

ESTIMATION OF MISSING VALUES IN REPLICATED FACTORIAL EXPERIMENT

BY

EKWUEME, CHINENYE LOVELYN

DEPARTMENT OF MATHEMATICS,

AHMADU BELLO UNIVERSITY,

ZARIA NIGERIA

MARCH, 2014

ESTIMATION OF MISSING VALUES IN REPLICATED FACTORIAL EXPERIMENT

BY

**EKWUEME CHINENYE LOVELYN
B.Sc (NAU 2008)
M.Sc/SCIE/00890/2009-2010**

**A THESIS SUBMITTED TO THE POSTGRADUATE SCHOOL,
AHMADU BELLO UNIVERSITY, ZARIA NIGERIA**

**IN PARTIAL FULFILMENT FOR THE AWARD OF MASTER OF SCIENCE (M.Sc)
DEGREE IN STATISTICS,
DEPARTMENT OF MATHEMATICS AHMADU BELLO UNIVERSITY, ZARIA
NIGERIA**

2014

DECLARATION

I hereby declare that the work in this thesis entitled ‘estimation of missing values in replicated factorial experiment under the supervision of Professor O.E Asiribo and Dr H.G. Dikko has been done by me. The information derived from the literature has been duly acknowledged in the text and a list of references provided. No part of this thesis has been previously presented for another degree or diploma at any university.

Ekwueme, Chinenye Lovelyn
Name of student

Signature

Date

CERTIFICATION

This thesis titled “**Estimation of Missing Values in Replicated factorial Experiment**” meets the regulations governing the award of the degree of Masters in Statistics of the Ahmadu Bello University Zaria, and is approved for its contribution to knowledge and literary presentation.

Prof O. E. Asiribo
Major Supervisor

Date

Dr. H. G. Dikko
Minor Supervisor

Date

External Examiner

Date

Dr. Babangida Sani
Head of Department

Date

Prof. A. Joshua
Dean, School of Postgraduate Studies

Date

DEDICATION

This Thesis is dedicated to GOD through his son Jesus Christ and my family

ACKNOWLEDGMENT

All praise be to God Almighty for sparing my life to attain this academic height.

My profound and sincere gratitude goes to my supervisor, Prof. O.E. Asiribo for his detailed comments, valuable suggestions, constructive criticisms and encouragement that led to the success of the thesis work.

I have to sincerely thank my second supervisor, Dr. H.G. Dikko for his immense contributions, corrections and consistent encouragement that, with no doubt, speed up the write-up. To the entire members of staff of the Department I say thank you for your guidance and contribution towards my study.

To my father, Chief Timothy Ekwueme, saying thank you is an understatement in showing my appreciation for being a pillar to me since my birth. My mother, Mrs. Veronica Ekwueme, needs a special thanks for the care, love and support giving to me at all the times.

I have to also show appreciation to my siblings, Lilian, Kenechukwu, Obinna and Chidimma for their patience and support given to me in the course of my study.

I cannot forget mentioning some of my friends/course mates whose company I enjoyed so much throughout this program; Chidomere Uchenna Emmanue, Gwafan Joy and Melekwe Joy.

Special thanks to the Group General Manager (Human Resources) of NNPC, Mr. Chris Osarumwense, who motivated me to go for my masters.

Lastly, I would like to thank Nigeria-Sao Tome & Principe Joint Development Authority for the Post Graduate Scholarship Award being given to me and all that have contributed in one way or the other towards the success of my study.

ABSTRACT

This study examined the power of Pairwise Deletion (PD), Multiple Imputation (MI) and Expectation Maximization (EM) methods in estimating missing values in cases where the data are missing at random. The data used is a replicated 2 x 3 x 4 factorial experiment in a randomized complete block design (RCBD) and a simulated data set (SMDS) in which data points were randomly selected as missing were used to examine the methods. The result shows that the missing values which are missing at random can be determined using EM method because the estimated values obtained in terms of means, standard error and P-values for all the variables considered were consistent and approximately similar. Also, the results obtained using this method were approximately similar to that of the real-life data set (RLDS) and simulated data set. The study therefore recommends that in a replicated factorial analysis with missing values, EM method has been shown to give better and appropriate results.

