigma^{-2}. \]  

(1.8)

Using the prior (1.8), Box and Tiao (1973) show that the posterior distribution of \( \theta \), conditional on \( \sigma^2 \) (written as \([\theta \mid \sigma^2, Y])\), is a multivariate normal

\[ [\theta \mid \sigma^2, Y] \sim N_t((Y_1, \ldots, Y_t)', \sigma^2 I_t). \]  

(1.9)
Box and Tiao also show that the posterior distribution of $\sigma^2$, given $Y$, is

$$[\sigma^2 \mid Y] \sim S^2 \chi_{(n-1)}^{-2}. \quad (1.10)$$

The prior defined in (1.8) leads to the same estimators as those in the frequentist perspective. The posterior distributions will be used in the imputation routine, and the estimators from the frequentist perspective are the estimators used in the multiple imputation procedure described in Chapter 2.
CHAPTER II

Background on the method of multiple imputation

In this chapter, we introduce and discuss the method of multiple imputation. Multiple imputation was first developed in the survey sampling context, and has been extended to other fields such as fixed-effects linear models and linear regression. We will discuss different multiple imputation procedures which have been developed for various models. We will also discuss how multiple imputation can be used in the one-way layout with missing data and calculate the expectations of the multiple imputation estimators.

2.1 Introduction

In complex survey designs, analyzing data can be very difficult with the presence of missing data. If the missing data are item non-response—that is, a returned survey is incomplete (not fully filled out)—imputation, which is described below, can help the data analyst around this problem. Multiple imputation has been developed in the survey sampling literature (e.g. Rubin (1986), Heitjan and Little (1988), and Schenker, Treiman, and Weidman (1988)), but has been extended to fixed-effects linear models by Schenker and Welsh (1988). We will continue in this direction by
developing a multiple imputation procedure for the one-way repeated measures model.

The notion of single imputation is a simple one. Replace every missing value with some “good guess” and analyze the data using standard complete-data methods. A good guess is “a revised record close to what a respondent would have reported were there no errors” (Greenberg and Surti, 1984).

The use of single imputation is common in survey sampling. In the 1982 Economic Census conducted by the Bureau of the Census, a complex “hot-deck” procedure was used to obtain values to impute in place of the missing data (Greenberg and Petkunos, 1987). Hot-deck imputation is a method which substitutes data from individuals with similar attributes for data of non-respondents. Another common imputation method is mean imputation. Mean imputation substitutes the mean of the responding units in place of each of the missing units. Rubin and Little (1987) discuss these and many other types of imputation procedures.

In the context of analysis of variance, imputation techniques have been developed so that balanced-design methods may be used to analyze the data. Yates (1933) proposed a method which yields correct estimates of the residual sums of squares. Bartlett (1937) developed a method which finds the values to impute that yield the correct least squares estimates. Rubin (1972) developed a method which finds the same values as Bartlett’s Method which can be implemented using existing full data computer packages. However, these techniques are only applicable for fixed-effects models.

The main disadvantage to single imputation is that it tends to underestimate the
variance of the estimator. Single imputation treats the imputed values as known, leading to less variability in the data.

Multiple imputation is a method introduced by Rubin (1978) which tries to solve the problem of underestimating the variance by including the variability of the imputation. Although the application of multiple imputation is straightforward, the underlying theory can be quite complicated. The idea of multiple imputation is to use a non-deterministic, single imputation method a number of times and to take the average of the estimates over all of the single imputations. In order to discuss multiple imputations in the context of survey sampling where it was developed, some background information and notation are needed.

### 2.2 Missing data mechanism

Little and Rubin (1987) discuss different types of sampling and response mechanisms in survey sampling. In multiple imputations, assumptions about the response mechanism partly determines the multiple imputation procedure used.

Let $X_1, X_2, \ldots, X_N$ be the vectors of covariates which are fully known to the analyst, and $Y_1, Y_2, \ldots, Y_N$ be the vectors of data of interest of the $N$ members in the population not known to the analyst. In the context of survey sampling, the covariates could be geographical area, stratum indicators, job industry codes, etc., which are typically known before the sample is taken. To simplify notation, let $X = (X_1, X_2, \ldots, X_N)$ and $Y = (Y_1, Y_2, \ldots, Y_N)$. From this population, we wish to take our sample.

In the case where $Y_i$ is scalar, define an indicator variable $I = (I_1, I_2, \ldots, I_N)$,
such that $I_i = 1$ if the $i$th member of the population is chosen to be in the sample, and $I_i = 0$ if the $i$th member is not chosen to be in the sample. Also, let $\mathbf{R} = (R_1, R_2, \ldots, R_N)$ be an indicator variable such that $R_i = 1$ if the $i$th member will respond to the survey, and $R_i = 0$ if the $i$th member will not respond to the survey. The distribution $[I | \mathbf{X}, \mathbf{Y}, \mathbf{R}]$ is called the sampling mechanism and the distribution $[\mathbf{R} | \mathbf{X}, \mathbf{Y}]$ is called the response mechanism where, for example, $[I | \mathbf{X}, \mathbf{Y}, \mathbf{R}]$ represents the conditional distribution of $I$ given $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{R}$. Note that if $Y_i$ is vector valued, then the notion of $I_i$ and $R_i$ can be extended to vectors of indicators.

The sampling mechanism is said to be unconfounded with $(\mathbf{Y}, \mathbf{R})$ if and only if

$$[I | \mathbf{X}, \mathbf{Y}, \mathbf{R}] = [I | \mathbf{X}].$$

(2.1)

An example of an unconfounded sampling mechanism is stratified random sampling. In stratified random sampling, the population is partitioned by some covariate, and a simple random sample is taken from each partition. Thus, the probability that an individual is chosen to be in the sample depends only on the number of individuals in the same stratum. In survey sampling, the sampling mechanism usually drives the form of the estimators. This is accomplished by using known covariates to establish the probability of inclusion. So, typically the sampling mechanism is unconfounded with $\mathbf{Y}$ and $\mathbf{R}$ and is usually only indirectly involved in the multiple imputation method through the estimators.

A response mechanism is said to be unconfounded if $[\mathbf{R} | \mathbf{X}, \mathbf{Y}] = [\mathbf{R} | \mathbf{X}]$. Rubin (1976) refers to an unconfounded response mechanism as missing at random (MAR), which is a special case of ignorable nonresponse. If the response mechanism is not
unconfounded, it is known as *nonignorable nonresponse*. The multiple imputation theory usually assumes that the response mechanism is MAR, but adjustments can be made to the method of imputation to allow for nonignorable nonresponse (see Rubin, 1987).

Define $Y_{inc} = \{Y_i : I_i = 1\}$ as the set of $Y_i$ which are included in the sample. $Y_{inc}$ can be broken into two subsets: $Y_{obs} = \{Y_i : I_i = 1, R_i = 1\}$ and $Y_{mis} = \{Y_i : I_i = 1, R_i = 0\}$, so that $Y_{inc} = (Y_{obs}, Y_{mis})$. $Y_{obs}$ is the set of $Y_i$ which are actually observed, and $Y_{mis}$ is the set of $Y_i$ which are included in the sample but not observed because of nonresponse. Also, define $R_{inc} = \{R_i : I_i = 1\}$ as the set of response indicators for the individuals in the sample. Note that once the survey is completed, $Y_{obs}$, $R_{inc}$, and $I$ are known.

### 2.3 The method of multiple imputation

The theory behind multiple imputation can be seen most easily from the Bayesian perspective but is used to adjust survey sampling estimators from frequentist theory when missing data are present. Hence, both perspectives are used in the following discussion. The idea is to randomly draw values from the posterior distribution, $[Y_{mis} | X, Y_{obs}, R_{inc}, I]$, many times to "average out" $Y_{mis}$ from the posterior of the parameter of interest.

Using Rubin’s (1987) notation, let $Q$ be a $t$-dimensional vector of parameters of interest and let $\hat{Q}$ be the statistic which estimates $Q$ in the absence of missing data, such that

$$ (Q - \hat{Q}) \sim N(0, \Omega). \quad (2.2) $$
Let \( \hat{U} \) be a statistic which estimates the variance-covariance matrix of \((Q - \hat{Q})\). When no data are missing, we call \( \hat{Q} \) and \( \hat{U} \) the *complete-data* estimates. In the case of missing data, when values are imputed for the missing values, the resulting data set is called the *completed-data set*. The corresponding estimates, \( \hat{Q}_* \) and \( \hat{U}_* \), are called the *completed-data* estimates.

The multiple imputation method can be reduced to two steps. First, create \( m \) completed-data sets by generating values from the posterior distribution of the missing data, \( [Y_{mis} | X, Y_{obs}, R_{inc}, I] \). Second, use the completed-data sets to compute estimates using the complete-data estimates. These estimates are then used to obtain a single estimate and its estimated variance, which includes the variability due to the missing data.

Define \( S_m = \{ (\hat{Q}_{*k}, \hat{U}_{*k}) : k = 1, \ldots, m \} \), as the set of estimates from the \( m \) completed-data sets, where \( \hat{Q}_{*k} \) and \( \hat{U}_{*k} \) have the same functional form as the complete-data estimates of \( \hat{Q} \) and \( \hat{U} \), respectively. The resulting estimate of \( Q \) from the multiple imputation procedure is

\[
\overline{Q}_m = \frac{1}{m} \sum_{k=1}^{m} \hat{Q}_{*k}, \tag{2.3}
\]

and the estimator of the variance of \( Q_m \) is

\[
T_m = \overline{U}_m + (1 + m^{-1})B_m, \tag{2.4}
\]

where

\[
\overline{U}_m = \frac{1}{m} \sum_{k=1}^{m} \hat{U}_{*k} \tag{2.5}
\]
and

\[ \mathbf{B}_m = \frac{1}{m-1} \sum_{k=1}^{m} (\hat{Q}_{.,k} - \bar{Q}_m)(\hat{Q}_{.,k} - \bar{Q}_m). \quad (2.6) \]

\( \bar{U}_m \) is called the \textit{within-repetition variance} and \( \mathbf{B}_m \) is called the \textit{between-repetition variance}. \( \mathbf{B}_m \) accounts for the variability due to the missing data.

From the Bayesian perspective, the posterior distribution is \([Q \mid X, Y_{\text{obs}}, R_{\text{inc}}, I]\). However, the estimators are developed assuming complete data and use the posterior distribution \([Q \mid X, Y_{\text{inc}}, R_{\text{inc}}, I]\). Rubin (1987) shows, by using conditional expectations and the strong law of large numbers, that if the posterior distribution for the complete-data case is \([Q \mid X, Y, R, I]\), then

\[ E[Q \mid X, Y_{\text{obs}}, R_{\text{inc}}, I] = \bar{Q}_\infty. \quad (2.7) \]

and

\[ \text{Var}(Q \mid X, Y_{\text{obs}}, R_{\text{inc}}, I) = \bar{U}_\infty + B_\infty = T_\infty, \quad (2.8) \]

where \( \bar{Q}_\infty, T_\infty, \bar{U}_\infty, \) and \( B_\infty \) are the limits, as \( m \to \infty \), of (2.3), (2.4), (2.5), and (2.6), respectively. By Bayesian asymptotic theory,

\[ [Q \mid \bar{Q}_\infty, \bar{U}_\infty, B_\infty] \sim N(\bar{Q}_\infty, \bar{U}_\infty + B_\infty). \quad (2.9) \]

Rubin (1987) also shows that in the frequentist perspective,

\[ \bar{Q}_\infty \sim N(Q, T), \quad (2.10) \]

where

\[ T = U + B, \quad (2.11) \]

where \( B \) is the true between-imputation variance and \( U \) is defined in (2.2).
2.4 Distributions for a small number of repetitions

Multiple imputation was developed in the context of large-scale sample survey data. For these large surveys, it is computationally difficult to generate and store a large number of completed-data sets. The number of repetitions, \( m \), had to be kept small, such as in Clogg, Rubin, Schenker, Schultz, and Weidman (1991) who used multiple imputation to create just two completed-data sets for the 1970 census data. Therefore, it is difficult to justify inferences, such as confidence intervals, using (2.9). In contrast, ANOVA problems often have relatively few observations and the time needed to create each completed-data set is negligible. When possible, it is better to use a large number of imputations. Rubin (1987), along with others, assumes that the number of observations is sufficiently large so that a \( \overline{U}_m \) is a good estimate of \( U \). Different work has been done for \( Q \) scalar, so the discussion will focus on scalar values of \( Q \) separately from multivariate values of \( Q \).

2.4.1 \( Q \) scalar

For a finite number of repetitions, Rubin (1987) shows an approximation to the posterior distribution of \( Q \), when \( Q \) is scalar, to be

\[
[Q \mid S_m] \sim t_{\nu}(\overline{Q}_m, T_m),
\]

(2.12)

where

\[
\nu = (m - 1)(1 + r_m^{-1})^2,
\]

(2.13)

\[
r_m = (1 + m^{-1}) \frac{B_m}{U_m},
\]

(2.14)
and $t_n(\theta, \tau^2)$ is the $t$-reference distribution with $n$ degrees of freedom centered at $\theta$ with variance $\tau^2$.

Rubin and Schenker (1986) conducted a simulation study to find the coverage probabilities of confidence intervals produced by (2.12). In this simulation, five different methods of imputation (described below), four different sample sizes ($n=20, 50, 100, \text{ and } 1000$), and three different response rates ($r=.4, .6, \text{ and } .9$) were considered. Two different distributions (lognormal and Laplace) were used to simulate the data.

Let $n_{\text{obs}}$ be the number of individuals who responded and $n_{\text{mis}}$ be the number of individuals who did not respond. The five different types of imputation methods are:

1. **Simple Random Imputation.** This method draws values for imputation at random with replacement from the observed data.

2. **Bayesian Bootstrap Imputation.** Suppose each individual in the population had one of the following values, $d_1, \ldots, d_k$, with probability $\pi_1, \ldots, \pi_k$, respectively. Let $\pi = (\pi_1, \ldots, \pi_k)$. If $\pi$ has an improper Dirichlet prior proportional to $\prod_j \pi_j$, then the posterior of $\theta$ is proportional to $\prod_j \pi_j^{q_j-1}$, where $q_j$ is the number of times $d_j$ appears in the observed data. For each repetition, draw a value $\pi_*$ from the posterior of $\theta$. Then draw the values for imputation from $d_1, \ldots, d_k$, with probabilities from $\pi_*$.

3. **Approximate Bayesian Bootstrap Imputation.** Draw $n_{\text{obs}}$ values with replacement from the observed values. For the $n_{\text{mis}}$ values needed for the imputation, draw $n_{\text{mis}}$ values with replacement from the $n_{\text{obs}}$ previously sampled.
4. **Fully Normal Imputation.** Draw $\sigma_*^2$ from $(n_{\text{obs}} - 1)s^2/\chi^2_{n_{\text{obs}} - 1}$ and $\theta_*$ from $N(\bar{Y}, \sigma_*^2)$, where $\bar{Y}$ is the average and $s^2$ is the sample variance from the observed data. Draw values for imputation from $N(\theta_*, \sigma_*^2)$.

5. **Imputation Adjusted for Uncertainty in the Mean and Variance.** Draw $\theta_*$ and $\sigma_*^2$ as in the fully normal imputation method. Draw $n_{\text{mis}}$ values from the observed data with replacement and label them $(Y_1, \ldots, Y_{n_{\text{mis}}})$. Let

$$Z_j = \frac{Y_{j*} - \bar{Y}}{\sqrt{(n_{\text{obs}} - 1)s^2/n_{\text{obs}}}}.$$

(2.15)

The value used for the imputation is $\theta_* + \sigma_* Z_j$, for $j = 1, \ldots, n_0$.

The study showed that, for $m = 2$, inferences based on the $t$-distribution yielded fairly accurate results for all sample sizes when the response rate was 90%. But even with a sample size of $n = 1000$, the coverage when the response rate was 60% was lower than the nominal rate (nominal rates are the actual rates). Overall, the coverage probabilities tended to be lower than expected. The two imputation techniques which seemed to perform the best were the (1) fully normal and (2) imputation adjusted for means and variances methods. Theoretical coverage probabilities for finite $n$ and infinite $m$ show that, for normal data, the fully normal imputation and the imputation adjusted for means and variances methods were equivalent and better than the other three methods. The fully normal and imputation adjusted for means and variances methods had coverage rates at or within 1% of the actual rates.
2.4.2 Q Multidimensional

Rather than developing interval estimates for \([Q | X, Y_{\text{obs}}, R_{\text{inc}}, I]\) when \(Q\) is a \(t\)-dimensional vector, Rubin (1987) considers hypothesis testing. Let \(Q_0\) be the null-value of \(Q\). The associated \(p\)-value for \(Q_0\), given \((S_m, B_\infty)\), is

\[
p\text{-value}(Q_0 | S_m) = P \left\{ \chi^2_t > (Q_0 - \overline{Q}_m) \left[ \overline{U}_m + (1 + m^{-1})B_\infty \right]' \right\}.
\] (2.16)

This is much easier to analyze than confidence sets because the confidence sets would be \(t\)-dimensional. This case is reduced to one dimension.

Since \(B_\infty\) is unknown, the \(p\)-value associated with \([Q_0 | S_m]\) can be found by integrating with respect to \(B_\infty\).

\[
p\text{-value}(Q_0 | S_m) = \int P \left\{ \chi^2_t > (Q_0 - \overline{Q}_m) \left[ \overline{U}_m + (1 + m^{-1})B_\infty \right]' \right\} [B_\infty | S_m]dB_\infty.
\] (2.17)

However, \([B_\infty | S_m]\) is a Wishart distribution and performing the integration is difficult.

In the special case when \(B_\infty\) is proportional to \(T_\infty\) (which implies that the proportion of missing data is the same for each component of \(Q\)), Rubin (1987) shows that the \(p\)-value can be reduced to a single integration.

\[
p\text{-value}(Q_0 | S_m) = \int P \{ \chi^2_t > [1 + (1 + m^{-1})r_\infty]^{-1} \times (Q_0 - \overline{Q}_m)\overline{U}_m^{-1} (Q_0 - \overline{Q}_m)' \} [r_\infty | S_m]dr_\infty,
\] (2.18)
where \( r_\infty = B_\infty / \overline{U}_\infty \). Using a similar argument as in the scalar case, the p-value is approximately

\[
p\text{-value}(Q_0 \mid S_m) \approx P \left\{ \chi_t^2 > \chi^2_{\nu}(1 + r_m)^{-1}(Q_0 - \overline{Q}_m)\overline{U}_m^{-1}(Q_0 - \overline{Q}_m)' \right\}, \quad (2.19)
\]

where

\[
\nu = (m - 1)(1 + r_m)^2 \quad (2.20)
\]

and

\[
r_m = (1 + m^{-1})Tr(B_m \overline{U}_m^{-1})/t . \quad (2.21)
\]

Let

\[
\hat{D}_m = (1 + r_m)^{-1}(Q_0 - \overline{Q}_m)\overline{U}_m^{-1}(Q_0 - \overline{Q}_m)'/t \quad (2.22)
\]

be the test statistic when \( B_\infty \) is proportional to \( T_\infty \). (2.19) can be written as

\[
p\text{-value}(Q_0 \mid S_m) \approx P\{F_{t, \nu} > \hat{D}_m\}. \quad (2.23)
\]

When \( B_\infty \) is not proportional to \( T_\infty \), Rubin (1987) suggests using

\[
D_m = \frac{(Q_0 - \overline{Q}_m) \left[ \overline{U}_m + (1 + m^{-1})B_m \right]^{-1}(Q_0 - \overline{Q}_m)'}{t}, \quad (2.24)
\]

with the corresponding p-value for \( D_m \)

\[
p\text{-value}(Q_0 \mid S_m) \approx P\{F_{t, \nu} > D_m\}. \quad (2.25)
\]

Li, Raghunathan and Rubin (1991) studied the accuracy of significance levels for the coefficients in multiple regression with various levels of missing data. Through simulation studies, they found that single imputation methods yield poor results,
whereas multiple imputations yield good results. They also show that when the number of coefficients is large but the fraction of missing data is less than 0.2, only small values of \( m \) are necessary to yield good results. Rather than using \( D_m \) defined in (2.24), Li et al. suggest using the statistic

\[
D_m^* = \frac{(\beta_m - \beta_0)' U_m^{-1}(\beta_m - \beta_0)}{t(1 + r_m)},
\]

where \( t \) is the dimension of \( \beta \) and \( \beta \) is the vector of regression coefficients and

\[
r_m = (1 + m^{-1}) Tr(B_m U_m^{-1})/(m - 1).
\]

The associated p-value for \( D_m^* \) is

\[
P_m^* = P \{ F_t, w > D_m^* \},
\]

where

\[
w = 4 + (s - 4)[1 + (1 - 2s^{-1})/r_m]^2
\]

and \( s = t(m - 1) \). The results of Li et al. suggest that this statistic is more accurate than \( D_m \) for larger values of \( t \).