TABLE OF CONTENTS

Title page.....	i
Declaration.....	ii
Certification.....	iii
Dedication.....	iv
Acknowledgement	v
Table of Contents.....	vi
List of Tables.....	vii
List of Appendices.....	viii
CHAPTER ONE.....	1
INTRODUCTION.....	1
1.1 Background.....	1
1.2 Problems with Missing Data.....	1
1.3 Implications of Missing Data.....	3
1.4 Missing Data Mechanism.....	3
1.4.1 Missing Completely at Random (MCAR)	4
1.4.2 Missing at Random (MAR)	4
1.4.3 Missing Not at Random (MNAR)	5
1.5 Patterns of Missing Data.....	5
1.6 Motivation of Study.....	6
1.7 Objectives of Study.....	6
1.8 Significance of the Study.....	7
CHAPTER TWO.....	8
2.1 LITERATURE REVIEW.....	8
2.2 Causes and Problems of Missing Observation.....	17
2.3 Precision of Missing Estimate.....	17
2.4 Missingness in RCBD.....	18
2.5 Brief Review on the Methods Used.....	19
2.5.1 Pairwise Deletion.....	19

2.5.2	Expectation Maximization (EM)	20
2.5.3	Multiple Imputation (MI)	23
CHAPTER THREE.....		26
METHODOLOGY.....		26
3.1	Introduction.....	26
3.2	The Data.....	27
3.3	Factorial Experiment.....	27
3.4	Testing the Assumptions for Anova	27
3.5	Advantages of Factorial Design.....	28
3.6	Brief Review on Simulation.....	30
3.6.1	Importance of Simulation.....	30
3.6.2	Description of how the Simulation of the Simulated Data was Obtained.....	30
3.7	Description of how Missing Values were Obtained.....	31
3.8	Methods for Replacing Missing Values.....	31
3.8.1	Pairwise Deletion.....	31
3.8.2	Expectation Maximization.....	32
3.8.2.1	Execution of EM using IBM SPSS.....	33
3.8.3	Multiple Imputation.....	34
3.8.3.1	Step 1: Imputation.....	35
3.8.3.2	Step 2: Statistical Analysis.....	35
3.8.3.3	Step 3: Combining Results.....	35
CHAPTER FOUR.....		38
DATA ANALYSIS AND INTERPRETAION.....		38
4.1	Introduction.....	38
4.2	Presentation and Interpretation of Results.....	38
4.2.1	Presentation and Interpretation of Results for the Real-Life Data and Simulated Data Sets.....	39
CHAPTER FIVE.....		66
SUMMARY, CONCLUSION AND RECOMMENDATION.....		66
5.1	Introduction.....	66
5.2	Summary and Conclusion.....	66
5.3	Recommendation.....	67
5.4	Contribution to Knowledge.....	68

5.5	Area for further Research.....	68
	REFERENCES.....	70
	APPENDICES.....	76

LIST OF TABLES

Table 3. 1:Anova Table Of $a \times b \times c$ Factorial Design	29
Table 4. 1:Estimated Means and Standard Errors for Plant Height Based On the Rlds and the Three Methods (PD, MI and EM).	39
Table 4. 2:Estimated Means and Standard Errors for Plant Height Based on the SIMDS and the Three Methods (PD, MI AND EM).	43
Table 4. 3:Estimated Means and Standard Errors for Plant Leave Based on the RLDS and the Three Methods (PD, MI AND EM).	47
Table 4. 4:Estimated Means and Standard Errors For Plant Leave Based on the SIMDS and the Three Methods (PD, MI AND EM).	51
Table 4. 5: Estimated Means and Standard Errors for Plant Branches Based on the RLDS and the Three Methods (PD, MI AND EM).	55
Table 4. 6: Estimated Means and Standard Errors for Plant Branches Based on the SIMDS and the Three Methods (PD, MI AND EM).	59
Table 4. 7:P-Values Obtained with the RLDS, Simds, Pw, MI And EM For Plant Heights.	63
Table 4. 8:P-Values Obtained with the RLDS, SIMDS, PW, MI and EM for Plant Leaves $\alpha = 0.05$	64
Table 4. 9:P-Values Obtained With the RLDS, SIMDS, PW, MI and EM for Plant Branches $\alpha = 0.05$	65

LIST OF APPENDICES

APPENDIX I: REAL LIFE DATA SET	76
APPENDIX II: REAL LIFE DATA SET MISSING AT RANDOM.....	78
APPENDIX III:COMPLETE SIMULATED DATA SET	80
APPENDIX IV:INCOMPLETE SIMULATED DATA	82
APPENDIX V:REAL LIFE DATA SET.....	84
APPENDIX VI:P-VALUES OBTAINED WITH THE RLDS, PW, MI AND EM FOR PLANT HEIGHT	86

CHAPTER ONE

INTRODUCTION

I.1 Background

Missing data, or missing values, occur when no data or value is stored for the variable in the current observation. Missing data can occur because of non-response when no information is provided for several items or no information is provided for a whole unit.

A growing literature suggests that how researchers deal with missing data can affect model estimate and standard error (Schafer, 1997; Vriens and Melton, 2002; Schafer and Graham, 2002; Little and Rubin, 2002; and Raghunathan, 2004). Frequently, researchers delete observations with missing items with no attention to the consequences attached to the model estimates and standard errors. The intent of any analysis is to make valid inferences regarding the population of interest. Therefore, it is important to respond to a missing data problem in a manner which reflects the population of interest.

1.2 PROBLEMS WITH MISSING DATA

Why are missing data a problem? The most serious concern is that missing data can introduce bias into estimates derived from a statistical model (Becker & Powers, 2001; Holt, 1997; Rubin, 1987). For example, it is possible that non-respondents might have different response profiles compared to those who responded completely. Thus, the remaining sample is no longer random or representative of the population from which it was randomly drawn. If the researcher chose to draw their conclusions based solely on those who responded, the conclusions would be biased.

Furthermore, missing data result in a loss of information and statistical power (Anderson, Basilevsky & Hum, 1983; Kim & Curry, 1977). The elimination of subjects with missing information on one or more variables from the statistical analysis in listwise deletion decreases the error degrees of freedom (*df*) in statistical tests such as the *t-test*. This decrease in turn leads to reduced statistical power and larger standard errors compared to those obtained from complete random samples (Cohen & Cohen, 1983; Cool, 2000).

Similar loss of *df* and statistical power occurs with pairwise deletion, where standard deviations, correlations and covariances are calculated on the basis of available data on each variable (Glasser, 1964; Raymond & Robert, 1987). As a result, the sample composition differs from variable to variable, and the population to which the results are generalized is no longer clearly defined.

Another problem with missing data is that they make common statistical methods inappropriate or difficult to apply (Rubin, 1987). For example, when missing data are present in a factorial analysis of variance the design is unbalanced. Consequently, the standard statistical analysis that is appropriate for balanced designs is no longer appropriate under this condition. Even if data are assumed to be missing in a completely random fashion, the proper analysis is complicated. Multivariate statistical methods, as they are programmed into commercial statistical software, are applicable to complete data sets by default.

Finally, valuable resources are wasted as a result of missing data. Time and funds spent on subjects who subsequently leave a study and produce missing data represent a loss (Buu, 1999; Holt, 1997). Such loss is a particular concern in longitudinal research, large-scale assessments, high-stake studies, and/or surveys that ask sensitive information or target respondents who are

not accustomed to responding to opinion surveys (such as the first generation among immigrants). Efforts to achieve higher response rates and complete profiles from respondents require researchers to allocate additional time and resources to trace cases who failed to respond or those whose responses were incomplete or unusable. These efforts may not always pay off.

1.3 IMPLICATIONS OF MISSING DATA

There are several reasons why researchers should be concerned with missing data. First, data are difficult to collect, and so researchers should use every piece of information they collect. Second, failing to adequately address issues of missing data can lead to biased results and incorrect conclusions. Finally, studies with missing data are more common than studies without them. Therefore, researchers should know what their best available options are in the very likely event that their study involves missing data.

1.4 MISSING DATA MECHANISM

The missing data mechanism is the process that generates missing values, that is, what predicts whether a given value is missing or not. Missing data occur to varying degrees and in various patterns (Cohen & Cohen, 1983). The impact of missing data on the validity of research findings depends on the mechanisms that led to missing data, the pattern of missing data, and the proportion of data missing (Tabachnick & Fidell, 2001), as discussed below.

It has been shown that the mechanism and the pattern of missing data have greater impact on research results than does the amount of data missing (Tabachnick & Fidell, 2001). Both are critical issues a researcher must address before choosing an appropriate procedure to deal with missing data. According to Little and Rubin (1987), mechanisms that lead to missing data can be classified as: missing completely at random, missing at random, and missing not at random

1.4.1 Missing Completely At Random (MCAR)

The idea of values missing completely at random appears in almost every technical paper on missing values. Adam and Jyoti (2010) stated that data are said to be missing completely at random (MCAR) when the probability that an observation is missing (r) does not depend on either the observed (y_{obs}) or the unobserved (y_{miss}) values. Mathematically, this can be expressed as $\Pr(r|y_{\text{obs}}, y_{\text{miss}}) = \Pr(r)$. In other words, the probability that an observation is missing is not associated with any variable you have measured or with any variable that are not measured. According to Ula (2007), MCAR occur when there is no systematic difference between complete and incomplete records. Finally, data are missing completely at random (MCAR) if the missing data mechanism is independent of both the observed and actual missing values.