Here, the degrees of freedom, \( \nu \), in (2.20) is different from \( w \). Li, Raghunathan and Rubin claim that this \( w \) outperforms \( \nu \) when \( t \) is large. We include here a description of how Li et al. arrived at the value for \( w \) because we will use the same technique to approximate a statistic in this thesis. Li et al. show that \( D_m^* \) has the distribution

\[
D_m^* \sim \frac{t^{-1} \chi_t^2}{(1 + at^{-1} \chi_s^2)/(1 + a)},
\]

where

\[
a = (1 + m^{-1}) \lambda.
\]
\( \lambda \) is the ratio of missing to observed information which, in this case, is assumed to be equal for each component. The Chi-square distributions in the numerator and denominator are independent. Li et al. assume that the approximate distribution of (2.30), is a scaled \( F \)-distribution, \( \delta F_{t,w} \). The values of \( \delta \) and \( w \) are obtained by equating the first two central moments of \( D^*_m \) with the first two moments of \( \delta F_{t,w} \). Solving for \( \delta \) and \( w \) yields

\[
\delta = \frac{(1 - 2/w)[1 + as/(s - 2)]}{(1 - a)} \quad (2.32)
\]

and

\[
w = 4 + (s - 4)[1 + (1 + (1 - 2s^{-1})/a)^2], \quad (2.33)
\]

if \( s > 4 \). Li et al. note that for small values of \( a \) (which corresponds to a large number of imputations, \( m \)) and a small fraction of missing data, \( \lambda, \delta \) is close to 1, so that \( D^*_m \sim F_{t,w} \), where \( w \) is defined in (2.33). This implies that (2.33) can be rewritten as (2.29).

They also found that, as \( m \to \infty \), \( D^*_m \) has approximately the same power and level as the likelihood ratio test based on the observed data. To conduct an \( \alpha \)-level test of the hypothesis \( H_0 : \beta = \beta_0 \) vs. \( H_0 : \beta \neq \beta_0 \), one could use \( P_m < \alpha \) for the rejection criterion.

Through simulations, Li et al. found that the power of the test, when \( m = 4 \), is comparable to the power when \( m = \infty \), even for large values of \( t \) (they looked at \( t \leq 35 \)). They also showed that for various levels of (1) the dimension, \( t \), (2) missing information, and (3) the coefficient of variation (which measures the difference between the components of \( Q \)), the exact levels for \( D^*_\infty \) are within 1% of the nominal
level when the fractions of missing information are unequal.

2.4.3 Multiple imputation procedure for a multivariate normal distribution.

Many multiple imputation procedures for Q, both scalar and multidimensional, assume that observations are independent and univariate so that the imputed values can be drawn from a univariate distribution. However, the multivariate normal distribution can sometimes be an appropriate model for data which have a covariance structure. Rubin and Schafer (1990) developed an efficient multiple imputation procedure for models which assume the data to be distributed as a multivariate normal with an arbitrary variance-covariance matrix. We include this section because in Chapter 5 we will develop a different multiple imputation procedure for a multivariate normal distribution with a specific variance-covariance matrix. We will discuss the advantages and disadvantages of both procedures and propose another possible multiple imputation procedure which can be investigated in the future.

Assume that $Y$ is distributed as a $t$-dimensional multivariate normal with mean $\mu$ and variance-covariance $\Sigma$. Let $\theta = (\mu, \Sigma)$. The procedure arranges data such that (1) the data have a monotone pattern of missing data (as described in Little and Rubin, 1986), (2) uses Gibbs sampling to produce a sequence of $\theta_1, \theta_2, \ldots, \theta_M$, where $M < m$, (3) uses Sampling-Importance Resampling (discussed by Rubin, 1988) to randomly select $m$ values of $\theta$ from the $M$ produced in the previous step, and (4) creates values for imputation from a conditional multivariate normal distribution.

Rubin and Schafer (1990) rearrange the data to have a monotone pattern of miss-
ing data to increase the overall performance of the algorithm. For the Gibbs sampler, they transform $\theta$ into regression coefficients $(\phi_1, \ldots, \phi_t)$ using the sweep operator. The regression coefficients are defined as $\phi_j = (\beta_j, \sigma_j^2)$, where $\beta_j$ is a $j$-dimensional vector of regression coefficients and $\sigma_j^2$ is the corresponding variance of the residual.

Values from the Gibbs sampler can be drawn from either one long run or from many short runs. Rubin and Schafer do not explicitly choose one long run or many short runs, but the paper seems to imply that the values of $\theta$ are taken from one long run.

The Jeffries prior, $\pi_j(\phi_j) \propto \sigma_j^{-2}$ is applied to each $\phi_j$. The monotone pattern of missing data leads to a joint prior distribution as the product of the individual priors.

The Sampling-Importance Resampling part of the procedure allows the imputer to change the prior distribution to $\pi^*(\theta)$ by applying weights to each of the $M$ values of $\theta_i$. The weight for each $\theta_i$ is the ratio $\pi^*(\theta_i)/\pi(\theta_i)$.

The imputed values are drawn from the conditional normal distribution given $\theta_i$ and the observed data.

### 2.5 Multiple imputation in linear models

Although most of the literature for multiple imputations focuses on the survey sampling setting, a few studies of multiple imputation procedures for use in regression and fixed effects linear models have been developed. For fixed effect linear models, Schenker and Welsh (1988) develop asymptotic distributions of the multiple imputation estimators when the completed-data estimators meet certain asymptotic assumptions. Lazzeroni, Schenker and Taylor (1990) study the robustness of the estimators from four specific multiple imputation procedures for linear regression.
Schenker and Welsh (1988) studied the large-sample properties of multiple imputation methods in the context of the general linear model. They show that if the imputed values meet certain criteria, then the multiple imputation procedure will yield an estimator for the fixed effects which is consistent. Furthermore, they show that the between-imputation variance, $\hat{B}$, asymptotically converges to a Weibull distribution. Schenker and Welsh show, as examples, four different multiple imputation procedures which satisfy the necessary criteria. In particular, Schenker and Welsh introduce the Normal Imputation procedure which draws values from the posterior distributions of the parameters given a Jeffreys prior. This procedure, an extension of Fully Normal imputation discussed in Rubin and Schenker (1986), is described in Section 2.4.1. We will use a variation of the normal imputation procedure for the repeated measures normal imputation procedure discussed in Chapter 5.

Lazzeroni et al. (1990) evaluate the robustness of four different imputation procedures used in a multiple linear regression setting. All four models are based on the $t$-multivariate regression model

$$Y_i = X_i'\beta + e_i, \quad (2.34)$$

where

$$e_i \overset{iid}{\sim} N(0, \sigma^2). \quad (2.35)$$

Let $n_{obs}$ be the number of observed values. All four methods generate

$$\sigma^2 = \frac{SSE}{w}, \quad (2.36)$$

where $w \sim \chi^2_{n_{obs}-1}$ and $SSE$ is the residual sums of squares from the observed data,
and then generate

\[ \beta^* \sim N(\hat{\beta}, \sigma^* \sigma^{*2}(X'X)^{-1}), \]  

(2.37)

where \( \hat{\beta} \) is the vector of estimated parameters from the observed data. For each replication, \((\beta^*, \sigma^*)\) are drawn and imputation is conducted in four different ways:

1. **Model Based (MB).** For each missing value, impute

   \[ Y_i^* \sim N(X_i'\beta^*, \sigma^*). \]  

   (2.38)

2. **Residual Draw (RD).** For each missing value, draw, using simple random sampling with replacement, one value \((Y_0, X_0)\) from the observed data and let

   \[ Y_i^* = Y_0 + (X_i' - X_0')\beta^*. \]  

   (2.39)

3. **Predictive Mean Matching (PMM).** For each missing value, create a set of values from the observed data which have predictive means close to the predictive mean of the missing value. Draw one value from the set to be used for the imputation. This can be thought of as a hot-deck procedure.

4. **Local-Residual Draw (LRD).** Create a completed-data set as in PMM and, for each missing value, draw one value from the completed-data set, \(Y_0\), and, with the corresponding covariate, \(X_0\), use

   \[ Y_i^* = Y_0 + (X_i' - X_0')\beta^*. \]  

   (2.40)

   as in RD.
Using a sample size of \( n = 250 \) and number of repetitions, \( m \), of 5, 20, and 50, Lazzeroni et al. recorded the absolute value of the bias, variance, mean square error, and error coverage rate. Rather than using the multiple linear regression model, the following model was used to generate the data for the simulation,

\[
Y_i^\lambda = X_i'\beta + e_i, \tag{2.41}
\]

with \( \sigma^2 = 1, 2 \) and \( \lambda = 1, 1/4 \). When \( \zeta = 1, \beta = (10,1,1)' \) and when \( \zeta = 1/4, \beta = (0,1,1)' \). When \( \lambda = 1 \), this model is a multiple linear regression model.

The simulations showed that MB and RD perform better than PMM and LRD in the linear case (\( \zeta = 1 \)) and had an error in the coverage rate of less than one percent. For the non-linear case, (\( \zeta = 1/4 \)), PMM and LRD perform better than MB and RD. Overall, LRD is the best imputation method in this situation. Lazzeroni et al. (1990) claim that LRD and PMM are more robust than the other two methods, but they suggest that these methods should be refined.

### 2.6 Multiple imputation for the one-way layout.

In this section, we apply the multiple imputation procedure to the one-way layout. We present this material because it is possible to study certain attributes of multiple imputation for the one-way model which cannot be studied in the repeated measures case because of the complexity of the multiple imputation procedure used. Since the two procedures are similar, we hope that similar properties, which can only be explored through simulation, exist.

Assume that the data analyst wishes to conduct an experiment for which the \( t-\)
treatment balanced one-way layout, defined in (1.1), is an appropriate model. But, due to missing data, the number of observations recorded for each treatment is different. One could analyze the data by treating the design as an unbalanced one-way layout and computing the estimators \( \hat{Y}_i \) and \( \hat{s}^2 \) which have equations defined in (1.2) and (1.3), respectively. Let \( n_i \) be the number of observed values for treatment \( i \) and \( n \) be the number of observations which should have been taken for each treatment. The number of missing observations for treatment \( i \) is \( a_i = n - n_i \).

We will consider a method similar to the Normal Imputation method described in Schenker and Welsh (1988). The only change is in the generation of \( \sigma^2 \). In our procedure, \( \sigma^2 \) from the normal imputation method is scaled so that the generated variance is conditionally unbiased with respect to the observed data.

We propose the following steps as a multiple imputation procedure for the balanced one-way layout.

Step 1: Using the observed data, calculate \( \hat{Y}_i \) and \( \hat{s}^2 \).

Step 2: Let \( n_{\text{obs}} = \sum_i n_i \) and draw
\[
x \sim \chi^2_{n_{\text{obs}}},
\]
(2.42)
to yield the value for the variance
\[
\sigma_k^2 = (n_{\text{obs}} - 2)\hat{s}^2/x.
\]
(2.43)

Step 3: For each \( i = 1, \ldots, t \), draw
\[
\theta_{ik} \sim N(\hat{Y}_i, \sigma_k^2/n_i).
\]
(2.44)
Step 4: For each missing value, draw

\[ Y_{ijk} \sim N(\theta_{ik}, \sigma_k^2). \quad (2.45) \]

Impute these values to create the \( k \)th completed-data set.

(Note that \( Y_{ijk} = Y_{ij} \) for observed values.)

Step 5: Repeat Steps 2 through 4 for \( k = 1, \ldots, m \) so that \( m \) completed-data set are formed.

Step 6: Using the \( k \)th completed-data set, compute \( Y_{i,k} \) and \( S_k^2 \) using the complete-data estimators, (1.2) and (1.3).

After these steps are completed \( m \) times, compute

\[ Y_{i,*} = \frac{1}{m} \sum_{k=1}^{m} Y_{i,k} \quad (2.46) \]

and

\[ S_*^2 = \frac{1}{m} \sum_{k=1}^{m} S_k^2. \quad (2.47) \]

Define the following vectors of the mean estimators

\[ \mathbf{Y}_* = (Y_{1,*}, \ldots, Y_{l,*}) \quad (2.48) \]

and

\[ \mathbf{Y}_k = (Y_{1,k}, \ldots, Y_{l,k}), \quad (2.49) \]

for \( k = 1, \ldots, m \). The between-imputation variance estimate is

\[ B_m = \frac{1}{m - 1} \sum_{k=1}^{m} (\mathbf{Y}_k - \mathbf{Y}_*)(\mathbf{Y}_k - \mathbf{Y}_*). \quad (2.50) \]
so that the total imputation variance is

\[ T_m = \frac{S^2}{n} \mathbf{I}_t + \left( 1 + m^{-1} \right) \mathbf{B}_m. \]  

We wish to compare the expectations of the estimators from the multiple imputation procedure with the estimators, \( \hat{Y}_i \) and \( \hat{\sigma}^2 \), from the unbalanced one-way layout. These estimates of the mean and variance in the unbalanced one-way layout are unbiased. To find the expectations in the multiple imputation case, iterated expectations are necessary over each step in the procedure.

To simplify the notation for the iterated expectations, the following notation will be used for the following distributions.

\[ \sigma^2_{\kappa_k} \equiv [\sigma^2_k \mid \mathbf{Y}_{\text{obs}}] \]  

\[ \theta_{\kappa_k} \equiv [\theta_{1k}, \ldots, \theta_{tk} \mid \mathbf{Y}_{\text{obs}}, \sigma^2_k] \]  

\[ \mathbf{Y}_{\text{imp}_k} \equiv [\mathbf{Y}_{\text{mis}} \mid \mathbf{Y}_{\text{obs}}, \sigma^2_k, \theta_k] \]  

By using the distribution of \( \chi^2_{n_{\text{obs}} - 1} \), it can easily be shown that

\[ E[\sigma^2_k \mid \mathbf{Y}_{\text{obs}}] = (n_{\text{obs}} - 2)\hat{S}^2/(n_{\text{obs}} - 2) = \hat{S}^2. \]  

To find the expectations of \( Y_{i\cdot} \) and \( S^2_{\cdot} \), the expectations for different relationships between the \( Y_{ijk} \) are needed. We derive the necessary expectations used in Theorem 3 in Theorems 1 and 2 below.

**Theorem 1** Given the multiple imputation procedure defined above, if \( Y_{ijk} \) and \( Y_{ilk} \) are imputed values from the \( k \)th repetition and \( Y_{ilk} \) is an observed value, then

\[ E[Y_{ijk}] = \theta_i, \]  

(2.56)
\[
\left[\left[\frac{\gamma_0 \gamma_\sigma}{\gamma_\sigma} \mathbf{A} \mid \gamma_\sigma \gamma_0\mathbf{A}\right]^{\gamma_0 \gamma_\sigma} \mathbf{A}\right]^{\gamma_0 \gamma_\sigma} \mathbf{A} = \\
\left[\left[\gamma_0 \gamma_\sigma \mathbf{A} \mid \gamma_0 \gamma_\sigma \mathbf{A}\right]^{\gamma_0 \gamma_\sigma} \mathbf{A}\right]^{\gamma_0 \gamma_\sigma} \mathbf{A} = [\gamma_0 \gamma_\sigma \mathbf{A}] \mathbf{A}
\]

(19.7)
\[
\dot{\mathbf{r}} + \mathbf{r} \left( \frac{1}{\mathbf{r}} + 1 \right) = \\
\dot{\mathbf{r}} + \frac{1}{\mathbf{r}} \gamma_0 + \gamma_0 (\frac{1}{\mathbf{r}} + 1) = \\
[\mathbf{r} \gamma_0 + \gamma_0 (\frac{1}{\mathbf{r}} + 1)]^{\gamma_0 \gamma_\sigma} \mathbf{A} = \\
\left[\left[\gamma_0 \mathbf{A} \mid \gamma_\sigma \gamma_0 \mathbf{A}\right]^{\gamma_0 \gamma_\sigma} \mathbf{A}\right]^{\gamma_0 \gamma_\sigma} \mathbf{A} = \\
\left[\left[\gamma_0 \mathbf{A} \mid \gamma_0 \mathbf{A}\right]^{\gamma_0 \gamma_\sigma} \mathbf{A}\right]^{\gamma_0 \gamma_\sigma} \mathbf{A} = [\gamma_0 \gamma_\sigma \mathbf{A}] \mathbf{A}
\]

(69.7)
\[
\mathbf{r} = \\
[\gamma_0 \gamma_\sigma \mathbf{A}] \mathbf{A} = \\
\left[\left[\gamma_0 \mathbf{A} \mid \gamma_\sigma \gamma_0 \mathbf{A}\right]^{\gamma_0 \gamma_\sigma} \mathbf{A}\right]^{\gamma_0 \gamma_\sigma} \mathbf{A} = \\
\left[\left[\gamma_0 \mathbf{A} \mid \gamma_0 \mathbf{A}\right]^{\gamma_0 \gamma_\sigma} \mathbf{A}\right]^{\gamma_0 \gamma_\sigma} \mathbf{A} = [\gamma_0 \gamma_\sigma \mathbf{A}] \mathbf{A}
\]

\[\texttt{Proof} \]
\[
\dot{\mathbf{r}} + \mathbf{r} \left( \frac{1}{\mathbf{r}} \right) = [\gamma_0 \gamma_\sigma \mathbf{A}] \mathbf{A}
\]

puv

(89.7)
\[
\dot{\mathbf{r}} + \mathbf{r} \left( \frac{1}{\mathbf{r}} \right) = [\gamma_0 \gamma_\sigma \mathbf{A}] \mathbf{A}
\]

(19.7)
\[
\dot{\mathbf{r}} + \mathbf{r} \left( \frac{1}{\mathbf{r}} + 1 \right) = [\gamma_0 \gamma_\sigma \mathbf{A}] \mathbf{A}
\]
\[ E[Y_{ij} | \theta_{ik}] = \frac{1}{n_i} \sum_{obs} Y_{ijk} \]

where \( \sum_{obs} \) means the sum over the observed values for that particular treatment.

To simplify notation, in the following theorem, let \( Y_{i(k)} \) be the average of the \( a_i \) imputed values for treatment \( i \) for the \( k \)th repetition,

\[ Y_{i(k)} = \frac{1}{a_i} \sum_{imp} Y_{ijk}, \] (2.64)

where \( \sum_{imp} \) means the sum over the imputed values.

**Theorem 2** Given the multiple imputation procedure defined above,

\[ E[Y_{i(k)}] = \theta_i, \] (2.65)
\[ E[Y_{ik}^2] = \frac{n + a_i}{a_i n_i} \sigma^2 + \theta_i^2, \quad (2.66) \]

and

\[ E[Y_{ik}^2 Y_{ij}] = \sigma^2/n_i + \theta_i^2. \quad (2.67) \]

**Proof:** Note that for all \( k = 1, \ldots, m, \)

\[ [Y_{ik} | Y_{\text{obs}}, \sigma_k^2, \theta] \sim N(\theta, \sigma_k^2/a_i). \quad (2.68) \]

Using similar arguments as in Theorem 1, the result follows. \( \blacksquare \)

**Theorem 3** Given the multiple imputation procedure defined in this section,

\[
E[Y_{i,\ast}] = \theta_i, \quad 1 \leq i \leq t, \quad (2.69)
\]

\[ E[S_{\ast}^2] = \sigma^2, \quad (2.70) \]

and

\[ E[B_m] = \begin{cases} \frac{\lambda_i \sigma^2}{n_i} & \text{for the } i\text{th diagonal element} \\ -\frac{n_{\text{obs}} - 5}{n_{\text{obs}} - 4} \frac{\tau^2}{n_i} & \text{for any off-diagonal element} \end{cases}, \quad (2.71) \]

where \( \lambda_i = a_i/n_i \) and \( \tau^2 = E[(S_{\ast}^2)^2] \).

**Proof:** By using (2.56) in Theorem 1 and that \( E(Y_{ij}) = \theta_i \) for the observed values,

\[
E(Y_{i,\ast}) = E \left( n^{-1} \sum_{k=1}^m \left( n^{-1} \sum_{j=1}^n Y_{ijk} \right) \right)
\]

\[ = n^{-1} m^{-1} \sum_{k=1}^m \sum_{j=1}^n E(Y_{ijk}) \]

\[ = n^{-1} m^{-1} \sum_{k=1}^m \sum_{j=1}^n \theta_i \]

\[ = \theta_i. \quad (2.72) \]
By (2.57), (2.58), and (2.59),
\[
E \left[ \sum_{j=1}^{n} (Y_{ijk} - Y_{i,k})^2 \right] = (1 - \frac{1}{n}) \sum_{j} E[Y_{ijk}^2] - \frac{1}{n} \sum_{j \neq l} E[Y_{ijk} Y_{ilk}]
\]
\[
= (1 - \frac{1}{n}) [n_i(\theta_i^2 + \sigma^2) + a_i ((1 + 2/n_i)\sigma^2 + \theta_i^2)]
\]
\[
- \frac{1}{n} [n_i(n_i - 1)\theta_i^2 + 2n_i a_i (\sigma^2/n_i + \theta_i^2)]
\]
\[
+ a_i(a_i - 1) (\theta_i^2 + 2\sigma^2/n_i)]
\]
\[
= (n - 1)\sigma^2,
\]
(2.73)

so that
\[
E[S_i^2] = E \left[ \frac{1}{(n-1)(1)} \sum_{i=1}^{t} \sum_{j=1}^{n} (Y_{ijk} - Y_{i,k})^2 \right]
\]
\[
= \frac{1}{(n-1)} \sum_{i=1}^{t} E \left[ \sum_{j=1}^{n} (Y_{ijk} - Y_{i,k})^2 \right]
\]
\[
= \frac{1}{(n-1)} \sum_{i=1}^{t} (n - 1)\sigma^2
\]
\[
= \sigma^2.
\]
(2.74)

For the \(i\)-th diagonal element of \(B_m\), first consider
\[
\sum_{k=1}^{m} (Y_{i,k} - Y_{i,\star})^2 = \sum_{k=1}^{m} \left[ \frac{1}{n_{obs}} \sum_{i} Y_{ijk} + \frac{1}{n_{imp}} \sum_{i} Y_{ijk}
\right.
\]
\[
\left. - \frac{1}{m} \sum_{l=1}^{m} \frac{1}{n} \left( \sum_{i} Y_{ijl} + \sum_{i} Y_{ijl} \right) \right]^2
\]
\[
= \left( \frac{\bar{a}_i}{n} \right)^2 \sum_{k=1}^{m} (Y_{i,k} - Y_{i,\star})^2,
\]
(2.75)

where
\[
Y_{i,\star} = \frac{1}{m} \sum_{k} Y_{i,k}.
\]
(2.76)
Using (2.66) and (2.67), the expectation of the diagonal elements can be reduced to

\[
E \left[ \frac{1}{m-1} \sum_{k=1}^{m} (Y_{i,k} - Y_{i,*})^2 \right]
\]

\[
= \left( \frac{a_n}{n(m-1)} \right)^2 E \left[ (1 - m^{-1}) \sum_{k \neq l} Y_{i,k} - m^{-1} \sum_{k \neq l} Y_{i,l} \right]
\]

\[
= \left( \frac{a_n}{n(m-1)} \right)^2 \left[ (1 - m^{-1}) \sum_{k=1}^{m} \left[ \frac{n + a_i}{a_i n_i} \sigma^2 + \theta_i^2 \right]ight.