1.4.2 Missing At Random (MAR)

Nancy (2004) pointed out that data are said to be Missing at Random (MAR) in analysis of Longitudinal Data if the occurrence of missing data relates to the values of the observed responses, but not to the values of the missing observations. A more realistic assumption under these circumstances is that data are missing at random (MAR), if the probability that an observation is missing depends only on the values of the observed data. In other words $\Pr(r|y_{\text{obs}}, y_{\text{miss}}) = \Pr(r|y_{\text{obs}})$. That is, the probability that an observation is missing is completely accounted for by the variables measured and not on those that have not been measured. Under circumstances where data are MCAR or MAR, the mechanism that determines whether a particular value is observed or missing is said to be ignorable. Alan(2005) in his definition stated that the missing data for a variable are MAR if the likelihood of missing data on the variable is not related to the participant's score on the variable, after controlling for other variables in the study.

1.4.3 Missing Not at Random (MNAR)

Ula (2007) defined MNAR as when the pattern of missingness is non random and is not predictable from other variables in the data set. Also, in any situations where the probability that an observation is missing depends on the values of the unobserved variables, the data are said to be missing not at random (MNAR). Under these circumstances, the nonresponse is said to be informative. Mathematically, this can be expressed as $\Pr(r|y_{\text{obs}}, y_{\text{miss}}) = \Pr(r|y_{\text{miss}})$. Graham & Donaldson (1993) referred to missing data mechanisms as “accessible” and “inaccessible.” An accessible mechanism is one where the cause of missingness can be accounted for. These situations encompass MCAR and most MAR circumstances. An inaccessible mechanism is one where the missing data mechanism cannot be measured. These situations include nonignorable mechanisms and MAR mechanisms where the cause of missingness is known, but is not measured.

1.5 PATTERNS OF MISSING DATA

Little and Rubin (1987) defined two ways of missing data patterns. These are arbitrary and monotone missing pattern. In arbitrary missing data, missing observation may occur anywhere and the ordering of variables is unimportant (Rubin 1978). In monotone missing pattern, the ordering of variables is important. In monotone, a data set with variables $X_{1+1}, X_{1+2}, X_{1+3} \dots X_{1+n}$ in the order is said to have a monotone missing pattern, if a variable x_j is observed for a particular individual it implies that all previous variables x_k , where $k < j$, are also observed for that individual.

1.6 MOTIVATION OF STUDY

The topic of missing data has obtained considerable attention in the last decade, as evidenced by several recent trends.

1. It has become difficult to publish empirical work in design without discussion of how missing data was handled.
2. More and more methods for handling missing data have sprouted-up over the last few years.

Although missing data has received a growing amount of attention, there are still some key misunderstandings regarding the problems that missing data generate, as well as acceptable solutions. Missing data are important to consider because they may lead to substantial bias in analyses.

This work will go a long way towards reducing the problem of subjective choice of methods being faced by researchers, which will automatically affect the standard error and its model estimate.

1.7 OBJECTIVES OF STUDY

The objectives of the study are as follows;

1. To describe the pattern of missing data.
2. Fill in (impute) missing values with estimated values using multiple imputation and expectation maximization.
3. Compare the estimated means, standard error and P-Values for the different missing value methods; pairwise deletion, multiple imputation and expectation maximization.

1.8 SIGNIFICANCE OF THE STUDY

Experts and Professionals in the field have devised several methods for estimating missing data in univariate and multivariate statistics. Several methods have been and continue to be developed to draw inferences from data sets with missing values (Little and Rubin 1987). This study will be of tremendous relevance to researchers and Statisticians who encounter problems in choosing which estimation method gives a more concise account of the missing observations under consideration with minimal error.

CHAPTER TWO

2.1 LITERATURE REVIEW

It is obvious that the problem of missing data is a common phenomenon in statistical literature. A lot has been said and done by different researchers on missing values. Trivellore *et al.*, (2001) stated that incomplete data is a pervasive problem faced by most applied researchers. Schafer and Graham (2002) see data missing completely at random as a special case of a more general category of missingness called missing data at random (MAR). They added that data are missing at random if the probability that the variable Y is missing is not related to the value of Y itself, after controlling for all other variables in the analysis. They pointed out that the advantage of the MAR assumption is that it allows the analyst to estimate missing values without explicitly modeling the probability that an item is missing. While, the disadvantage is that for practical purposes the assumption cannot be tested unless the missing values can somehow be obtained by the researcher.

Most of the techniques presently available for creating multiple imputation (MI) assume that the missing values are 'missing at random' (MAR) in the sense defined by Rubin (1976) and Little and Rubin (1987). That is, they assume that missing data values carry no information about probabilities of missingness. This assumption is mathematically convenient because it allows one to eschew an explicit probability model for nonresponse. In some applications, however, ignorability may seem artificial or implausible. With attrition in a longitudinal study, for example, it is possible that subjects drop out for reasons related to current data values. It is important to note that the MI paradigm does not require or assume that nonresponse is ignorable. Imputations may in principle be created under any kind of assumptions or model for the missing-data mechanism, and the resulting inferences will be valid under that mechanism.

Ravindra *et al.*, (2006), in their own view said that missing values are common when working with large data sets. They added that in the last few decades, researchers have applied several methods for imputing missing values, including various ad-hoc methods as well as advanced model-based approaches. One of the easiest techniques used is to fill in the missing value with the mean of non- missing values. While this technique is simple and easy to apply, it causes underestimation of standard deviations and standard errors, since there is no variation in the imputed values. Also, it ignores correlations that often occur in spatially and temporally varying data .

Wothke (1998) in his research, examined listwise, pairwise, mean imputation and maximum likelihood methods for growth curve modeling for cases where the data were MCAR and MAR. For the MCAR data, estimates of the model parameters were unbiased for Full information Maximum Likelihood (FIML), listwise deletion (LD) and pairwise deletion (PD), while mean imputation showed no bias in it means but exhibited strongly biased variance and covariance of the estimates. For the MAR data, FIML produced unbiased estimates while PD estimates exhibited a small negative bias. Listwise deletion and mean imputation methods resulted in sampling distributions that did not include the parameter value.

According to Leila (2009) handling missing values when tackling real life data set is a great challenge arousing, the interest of many scientific communities. He also pointed out that most statistical analyses and softwares are traditionally developed for complete observations only. Incomplete observations with missing values cannot be easily handled and therefore are routinely excluded from the statistical analysis. This results in the loss of valuable information that could have led to more reliable statistical estimation. This problem becomes more serious when the

total number of observations in the data set is not very large and the researcher needs to use every bit of precious information from the data.

In social science and psychological research involving important and sensitive issues pertaining to personal habit and attitudes, items in the questionnaire might embarrass the respondent and can cause anxiety and discomfort in response. Thus missing data are associated with these kind of questions because the respondents may fail to respond positively or may willfully give false answers or may decide not to return questionnaire leading to distortion of the result of the survey. The performance of multiple imputation in a variety of missing data situations has been well-studied and it has been shown to perform favorably (Graham *et al.*, 1997; Graham & Schafer, 1999; Schafer & Graham, 2002).

Ula (2007) listed some factors responsible for missing data in experimental block designs, which include

1. Loss of information, efficiency or power due to loss of data.
2. Problems in data handling, computation and analysis due to irregularities in the data patterns and non-applicability of standard software.
3. Serious bias if there are systematic differences between observed and the unobserved data.

Teel (1960) noted that a missing value or variable introduces a new problem into the body of the analysis since treatment and block effects are no longer orthogonal in two way classification. According to him the estimated value is entered into the table with the observed values and the analysis of variance is performed as usual with one degree of

freedom being subtracted from both total and error degrees of freedom for the estimated missing value.

As Graham and Hofer (2000) stated, the missing data mechanism is rarely completely inaccessible. Often, the mechanism is actually made up of both accessible and inaccessible factors. Thus, although a researcher may not be confident that the data present a purely accessible mechanism, covering as much of the mechanism possible will usually produce sound results (Graham *et al.*, 1997; Little, 1995; Rubin, 1996). A sensitivity analysis conducted by Graham *et al.*, (1997) showed that the effects of an inaccessible mechanism are often surprisingly minimal in the implementation of multiple imputation. Thus, encountering a situation where a portion of the missing data is inaccessible should not discourage the researcher from applying a statistically principled method. Rather, traditionally, missing data problems are approached by use of mean imputation, listwise deletion (LD) or pairwise deletion (PD) also called available case analysis.