\]

\[
- m^{-1}(m)(m-1) \left( \frac{\sigma^2}{n_i} + \theta_i^2 \right) \]

\[
= \left( \frac{a_n}{n(m-1)} \right)^2 \left[ (1 - m^{-1})(m) \left[ \frac{n + a_i}{a_i n_i} \sigma^2 + \theta_i^2 \right] (m - 1) \left( \frac{\sigma^2}{n_i} + \theta_i^2 \right) \right]
\]

\[
= \left( \frac{a_n}{n} \right)^2 \left[ \frac{n}{a_i n_i} \right] \sigma^2
\]

\[
= \frac{\lambda_i}{n} \sigma^2.
\]  

(2.77)

A similar argument can be used to show that all of the off–diagonal elements of $B_m$ have expectation

\[
E \left[ \frac{1}{m-1} \sum_{k=1}^{m} (Y_{i,k} - Y_{i,l}) (Y_{i,k} - Y_{i,l}) \right] = \frac{n_{obs} - 5}{n_{obs} - 4} \frac{\tau^2}{n^2}.
\]  

(2.78)

By the above theorem, the estimator of the variance, $T_m$, is a biased estimator of $\sigma^2$. However, as the sample size increases, the covariance decreases. For large samples, a diagonal matrix can approximate $T_m$. By (2.70), the within-repetition variance, $S^2_w$, is unbiased with respect to $\sigma^2$ and the diagonal elements of $B_m$ are functions of the fractions of missing data.

From (2.70) and (2.71), the estimate of the variance for the estimator $Y_{i,*}$ has expectation $(1 + \lambda_i)n^{-1} \sigma^2$ which is equivalent to $\sigma^2/n_i$. The variance of the estima-
tor $\hat{Y}_i$ is also $\sigma^2/n_i$. This suggests that the estimators for the multiple imputation procedure are comparable to the unbalanced one-way layout estimators. For the unbalanced one-way layout, good multiple comparison procedures exist [e.g., Tukey-Kramer, see Hochberg and Tamhane (1987)], but this is not the case for the repeated measures model. We will use a multiple imputation procedure for repeated measures similar to the procedure shown here which will yield valid inferences and may have similar strength.
CHAPTER III

The one-way repeated measures model

Both frequentist- and Bayesian-based estimation are necessary for multiple imputation. In this chapter, we will describe the balanced one-way repeated measures model and discuss properties using both the frequentist- and Bayesian-based perspectives. The frequentist-based perspective will be discussed in the context of the estimators needed for Dunnett’s multiple comparison procedure. The Bayesian-based perspective will be discussed in the context of the posterior distributions needed for the multiple imputation procedure.

3.1 Introduction

The simplest repeated measures model is a two-way mixed model in which one of the components is a fixed effect and the other component is a random effect with no interactions between components. The fixed effect is typically called the treatment effect and is usually the primary interest. The random effect is usually the result of sampling from a population. For notational purposes, assume that there are \( n \) subjects, each of whom receives \( t \) treatments. The model can be written in the form

\[
Y_{ij} = \theta_i + e_{ij}, \quad i = 1, \ldots, t, \quad j = 1, \ldots, n, \tag{3.1}
\]
where the $Y_{ij}$ are the observed values, $\theta_i$ are the fixed treatment effects, $c_j$ is the random subject effect, and $e_{ij}$ is the random error. The one-way repeated measures model is called balanced if the generalized least squares estimates are equivalent to the ordinary least squares estimates. The model described above is a case where the design is balanced, thus we call it the balanced one-way repeated measures model.

For each observation, $c_j$ and $e_{ij}$ are independent normal random variables with mean equal to zero and variances $\sigma^2_c$ and $\sigma^2_e$. That is,

$$c_j \sim N(0, \sigma^2_c)$$  

and

$$e_{ij} \sim N(0, \sigma^2_e).$$  

Also, assume that $c_j$ and $c_{j'}$ are independent for all $j \neq j'$ (i.e., the subjects are independent). The $e_{ij}$ are assumed to be independently distributed.

### 3.2 Frequentist-based inferences for the one-way repeated measures model

The frequentist-based estimates will be used for making inferences in the complete-data and completed-data cases. For Dunnett’s multiple comparison procedure, estimates of the treatment effects, $\theta_1, \ldots, \theta_t$, and the random error, $\sigma^2_e$, are necessary. In this section, we will discuss these estimates and their corresponding distributions.

Frequentist-based estimates for the parameters of the balanced one-way repeated measures model are based on the joint distribution of the $Y_{ij}, i = 1, \ldots, t, j = 1, \ldots, n$. The model (3.1), with (3.2), (3.3), and the independence assumption, implies
that $Y_{ij}$ has a normal distribution

$$Y_{ij} \sim N(\theta_i, \sigma_r^2 + \sigma_c^2). \quad (3.4)$$

By the independence assumptions on (3.1), the covariance between subjects $j$ and $j'$ is

$$Cov(Y_{ij}, Y_{ij'}) = 0, \quad (3.5)$$

and the covariance between treatments $i$ and $i'$ for a single subject is

$$Cov(Y_{ij}, Y_{ij}) = \sigma_r^2. \quad (3.6)$$

Let $Y_j = (Y_{1j}, Y_{2j}, \ldots, Y_{ij})'$ be the $t$-dimensional vector of observed values for the $j$-th subject. Equations (3.4) and (3.6) imply that

$$Y_j \sim \mathcal{N}_t(\theta, \mathbf{V}), \quad (3.7)$$

where

$$\theta = (\theta_1, \ldots, \theta_t), \quad (3.8)$$

and

$$\mathbf{V} = \sigma_r^2 \mathbf{I}_t + \sigma_c^2 \mathbf{J}_t, \quad (3.9)$$

where $\mathbf{I}_t$ is the $t \times t$ identity matrix and $\mathbf{J}_t$ is the $t \times t$ matrix of ones. For notational purposes later, let $\mathbf{J}_{r \times s}$ be the $r \times s$ matrix of ones.

Box and Tiao (1973) show that the estimator of $\theta_i$ is

$$\hat{Y}_i = n^{-1} \sum_{j=1}^{n} Y_{ij}, \quad 1 \leq i \leq t. \quad (3.10)$$
Let $Y_i = (Y_{i1}, \ldots, Y_{it})$. The distribution of $Y_i$ is

$$Y_i \sim N(\theta, n^{-1}V).$$  \hspace{1cm} (3.11)

The estimator of $\sigma^2$ is given by the mean square error, $MSE = SSE/\nu_c$, where $\nu_c = (t-1)(n-1)$ and

$$SSE = \sum_{i=1}^{t} \sum_{j=1}^{n} (Y_{ij} - Y_{i\cdot} - Y_{\cdot j} + Y_{\cdot \cdot})^2.$$  \hspace{1cm} (3.12)

Hochberg and Tamhane (1987) show that $MSE$ has a Chi-square distribution with $\nu_c$ degrees of freedom and is independent of $Y_i$.

### 3.3 Review of Bayesian-based one-way repeated measures model

The Bayesian-based inferences—namely the posterior distributions—are needed to draw values for imputation. The imputation procedure developed in Chapter 5 will use the posterior distributions of parameters which are one-to-one functions of $\theta$, $\sigma^2$, and $\sigma^2_c$. In this section, we will review the work of Box and Tiao (1973) who develop the posterior distributions for these parameters using a Jeffries prior. We will first discuss the form of the likelihood function and then, using the Jeffries prior, find the posterior distributions.

#### 3.3.1 The likelihood function

Since (3.5) holds for all $j$, the $Y_j$ are mutually independent. This implies that the likelihood function can be written as

$$f(\theta, \sigma^2, \sigma^2_c | Y) = \prod_{j=1}^{n} f(Y_j),$$  \hspace{1cm} (3.13)
which has the form

\[
\ell(\theta, \sigma^2, \sigma_r^2 \mid Y) \propto (\sigma_r^2)^{-\frac{n(t-1)}{2}} \left( \sigma^2 + t\sigma_r^2 \right)^{-\frac{n}{2}} \exp \left\{ -\frac{1}{2} \left[ \frac{t\sum_j (Y_{ij} - Y_i)^2}{\sigma_r^2 + t\sigma^2} + \sum_i \sum_j (Y_{ij} - Y_i, Y_j, Y_{ij})^2 \right] + Q(\theta) \right\},
\]

(3.14)

where

\[
Q(\theta) = (Y - \theta)'V^{-1}(Y - \theta).
\]

(3.15)

From Box and Tiao (1973), the inverse of \( V \) is

\[
V^{-1} = \sigma_r^{-2} \left[ I - \frac{\sigma^2}{\sigma_r^2 + t\sigma^2} J \right].
\]

(3.16)

Box and Tiao (1973) use the transformation

\[
\phi_i = \theta_i - \bar{\theta}, \quad i = 1, \ldots, t - 1,
\]

(3.17)

where

\[
\bar{\theta} = \frac{1}{t} \sum_{i=1}^t \theta_i
\]

(3.18)

and \( \phi_t = -(\phi_1 + \cdots + \phi_{t-1}) \). This transformation, and setting \( \sigma_r^2 = \sigma^2 + t\sigma^2 \), makes it possible to divide the likelihood function into two separate functions. The two functions have the functional forms of a normal distribution. Now the likelihood can be written as

\[
\ell(\bar{\theta}, \phi, \sigma_r^2, \sigma^2 \mid Y) = \ell_1(\bar{\theta}, \sigma_r^2 \mid Y)\ell_2(\phi, \sigma^2 \mid Y),
\]

(3.19)

where

\[
\ell_1(\bar{\theta}, \sigma_r^2 \mid Y) \propto (\sigma_r^2)^{-\frac{t+1}{2}} \exp \left\{ -\frac{1}{2\sigma_r^2} \left[ \nu_rM\sigma + nt (\bar{\theta} - p)^2 \right] \right\},
\]

(3.20)
and

$$\ell_2(\phi, \sigma_c^2 \mid Y) \propto (\sigma_c^2)^{-\frac{\nu_c+1}{2}} \exp \left\{ -\frac{1}{2\sigma_c^2} [S(\phi) + \nu_cMSE] \right\},$$

where \(\nu_c\) and \(MSE\) are defined in the previous section,

$$S(\phi) = n \sum_{i=1}^{t} \left[ \phi_i - (Y_i - \bar{Y}_i) \right]^2,$$

$$MSC = SSC/\nu_c,$$

where

$$SSC = t \sum_{j} (Y_{ij} - \bar{Y}_i)^2$$

and \(\nu_c = t - 1\). The separation of the likelihood function implies that \(\sigma_e^2\) is independent of \(\sigma_c^2\) and \(\phi\).

### 3.3.2 The prior and posterior distributions

Different prior distributions for the parameters \(\theta\), \(\sigma_e^2\), and \(\sigma_r^2\) have been proposed for the mixed model. Broemeling (1985) defines a normal distribution prior conditioned on \(\sigma_r^2\) for the treatment effects and a gamma prior distributions for \(\sigma_e^2\) and \(\sigma_c^2\) to find approximate posterior distributions. Box and Tiao (1973) define two prior distributions based on a Jeffries prior. The first prior is defined as

$$\pi(\theta, \phi, \sigma_e^2, \sigma_c^2) \propto \sigma_r^{-2} \sigma_c^{-2}.$$  

(3.24)

The second prior has the same form as (3.24), but adds the condition that \(\sigma_r^2 < \sigma_c^2\). This assumption assures that \(\sigma_c^2 > 0\). (3.24) is referred to as the unconstrained prior and the latter case is called the constrained prior. The unconstrained prior can lead to results where the estimate of the random effects variance is negative. This will be
discussed later. The constrained prior yields posterior distributions which do not have familiar distributional forms, although approximations to the posteriors do exist.

We will use Box and Tiao's unconstrained prior (3.24) for the imputation since this prior leads to estimates which are analogous to the frequentist-based approach and the posteriors have well-known distributions from which samples can be drawn. This prior suffers from the chance that $\sigma_c^2$ will be negative. The use of other prior distributions to develop other multiple imputation procedures will be left to further research. The following development of the posterior distributions for $\tilde{\theta}$, $\phi$, $\sigma_c^2$, and $\sigma_{cc}^2$ is found in Box and Tiao (1973).

Using the unconstrained prior with (3.19), the posterior distribution has the form

$$
\pi \left( \tilde{\theta}, \phi, \sigma_c^2, \sigma_{cc}^2 \mid \mathbf{Y} \right) \propto \sigma_c^{-2} \sigma_{cc}^{-2} \ell_1(\tilde{\theta}, \sigma_c^2 \mid \mathbf{Y}) \ell_2(\phi, \sigma_c^2 \mid \mathbf{Y}).
$$

(3.25)

The functional part of the posterior containing $\phi$ can be written as

$$
\sigma_c^{-2} \ell_2(\phi, \sigma_c^2 \mid \mathbf{Y}) = g(\sigma_c^2) \exp \left\{ -\frac{S(\phi)}{2\sigma_c^2} \right\},
$$

(3.26)

for some function $g(\sigma_c^2)$, which implies that, conditional on $\sigma_c^2$, the posterior distribution for $\phi$ is a $(t - 1)$-dimensional multivariate normal

$$
N_{t-1} \left( \hat{\phi}, \sigma_c^2 \Sigma \right),
$$

(3.27)

where

$$
\Sigma = n^{-1} \left[ \mathbf{I}_{t-1} - t^{-1} \mathbf{J}_{t-1} \right]
$$

(3.28)

and $\hat{\phi} = (\hat{\phi}_1, \ldots, \hat{\phi}_{t-1})'$, in which each component is

$$
\hat{\phi}_i = Y_i - \bar{Y}, \quad 1 \leq i \leq t - 1.
$$

(3.29)
Box and Tiao (1973) also show, using (3.26), that the posterior distribution of $\sigma^2_e$ is an inverse Chi-square distribution with $\nu_e$ degrees of freedom,

$$[\sigma^2_e | \mathbf{Y}] \sim \nu_e \text{MSE} \chi^-2_{\nu_e}.$$ (3.30)

Using the functional part of the posterior containing $\sigma^2_{\epsilon^c}$, Box and Tiao show that the posterior distribution of $\sigma^2_{\epsilon^c}$ is

$$[\sigma^2_{\epsilon^c} | \mathbf{Y}] \sim \nu_e \text{MSC} \chi^-2_{\nu_e}.$$ (3.31)

The corresponding estimate for $\sigma^2_{\epsilon^c}$ is $\text{MSC}$ and the estimate of $\sigma^2_e$ is $\text{MSE}$. One can use these two estimates to obtain the estimate of $\sigma^2_e$ as

$$\hat{\sigma}^2_e = t^{-1}(\text{MSC} - \text{MSE}).$$ (3.32)

Scheffé (1959) notes that this estimate can be negative with positive probability. Scheffé suggests using

$$\hat{\sigma}^2_e = \max(0, \hat{\sigma}^2_e).$$ (3.33)

The posterior distributions developed in this section are not for $\mathbf{\theta}$ and $\sigma^2_e$. However, we will generate values from the posterior distributions developed here and transform the generated values to get values for $\mathbf{\theta}$ and $\sigma^2_e$. The implementation of the multiple imputation procedure presented in Chapter 5 will make use of the posterior distributions presented above.
CHAPTER IV

Multiple comparison procedures

In this chapter, we will first discuss three different multiple comparison procedures for fixed effects linear models. We will discuss (1) Scheffé’s method which can be applied to many different models, both balanced and unbalanced, (2) Tukey’s method, a procedure which was developed for the balanced one-way layout but has been modified for the unbalanced case, and (3) Dunnett’s method which compares a control to the treatment effects. We will only focus on two-sided tests for all three methods. We will then discuss how Dunnett’s multiple comparison procedure can be applied to the balanced one-way repeated measures model and what difficulty arises when the design is unbalanced.

4.1 Introduction

In linear models, the traditional method for analyzing the group effects is a test for homogeneity (e.g., Analysis of Variance). However, this type of procedure only tests for equality across all of the groups and cannot make individual comparisons between groups. For example, the F-test is used in the one-way layout to test if all of the fixed effects are equal but cannot reveal which of the fixed effects, if any, are significantly different from the others. Multiple comparison procedures have been
developed to make more detailed inferences about the linear contrasts of a group while controlling the overall error rate. The majority of the following discussion about multiple comparison procedures is taken from Hochberg and Tamhane (1987).

Let $\mathcal{F}$ denote a family of inferences and $\mathcal{P}$ a multiple comparison procedure for $\mathcal{F}$. Let $M(\mathcal{F}, \mathcal{P})$ be the random number of wrong inferences. The familywise error rate (FWE) is defined as

$$\text{FWE}(\mathcal{F}, \mathcal{P}) = P\{ M(\mathcal{F}, \mathcal{P}) > 0 \}. \quad (4.1)$$

$\mathcal{P}$ is called an exact $(1 - \alpha)$ procedure if the FWE is equal to $\alpha$. The procedure is said to be conservative if $\text{FWE} < \alpha$ and liberal if $\text{FWE} > \alpha$. Multiple comparison procedures have been developed to make simultaneous inferences about linear contrasts while controlling the FWE.

### 4.2 Fixed-effects linear models

For fixed-effects linear models, the parameters of interest are usually the $t$ fixed effects, $\theta_1, \ldots, \theta_t$. Let $\hat{\theta}$ be the least squares estimator of $\theta = (\theta_1, \ldots, \theta_t)$, with distribution

$$\hat{\theta} \sim N(\theta, \sigma^2 V), \quad (4.2)$$

where $V = (v_{ij})$ is a $t \times t$ matrix. Let $S^2$ be an estimator of $\sigma^2$, independent of $\hat{\theta}$, such that

$$S^2 \sim \sigma^2 \chi^2_\nu / \nu. \quad (4.3)$$

If the components of $\hat{\theta}$ have the same variance (i.e., if $v_{ii} = v_{jj}$ for $i \neq j$) then Hochberg and Tamhane (1987) call the corresponding design balanced. If this condition does not hold, the design is called unbalanced. A balanced design, in most
cases, corresponds to having the same number of observations for each of the groups. Exact multiple comparison procedures are more easily attainable when the design is balanced. For example, even the unbalanced one-way layout has no known exact multiple comparison procedure for pairwise comparisons of the fixed effects.

Many different multiple comparison procedures exist depending on the particular comparisons and designs. There are single stage, stepwise up, and stepwise down procedures which test for all linear effects, paired differences, comparisons with the control, or comparisons with the best. The method of comparison for different multiple comparison procedures is to compare the half-widths of the confidence intervals. We will now consider three multiple comparison procedures in detail.

4.2.1 Scheffé's multiple comparison procedure

Scheffé's procedure provides exact simultaneous confidence intervals for all linear combinations of the $\theta_i$. Let $\mathcal{L}$ be a subspace of $\mathbb{R}^t$ with dimension $t' \leq t$. The $(1 - \alpha)$-level simultaneous confidence intervals have the form

$$\ell'\hat{\theta} \in \left[ \ell'\hat{\theta} \pm \left( \ell'F^\alpha_{t', \nu} \right)^{1/2} \left( \ell'V\ell \right)^{1/2} \right] \quad \forall \ell \in \mathcal{L}, \quad (4.4)$$

where $F^\alpha_{t', \nu}$ is the upper $\alpha$-th percentile of the $F$-distribution with $t'$ and $\nu$ degrees of freedom. Notice here that the critical value, $F^\alpha_{t', \nu}$, is one-dimensional so that the critical value is easily computed.

Define $\mathcal{L}_c$ to be the contrast space $\{c \in \mathbb{R}^t : \sum_i c_i = 0\}$. For the unbalanced one-way layout, Scheffé's multiple comparison procedure produces $(1 - \alpha)$-level si-
multaneous confidence intervals of all contrasts of $\theta$ of the form

$$c'\theta \in \left( \sum_{i=1}^{t} c_i Y_i, \pm \left( (t - 1) F_{\alpha}^{0.1, n_{obs}} \right)^{1/2} S \left( \sum_{i=1}^{t} \frac{c_i^2}{n_i} \right)^{1/2} \right) \quad \forall c \in \mathcal{L}_c,$$

where $n_{obs} = \sum_i (n_i - 1)$ and $Y_i$ is defined in (1.4).