Schafer and Olsen (1998) noted that multiple imputation methods resemble other methods of ad hoc case deletion because it addresses the missing-data issue at the beginning, before substantive analyses are run. They argue that unlike the other ad hoc methods, multiple imputations do not have to be MCAR but instead need only meet the less rigorous assumption that the missing data are missing at random (MAR). They also stated that multiple imputation techniques are statistically defensible and incorporate missing-data into all summary statistics. They however suggested that the direct maximum likelihood methods may be more efficient than multiple imputations because they do not rely on simulation.

According to David (2008), The two most important treatments of missing data in the recent literature are expectation maximization (known as the EM algorithm) (Dempster *et al.*, 1977) and multiple imputation (MI) (Rubin, 1978). These are not distinct models, and EM is often used as a starting point for MI.

Peugh and Enders (2004) carried out a methodological review of educational research journals and found that 96% of the 160 studies with missing data used these traditional methods. Nevertheless, statistical researchers do not recommend their use for parameter estimation in missing data situations, since such methods tend to bias the estimates of means, variances and correlations (Baraldi & Enders, 2010; Little and Rubin, 2002, Wilkinson & Task Force on Statistical Inference, 1999). Multiple imputations (MI) require the missing mechanism to be MAR although Thijs *et al.*, (2002) have used it in an MNAR setting. Finch (2008) has compared MI and expectation maximization (EM) under MAR and MNAR and has suggested that EM results in greater bias for parameter estimates. David (2007) states that Most researchers who use survey data must grapple with the problem of how best to handle missing information. He also concluded in his research that multiple imputation allows a researcher to use more of the available data, thereby reducing biases that may occur when observations with missing data are simply deleted.

Craig and Deborah (2001) reported a study in which a Monte Carlo simulation was used to examine the performance of four missing data methods in structural equation models: full information maximum likelihood (FIML), listwise deletion, pairwise deletion, and similar response pattern imputation. The effects of three independent variables were examined (factor loading magnitude, sample size, and missing data rate) on four outcome measures: convergence failures, parameter estimate bias, parameter estimate efficiency, and model goodness of fit.

Results indicated that FIML estimation was superior across all conditions of the design. Under ignorable missing data conditions (missing completely at random and missing at random), FIML estimates were unbiased and more efficient than the other methods. In addition, FIML yielded the lowest proportion of convergence failures and also provided near-optimal Type 1 error rates across both simulations.

David and Rebekah (2011) urges counseling psychology researchers to recognize and report how missing data are handled, because consumers of research cannot accurately interpret findings without knowing the amount and pattern of missing data or the strategies that were used to handle those data. The authors reviewed patterns of missing data and also described some of the common strategies for dealing with them. They also provided an illustration in which data were simulated and they evaluated three methods of handling missing data: mean substitution, multiple imputation, and full information maximum likelihood. Results suggested that mean substitution is a poor method for handling missing data, whereas both multiple imputation and full information maximum likelihood were recommended as alternatives to the approach.

Enders and Bandalos (2001) conducted a simulation study on the relative performance of the FIML estimation in structural equation models. They compared the statistical behavior of the FIML method with the LD, PD, and single imputation methods. Their simulation results indicate that the FIML estimation is superior to all those ad hoc methods for treating incomplete observations in all conditions studied. The FIML estimates are unbiased and more efficient than the ad hoc methods. In addition, the FIML estimation produces the lowest proportion of convergence failures during optimization.

Rufus (2006) reported an algorithm developed by Dempster *et al.*, (1977) presented an algorithm for computing maximum likelihood estimates from missing data sets. Each iteration of their algorithm consists of an expectation step followed by a maximization step. They assumed a family of sampling densities $f(x|\phi)$ depending on parameters ϕ and they then derived their corresponding family of sampling densities $g(y|\phi)$. They assumed that the EM algorithm attempts to find a value of ϕ which maximizes $g(y|\phi)$ given an observed y , but it does this by making use of the related family $f(x|\phi)$. Schafer and Olsen (1998) stated that with the development of the EM algorithm, statisticians have stopped viewing missing data as a “nuisance” and have reevaluated it as a source of variability to be averaged over.

Schafer and Olsen (1998) described a technique developed by Rubin (1987) where each value is replaced with a set of $m > 1$ plausible values which allows the variances to be averaged by simulation.