Scheffé (1959) shows that this method is equivalent to the $F$-test for the one-way layout.

### 4.2.2 Tukey's multiple comparison procedure

Tukey's multiple comparison procedure yields exact simultaneous confidence intervals for pairwise comparisons of the fixed effects in the balanced one-way layout. The $(1 - \alpha)$-level simultaneous confidence intervals of the pairwise differences have the form

$$\theta_i - \theta_k \in \left( Y_i - Y_k, \pm Q_{\alpha}^{\ast, (n-1)} S n^{-1/2} \right) \quad 1 \leq i < k \leq t,$$

where $Q_{\alpha}^{\ast, (n-1)}$ is the upper $\alpha$-th percentile of the Studentized Range distribution with parameters $t$ and $t(n - 1)$. Tukey's procedure, among others, can be extended from differences in the fixed effects to all contrasts in $\mathcal{L}_c$ by the following result.

**Theorem 4** (Hochberg and Tamhane, 1987) Let $x = (x_1, \ldots, x_t)'$, $x \in \mathbb{R}^t$ and $\xi \geq 0$, then

$$|x_i - x_k| \leq \xi \; \text{if and only if} \; |c'x| \leq \xi \sum \frac{|c_i|}{2} \forall c \in \mathcal{L}_c.$$

Using Theorem 4, Tukey's multiple comparison procedure can be modified by extending (4.6) to all contrasts by using the confidence intervals

$$c'\theta \in \left( \sum_{i=1}^{t} c_i Y_i, \pm Q_{\alpha}^{\ast, (n-1)} S n^{-1/2} \sum_{i=1}^{t} \frac{|c_i|}{2} \right) \quad \forall c \in \mathcal{L}_c.$$


Note that Scheffé's procedure in (4.4) yields a similar form of contrast confidence intervals for the one-way layout to that of Tukey's in (4.8). However, Tukey's procedure has been shown to have confidence intervals whose half-width is the shortest possible for all confidence intervals of the pairwise differences (Hochberg and Tamhane, 1987). Hence, for the balanced one-way layout, Tukey's multiple comparison procedure is better than Scheffé's multiple comparison procedure for finding differences in the treatment effects.

For the unbalanced one-way layout, a procedure with an exact confidence level does not exist. However, conservative procedures, such as Tukey-Kramer, and others are available (Hochberg and Tamhane, 1987). The Tukey-Kramer procedure has been shown to be conservative and to yield shorter confidence intervals than Scheffé's procedure.

### 4.2.3 Dunnett’s multiple comparison procedure

Dunnett's procedure was developed for the one-way fixed effects linear model to compare different treatments to a control. For notational purposes, let the last effect, $\theta_t$, be the control. The $(1 - \alpha)$-level simultaneous confidence intervals associated with all treatments versus a control have the form

\[
\theta_i - \theta_t \in \left( \hat{\theta}_i - \hat{\theta}_t \pm |T|_{\nu_{(\nu_j)}}^{\alpha} \sqrt{v_{ii} + v_{tt} - 2v_{it}} \right), \quad i = 1, \ldots, t-1,
\]

where $v_{ij}$ is the $(i,j)$ element of $V$ in (4.2), and $|T|_{\nu_{(\nu_j)}}^{\alpha}$ is the $\alpha$-th percentile of the $(t-1)$-dimensional multivariate $T$-distribution with $\nu$ degrees of freedom and correlation structure defined by $\{\rho_{ij}\}$ (Hochberg and Tamhane, 1987). The correlation
coefficient, \( \rho_{ij} \), is the correlation between \( \hat{\theta}_i - \hat{\theta}_l \) and \( \hat{\theta}_j - \hat{\theta}_l \) and has the form

\[
\rho_{ij} = \frac{\nu_{ij} - \nu_{ll} - \nu_{jl} + \nu_{tt}}{[(\nu_{ii} + \nu_{tt} - 2\nu_{il})/(\nu_{jj} + \nu_{tt} - 2\nu_{jl})]^{1/2}} \quad \forall i \neq j. \tag{4.10}
\]

For the balanced one-way layout, \( \nu_{ii} = 1/n, \nu_{ij} = 0 \), thus making \( \rho_{ij} = 1/2 \) for all \( i \neq j \), so that Dunnett’s procedure reduces to

\[
\theta_i - \theta_l \in \left( Y_i - Y_l \pm |T|_{l-1, \nu_{ii}(1/2)}^{t-1, \nu_{jj}(\rho_{ij})} S/\sqrt{n} \right), \quad i = 1, \ldots, t - 1. \tag{4.11}
\]

The cumulative distribution of the multivariate T-distribution cannot be expressed in a closed form. It is necessary to use numerical integration to compute \( |T|_{l-1, \nu_{ii}(\rho_{ij})}^{t-1, \nu_{jj}(\rho_{ij})} \). When all of the \( \rho_{ij} \) are the same, the design is called balanced with respect to treatments and the cumulative distribution reduces to a double integration which can be approximated using numerical methods on a computer in real time (Hochberg and Tamhane, 1987). When the design is unbalanced with respect to treatments, a \((t - 1)\)-dimensional integral must be evaluated to find \( |T|_{l-1, \nu_{ii}(\rho_{ij})}^{t-1, \nu_{jj}(\rho_{ij})} \), which is computationally prohibitive.

For the unbalanced case, Hochberg and Tamhane (1987) suggest using (4.9), but replace each of the \( \rho_{ij} \) by a single value, \( \rho \), so that the cumulative distribution reduces to a double integration. Suggestions for the value of \( \rho \) are discussed in detail by Hochberg and Tamhane.

### 4.3 Multiple comparison procedures for the one-way repeated measures model.

Little literature exists regarding multiple comparison procedures for the one-way repeated measures model defined in (3.1). Scheffé’s procedure is available for both
balanced and unbalanced repeated measures, but, as in the one-way layout, it is a
conservative procedure.

Multiple comparisons are difficult to develop when the covariance matrix of the
vector of estimators of \( \theta \) has an arbitrary form. Procedures are easier to develop
if more is assumed about the structure of the covariance matrix. If the covariance
matrix of the vector of estimators, \( \Sigma = (\sigma_{ij}) \), has the form

\[
\sigma_{ij} = \begin{cases} 
2\lambda_i + \tau^2 & \text{if } i = j \\
\lambda_i + \lambda_j & \text{if } i \neq j
\end{cases}
\]  

(4.12)

for some \( \tau^2 > 0 \) and \( \lambda_i \in \mathbb{R} \), then the model is called spherical. Multiple comparison
procedures exist for spherical models. One such method can be derived from the
following theorem.

**Theorem 5** [Hochberg and Tamhane (1987)] *If the model is spherical, then*

\[
\max_{1 \leq i < k \leq t} \frac{|Y_i - Y_k - (\theta_i - \theta_k)|}{\sqrt{MSE/n}}
\]  

(4.13)

where \( Y_i \) and \( MSE \) are defined as in Chapter 3, has a Studentized range distribution
with parameter \( t \) and degrees of freedom \((t - 1)(n - 1)\).

Using Theorem 4 and Theorem 5, Hochberg and Tamhane (1987) show that if the
model is spherical, then exact \((1 - \alpha)\)-level simultaneous confidence intervals for all
contrasts \( c'\theta \) have the form

\[
c'\theta \in \left( \sum c_i Y_i \pm Q_{t(1-\alpha)(n-1)}^{\alpha} \sqrt{\frac{MSE}{n}} \sum \frac{|c_i|}{2} \right) \forall c \in \mathcal{L}_c.
\]  

(4.14)

This multiple comparison procedure is similar to Tukey's procedure.
A special case of sphericity is when all of the terms of the off-diagonal elements of the covariance matrix are the same. This is referred to as the uniform model. The covariance matrix under a uniform model can be written as

$$\Sigma = \sigma_0^2 [(1 - \rho)I_t + \rho J_t]$$  \hspace{1cm} (4.15)

for some $\sigma_0^2 > 0$ and $0 \leq \rho \leq 1$. The one-way repeated measures model is an example of a uniform model.

The balanced one-way repeated measures model defined in (3.1) yields the covariance matrix (3.9) which can be rewritten in two ways,

$$\Sigma = \frac{1}{n} \left[ \sigma_c^2 I_t + \sigma_e^2 J_t \right]$$

$$= \sigma_c^2 \left[ \frac{1}{n} I_t + \frac{\sigma_e^2}{n \sigma_c^2} J_t \right]$$

$$= \frac{\sigma_c^2 + \sigma_e^2}{n} \left[ \left( 1 - \frac{\sigma_c^2}{\sigma_e^2 + \sigma_c^2} \right) I_t + \frac{\sigma_e^2}{\sigma_e^2 + \sigma_c^2} J_t \right].$$  \hspace{1cm} (4.16)

$$= \left[ 1 - \frac{\sigma_c^2}{\sigma_e^2 + \sigma_c^2} \right] I_t + \frac{\sigma_e^2}{\sigma_e^2 + \sigma_c^2} J_t.$$  \hspace{1cm} (4.17)

(4.17) shows that the one-way repeated measures model is a uniform model so that (4.14) is a multiple comparison procedure for all linear contrasts for the balanced one-way repeated measures model.

Although Dunnett’s treatments versus the control multiple comparison procedure was developed for the balanced fixed-effects model, the procedure can be applied to the balanced one-way repeated measures model because the covariance structure is uniform. This uniformity in the variance-covariance matrix causes the random effects variance, $\sigma_c^2$, to cancel in (4.10). This can be seen by using (4.10) with (4.16) to yield
the value of \( p_{ij} \) as

\[
p_{ij} = \frac{\sigma^2_{\epsilon \sigma^2_e} - \sigma^2_{\epsilon \sigma^2_e} + \left( \frac{1}{n} + \frac{\sigma^2_{\epsilon \sigma^2_e}}{n\sigma^2_e} \right)}{\left( \frac{1}{n} + \frac{\sigma^2_{\epsilon \sigma^2_e}}{n\sigma^2_e} \right) + \left( \frac{1}{n} + \frac{\sigma^2_{\epsilon \sigma^2_e}}{n\sigma^2_e} \right) - 2 \left( \frac{\sigma^2_{\epsilon \sigma^2_e}}{n\sigma^2_e} \right)}
\]

\[
= \frac{1}{2}. \quad (4.18)
\]

Also note that, for all \( i = 1, \ldots, t - 1 \),

\[
v_{it} + v_{it} - 2v_{it} = \left( \frac{1}{n} + \frac{\sigma^2_{\epsilon \sigma^2_e}}{n\sigma^2_e} \right) + \left( \frac{1}{n} + \frac{\sigma^2_{\epsilon \sigma^2_e}}{n\sigma^2_e} \right) - \frac{2\sigma^2_{\epsilon \sigma^2_e}}{n\sigma^2_e}
\]

\[
= \frac{2}{n}. \quad (4.19)
\]

Using (4.18) and (4.19) in (4.9) implies that Dunnett’s multiple comparison procedure confidence intervals for the balanced one-way repeated measures model can be written as

\[
\theta_i - \theta_t \in \left( Y_i - Y_t \pm |T|_{t-1, (t-1)(n-1), 1/2} S\sqrt{\frac{2}{n}} \right) \quad i = 1, \ldots, t - 1. \quad (4.20)
\]

Notice that in (4.20) the random effects variance, \( \sigma^2_{\epsilon \sigma^2_e} \), drops out of \( p_{ij} \). In the presence of missing data, the generalized least squares estimates will most likely not be equal to the ordinary least squares estimates so the design will be unbalanced. In the unbalanced case, the generalized least squares estimates are the best linearly unbiased estimates and should be used. However, as a simple example illustrating why Dunnett’s multiple comparison procedure is not appropriate for the unbalanced one-way repeated measures model, we will use the ordinary least squares estimates (which are the treatment means) since they are more likely to be used by the typical user.
Assume that you have a balanced repeated measures model with \( t = 3 \) and \( n = 4 \). Furthermore, one value is missing from the third treatment. Using the ordinary least squares estimates,

\[
\rho_{12} = \frac{4\sigma_r^2 + \sigma_c^2}{7\sigma_r^2 + \sigma_c^2}
\]  

and

\[
v_{11} + v_{33} - 2v_{13} = \frac{1}{12} \left(7 + \frac{\sigma_c^2}{\sigma_r^2}\right).
\]  

Notice that both \( \rho_{12} \) and \( v_{11} + v_{33} - 2v_{13} \) are functions of both variances so that the random effects variance must be introduced into the confidence interval. This implies that an estimator of \( \sigma_r^2 \) is needed to calculate the confidence intervals and thus the distributions needed for the confidence intervals must be investigated.

In general, if the design is unbalanced, \( \sigma_r^2 \) will remain in \( \rho_{ij} \) and \( v_{ii} + v_{ll} - 2v_{il} \) so that the estimator of \( \sigma_r^2 \) would have to be introduced into the problem.

In the next chapter, we will develop a multiple imputation procedure which will impute values for the missing data in a balanced one-way repeated measures model so that confidence intervals based on Dunnett's multiple comparison procedure can be calculated with the approximately correct \( \alpha \)-level without involving the estimation of \( \sigma_r^2 \).
CHAPTER V

An imputation procedure for the repeated measures model

In this chapter, we propose a multiple imputation procedure which we have developed to impute values into a one-way repeated measures model. We will use the procedure to produce estimates which we will use in a modified version of Dunnett’s multiple comparison procedure presented in Chapter 6.

5.1 One-way repeated measures multiple imputation procedure

Suppose a repeated measures experiment was to be conducted where each of the \( n \) subjects was given each of the \( t \) treatments exactly one time. However, some of the observations were not recorded and may be considered missing. Let \( n^* \) be the number of subjects with no missing observations. We must assume that \( n^* \) is large enough to get decent initial estimates. Using the same notation as in Chapter 3, the repeated measures model has the form

\[
Y_{ij} = \theta_i + \epsilon_j + \epsilon_{ij}, \quad i = 1, \ldots, t, \quad j = 1, \ldots, n. \tag{5.1}
\]

From Chapter 3, we know that in the complete-data case, the vector of fixed
effects estimate is

$$Y_i \sim N(\theta, n^{-1}V).$$  \hfill (5.2)

From Mardia, Kent and Bibby (1989), if $A$ is a $t \times p$ matrix of constants, for some $p \geq 1$, then $A Y_i$ is distributed as a $p$-dimensional multivariate normal

$$A Y_i \sim N(A \theta, AVA').$$  \hfill (5.3)

This implies that multiple imputation procedures can be employed to create completed-data sets and that inferences drawn from these completed-data sets will be valid for the linear combinations of interest.

Various different imputation procedures have been discussed in Chapter 2, but some of these procedures would not be appropriate for the repeated measures model. Procedures which use other observations from the same experiment, such as hot deck and Bayesian bootstrap, would be difficult to implement since only one datum point is observed for a given individual and treatment combination. We will consider an imputation procedure which extends the Fully Normal procedure used for the one-way layout to the repeated measures model. This procedure is attractive since it uses information available from the design of the model and imputed values can be randomly drawn using standard computer packages. We will call this procedure the “repeated measures normal imputation (RMNI)” procedure. To create a completed-data set, this procedure takes random values from the Bayesian posterior distributions for the unknown parameters and generates the imputed values from the resulting likelihood function.
5.2 The repeated measures normal imputation procedure

For the RMNI procedure, we will assume that the missing-data mechanism is ignorable. An adjustment can be made to the procedure if the missing data mechanism is nonignorable and will not be considered here. The assumption of ignorable nonresponse makes it possible to develop the procedure directly from the posteriors and the likelihood function without further assumptions concerning the missing-data mechanism.

This procedure can be viewed as a two-stage process. First, given the completely observed data, we randomly draw values for the parameters, $\theta$, $\sigma_r^2$, and $\sigma_c^2$, from their respective posterior distributions. Second, given these randomly drawn parameter values and the observed data, we randomly draw values for the missing-data from their conditional distributions. The details of the two steps are given below.

The data from the $n_*$ completely observed subjects create a balanced data set so that posterior distributions for the repeated measures model with no missing data can be used to generate random values for $\theta$, $\sigma_r^2$, and $\sigma_c^2$, assuming that there are sufficient complete observations to get a good initial estimate. The posterior distributions from Chapter 3 necessary for the RMNI procedure are

\begin{align}
  [\sigma_r^2|Y] &\sim (t - 1)(n_* - 1)\text{SSE}\chi_{(t-1)(n_*-1)}^2, \\
  [\sigma_c^2|Y] &\sim (n_* - 1)\text{SSC}\chi_{n_*-1}^2, \quad \text{and} \\
  [\phi|Y, \sigma_r^2] &\sim N_{t-1} \left( \phi, \sigma_r^2 \Sigma \right),
\end{align}

(5.4) 
(5.5) 
(5.6)
where SSE, SSC, \( \phi \), \( \hat{\phi} \), \( \sigma_{rr}^2 \), and \( \Sigma \) are defined in Chapter 3 and the statistics are computed using the data from the \( n \) completely observed subjects.

The random value for the parameter \( \sigma_{rr}^2 \) can be drawn from the distribution in (5.4), call it \( \sigma_{rr}^2 \). Let \( \sigma_{rr}^2 \) be the random value for the parameter \( \sigma_{rr}^2 \) drawn from the posterior distribution in (5.5). Then the random value for the parameter \( \sigma_{rr}^2 \) will be

\[
\sigma_{rr}^2 = \max \left\{ 0, \frac{1}{t} \left[ \sigma_{rr}^2 - \sigma_{rr}^2 \right] \right\}. \tag{5.7}
\]

Let \( \phi \) be the random vector generated from (5.6). Transform \( \phi_{1}, \ldots, \phi_{(t-1)} \) into \( \theta_{1}, \ldots, \theta_{t} \) by using

\[
\theta_{i} = \phi_{i} + \eta_{i}, \quad i = 1, \ldots, t. \tag{5.8}
\]

The second stage of the procedure is to draw values from the conditional distribution of the missing values, given the observed values and the parameters drawn in the first stage. In order to do this, the conditional distribution of a partitioned multivariate normal is needed. For a partitioned vector having a multivariate normal distribution, the conditional distribution of one member of the partition, given the other is well known, is presented here as a theorem.

**Theorem 6** [Mardia et al. (1989)] Let \( X \) be a \( p \)-dimensional multivariate normal random variable which can be partitioned as

\[
X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_p \left( \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right), \tag{5.9}
\]

where \( X_1 \) is a \( p_1 \)-dimensional vector. Then

\[
X_2 | X_1 \sim N_{p-p_1} \left( \mu_2 + \Sigma_{21} \Sigma_{11}^{-1} (X_1 - \mu_1), \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} \right). \tag{5.10}
\]
Since the between-subjects covariance in the repeated measures model is zero, the conditional distribution of each missing datum, given the parameters from $\sigma_{\epsilon}^2$, $\sigma_{\epsilon}^2$, and $\theta_{t_{\epsilon}}$, can be considered independently. Let $Y_j$ be the vector of responses, missing or observed, for subject $j$. Furthermore, let $Y_j$ be partitioned into missing and observed components, $Y_{\text{mis}_j}$ and $Y_{\text{obs}_j}$, with dimensions $a_j$ and $t - a_j$, respectively. For notational convenience, let $Y_{\text{mis}_j}$ have mean $\theta_2$ and let $Y_{\text{obs}_j}$ have mean $\theta_1$, with the corresponding partitioned covariance matrix with the form defined in Theorem 6.

From Chapter 3, $Y_j$ has a covariance matrix of the form

$$\Sigma = [\sigma_{\epsilon}^2I_t + \sigma_{\epsilon}^2J_t].$$

(5.11)

The partitioned components of the covariance matrix are,

$$\Sigma_{11} = \left[\sigma_{\epsilon}^2I_{t-a_j} + \sigma_{\epsilon}^2J_{t-a_j}\right],$$

(5.12)

$$\Sigma_{22} = \left[\sigma_{\epsilon}^2I_{a_j} + \sigma_{\epsilon}^2J_{a_j}\right],$$

(5.13)

$$\Sigma_{12} = \sigma_{\epsilon}^2J_{(t-a_j)\times a_j},$$

(5.14)

and

$$\Sigma_{21} = \sigma_{\epsilon}^2J_{a_j\times(t-a_j)}.$$ 

(5.15)

The inverse for $\Sigma_{11}$ is found using (3.16) and can be written as

$$\Sigma_{11}^{-1} = \sigma_{\epsilon}^{-2}\left[I_{(t-a_j)} - \frac{\sigma_{\epsilon}^2}{\sigma_{\epsilon}^2 + (t-a_j)\sigma_{\epsilon}^2}J_{(t-a_j)}\right].$$

(5.16)

Using (5.16) and the property $J_{q\times p}J_{p\times r} = pJ_{q\times r}$,

$$\Sigma_{21}\Sigma_{11}^{-1} = \sigma_{\epsilon}^2J_{a_j\times(t-a_j)}\left(\sigma_{\epsilon}^{-2}\right)\left[I_{(t-a_j)} - \frac{\sigma_{\epsilon}^2}{\sigma_{\epsilon}^2 + (t-a_j)\sigma_{\epsilon}^2}J_{(t-a_j)}\right]$$
Using (5.17),

\[ \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} = \frac{\sigma_c^2}{\sigma_c^2 + (t - a_j)\sigma_c^2} J_{a_j \times (t - a_j)} \left( \sigma_c^2 J_{(t - a_j) \times a_j} \right) \]

\[ = \frac{(t - a_j)\sigma_c^2}{\sigma_c^2 + (t - a_j)\sigma_c^2} J_{a_j}. \quad (5.18) \]

Using Theorem 6, (5.13), and (5.18), the conditional distribution of \( \mathbf{Y}_{mix_j} \), given \( \mathbf{Y}_{obs_j} \), has the mean

\[ \theta_2 + \Sigma_{21} \Sigma_{11}^{-1} (\mathbf{Y}_{obs} - \theta_1) = \theta_2 + \frac{\sigma_c^2}{\sigma_c^2 + (t - a_j)\sigma_c^2} \sum_{i \in obs} (y_{ij} - \theta_i) \mathbf{1}_{a_j}, \quad (5.19) \]

where \( \mathbf{1}_{a_j} \) is a \( a_j \)-dimensional vector of ones.

From Theorem 6, (5.13), and (5.18), the conditional distribution of \( \mathbf{Y}_{mix_j} \), given \( \mathbf{Y}_{obs_j} \), has the covariance

\[ \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} = \sigma_c^2 \mathbf{I}_{a_j} + \sigma_c^2 \mathbf{J}_{a_j} - \frac{(t - a_j)(\sigma_c^2)}{\sigma_c^2 + (t - a_j)\sigma_c^2} \mathbf{J}_{a_j}, \]

\[ = \sigma_c^2 \mathbf{I}_{a_j} + \frac{\sigma_c^2 \sigma_c^2}{\sigma_c^2 + (t - a_j)\sigma_c^2} \mathbf{J}_{a_j}. \quad (5.20) \]

This gives us all of the distributions necessary for our imputation procedure. The RMNI procedure creates \( m \) completed-data sets and can be summarized in the following five steps. For \( k = 1, \ldots, m \),
1. Use the \( n_* \) completely observed data to compute SSE, SSC, and \( \hat{\phi} \) as defined in Chapter 3.

2. Draw \( \sigma^2_{r*} \sim (t - 1)(n_* - 1)\text{SSE}\chi^2_{(t-1)(n_*-1)} \).

3. Draw \( \sigma^2_{e*} \sim (n_* - 1)\text{SSC}\chi^2_{n_*-1} \), and let

\[
\sigma^2_{e*} = \max\{0, t^{-1}(\sigma^2_{e*} - \sigma^2_{r*})\}.
\]

4. Draw \( \theta_* \sim N_{t-1}(\hat{\phi}, \sigma^2_{r*}\Sigma_*), \) where \( \Sigma_* = n_*^{-1}[I_{t-1} - t^{-1}J_{t-1}] \), and let

\[
\theta_{i*} = \phi_{i*} + Y_{i*}, \quad i = 1, \ldots, t.
\]

5. Let \( a_j \) be the number of missing values for the \( j \)-th subject. If \( a_j > 0 \), then draw \( Y_{\text{mis}_j} \) from

\[
Y_{\text{mis}_j} \sim N_{a_j}\left(\theta_{\text{mis}_j} + \frac{\sigma^2_{e*}}{\sigma^2_{e*} + (t - a_j)\sigma^2_{e*}} \sum_{i \in \text{obs}_j} (Y_{ij} - \theta_{i*})1_{a_j}, \right.
\]

\[
\sigma^2_{e*}1_{a_j} + \frac{\sigma^2_{e*}\sigma^2_{r*}}{\sigma^2_{e*} + (t - a_j)\sigma^2_{r*}} J_{a_j} \right), \quad (5.21)
\]

where \( \theta_{\text{mis}_j} \) is a vector of \( \theta_{i*} \) that corresponds to the missing treatments for the \( j \)-th subject. For example, if the fourth subject is missing the first and third treatments, then \( \theta_{\text{mis}_j} = (\theta_{1*}, \theta_{3*})' \). Use \( Y_{\text{mis}_j} \) to impute the appropriate missing values of the \( j \)-th subject to create a vector of \( t \) completed observations for the \( k \)-th completed-data set

\[
Y_{jk} = (Y_{1jk}, \ldots, Y_{ijk})'. \quad (5.22)
\]
The RMNI procedure is not the only multiple imputation procedure possible for this model. Other procedures can be developed. An alternate procedure could be a variation of the procedure introduced by Schafer and Rubin (1990). However, the procedure discussed in their paper relies on Gibbs sampling to draw the imputed values. To use Gibbs sampling, the imputer would need to evaluate the convergence of the Gibbs sampler. Gibbs sampling can also be computationally expensive. We prefer to explore the proposed procedure which has \( m \) fixed so that the process can be performed in real time without extensive user intervention.

### 5.3 Simulation of a repeated measures data set with missing data

A simulation was conducted using the RMNI procedure to evaluate the properties of the imputation procedure. Two properties which concern us the most are the accuracy of the estimators for \( \theta, \sigma_r^2, \) and \( \sigma_e^2 \), and the preservation of the covariance structure.

The simulation assumes a repeated measures model with normal errors with five treatments and 30 subjects. The values for the parameters are taken as \( \theta = (1, 3, 5, 7, 7)' \), \( \sigma_r^2 = 0.8 \), and \( \sigma_e^2 = 0.2 \), so that the vector of observations of each subject, \( Y_j \), has a multivariate normal distribution

\[
Y_j \sim N_5 \left( (1, 3, 5, 7, 7)', 0.8I_5 + 0.2J_5 \right).
\]  

(5.23)

The simulation randomly creates 5000 complete-data sets for which the exact analysis is carried out. The exact analysis will be used as a reference for comparing
the multiple imputation results. The same randomly assigned missing data pattern is used to create the missing data for all of the data sets. The missing data pattern is selected by drawing a Bernoulli(0.10) random number for each observation. If the Bernoulli random value is 0, then the observation is considered missing. There were a total of 16 missing values from 14 different subjects, leaving 16 fully observed subjects. Treatments 1 through 5 are missing 3, 4, 0, 6, and 3 observations, respectively. Two subjects have two missing observations. One has missing values at treatment 1 and 5. The other has missing values at treatments 1 and 4. The other subjects with missing data have only one missing observation each.

For each complete-data set, the estimates for $\theta$, $\sigma^2_\epsilon$, $\sigma^2_c$, and the covariance matrix are calculated. These estimates will be compared to the estimates from the RMNI procedure. To compare the covariance structure of the complete-data to the covariance structure completed-data from the RMNI procedure, an estimator of the covariance matrix is used and is written as

$$\hat{\gamma}_{i'j'} = \frac{1}{29} \sum_{j=1}^{n} (Y_{ij} - Y_{i.})(Y_{i'j} - Y_{i'.}), \quad 1 \leq i, i', \leq t. \quad (5.24)$$

For the RMNI procedure, $m = 100$ repetitions are taken so that 100 completed-data sets are created. The 16 completely observed subjects are used to calculate the initial estimates defined in Step 1 of the procedure. For each completed-data set, the estimates for $\theta$, $\sigma^2_\epsilon$, $\sigma^2_c$, and the covariance matrix are calculated using the same formulas as in the complete-data case. These estimates are averaged over the 100
repetitions, which yields the estimators

\[
\hat{Y}_{ij} = \frac{1}{(100)(30)} \sum_{k=1}^{100} \sum_{j=1}^{30} Y_{ijk},
\]

\[
\hat{\sigma}_c^2 = \frac{1}{(100)(29)(4)} \sum_{k=1}^{100} \sum_{i=1}^{5} \sum_{j=1}^{30} (Y_{ijk} - Y_{i,k} - Y_{.j,k} + Y_{..k})^2,
\]

\[
\hat{\sigma}_{ce}^2 = \frac{1}{(100)(4)} \sum_{k=1}^{100} \sum_{j=1}^{30} (Y_{.jk} - Y_{..k})^2,
\]

making

\[
\hat{\sigma}_c^2 = \max \left(0, 0.2 \left(\hat{\sigma}_{ce}^2 - \hat{\sigma}_c^2\right)\right).
\]

The estimate of the covariance matrix has the same formula as (5.24) and is averaged over the 100 repetitions.

The results of the simulation are presented in Figures 1, 2, and 3 which plot the point estimates from the actual data (on the horizontal axis) versus the point estimates from the RMNI procedure (on the vertical axis) for the parameters \(\theta_1, \ldots, \theta_5\), \(\sigma_c^2\), and \(\sigma_e^2\), respectively. The line in each plot is the diagonal. If a point lies on the diagonal, then the estimator for the actual data and the estimator produced by the RMNI procedure are equivalent. Moreover, if the estimators from the imputed data sets were overestimating their respective parameters, the points would tend to be above the diagonal. If the estimators from the imputed data sets were underestimating their respective parameters, the points would tend to be below the diagonal.

The plots of the estimators for the treatment effects show the imputed data sets yield estimates close to the actual data. More importantly, the plots indicate that the estimators from the imputed data sets do not appear to under- or over-estimate the true \(\theta_i\). Since there are no missing data for \(\theta_3\), the plot for the estimators of \(\theta_3\)
Table 1: Average of the estimates for the variance-covariance matrix using the actual data and the imputed data.

<table>
<thead>
<tr>
<th></th>
<th>Actual Data</th>
<th></th>
<th>Imputed Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.997 0.202</td>
<td>0.201</td>
<td>0.202 0.200</td>
</tr>
<tr>
<td>2</td>
<td>0.202 1.004</td>
<td>0.203</td>
<td>0.201 0.203</td>
</tr>
<tr>
<td>3</td>
<td>0.201 0.203</td>
<td>1.002</td>
<td>0.203 0.199</td>
</tr>
<tr>
<td>4</td>
<td>0.202 0.201</td>
<td>0.203 1.002</td>
<td>0.202 0.198</td>
</tr>
<tr>
<td>5</td>
<td>0.201 0.203</td>
<td>0.199 0.202</td>
<td>1.000 0.198</td>
</tr>
</tbody>
</table>

shows that the estimators are the same. Also, note that the more missing data there are for a particular treatment, the wider the points are spread about the diagonal. For example, \( \theta_4 \), with six missing observations, shows the greatest spread.

Figure 2 shows that the estimator for \( \sigma^2_e \) from the imputed data sets is generally overestimating \( \sigma^2_e \). The departure from the diagonal is more pronounced as the value of the estimator departs from the true parameter value \( \sigma^2_e = 0.80 \). If the estimate for \( \sigma^2_e \) is larger than 0.80, then the estimate from the RMNI procedure is often greater than the complete-data estimate. This can also be seen in estimates of the variances listed in Table 1.

Figure 3 does not seem to indicate that the covariance terms are over- or underestimated. Table 1 offers more insight into the covariance structure.

Table 1 lists the estimate of the covariance matrix of the five treatments averaged over the 5000 replications. Notice that the variances for the imputed data sets are somewhat larger than those for the actual data sets and the covariances for the imputed data sets are slightly smaller than those for the actual data sets. This may
imply that the confidence intervals formed in the multiple comparison procedure will be conservative since they may have a larger variance estimate and, hence, a larger confidence width.

Although the estimate for $\sigma^2_c$ from the RMNI procedure is typically higher than the estimate from the actual data, increasing the number of repetitions, $m$, may help. In the next chapter, we will use the RMNI procedure to develop confidence intervals by modifying the Dunnett’s multiple comparison procedure. We will see, through simulations, that the overestimation of $\sigma^2_c$ does not seem to affect the confidence intervals.
Figure 1: Plots of the estimators for $\theta_i$, $i = 1, \ldots, 5$, using the actual data (horizontal) versus the imputed data (vertical) over the 5000 replications.
Figure 2: Plot of the estimator for $\sigma^2$ using the actual data (horizontal) versus the imputed data (vertical) over the 5000 replications.
Figure 3: Plot of the estimator for $\sigma^2_c$ using the actual data (horizontal) versus the imputed data (vertical) over the 5000 replications.
CHAPTER VI

Modified Dunnett's multiple comparison procedure using multiple imputation

In this chapter, we will use the completed-data sets created by the RMNI procedure in Chapter 5 to develop confidence intervals based on Dunnett's multiple comparison procedure. We will also conduct a series of simulations to assess how well our procedure works and to approximate the \( \alpha \)-level of the simultaneous confidence intervals.

Using the frequentist-based estimators derived in Chapter 3, we have established that \( Y_{i,k} \) is the estimate for treatment \( i, i = 1, \ldots, t \), and \( \hat{\sigma}_{\epsilon k}^2 \) is the MSE for the \( k \)th completed-data set. The multiple imputation estimators for \( \theta_i \) and \( \sigma^2 \) are

\[
Y_{i,*} = m^{-1} \sum_{k=1}^{m} Y_{i,k}, \quad i = 1, \ldots, t \tag{6.1}
\]

and

\[
\hat{\sigma}^2_{\epsilon,*} = m^{-1} \sum_{k=1}^{m} \hat{\sigma}^2_{\epsilon k}, \quad \tag{6.2}
\]

respectively. The estimate of the between-imputation variance-covariance matrix is defined by (2.6). From these three estimates, we wish to derive confidence intervals based on Dunnett's multiple comparison procedure. Let \( \mathbf{Y}_* = (Y_{1,*}, \ldots, Y_{t,*}) \). From (2.11), the variance-covariance matrix for \( \mathbf{Y}_* \) is approximately

\[
\mathbf{T} = n^{-1} \mathbf{V} + (1 + m^{-1}) \mathbf{B}, \tag{6.3}
\]

69
where \( V \) is defined in (3.9) and \((1 + m^{-1})B = (b_{ij})_{i\times t}\) depends on the amount and distribution of the missing data. The exact form of \( B \) is unknown—which is partially why multiple imputations are employed. \( T \) can be written as

\[
T = \sigma_t^2 \left[ \frac{1}{n} I_t + \frac{\sigma_c^2}{n\sigma_r^2} J_t \right] + (1 + m^{-1})B
\]

\[
= \sigma_t^2 \left[ \frac{1}{n} I_t + \frac{\sigma_c^2}{n\sigma_r^2} J_t + \frac{1}{\sigma_r^2} B \right].
\]

The correlation matrix for the multivariate T-distribution used in Dunnett’s multiple comparison procedure uses the variance-covariance matrix of the estimates of the fixed effects. The equation for this correlation is (4.10) and, using (6.4), \( \rho_{ij} \) can be written as

\[
\rho_{ij} = \frac{1}{\sigma_r^2} (b_{ij} - b_{it} - b_{jt} + b_{it}) + n^{-1} \left[ \frac{2}{n} + \frac{b_{it} + b_{jt} - 2b_{it}}{\sigma_r^2} \right]^{1/2} \left[ \frac{2}{n} + \frac{b_{it} + b_{jt} - 2b_{it}}{\sigma_r^2} \right]^{1/2}.
\]

The other part of Dunnett’s multiple comparison procedure is \( v_{ii} + v_{ii} - 2v_{it} \) used in (4.9). Using (6.4), this value can be written as

\[
v_{it} + v_{it} - 2v_{it} = \frac{2}{n} + \frac{b_{ii} + b_{it} - 2b_{it}}{\sigma_r^2}.
\]

(6.5) shows that \( \sigma_r^2 \) still drops out of \( \rho_{ij} \), but the values of \( \rho_{ij} \) are not all the same. The form of \( T \) is too complex to derive a modification to Dunnett’s multiple comparison procedure in the general case. Instead, we first derive the modified Dunnett’s multiple comparison procedure in the case when the amount of missing data is the same for each treatment. Through simulations, we will show how well this modification works when the number of missing values varies for each treatment. We
will then address the effect that $b_{ij}$ has on the confidence intervals when the $b_{ij}$ are not all the same.

### 6.1 Special case: one missing value per treatment

We will discuss the simplest case where there is only one missing value in each treatment and, at most, one missing value for each subject. Any pattern of missing data which leads to $B = a_1 I_t + a_2 J_t$, where $a_1$ and $a_2$ are constants, will have all the $p_{ij}$ equivalent. This special case is not likely to happen when the response mechanism is random, but this case is easier to analyze than the general case.

Since the missing values are missing for different subjects, the imputed values are drawn independently, which means that the covariances in $B$ are zero and drop out of (6.5).

For each repetition, the missing values will be drawn from a normal distribution with different means, depending on the observed values for the given subject, but the variances will all, conditioned on the fully observed subjects, have the form

$$
\sigma^2 + \frac{\sigma^2 \sigma^2}{\sigma^2 + (t - 1) \sigma^2}.
$$

(6.7)

Since the imputed values will have the same variance, the variance due to multiple imputations will be the same. This implies that $b = b_{11} = \cdots = b_{tt}$, so that $B = b I_t$ and

$$
T = \sigma^2 \left[ \frac{1}{n} I_t + \frac{\sigma^2}{n \sigma^2} J_t \right] + (1 + m^{-1}) b I_t
$$

$$
= \sigma^2 \left[ \left( \frac{1}{n} + \frac{(1 + m^{-1}) b}{\sigma^2} \right) I_t + \frac{\sigma^2}{n \sigma^2} J_t \right].
$$

(6.8)
We can calculate the correlation, $\rho_{ij}$, for the multivariate $T$-distribution used in Dunnett’s multiple comparison procedure and for $v_{ii} + v_{tt} - v_{it}$ by substituting (6.8) into (6.5) and (6.6), respectively, to get

\begin{align*}
\rho_{ij} &= \frac{\frac{\sigma_i^2}{n\sigma_i^2} - \frac{\sigma_j^2}{n\sigma_j^2} - \frac{\sigma_{ij}^2}{n\sigma_{ij}^2} + \left(\frac{1}{n} + \frac{(1+m^{-1})b}{\sigma_i^2} + \frac{\sigma_{ij}^2}{n\sigma_{ij}^2}\right) - \left(\frac{1}{n} + \frac{(1+m^{-1})b}{\sigma_j^2} + \frac{\sigma_{ij}^2}{n\sigma_{ij}^2}\right)}{\sqrt{\left(\frac{1}{n} + \frac{(1+m^{-1})b}{\sigma_i^2} + \frac{\sigma_{ij}^2}{n\sigma_{ij}^2}\right) + \left(\frac{1}{n} + \frac{(1+m^{-1})b}{\sigma_j^2} + \frac{\sigma_{ij}^2}{n\sigma_{ij}^2}\right)} - 2\left(\frac{\sigma_{ij}^2}{n\sigma_{ij}^2}\right)} \\
&= \frac{1}{2}
\end{align*}

(6.9)

and

\begin{align*}
v_{ii} + v_{tt} - v_{it} &= \left(\frac{1}{n} + \frac{(1+m^{-1})b}{\sigma_i^2} + \frac{\sigma_{ij}^2}{n\sigma_{ij}^2}\right) + \left(\frac{1}{n} + \frac{(1+m^{-1})b}{\sigma_j^2} + \frac{\sigma_{ij}^2}{n\sigma_{ij}^2}\right) - 2\left(\frac{\sigma_{ij}^2}{n\sigma_{ij}^2}\right) \\
&= 2\left(\frac{1}{n} + \frac{(1+m^{-1})b}{\sigma_i^2}\right)
\end{align*}

(6.10)

Confidence intervals based on Dunnett’s multiple comparison procedure have the form (4.9). Substituting (6.9) and (6.10) into (4.9), the half-width of the confidence interval is

\begin{align*}
\left| T \right|_{t-1, w, \{1/2\}} \hat{\sigma}_{r^*} \sqrt{2\left(\frac{1}{n} + \frac{(1+m^{-1})b}{\hat{\sigma}_{r^*}^2}\right)} = \left| T \right|_{t-1, w, \{1/2\}} \sqrt{2\left(\frac{\hat{\sigma}_{r^*}^2}{n} + (1+m^{-1})\hat{b}\right)}
\end{align*}

(6.11)

for some degrees of freedom, $w$, and some estimate, $\hat{b}$, which will be defined later.

The statistic of interest from (6.11), call it $S^2_r$, is

\begin{align*}
S_r^2 &= 2n^{-1}\hat{\sigma}_{r^*}^2 + 2(1+m^{-1})\hat{b}
\end{align*}

(6.12)

If we assume that $\hat{\sigma}_{r^*}^2$ has an approximate Chi-square distribution with $\nu$ degrees of freedom, where

\begin{align*}
\nu = (n - 1)(t - 1) - \text{(Number of missing values)},
\end{align*}

(6.13)
and we take a sufficient number of repetitions so that \( \hat{b} \) is approximately constant (or that the distribution of the estimate of \( b \) is approximately degenerate), then \( S^2 \) approximately has a non-central Chi-square distribution. In this special case, \( \nu = (n-1)(t-1)-t \), but we will continue to use the form of \( \nu \) in (6.13) because the confidence intervals developed here will be used with differing amounts of missing data. We will approximate the distribution of \( S^2 \) with a Chi-square distribution by matching the first and second central moments. This is the same method that Li, Rathunathan, and Rubin (1991) used to find the degrees of freedom for an approximate \( F \)-distribution (see Chapter 2). Assume that

\[
S^2 = \delta X, \quad (6.14)
\]

where \( \delta \in \mathbb{R}^+ \) and \( X \sim \chi^2_w \).