Enders (2001) carried out another simulation study to compare the statistical behavior of the FIML method with the LD, PD, and mean imputation methods under nonnormal situations. His simulation results indicated that under MCAR and MAR, the FIML estimates involve less bias and are generally more efficient than those of the ad hoc methods.

David (2007) concluded that Multiple imputation allows a researcher to use more of the available data, thereby reducing biases that may occur when observations with missing data are simply deleted.

Stubbendick and Ibrahim (2003) employed maximum likelihood methods to non-ignorable missingness of both the response and covariates in normal random effects models and later modified the method to suit discrete longitudinal data (Stubbendick and Ibrahim 2006). Jeffrey

(2003) stated that multiple imputation (MI) almost always produces estimates which are more representative of the population than do the more popular methods of handling missing data, listwise deletion (LD) and mean substitution (MS). Means and standard errors were also computed using these methods in order to illustrate earlier points regarding multiple imputation:

Applications of EM-type algorithms have a long history (McLachlan & Krishnan, 1997). The popularity of the EM algorithm among statisticians is largely based on the fact that this approach allows many complex statistical problems to be reformulated as a missing data problem in a way that greatly simplifies parameter estimation (e.g., mixture models, random effects models, hierarchical linear models, unbalanced designs including repeated measures). In most applications, it is assumed that data follow a multivariate normal distribution. In addition to taking the missing data mechanism into account, the primary advantages of the EM algorithm are simplicity and ease of computing (Dempster, *et al.*, 1977; Little & Rubin, 1987).

Wu and Wu (2007) have also employed generalized linear mixed models for data with informative dropouts and missing covariates. In all these cases, parameter estimation was done via an extension of the EM algorithm although Finch (2008) notes that EM approaches rely heavily on the assumption of multivariate normality, which does not apply to dichotomous item responses. Some researchers have however used EM algorithm to impute polytomous categorical data (Enders, 2004, Bernaards & Sitjsma, 1999) with no major challenges. It is however important to note here, just like De Boeck and Wilson (2004) noted, that no modeling approach, whether for MAR or for MNAR can fully compensate for the loss of information that occurs due to incompleteness of the data. The aim of taking into account the missingness mechanism is to be able to produce estimates that are more efficient.

Muthén *et al.*, (1987) discussed how FIML method applies to structural equation modeling. They state that their method using LISREL allows for the latent variable model to include missingness. Their paper examined maximum likelihood estimation of the θ parameters. Wothke(1998) states that FIML assumes multivariate normality, and maximizes the likelihood of the model with the observed data. He also stated that two structural equation modeling programs, AMOS (Arbuckle, 1995) and Mx (Neale, 1994), implemented this FIML method for dealing with missing data. He critically examined other methods for estimation using FIML and states that those approaches are only practical when the data have just a few distinct patterns of missing data. In addition, he stated that using AMOS (Arbuckle, 1995) and Mx does not require the same level of technical expertise as the methods presented by Dempster *et al.*, (1977) and Muthén *et al.*, (1987). Wothke (1998) suggested that both AMOS and Mx maximize the case-wise likelihood of the observed data, computed by minimizing the function. He further states that both AMOS and Mx are not limited by the number of missing-data patterns, and do not require complex steps to accommodate missing data.

Newsom (2010) suggested that If there is a large amount of missing data and data are at least MAR, there are clear advantages to using modern missing data approaches (FIML, EM, or MI) compared with listwise deletion or older imputation methods. Regarding the question of how large a proportion of missing data can be tolerated by missing data methods, there are no firm guidelines agreed upon by statisticians at present. If only a few data values are missing in a random pattern from a large data set (i.e., the MCAR condition holds), the missing data problem is less serious and almost any procedure for handling missing data yields similar results. However, if a substantial amount of data is missing from a small to moderate sized data set, the problem can be very serious (Cohen & Cohen, 1983; Cool, 2000; Tabachnick & Fidell, 2001).

2.2 CAUSES AND PROBLEMS OF MISSING OBSERVATION

In many disciplines, missing data are relatively uncommon and often taken as an indicator of sloppy science or poor methodology, and as a result, techniques for dealing with missing data when they occur are generally frowned upon in these areas. In the social sciences, however, missing or incomplete data are a nearly ubiquitous aspect of research. Missing observations can occur when for example; the yield of a corner plot of a field partitioned into a latin square has been damaged by some influence external to the experiment, e.g. a tractor. Other examples include animals dying in the course of the experiment, loss of data due to natural disaster etc. Data collection from human beings in the real life poses considerably greater challenges than in the laboratory setting. Participants have other commitments, they move, they become sick or die, they may not wish to provide information of a sensitive nature, and any number of random or systematic forces may prevent data from being observed.