The mean and variance for \( S^2 \) from (6.12) are

\[
E \left( S^2 \right) = \frac{2\nu}{n} + 2(1 + m^{-1}) \hat{b} \quad (6.15)
\]

and

\[
Var \left( S^2 \right) = \frac{4}{n^2(2\nu)}. \quad (6.16)
\]

The mean and variance for \( S^2 \) from (6.14) are

\[
E \left( S^2 \right) = \delta w \quad (6.17)
\]

and

\[
Var \left( S^2 \right) = 2\delta^2 w. \quad (6.18)
\]

Equating the expectations and variances, and solving for \( \delta \) and \( w \), we find that

\[
\delta = \frac{2}{n} \left[ \frac{\nu}{\nu + (1 + m^{-1})n\hat{b}} \right] \quad (6.19)
\]
and
\[ w = \frac{(\nu + (1 + m^{-1})n\hat{b})^2}{\nu}. \] (6.20)

For Dunnett’s multiple comparison in the balanced one-way repeated measures model, the confidence intervals in (4.20) have a half-width of \( S\sqrt{2/n} \). In this complete-data case, \( \delta \) would correspond to \( 2/n \) and \( w \) would be \( \nu \). \( \delta \) and \( w \) for the missing data case are generalizations of the forms for the complete data case; the difference lies in the value of \( (1 + m^{-1})n\hat{b} \) (note that if the data set is complete, this value would be zero).

Consider the estimator of the between-repetition variance, \( B_m \), from (2.6). Let \( \hat{b}_{ii'} \) be the \((i, i')\) element of \( B_m \). A similar argument to that used in (2.75) to obtain the \( i \)th diagonal element of \( B_m \) can be used to show that
\[
\hat{b}_{ii'} = \frac{a_i a_{i'}}{n^2(m - 1)} \sum_{k=1}^{m} \left( Y_{i(.)k} - Y_{i(.)\#} \right) \left( Y_{i'i(.)k} - Y_{i'i(.)\#} \right),
\] (6.21)
where \( a_i \) and \( a_{i'} \) are the numbers of missing values in treatments \( i \) and \( i' \), respectively, and \( Y_{i(.)k} \) is the average of the \( a_i \) imputed values for treatment \( i \) in the \( k \)th repetition as defined in (2.64). In our special case, \( a_i = 1 \) for all \( i \), but we will leave the \( a_i \) in the formula since we will be extending this to other cases. Since we wish to estimate \( b \), we need only consider the case where \( i = i' \).

\[
\hat{b}_{ii} = \frac{a_i^2}{n^2(m - 1)} \sum_{k=1}^{m} \left( Y_{i(.)k} - Y_{i(.)\#} \right)^2. \] (6.22)

Since \( B = bI_t \), we can estimate \( b \) by
\[
\hat{b} = t^{-1} \sum_{i=1}^{t} \hat{b}_{ii}. \] (6.23)
In our special case, \( a_i = 1 \), so that the expectation of \( \hat{b} \) is

\[
E(\hat{b}) = t^{-1} \sum_{i=1}^{t} E(\hat{b}_{ij})
\]

\[
= t^{-1} \sum_{i=1}^{t} n^{-2} E \left[ \frac{1}{m - 1} \sum_{k=1}^{m} (Y_{i(k)} - Y_{i(*)})^2 \right]
\]

\[
= t^{-1} \sum_{i=1}^{t} n^{-2} \text{Var}(Y_{ijk})
\]

\[
= n^{-2} \text{Var}(Y_{ijk}), \quad (6.24)
\]

where \( Y_{ijk} \) is an imputed value. The marginal distribution of \( Y_{ijk} \) is difficult to calculate directly. One would expect that \( \sigma_r^2 \) and/or \( \sigma_x^2 \) would affect the variance of the imputed values. We will use simulations to find if \( \sigma_r^2 \) and/or \( \sigma_x^2 \) will affect the half-width of the confidence intervals.

Using the above results, the \( t - 1 \) multiply-imputed \((1 - \alpha)\) simultaneous confidence intervals based on Dunnett's multiple comparison procedure for the one-way repeated measures model have the form

\[
Y_{i,*} - Y_{l,*} \pm T \left|_{t-1,w,(1/2)} \right. \left( \sqrt{2 \left( \frac{\hat{\sigma}_r^2}{n} + (1 + m^{-1})\hat{b} \right)} \right), \quad (6.25)
\]

where \( T \left|_{t-1,w,(1/2)} \right. \) is the \( \alpha \)-th percentile of the \((t - 1)\)-dimensional multivariate \( T \)-distribution with \( w \) degrees of freedom defined in (6.20) and correlation matrix \( \{ \rho_{ij} = 1/2 \} \).

### 6.2 Simulations

We conducted four sets of simulations, each with a different pattern of missing data. The first set of simulations has a pattern of missing data corresponding to the pattern
of missing data described in Section (6.1)—namely, only one missing value for each treatment. One important reason to conduct this set of simulations is to see how accurate the confidence intervals are in the case for which our procedure was developed. We hope that the confidence intervals would be good approximations in this simple case. We will vary the number of multiple imputations, $m$, the number of treatments $t$, the number of subjects $n$, and the two variances, $\sigma_1^2$ and $\sigma_2^2$. Since $n$ will be set at both 20 and 30, the proportions of missing data will be 5\% and 3\frac{1}{3}\% respectively.

The next three sets of simulations have three different patterns of missing data and different amounts of missing data. For the second set of simulations, we use the random pattern of missing data generated in Section (5.3). This pattern of missing data is generated using a 10\% chance of any value being missing and has the fewest missing values of the three sets. The third set of simulations has a pattern of missing data which was developed from the pattern of missing data in the second set of simulations. This pattern of missing data mimics the type of censored data that might occur from an actual experiment and cannot be considered missing at random. The fourth set of simulations has almost as many missing values as the third set, but the missing values were generated randomly. This set of simulations has the same number of fully observed subjects as the second set of simulations, but the 14 subjects who have missing values have the response pattern drawn randomly with a 50\% chance of missingness. These patterns of missing data will be explained more fully in their respective sections.

For all four sets of simulations, we want to discover (1) how close the $\alpha$-levels of
the confidence intervals created by the multiple imputation procedure come to the nominal rate, (2) how the half-widths of the confidence intervals compare to the half-widths of the confidence intervals created with no missing data, and (3) the effects of \( \sigma_r^2 \) and \( \sigma_e^2 \) on the half-widths of the confidence intervals. The number of repetitions, \( m \), is briefly investigated to help determine if \( m \) is sufficiently large to estimate \( \mathbf{B} \). We will also try to determine if the number of treatments, \( t \), and the number of subjects, \( n \), have an effect on the RMNI procedure.

We hope that, even with different proportions of missing data for each treatment, the \( \alpha \)-levels for the multiply-imputed confidence intervals will be approximately the same as the nominal rates or indicate a conservative procedure. We will be able to create the multiply-imputed confidence intervals for each simulated data set and determine if the true means are within the confidence intervals. Simulating many data sets, we will be able to estimate the \( \alpha \)-level of the multiply-imputed confidence intervals.

We will compare the half-widths from the multiply-imputed confidence intervals with the half-widths from the complete-data confidence intervals. We expect the complete-data confidence intervals to be smaller because they are based on more observed data. We hope that the increase in the half-widths of the multiply-imputed confidence intervals will be somehow proportional to the amount of missing data.

We are more concerned with the effect that the random effects variance, \( \sigma_r^2 \), has on the widths of the multiply-imputed confidence intervals, rather than with the effect of \( \sigma_r^2 \), because \( \sigma_e^2 \) is not a factor in the complete-data confidence intervals. If \( \sigma_e^2 \)
causes the widths of the multiply-imputed confidence intervals to increase, then the multiply-imputed confidence intervals may suffer when the value of $\sigma_r^2$ is large relative to the size of $\sigma_k^2$.

All of the simulations are conducted on a DEC500/25 workstation; the computer code is written in double precision FORTRAN, and the program calls the subroutines RNSET, RNBIN, DCHFAC, DRNNOA, DRNCHI, and CTIME from the IML subroutine libraries. The computer program is listed in Appendix B. The subroutine CTIME is only used to estimate the amount of time the RMNI procedure takes to compute the multiple imputation estimates and is not necessary in the algorithm. Critical values for the multivariate $T$-distribution are calculated using a program written by Professor Jason Hsu in Splus 3.1. Since the value of the degrees of freedom, $w$, in (6.20) can be a real number, $w$ is rounded down to the nearest integer in the calculation of the critical value.

All four sets of simulations will assume that the data for each subject are being drawn from a multivariate normal distribution

$$Y_j \sim N(\theta, \sigma_r^2 I_t + \sigma_k^2 J_t),$$

(6.26)

where $\theta = (1, 1, \ldots, 1)'$. The values of $\sigma_r^2$ and $\sigma_k^2$ will vary within each set of simulations to help assess the effect of the variances. We will create 95% simultaneous confidence intervals for the $t - 1$ treatments.

For all the simulations, we will replicate 15,000 data sets. We chose 15,000 replications to get a good estimate of $\alpha$. 15,000 replications will allow us to estimate $\alpha$ within approximately $\pm 0.0035$. We would have liked to increase the number of repli-
cations, but because of the number of simulations we were conducting, we needed to limit the number of replications because of time constraints (a single run of 15,000 replications took 3 to 50 hours, depending on the value of $m$). We will now describe how we arrived at the error term of $\pm 0.005$. Let $\alpha$ be the probability of making at least one incorrect inference. We wish to find a 95% confidence interval for the estimate for $\alpha$. Assuming that the data sets are independent, if $X$ is the number of times at least one incorrect inference is made, then $X$ is distributed as a binomial$(15000, \alpha)$. If $\alpha$ is close to 0.05, a Clopper-Pearson 95% confidence interval for the probability of success listed in Hollander and Wolfe (1973) will be

$$\alpha \in \left( 0.05 \pm 1.96 \left[ \frac{0.05(1-0.05)}{15,000} \right]^{1/2} \right),$$

which reduces to

$$\alpha \in (0.0465, 0.0535).$$

### 6.2.1 Simulations for the special case: one missing value per treatment

For this set of simulations, we assume that the first $t$ subjects have one missing value each such that the first subject is missing the first treatment, the second subject is missing the second treatment, and so on until the $t$-th subject. For this set of simulations, we will first fix the values of $t = 5$, $n = 30$, $\sigma^2 = 0.8$, and $\sigma^2 = 0.2$, as in the the simulations in Section 5.3. We will set $m$ equal to 100, 500, and 1000. Using these three simulations, we will get an idea of when $m$ is sufficiently large. We will also be able to determine approximately how long the RMNI procedure is taking.
Table 2: Estimated half-widths and $\alpha$-level when $m = 100, 500, \text{ and } 1000$.

<table>
<thead>
<tr>
<th>$m$</th>
<th>Actual half-width</th>
<th>$\alpha$</th>
<th>Imputed half-width</th>
<th>$\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.5710</td>
<td>0.049</td>
<td>0.5831</td>
<td>0.048</td>
</tr>
<tr>
<td>500</td>
<td>0.5709</td>
<td>0.049</td>
<td>0.5830</td>
<td>0.049</td>
</tr>
<tr>
<td>1000</td>
<td>0.5701</td>
<td>0.049</td>
<td>0.5821</td>
<td>0.050</td>
</tr>
</tbody>
</table>

Using the results from the three simulations, we will choose a value of $m$ which we believe is sufficiently large and run more simulations by varying $t$, $n$, $\sigma_r^2$, and $\sigma_c^2$.

Table 2 lists the estimated half-widths and $\alpha$-levels when $t = 5$, $n = 30$, $\sigma_r^2 = 0.8$, $\sigma_c^2 = 0.2$, and $m$ is 100, 500, and 1000 for confidence intervals using estimates from the complete-data (Actual) and the estimates derived from the RMNI procedure (Imputed). Both the half-widths and $\alpha$-levels for all three values of $m$ are essentially the same. All of the estimates of $\alpha$ fall within the confidence interval (6.28). Although it seems that 100 imputations is a sufficiently large number of repetitions to get good estimates, we decided to use 500 imputations for the rest of the simulations in this set to be on the conservative side. Using such a large number of repetitions may be inefficient, but the amount of computer time necessary for each replication is small. The times for the RMNI procedure to make the imputations and calculate the estimates are approximately 0.6, 2.6, and 5.4 seconds for $m$ equal to 100, 500, and 1000, respectively.

Setting $m = 500$, we will vary $t$, $n$, $\sigma_r^2$, and $\sigma_c^2$. We chose to set $t = 3, 5, \text{ and } 7$ because we believe that these are reasonable sizes for the number of treatments and
larger values of $t$ cause the problem to be more computationally expensive. We chose to use $n$ equal to 20 and 30 to see how well the procedure will perform when the number of subjects is fairly small. The procedure needs to calculate initial estimates from the completely observed data. If $n$ is too small, then there would be few fully observed subjects to use in calculating the initial estimates. We believe that, as $n$ increases, the better the technique will work because a larger $n$ should yield more fully observed subjects and thus better initial estimates. We chose $\sigma_c^2$ and $\sigma_r^2$ in pairs so that the within-subject correlation would be easy to calculate. The first pair was chosen so that there would be a within-subject correlation of 0.2 and $Var(Y_{ij}) = 1$. We chose the second and third pairs of variances so that $\sigma_c^2(=9)$ is large in the second pair and $\sigma_r^2(=9)$ is large in the third pair. If $\sigma_c^2$ has an effect on the half-width of the confidence intervals, then the second pair should cause a significant increase in the half-width. If $\sigma_r^2$ has an effect on the half-width of the confidence intervals, then the third pair should cause a significant increase in the half-width.

Table 3 lists the estimated half-widths and Table 4 lists the estimated $\alpha$-levels for the simultaneous confidence intervals created by the RMNI procedure, given the pattern of missing data and the various values of $t$, $n$, and pairs of $\sigma_c^2$ and $\sigma_r^2$. Again, all of the estimated values of $\alpha$ fall within the bounds given in (6.28). Also, the ratios of the multiple imputation half-widths to the complete-data half-widths are in the range of 1.02 to 1.03. This would imply that $t$, $n$, $\sigma_c^2$, and $\sigma_r^2$ did not have much, if any, effect in changing the confidence intervals when the multiple imputation procedure is implemented.
Table 3: Estimated half-widths from simulations for $m = 500$ and $t = 3, 5, 7$. $n = 20, 30$, and $(\sigma^2_r, \sigma^2_c) = (0.8, 0.2), (9, 1), (1, 9)$.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$t$</th>
<th>$\sigma^2_r = 0.8, \sigma^2_c = 0.2$</th>
<th>$\sigma^2_r = 1, \sigma^2_c = 9$</th>
<th>$\sigma^2_r = 9, \sigma^2_c = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Actual</td>
<td>Imputed</td>
<td>Actual</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>0.052</td>
<td>0.048</td>
<td>0.5709</td>
</tr>
<tr>
<td>20</td>
<td>5</td>
<td>0.053</td>
<td>0.053</td>
<td>0.5298</td>
</tr>
<tr>
<td>30</td>
<td>5</td>
<td>0.049</td>
<td>0.049</td>
<td>0.050</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>0.051</td>
<td>0.051</td>
<td>0.050</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>0.051</td>
<td>0.051</td>
<td>0.050</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>0.051</td>
<td>0.051</td>
<td>0.050</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>0.051</td>
<td>0.051</td>
<td>0.050</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>0.051</td>
<td>0.051</td>
<td>0.050</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>0.051</td>
<td>0.051</td>
<td>0.050</td>
</tr>
</tbody>
</table>

Table 4: Estimated $\alpha$-levels of the simultaneous confidence intervals from simulations for $m = 500$ and $t = 3, 5, 7$, $n = 20, 30$, and $(\sigma^2_r, \sigma^2_c) = (0.8, 0.2), (9, 1), (1, 9)$.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$t$</th>
<th>$\sigma^2_r = 0.8, \sigma^2_c = 0.2$</th>
<th>$\sigma^2_r = 1, \sigma^2_c = 9$</th>
<th>$\sigma^2_r = 9, \sigma^2_c = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Actual</td>
<td>Imputed</td>
<td>Actual</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>0.052</td>
<td>0.048</td>
<td>0.051</td>
</tr>
<tr>
<td>20</td>
<td>5</td>
<td>0.053</td>
<td>0.053</td>
<td>0.053</td>
</tr>
<tr>
<td>30</td>
<td>5</td>
<td>0.049</td>
<td>0.049</td>
<td>0.050</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>0.051</td>
<td>0.051</td>
<td>0.050</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>0.051</td>
<td>0.051</td>
<td>0.050</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>0.051</td>
<td>0.051</td>
<td>0.050</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>0.051</td>
<td>0.051</td>
<td>0.050</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>0.051</td>
<td>0.051</td>
<td>0.050</td>
</tr>
</tbody>
</table>
The values of \( n \) and \( t \) did not seem to affect the \( \alpha \)-level or the half-widths. These two variables may have an effect on the minimum value of \( m \) necessary, but, since we took \( m = 500 \) for all of the simulations in Table 3 and Table 4, the effect would not be apparent.

It also seems that, even when \( \sigma_r^2 = 9 \), the half-width of the confidence interval is proportionally close to the complete-data confidence interval half-width. This implies that \( \sigma_r^2 \) has little or no effect on the confidence intervals. The effect of \( \sigma_r^2 \) will be investigated more closely at the end of this chapter.

### 6.2.2 Simulations for the random pattern of missing data

For the second set of simulations, we used the pattern of missing data randomly generated in Section 5.3. This pattern of missing data represents a randomly generated pattern, rather than a fixed pattern as in the first set of simulations, with a moderate amount of missing data (around 10%). We will use the same confidence intervals developed for the special case at the same \( \alpha \)-level to see (1) how close the \( \alpha \)-level for the confidence intervals from the multiple imputation estimates are to the nominal rate, when the assumption of the same amount of missing data per treatment is violated, and (2) if \( \sigma_r^2 \) has any effect on the half-width of the confidence intervals. \( \sigma_r^2 \) may have an effect since the assumption of the same number of missing values in the special case is violated. Just as in the first set of simulations, we will first fix \( \sigma_r^2 = 0.8 \), and \( \sigma_r^2 = 0.2 \), and vary \( m \) as 100, 500, and 1000. The values of \( t \) and \( n \) are fixed as 5 and 30, respectively.

Table 5 lists the estimated half-widths and \( \alpha \)-levels for the confidence intervals
Table 5: Estimated half-widths and $\alpha$-level when $m = 100, 500, \text{ and } 1000$.

<table>
<thead>
<tr>
<th>$m$</th>
<th>Actual half-width</th>
<th>$\alpha$</th>
<th>Imputed half-width</th>
<th>$\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.5706</td>
<td>0.051</td>
<td>0.6179</td>
<td>0.051</td>
</tr>
<tr>
<td>500</td>
<td>0.5709</td>
<td>0.047</td>
<td>0.6177</td>
<td>0.047</td>
</tr>
<tr>
<td>1000</td>
<td>0.5708</td>
<td>0.049</td>
<td>0.6178</td>
<td>0.048</td>
</tr>
</tbody>
</table>

Table 6: Estimated half-widths and $\alpha$-level when $(\sigma_r^2, \sigma_z^2) = (0.8, 0.2), (1, 9), \text{ and } (9, 1)$.

<table>
<thead>
<tr>
<th>$(\sigma_r^2, \sigma_z^2)$</th>
<th>Actual half-width</th>
<th>$\alpha$</th>
<th>Imputed half-width</th>
<th>$\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0.8, 0.2)</td>
<td>0.5709</td>
<td>0.047</td>
<td>0.6177</td>
<td>0.047</td>
</tr>
<tr>
<td>(1, 9)</td>
<td>0.6385</td>
<td>0.048</td>
<td>0.6969</td>
<td>0.047</td>
</tr>
<tr>
<td>(9, 1)</td>
<td>1.9149</td>
<td>0.046</td>
<td>2.0628</td>
<td>0.048</td>
</tr>
</tbody>
</table>

using the complete data and the RMNI procedure. The same results as in the first set of simulations are found here. Namely, the $\alpha$-levels are within the error bound, and the half-widths for the different values of $m$ are approximately the same for the multiply-imputed confidence intervals. To be on the conservative side, we will again use 500 repetitions for the rest of the simulations in this section.

Since $t$, $n$, and $m$ are fixed, the only parameters to vary are $\sigma_r^2$ and $\sigma_z^2$. We will again use the pairs (0.8, 0.2), (1, 9), and (9, 1) for the same reasons as in the first set of simulations.

Table 6 lists the estimated half-widths and $\alpha$-levels when the variances are varied. The estimated values of $\alpha$ for the multiple imputation confidence intervals are within
the bounds in (6.28). Once again, the multiple imputation confidence intervals do not seem to be affected by the random effects variance, $\sigma_e^2$. The multiple imputation confidence intervals developed with the assumption of equal amounts of missing data for each of the treatments still seem to give reasonable results for this pattern of missing data.

6.2.3 Simulations for a specific pattern of missing data

In this section, we will modify the pattern of missing data in Section 6.2.2 so that it mimics a pattern of censored data that might occur in practice by first switching the response indicators for Treatment 3 with Treatment 1; we do this because Treatment 3 has no missing observations and we want at least one observed value for each subject. This would imply that the first treatment must have observations for every subject. Now, given this new matrix of response indicators, if subject $j$ has a missing value for treatment $i$, then all subsequent treatments are also missing. This pattern of missing data corresponds to subjects dropping out of the study after at least one observation is taken from them. This pattern of missing data is no longer random nonresponse and causes a great deal of missing values for the control (Treatment 1). This pattern of missing data is far from the case for which the RMNI procedure and the multiply-imputed confidence intervals were developed. We will use the RMNI procedure here to see how well it works in a situation far from ideal.

The pattern of missing data can be described by listing the number of subjects that have their first missing value at each of the treatments. The number of subjects (out of the 30 subjects in each treatment) that have their first missing value at treatments
Table 7: Estimated half-widths and $\alpha$-level when $(\sigma_r^2, \sigma_c^2) = (0.8, 0.2)$, $(1,9)$, and $(9,1)$ for the specific pattern of missing data.

<table>
<thead>
<tr>
<th>$m$</th>
<th>Actual half-width $\alpha$</th>
<th>Imputed half-width $\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.5706 0.050</td>
<td>0.7039 0.063</td>
</tr>
<tr>
<td>500</td>
<td>0.5705 0.048</td>
<td>0.7039 0.064</td>
</tr>
<tr>
<td>1000</td>
<td>0.5706 0.050</td>
<td>0.7032 0.061</td>
</tr>
</tbody>
</table>

2, 3, 4, and 5 are 4, 3, 6, and 1, respectively. Thus, there are four missing values in treatment 2, seven missing values in treatment 3, 13 missing values in treatment 4, and 14 missing values from treatment 5—38 missing values in all. This amounts to 25.3% of the data missing overall, but treatment 5 has 46.7% of its data missing and treatment 1 has no missing data, so the proportions of missing data on each of the treatments vary tremendously.

For this set of simulations, we once again fix $t = 5$, $n = 30$, $\sigma_r^2 = 0.8$, and $\sigma_c^2 = 0.2$, and vary $m$ as 100, 500, and 1000. Table 7 lists the estimated half-widths and $\alpha$-levels for the confidence intervals using the complete-data and the RMNI procedure. In this case, the $\alpha$-levels for all three values of $m$ are significantly above the nominal rate of 0.05. Increasing the number of repetitions does not seem to be a solution since the half-widths for all three values of $m$ are approximately the same.

Next, we fixed $m = 100$ and varied $\sigma_r^2$ and $\sigma_c^2$ just as in the second set of simulations. We varied the variances to see if they affected the half-widths. We chose $m = 100$ for this set of simulations since 100 repetitions seemed like a sufficient num-
Table 8: Estimated half-widths and $\alpha$-level when $(\sigma_c^2, \sigma_r^2) = (0.8, 0.2), (1, 9), \text{ and } (9, 1)$ for the specific pattern of missing data.

<table>
<thead>
<tr>
<th>$\sigma_c^2, \sigma_r^2$</th>
<th>Actual half-width</th>
<th>$\alpha$</th>
<th>Actual half-width</th>
<th>$\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0.8, 0.2)</td>
<td>0.5706</td>
<td>0.050</td>
<td>0.7039</td>
<td>0.063</td>
</tr>
<tr>
<td>(1, 9)</td>
<td>0.6380</td>
<td>0.050</td>
<td>0.8357</td>
<td>0.065</td>
</tr>
<tr>
<td>(9, 1)</td>
<td>1.9138</td>
<td>0.048</td>
<td>2.3350</td>
<td>0.050</td>
</tr>
</tbody>
</table>

To get a good estimate of $b$ and it greatly reduced the amount of computing time. Table 8 lists the half-widths and $\alpha$-levels for the different variances. The half-widths still do not seem to be affected by $\sigma_r^2$. Once again we see that the estimated $\alpha$-levels for the multiply-imputed confidence intervals seem different from the nominal rate, except in the case when $\sigma_c^2 = 9$ and $\sigma_r^2 = 1$. We have no explanation for why this case would be different from the rest.

Since the pattern of missing data is heavier for the later treatments, the variance-covariance matrix, $\mathbf{T}$, in (6.4) does not have a form which would result in $\rho_{ij} = 1/2$. When we use the confidence intervals developed in the special case, we assume that $\rho_{ij} = 1/2$. In this case, this assumption is violated. Since $\sigma_c^2$ does not seem to affect the confidence intervals, it may be possible to use techniques developed for the unbalanced fixed-effects linear models by adjusting the critical value to obtain a more appropriate interval for this pattern of missing data. This point will be discussed further in the future research section of Chapter 7.
Table 9: Bernoulli random numbers generated for the pattern of missing data (1 = missing, 0 = observed).

<table>
<thead>
<tr>
<th>Sub.</th>
<th>Treatment 1 2 3 4 5</th>
<th>Sub.</th>
<th>Treatment 1 2 3 4 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0 0 1 1 0</td>
<td>8</td>
<td>1 0 1 1 1</td>
</tr>
<tr>
<td>2</td>
<td>0 0 0 0 1</td>
<td>9</td>
<td>0 1 0 1 0</td>
</tr>
<tr>
<td>3</td>
<td>1 0 1 0 0</td>
<td>10</td>
<td>0 0 1 1 1</td>
</tr>
<tr>
<td>4</td>
<td>0 0 0 1 0</td>
<td>11</td>
<td>1 1 0 1 1</td>
</tr>
<tr>
<td>5</td>
<td>1 0 0 1 1</td>
<td>12</td>
<td>0 0 1 1 0</td>
</tr>
<tr>
<td>6</td>
<td>0 0 0 1 0</td>
<td>13</td>
<td>1 0 0 1 1</td>
</tr>
<tr>
<td>7</td>
<td>1 1 1 0 0</td>
<td>14</td>
<td>1 1 0 1 0</td>
</tr>
</tbody>
</table>

6.2.4 The fourth set of simulations

Since the third set of simulations yielded estimated α-levels above the nominal rate, we decided to run a fourth set of simulations with a similar percentage of missing data to that of the third set, but where the missing values are drawn at random. This would help us decide whether the percentage of missing data or the pattern of missing data was affecting the α-level.

As in the second and third sets of simulations, we let \( t = 5 \), \( n = 30 \), and the number of fully observed subjects remained at 16. Each possible observation for the remaining 14 subjects was drawn randomly from a Bernoulli(0.5). If the random number generated from the Bernoulli was one, then that observation was considered missing. Table 9 lists the 14 subjects and the value of the Bernoulli random variable generated for each of the five treatments. Since \( m = 500 \) has been more than sufficient
Table 10: Estimated half-widths and $\alpha$-level when $(\sigma_c^2, \sigma_2^2) = (0.8,0.2), (1,9), \text{ and } (9,1)$ for the fourth set of simulations.

<table>
<thead>
<tr>
<th>$(\sigma_c^2, \sigma_2^2)$</th>
<th>Actual half-width</th>
<th>$\alpha$</th>
<th>Imputed half-width</th>
<th>$\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0.8, 0.2)</td>
<td>0.5706</td>
<td>0.051</td>
<td>0.6808</td>
<td>0.047</td>
</tr>
<tr>
<td>(1, 9)</td>
<td>0.6383</td>
<td>0.051</td>
<td>0.7935</td>
<td>0.047</td>
</tr>
<tr>
<td>(9, 1)</td>
<td>1.9137</td>
<td>0.051</td>
<td>2.2599</td>
<td>0.041</td>
</tr>
</tbody>
</table>

for the previous sets of simulations, we set $m = 500$ and varied $\sigma_c^2$ and $\sigma_2^2$ using the same values as in the previous sets of simulations.

Table 10 lists the estimated half-widths and $\alpha$-levels for the three different sets of $\sigma_c^2$ and $\sigma_2^2$. The data in the table seem to indicate that the RMNI procedure is doing better than in the previous sets of simulations but is still yielding $\alpha$-levels not equal to the nominal rate. As in the other sets of simulations, the value of $\sigma_c^2$ does not seem to affect the overall half-widths of the multiply-imputed confidence intervals.

Although this set of simulations seems to indicate that the multiply imputed confidence intervals are conservative, the opposite inference would be made for the third set of simulations. In the setting where an analyst would have a pattern of missing data for a single data set, it would be difficult to determine whether or not the multiply-imputed confidence intervals are conservative or liberal. It may be that the liberal confidence intervals in the third set of simulations is due to the extreme unevenness of the missing data, but further investigations must be done to confirm this. As mentioned before, this problem may be eliminated by using known
Table 11: Ratio of $S'^2$ to $2MSE/n$ when $n = 30$, $t = 5$, and $(\sigma^2_c, \sigma^2_e) = (0.8,0.2), (9,1), (1,9)$.

<table>
<thead>
<tr>
<th>$(\sigma^2_c, \sigma^2_e)$</th>
<th>Sim 1</th>
<th>Sim 2</th>
<th>Sim 3</th>
<th>Sim 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0.8, 0.2)</td>
<td>1.04</td>
<td>1.17</td>
<td>1.50</td>
<td>1.41</td>
</tr>
<tr>
<td>(9, 1)</td>
<td>1.04</td>
<td>1.16</td>
<td>1.47</td>
<td>1.38</td>
</tr>
<tr>
<td>(1, 9)</td>
<td>1.05</td>
<td>1.19</td>
<td>1.69</td>
<td>1.53</td>
</tr>
</tbody>
</table>

adjustments to the critical value.

To investigate the effect of $\sigma^2_c$ on the width of the confidence intervals, we will compare $S'^2$ from (6.12) to $2MSE/n$ for data from the four sets of simulations. If $\sigma^2_c$ has an effect on the confidence widths, it would be found in $S'^2$. Since we know that the $MSE$ is not affected by $\sigma^2_c$, a comparison of $S'^2$ to $2MSE/n$ can give us insight into the effects of $\sigma^2_c$. Table 11 lists the ratios of $S'^2$ to $2MSE/n$ from the four different sets of simulations when $t = 5$, $n = 30$, and for the three pairs of variances. For each set of simulations, the proportion of missing data is different, so we expect the ratios to be different between simulations. For all four simulations, ratios are larger when $\sigma^2_c = 9$. The effect of $\sigma^2_c$ also seems to vary with the proportion of missing data since the differences in the ratios increase as the proportion of missing data increases. It seems that $\sigma^2_c$ may have an effect on the width of the confidence interval, but we have seen that the overall effect to the half-widths of the multiply-imputed confidence intervals is small.

In this chapter, we have presented a modified Dunnett's multiple comparison procedure for the one-way repeated measures model with missing data. Through
simulations, we have shown that the $\alpha$-level is close to the nominal rate when the pattern of missing data has a special form as in the first set of simulations. The second set of simulations shows that this technique works well when the proportion of missing data is small (around 10%), and the proportions of missing data for each of the treatments are close to the same. The third and fourth sets of simulations show that a modification may be necessary to account for the unbalancedness of the missing data. Most importantly, all four sets of simulations seem to indicate that the random effects variance, $\sigma^2$, has only a minor effect on the half-widths of the multiply-imputed confidence intervals. In the next chapter, we will discuss the overall results of the simulations conducted here and what may be done to improve them in situations similar to simulation sets three and four.
CHAPTER VII

Conclusions and further research

In this dissertation, we have taken techniques from three different disciplines—survey sampling, Bayesian linear models, and multiple comparison procedures—to develop a set of simultaneous confidence intervals for comparing the treatment effects to a control for the one-way repeated measures model with missing data. We used posterior distributions from the Bayesian one-way repeated measures model to construct the multiple imputation procedure (called the RMNI procedure) and used the estimates from the RMNI procedure, along with Dunnett’s multiple comparison procedure, to construct the multiply-imputed confidence intervals.

In this chapter, we will discuss our conclusions concerning the development and use of the RMNI procedure to create the multiply-imputed confidence intervals. These conclusions are based on Chapters 5 and 6. We will also discuss how this technique may be improved, and related areas which can be investigated in the future.

7.1 Conclusions

In Chapter 5, we developed the RMNI procedure for the one-way repeated measures model. This procedure relies on the assumptions of the one-way repeated measures model in Chapter 3, the randomness of the response mechanism, and the number of
fully observed subjects. If there were no observed values for a subject, the RMNI procedure would drop that subject from the analysis. We believe that, since there is no information being contributed for this type of missing data, there would be no reason to keep this subject in the analysis.

The RMNI procedure assumes that the pattern of missing data is random in the generation of the multiply-imputed data. If a different response mechanism is suspected, the RMNI procedure could be modified to account for the different response mechanism. This point is further discussed in the following section.

The RMNI procedure also assumes that there is a sufficient number of fully observed subjects to generate good initial estimates of $\theta$, $\sigma^2_e$, and $\sigma^2_r$. In the simulations, we assumed that there were 20 or 30 subjects in each treatment and that at least 12 of them were fully observed. For these simulations, 13 (or 16, which was used in simulation sets two, three, and four) seemed to be a sufficient number of fully observed subjects, but the minimum number of fully observed subjects needed is unknown.

In Chapter 6, we developed confidence intervals based on Dunnett's multiple comparison procedure. To develop the confidence intervals, we looked at a special pattern of missing data which allowed us to simplify the structure of the between-imputation variance-covariance matrix. We assumed that there was only one missing value for each treatment and that each subject had, at most, one missing value. We showed that this assumption implied that all of the between-imputation variances were equal and that all of the covariances were zero. This led to a scalar estimator for the between-imputation variance-covariance, $\hat{b}$, rather than a $t \times t$ matrix. We use the
confidence intervals developed for this special case when the pattern of missing data differs from that of the special case in hopes that the change in the structure of the between-imputations variance-covariance matrix does not change significantly as the pattern of missing data changes.

The simulations in Chapter 6 were conducted to discover (1) how close the α-level of the confidence intervals created using the multiple imputation estimates come to the nominal rate, (2) how the half-widths of the confidence intervals compare to the half-widths of the confidence intervals created with no missing data, and (3) the effects of $\sigma^2_1$ and $\sigma^2_2$ on the half-widths of the confidence intervals. The amount of time the RMNI procedure needed to compute the multiple imputation estimates was never an issue since the algorithm only took 2-12 seconds for the RMNI procedure to calculate the estimates for any one of the simulations.

From the second set of simulations, we saw that when the proportion of missing data is small (around 10%), the α-levels of the confidence intervals created from the multiple imputation estimates are close to the nominal rates. As the proportion of missing data increases, the estimated α-levels seem to stray from the nominal rate. In the third set of simulations (the censored pattern of missing data), we saw that the α-level increased, making the confidence intervals liberal. In the fourth set of simulations (the random pattern of missing data with the larger amount of missing data), we saw that the α-level decreased, making the confidence intervals conservative. The trouble may lie in the value of $T$, the overall variance-covariance matrix of the estimator $\mathbf{Y}$, in (6.3) and, more specifically, in the value of $\mathbf{B}$, the between-imputation
Table 12: The largest ratio of half-widths and the ratio of the number of observations for the four sets of simulations.

<table>
<thead>
<tr>
<th>Simulation Set</th>
<th>Ratio of half-widths</th>
<th>Ratio of observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.02</td>
<td>1.034</td>
</tr>
<tr>
<td>2</td>
<td>1.09</td>
<td>1.119</td>
</tr>
<tr>
<td>3</td>
<td>1.31</td>
<td>1.339</td>
</tr>
<tr>
<td>4</td>
<td>1.24</td>
<td>1.304</td>
</tr>
</tbody>
</table>

We can get a good feel for the value of B by looking at the estimator of B. The estimate for the between-imputation covariances (or the off-diagonal elements of B) is (6.21) and the estimate for the variance is (6.22). If the numbers of missing values for the treatments are different, the estimated variances, (6.22), will be different from each other. The same is true for the covariances. It may be possible to correct for this by applying techniques used for Dunnett’s multiple comparison procedure when the design is an unbalanced fixed-effects linear model as discussed in Hochberg and Tamhane (1987) or Hsu (1992). This will be discussed further in the next section.

The half-widths (or the widths) of the confidence intervals define the relative strength of the multiple comparison procedure. One wants the shortest possible confidence intervals which have the correct overall α-level. One would expect the multiply-imputed confidence intervals would have a larger half-width than the complete-data confidence intervals since they are based on fewer observed data. From the first set of simulations with t = 5 and n = 30, we see that the half-widths of the multiply-
imputed confidence intervals are at most 1.02 times larger than the half-widths of the complete-data confidence intervals. The multiply-imputed confidence intervals are constructed using 145 observations, whereas the complete-data confidence intervals are based on 150 observations. The ratio of complete to observed data is $150/145 = 1.034$. Table 12 lists the ratios of half-widths for the multiply-imputed data to the half-widths for the complete-data and the corresponding ratios of the number of observations in the missing data case to the complete-data case for the four sets of simulations. The half-width ratios represent the largest ratio for each set of simulations. The largest ratios occurred when $\sigma_c^2 = 9$, which makes sense since we learned in Chapter 6 that the ratio of the MSE to $S_\star^2$ is larger for larger values of $\sigma_c^2$. Hence, we feel that the random-effects variance, $\sigma_c^2$, has some effect on the width of the confidence interval, but this effect is fairly small.

Since the estimate of the random-effects variance, $\sigma_c^2$, does not appear in the Dunnett's multiple comparison confidence intervals for the one-way repeated measures model, we hoped that the random-effects variance would not affect the confidence intervals based on the multiple imputation estimates. All four sets of simulations suggest that $\sigma_c^2$ does not play a large role in the confidence intervals based on the multiple imputation estimates. In Chapter 6, we saw that $\sigma_c^2$ does have some effect on $S_\star^2$, but that the effect is minor. The estimated overall half-widths from the simulations did not increase significantly when $\sigma_c^2 = 9$ versus when $\sigma_c^2 = 1$. This implies that when $\sigma_c^2$ is even nine times larger than $\sigma_c^2$, the widths of the confidence intervals do not increase proportionally and the width of the intervals is more controlled by
the proportion of missing data.

We have seen that the RMNI procedure can compute the estimates for the confidence intervals in a relatively short period of time and that these confidence intervals yield approximately the correct inference when the proportion of missing data is 10% or less. However, there is room for improvement. When the proportion of missing data is larger, the effect of the pattern of missing data seems more important. We will discuss one possible way to improve the confidence intervals in the next section.

7.2 Further research

Several different aspects of this research can be investigated further. In this section, we will suggest the following areas for future research: (1) the effects that \( m \), the proportion of missing data, and the missing data mechanism have on the estimates from the RMNI procedure, (2) changing the critical value to reflect the different values of \( \rho_{ij} \) based on the variance-covariance matrix \( \mathbf{T} \), (3) the use of the RMNI procedure in other multiple comparison procedures such as Tukey's multiple comparison procedure, (4) the development of other multiple imputation procedures for the one-way repeated measures model, and (5) the use of multiple imputation on other, more complex random-effects or mixed models.

The necessary value of \( m \) is important to discover if this technique is used in practice, since the time it takes to calculate the estimates is related to the number of multiple imputations. If the computation time has to be limited, it is important to know how small \( m \) can be to yield good multiple imputation estimates. This value should be affected by \( n \), \( t \), and the proportion and pattern of missing data. The
survey sampling literature suggests that if the data set is large (e.g., thousands of observations), \( m \) can be as small as 5. The simulations conducted here would seem to suggest that \( m = 100 \) is sufficient but may not be necessary. It may be possible to modify the RMNI procedure so that \( m \) can be determined by the computer. One could set a tolerance level (or error term) and let \( m \) be chosen iteratively. As the RMNI procedure creates each completed-data set, estimates could be calculated and compared to the previous estimates. If the differences between the estimates are within the tolerance level, the computer would use the last estimates calculated, or else the next completed-data set could be created. Using this scheme, the value of \( m \) would not have to be preset.

Future research can be conducted to determine what can be done if the response mechanism is nonignorable. Cases of nonignorable nonresponse are usually handled on a case-by-case basis since the response mechanism is usually specific to the problem at hand. A "generic" multiple imputation procedure for nonignorable nonresponse does not seem likely, but we are confident that multiple imputation procedures can be developed on a case-by-case basis to handle nonignorable nonresponse in the one-way repeated measures model in a similar manner to the RMNI procedure. The RMNI procedure could even be modified to account for the hypothesized response mechanism.

The simulation studies suggest that the multiply-imputed confidence intervals have an \( \alpha \)-level close to or at the nominal rate, but as the proportion of missing data increases, the \( \alpha \)-levels seem to stray from the nominal rate. Simulation sets three
and four of Chapter 6 suggest that the confidence intervals formed from the multiple imputation estimates have $\alpha$-levels different from the nominal rate. The theoretical development for these confidence intervals is based on the assumption of one missing value from each treatment. In this special case, $\rho_{ij} = 1/2$ for all $i \neq j$. This special case does not hold for the patterns of missing data in the third and fourth sets of simulations. An adjustment to the critical value, achieved by changing the $\rho_{ij}$, similar to adjustments made for Dunnett’s multiple comparison for unbalanced fixed-effect linear models may be needed. Two possible methods for unbalanced fixed-effects linear models can be attributed to Hochberg and Tamhane (1987) and Hsu (1993). Hochberg and Tamhane (1987) discuss estimating $\pi_{ij}$ with a single value derived from the data. Since $\sigma_c^2$ does not seem to be a factor in the estimator of variance, it may be possible to use estimates developed for the unbalanced fixed-effects linear models to approximate the critical value. We could use factor analytic approach described by Hsu (1992) to derive a matrix which would be “close to” the correlation matrix corresponding to the matrix $T'$, but would have a form such that all of the $\rho_{ij}$ would be equal.

Since the RMNI procedure can be utilized to create confidence intervals based on Dunnett’s multiple comparison procedure, it also may be possible to use the RMNI procedure to derive confidence intervals based other multiple comparison procedures, such as Tukey’s. This would be useful since Dunnett’s multiple comparison procedure only compares the treatments with one control. A modification to Tukey’s multiple comparison procedure would allow one to make all pairwise comparisons.
The two possible problems that we see in the RMNI procedure are (1) the procedure relies on initial estimates from the completely observed data and (2) the posterior distributions use to draw values for $\sigma_c^2$ and $\sigma_{cc}^2$ can lead to values of $\sigma_c^2$ which are negative.

The RMNI procedure relies on the initial estimates from the completely observed data. When $t$ is large, it may happen that most of the subjects have at least one missing value. One way to avoid this problem is to develop a procedure which uses all of the observed data, such as the procedure devised by Rubin and Schafer (1990) discussed in Chapter 2. Rubin and Schafer’s multiple imputation procedure for an arbitrary multivariate normal can be modified to account for the covariance structure of the one-way repeated measures model. Their procedure uses Gibbs sampling to generate the multiply-imputed data; this raises the question of the convergence of the Gibbs sampler. If the procedure is to be implemented in general, a convergence criterion for the Gibbs sampler would need to be addressed.

Since the value of $\sigma_c^2$ is calculated as $t^{-1}(\sigma_c^2 - t\sigma_{cc}^2)$, this value may be negative. To avoid this problem, we took the $\sigma_c^2 = \max(0, t^{-1}(\sigma_c^2 - t\sigma_{cc}^2))$. Box and Tiao (1973) also find the approximate posterior distributions under the constraint that $\sigma_c^2 < \sigma_{cc}^2$, so that $\sigma_c^2$ is nonnegative. These approximate posterior distributions can be used in place of the posteriors used in the RMNI procedure.

The results from this research imply that multiple imputation may be a useful tool for more complex random or mixed-effects linear models with missing data. In these complex designs, estimates are more easily derived when the design is balanced. In the
presence of missing data, estimates may not be as easily derived. Multiple imputation may be used in a fashion similar to what we have used here to find estimates in these more complex designs.
Appendix A

Uncommon distributions

In this appendix, we will describe two distributions not commonly found in the literature—the Studentized range distribution used in Tukey's multiple comparison procedure and the multivariate T-distribution used in Dunnett's multiple comparison procedure. Both distributions are introduced in Chapter 4. Since the percentiles of the distributions are important to the multiple comparison procedures, they are also presented here.

A.1 The Studentized range distribution

From Hochberg and Tamhane (1987), the Studentized range distribution is defined as a function of random variables. If $Z_1, Z_2, \ldots, Z_k$ are independent and identically distributed standard normal random variables and $U$ is a Chi-square random variable with $\nu$ degrees of freedom independent from the normal random variables, then the random variable

$$Q_{k,\nu} = \frac{\max_{1 \leq i < k} |Z_i - Z_\nu|}{\sqrt{U/\nu}}$$

(A.1)

is called the the Studentized range distribution with parameter, $k$, and degrees of freedom, $\nu$. 
The cumulative distribution of $Q_{k,\nu}$ can be written as

$$P(Q_{k,\nu} < q) = k \int_0^\infty \int_{-\infty}^{\infty} [\Phi(y) - \Phi(y - q_x)]^{k-1} d\Phi(y) dF_{\nu}(x), \quad (A.2)$$

where $\Phi(y)$ is the cumulative distribution of the standard normal distribution and $F_{\nu}(x)$ is the cumulative distribution of the Chi-square distribution with $\nu$ degrees of freedom. The upper $\alpha$-th percentile for the Studentized range distribution is obtained by setting (A.2) to $1 - \alpha$ and solving for $q$.

### A.2 The multivariate $T$-distribution

From Hochberg and Tamhane (1987), the multivariate $T$-distribution starts with similar random variables to those used in the Studentized range distribution, except that the normal random variables for the multivariate $T$-distribution are correlated. Let $Z = (Z_1, Z_2, \ldots, Z_k)$ have a $k$-dimensional multivariate normal distribution with mean $0$, unit variances, and the correlations $\text{Corr}(Z_i, Z_{i'}) = \rho_{ii'}$ for $i \neq i'$. Let $U$ be a Chi-square random variable with $\nu$ degrees of freedom independent of $Z$. Then the random variable $T = Z/\sqrt{U/\nu}$ is called the multivariate $T$-distribution with $\nu$ degrees of freedom and associated correlation matrix $R = (\rho_{ii'})_{k \times k}$. If $R$ is positive definite, then the density function of $T = (T_1, T_2, \ldots, T_k)$ is

$$f(t) = \frac{\Gamma((k + \nu)/2)(det R)^{-1/2}(\nu \pi)^{k/2}}{\Gamma(\nu/2)} \left[1 + \frac{t'R^{-1}t}{(k + \nu)/2}\right]^{-(k + \nu)/2}. \quad (A.3)$$

The critical point used in Dunnett's multiple comparison procedure uses the distribution of $\max_{1 \leq i \leq k} |T_i|$. As discussed in Chapter 4, finding the critical value of the multivariate $T$-distribution is computationally prohibitive, except when all of the $\rho_{ii'}$
are the same for all \( i \neq i' \). We will present the cumulative distribution of \( \max |T_i| \) for the special case of \( \rho_{ii'} \equiv \rho \) for all \( i \neq i' \).

\[
P(\max |T_i| < t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left[ \Phi^k \left( \frac{\sqrt{\rho y + t x}}{\sqrt{1 - \rho}} \right) - \Phi^k \left( \frac{\sqrt{\rho y - t x}}{\sqrt{1 - \rho}} \right) \right] d\Phi(y) dF_{\nu}(x),
\]

(A.4)

where \( \Phi(y) \) is the cumulative distribution of the standard normal and \( F_{\nu}(x) \) is the cumulative distribution of a Chi-square distribution with \( \nu \) degrees of freedom. The upper \( \alpha \)-th percentile for the multivariate \( t \)-distribution is obtained by setting (A.4) to \( 1 - \alpha \) and solving for \( t \).
In this appendix, we will list the two programs we used for all of the simulations. The program `dunnett.f` generates the pattern of missing data and the simulated data sets and then calculates the estimates for both the complete-data and multiply-imputed data for a specified number of replications, repetitions, treatments, and subjects. The program `conf.f` reads the estimates generated from `dunnett.f` and computes the estimated half-width and α-level for the simulation. We could not combine the two programs because we did not know the degrees of freedom for the multiply-imputed confidence intervals.

We used an Splus program written by Professor Jason Hsu (which is not included here) to calculate the critical values for the multivariate T-distribution. Professor Hsu’s program calculates the critical values when the degrees of freedom is an integer value. Since the degrees of freedom for the multiply-imputed confidence intervals may not be an integer, the value of the degrees of freedom is rounded down to the nearest integer to find the critical value, `criti`. Rounding down increases the width slightly so that the confidence interval will be more conservative rather than more liberal.

Both programs were compiled and executed using a Digital Equipment Company (DEC) 5000 model 25.
B.1 The dunnett.f program

The dunnett.f program is written in FORTRAN and uses subroutines from the IMSL libraries. These subroutines are used to set the random number generator seed and to generate the random numbers needed for the simulations. The seed is always equal to one. The program requires the following variables to be set in advance.

\begin{itemize}
  \item \textit{num} = number of replicated data sets
  \item \textit{imp} = number of repetitions
  \item \textit{trt} = number of treatments
  \item \textit{sub} = number of subjects
  \item \textit{sige} = error variance
  \item \textit{sigc} = random effects variance
\end{itemize}

All of the treatment means have their values set at one. The estimates produced by the program are saved in various files. Descriptions of these files are listed in the initial comment statements of the program.

The sample program listed below is used in the second set of simulations in Chapter 6. For the other sets of simulations, only the creation of the pattern of missing data had to be changed.

\begin{verbatim}
  file: dunnett.f  last modified: June 10, 1993
  This program implements the RMNI procedure and calculates the
\end{verbatim}
estimates that will be used in the multiply imputed confidence intervals.

input files: none

screen output: time in seconds it takes to compute estimates for each replication using the RMNI proc.

output files: 5

dunnett = contains the estimates of the degrees of freedom, the variance for the conf. int's, and b.

matrix = contains estimated variance-covariance matrix for complete data and imputed data for all data sets averaged together.

impute = contains estimates of the treatment means and the two variances for the multiply imputed data.

real = contains estimates of the treatment means and the two variances for the complete data.

response = contains the pattern of missing data.

integer i, j, k, l,
& num, trt, sub, irank,
& imp, obs0, add
parameter ( num=15000, trt=5, sub=30, imp=500)
integer r(trt,sub), mis(sub), imis, iobs
double precision y(trt,sub), sigc, th(trt),
& z(trt,sub), sig(trt,trt), sige,
& mu(trt), tri(trt,trt), sige1,
& varmis(trt,trt), normal(trt), sigc1,
& trimis(trt,trt), gamma,
& mumis(trt), sse2,
& theta(trt), theta0, muobs,
& offdiag, diag,
& xi, dof,
& x(trt), x2, sse0,
& estimate, dobs, dsub,
& muimp(trt), dimp,
& ssc2, dtrt,
& mean(imp,trt), sigeimp, ssc0,
& sigcimp,
& phi0(trt-1), phi1(trt), theta2,
& sig1(trt-1,trt-1), se(imp), a,
& tri1(trt-1,trt-1), sc(imp), b,
& norm0(trt), phi(trt-1),
** Set parameters ****************************
call rnset(l)
tol = 100.*amach(4)
missing = .ld0
dsub = sub
dtrt = trt
dimp = imp
add = 0
dnum = num
dm1 = 1d0/(dimp-ld0)

** Set true parameters ****************************
data mu/1,1,1,1,1/
sigc = .2d0
sige = .8d0
call matrix(trt,sige,sigc,sig)
call dchfac(trt,sig,trt,tol,irank,tri,tri)
c ****** Create pattern of missing data ******************************
50 do 70 j = 1, sub
   mis(j) = 0
   do 60 i = 1, trt
      c  call rnbin(1,1,missing,r(i,j))
      if (i.eq.j) then
         r(i,j) = 1
      else
         r(i,j) = 0
      endif
      mis(j) = mis(j) + r(i,j)
   60 continue
   add = add + mis(j)
70 continue
   dofs = (dsub-ld0)*(dtrt-ld0) - add
   c ****** if no values are missing, go back and generate more ****
   if (add.eq.0) goto 50
   do 80 j = 1, sub
      write(9,610) (r(i,j), i=1,trt)
   80 continue

c ****** Initialize matrices ***************************************
65 do 90 i = 1, trt
   do 90 j = 1, trt
      matr(i,j) = 0d0
      mati(i,j) = 0d0
90 continue


c ****** Main program **********************************************
c

do 500 k = 1, num

c ****** Draw sample data points ************************************
do 110 j = 1, sub
   call drnnoa(trt,norm0)
   call dmult(trt,tri,norm0,normal)
do 100 i = 1, trt
   y(i,j) = mu(i) + normal(i)
100 continue
```
110    continue
    call corr(trt,sub,y,matt)
    do 105 i = 1, trt
       do 107 j = 1, trt
          matr(i,j) = matr(i,j) + matt(i,j)/dnum
107    continue
105    continue

    c      ***** Calculate complete data estimates ********************
    call estimate(trt,sub,y,real,theta0,sse0,ssc0)
    dssc0 = (ssc0/(dsub-1d0)-sse0/(dsub-1d0)/(dtrt-1d0))/dtrt
    if (dssc0.lt.0d0) dssc0 = 0d0
    write(10,600) (real(i),i=1,trt),
        & sse0/(dtrt-1d0)/(dsub-1d0),ssc0/(dsub-1d0)

    c      *** Compute fully observed estimates ********************
    time0 = ctime()
    obs0 = 0
    do 120 j = 1, sub
       if (mis(j).eq.0) then
          obs0 = obs0 + 1
          do 130 i = 1, trt
             z(i,obs0) = y(i,j)
130       continue
       endif
120    continue
    dobs = obs0
    call estimate(trt,obs0,z,th,theta0,sse0,ssc0)
    do 140 i = 1, trt-1
       phi0(i) = th(i) - theta0
140    continue

    c      ***** Generate imputed values ****************************
    do 200 l = 1, imp

    c      ***** generate values for the variances ********************
    dobs = obs0
    dof = (dtrt-1d0)*(dobs-1d0)
    call drnchi(l,dof,x1)
    dof = dobs-1d0
    call drnchi(l,dof,x2)
    sige1 = sse0/x1
    sigc1 = (ssc0/x2-sige1)/dtrt
    if (sigc1.lt.0) sigc1 = 0
    a = sige1/dobs
    b = -1d0*a/dtrt
```
c

****** create covariance matrix and generate the means ******
call matrix(trt-1,a,b,sig1)
call dchfac(trt-1,sig1,trt-1,tol,irank,tri1,trt-1)
call drnnoa(trt-1,phi)
call dmult(trt-1,tri1,phi,normal1)
phi1(trt) = 0d0
do 230 i = 1, trt-1
   phi1(i) = phi0(i) + normal1(i)
   phi1(trt) = phi1(trt) - phi1(i)
230 continue
do 240 i = 1, trt
   theta(i) = phi1(i) + theta0
240 continue

c

****** generate values from conditional normal ************
do 250 j = 1, sub
   if (mis(j).gt.0) then
      imis = 0
      iobs = 0
      muobs = 0d0
      do 260 i = 1, trt
         if (r(i,j).eq.1) then
            imis = imis + 1
            mumis(imis) = theta(i)
         else
            iobs = iobs + 1
            muobs = muobs + y(i,j) - theta(i)
         endif
      260 continue
      dobs = iobs
      gamma = sigc1/(dobs*sigc1+sige1)
diag = sige1
      offdiag = sige1*gamma
      muobs = gamma*muobs
      do 280 i = 1, mis(j)
         mumis(i) = mumis(i) + muobs
      280 continue
      call matrix(mis(j),diag,offdiag,varmis)
      call dchfac(mis(j),varmis,mis(j),tol,irank,trimis,mis(j))
call drnnoa(mis(j),norm0)
call dmult(mis(j),trimis,norm0,normal)
do 300 i = 1, mis(j)
x(i) = mumis(i) + normal(i)

continue
imis = 0
do 310 i = 1, trt
    if (r(i,j).eq.1) then
        imis = imis + 1
        y(i,j) = x(imis)
    endif
continue
endif
continue
continue
continue
****** Find summary statistics for completed data set ******
call corr(trt,sub,y,matt)
do 320 i = 1, trt
do 325 j = 1, trt
    mati(i,j) = mati(i,j) + matt(i,j)/dimp/dnum
continue
325 continue
call estimate(trt,sub,y,muimp,theta2,sse2,ssc2)
do 330 i = 1, trt
    mean(l,i) = muimp(i)
    muimp(i) = OdO
continue
330 continue
se(l) = sse2/(dtrt-ld0)/(dsub-ld0)
sc(l) = ssc2/(dsub-ld0)
continue
200 continue
c
****** Compute final statistics ****************************
do 400 i = 1, trt
do 410 l = 1, imp
    muimp(i) = muimp(i) + mean(l,i)
continue
410 continue
muimp(i) = muimp(i)/dimp
continue

sigeimp = OdO
sigcimp = OdO
do 420 l = 1, imp
    sigeimp = sigeimp + se(l)
sigcimp = sigcimp + sc(l)
420 continue
sigeimp = sigeimp/dimp
sigcimp = sigcimp/dimp
sigcimp = (sigcimp - sigeimp)/dtrt
if (sigcimp lt 0d0) sigcimp = 0d0

bhat = 0d0
do 430 i = 1, trt
    bhat(i) = 0d0
    do 440 l = 1, imp
        bhat(i) = bhat(i) + dml*(muimp(i)-mean(l,i))*2d0
    440 continue
    bhat = bhat + bhat(i)/dtrt
430 continue

sstar = (2d0/dsub*sigeimp + 2d0*(1d0+1d0/dimp)*bhat)**.5d0
dofdun = (dofobs + dsub*(1d0+1d0/dimp)*bhat)**2d0/dofobs

time1 = ctime()
print *, time1 - time0

write( 8,600) (muimp(i), i=1,trt), sigeimp, sigcimp
write(12,650) dofdun, sstar, bhat
500 continue

write(11,620) ( (matr(i,j), i=1,trt), j=1,trt)
write(11,630) ( (mati(i,j), i=1,trt), j=1,trt)

600 format(5(f6.3,1x),2(f9.4,1x))
610 format(5(i1,2x))
620 format(5(f6.4,1x))
630 format('')
640 format((3f8.5,1x),i2,1x,f9.5)
650 format(f12.6,1x,f12.5,1x,f12.5)

close(unit=8)
close(unit=9)
close(unit=10)
close(unit=11)
close(unit=12)

stop
subroutine matrix(n,a,b,mat)
   integer n, i, j
   double precision mat(n,n), a, b
   do 900 i = 1, n
      do 950 j = 1, n
         if (i.eq.j) then
            mat(i,i) = a+b
         else
            mat(i,j) = b
         endif
      950 continue
   900 continue
   return
end

subroutine estimate(t,n,q,mean,overall,sse,ssa)
   integer t, n, i, j
   double precision q(t,n), mean(t), sse, ssa,
      &              dn, overall, dt, meanb(1000)
   overall = 0d0
   dn     = n
   dt     = t
   do 900 i = 1, t
      mean(i) = 0d0
   do 910 j = 1, n
      mean(i) = mean(i) + q(i,j)
910 continue
overall = overall + mean(i)
mean(i) = mean(i)/dn

900 continue
overall = overall/dn/dt
do 920 j = 1, n
  meanb(j) = 0d0
  do 930 i = 1, t
    meanb(j) = meanb(j) + q(i,j)/dt
  930 continue
  do 940 j = 1, n
    ssa = ssa + (meanb(j)-overall)**2d0
    sse = sse + (q(i,j)-mean(i)-meanb(j)+overall)**2d0
  940 continue
ssa = ssa*dt
return
end

c ********************************************
c dmult : c = transpose(a)*b
c ********************************************
subroutine dmult(t,a,b,c)
+ + + -
c
integer i,j,t
double precision a(t,t), b(t), c(t)
c
do 900 i =1, t
  c(i) = 0d0
  do 910 j = 1, t
    c(i) = c(i) + a(j,i)*b(j)
  910 continue
  do 900 continue
return
end
c ********************************************
c computes the covariance matrix
subroutine corr(t,n,x,y)
   integer t, n, i, j, j2
   double precision x(t,n), y(t,t), bar(100), dt, dn
   
dt = t
   dt = 1d0/dt
   dn = n
   
do 900 i = 1, t
       bar(i) = 0d0
       do 910 j = 1, t
           y(i,j) = 0d0
           910 continue
       do 915 j = 1, n
           bar(i) = bar(i) + x(i,j)
       915 continue
       bar(i) = bar(i)/dn
   900 continue
   
do 950 i = 1, t
     do 960 j = i, t
       do 970 j2 = 1, n
          y(i,j) = y(i,j) + (x(i,j2)-bar(i))*(x(j,j2)-bar(j))
       970 continue
       y(i,j) = y(i,j)/(dn-1d0)
       if (i.ne.j) y(j,i) = y(i,j)
     960 continue
   950 continue
   
return
end

B.2 The conf.f program

The conf.f program was written in FORTRAN and requires that the number of replications, n, number of treatments, t, the number of subjects, sub, and the two critical values, critr and criti (for the complete data and multiply imputed data, respectively), be set in advance.
This program takes the output files from dunnett.f, computes the confidence intervals, and calculates if the ci's contain 0.

**input:** dunnett, impute, real

**screen output:** # of replications, estimated alpha level for complete data, estimated alpha level for imputed data, estimated half-width for complete data, and estimated half-width for imputed data.

```c
integer i, j, n, t, cr, ci, out, sub
parameter (n=15000,t=7,sub=20)
double precision trtr(t), trti(t), ser, sei, scr, sci, cutr, cuti,
  & sstar, r1, r2, critr, criti, temp, dsub,
  & lowr(t-l), lowi(t-l), highr(t-1), highi(t-1),
  & dn, dni, widthr, widthi

open(unit=7,file="real",status="old")
open(unit=8,file="impute",status="old")
open(unit=9,file="dunnett",status="old")

****** Set initial values ****************************************
** The critical values, critr and criti, have to be changed before using this program.
cr = 0
ci = 0
critr = 2.605926
criti = 2.608501
dsub = sub
dn = n
dni = 1d0/dn
widthr = 0d0
widthi = 0d0
do 50 i =1, n
  read(7,*) (trtr(j), j=1,t), ser, scr
  read(8,*) (trti(j), j=1,t), sei, sci
  read(9,*) r1, sstar, r2

  cutr = critr*(2d0/dsub*ser)**.5d0
  cuti = criti*sstar
```
widthr = widthr + cutr*dni
widthi = widthi + cuti*dni

out = 0

do 70 j = 1, t-1
  temp = trtr(j)-trtr(t)
  lowr(j) = temp - cutr
  highr(j) = temp + cutr
  if (temp.gt.cutr.or.temp.lt.-1d0*cutr) out = 1
70 continue
if (out.eq.1) cr = cr + 1
out = 0

do 80 j = 1, t-1
  temp = trti(j)-trti(t)
  lowi(j) = temp - cuti
  highi(j) = temp + cuti
  if (temp.gt.cuti.or.temp.lt.-1d0*cuti) out = 1
80 continue
if (out.eq.1) ci = ci + 1

50 continue
print *, n, cr, ci, widthr, widthi
close(unit=7)
close(unit=8)
close(unit=9)
90 format(8(f7.3,1x))
stop
end
BIBLIOGRAPHY


119