2.3 PRECISION OF MISSING ESTIMATE

Missing value analysis helps address several concerns caused by incomplete data. If cases with missing values are systematically different from cases without missing values, the results can be misleading. Also, missing data may reduce the precision of calculated statistics because there is less information than originally planned. Another concern is that the assumptions behind many statistical procedures are based on complete cases, and missing values can complicate the theory required.

2.4 MISSINGNESS IN RCBD

Sometimes, in carrying out a blocked experiment, an observation is missing for reasons extraneous to the experiment. For example, a plant dies because of an accident in the greenhouse, a subject leaves town or is ill and cannot complete the experiment (assuming the illness is not related to the treatment), or the data are lost or erased. According to Shirley *et al.*, (2003) one way to handle this situation is to remove the entire block that contains the missing value. The analysis is then carried out with $b - 1$ blocks.

Model is

$$y_{ij} = \mu + b_i + \tau_j + \varepsilon_{ij} \quad 2.1$$

$$i = 1, 2, \dots, r$$

$$j = 1, 2, \dots, t$$

Say the missing observation denoted by x is in block i for treatment j

Let

B = total of all observations for block i ,

T = total of all observations for treatment j , and

G = total of all observations .

The predicted value in block i , treatment j is

$$\hat{\mu} + \hat{b}_i + \hat{\tau}_j \quad 2.2$$

Where

$$\hat{\mu} = \frac{G+x}{rt}$$

$$\hat{b}_i = \frac{B+x}{t} - \frac{G+x}{rt}$$

$$\hat{\tau}_j = \frac{T+x}{r} - \frac{G+x}{rt}$$

$$\text{So } \hat{\mu} + \hat{b}_i + \hat{\tau}_j = \tag{2.6}$$

$$\frac{x(r+t-1) + rB + tT - G}{rt} \tag{2.7}$$

Estimation of the missing value is found by equating this to x , and solving for x .

$$x(rt - r - t + 1) = rB + tT - G \tag{2.8}$$

$$x = \frac{rB + tT - G}{(r-1)(t-1)} \tag{2.9}$$

Which is the well known missing value formula for the randomized block model.

2.5 BRIEF REVIEW ON THE METHODS USED

2.5.1 Pairwise Deletion

This method looks at pairs of analysis variables and uses a case only if it has nonmissing values for both of the variables. Frequencies, means, and standard deviations are computed separately for each pair. Because other missing values in the case are ignored, correlations and covariances for two variables do not depend on values missing in any other variables. In this method, the maximum amount of available data is retained, and so this method is sometimes referred to as

available case analysis (Pigott, 2001). Cases are excluded from only operations in which data are missing on a variable that is required (Bennett, 2001; Roth, 1994). In a correlation matrix, for example, a case that has missing data on one variable would not be used to calculate the correlation coefficient between that variable and another but would be included in all other correlations. This means that different cases are used to calculate the different bivariate correlations.

According to David and Rebekah (2011), the problems with pairwise deletion come from the use of different cases for each correlation, which results in difficulty in comparing correlations and oftentimes the inability to use these correlations in multivariate analyses (the resulting correlation matrix can be inadmissible for the underlying matrix algebra).

2.5.2 Expectation Maximization (EM).

The EM algorithm, proposed by Dempster *et al.*, (1977) to solve the problems faced in maximum likelihood methods, combines statistical methodology with algorithmic application and it receives attention for the solution of various missing value problems (Dempster *et al.*, 1977). The EM algorithm is a general method for incomplete data and it increases the relation between the missing data and the unknown parameters of the data model. Finding the model parameters is easy when the missing values are known. Similarly, when the parameters are known, it is possible to make estimations for the missing values. The EM algorithm, which is an iterative method, was proposed based on the reciprocal dependence between the model parameters and the missing values. If the data space is properly chosen, the EM algorithm can be estimated effectively the missing data values. The Expectation maximization (EM) approach is an iterative procedure that proceeds in two steps (Little and Rubin 1987).

The first step called the expectation (E) step computes the expected value of the complete data log likelihood based upon the complete data cases and the algorithm's best guess as to what the sufficient statistical functions are for the missing data based upon the model specified and the existing data points.

In the second step called the maximization (M) step, it substitutes the expected values for the missing data obtained from the E step and then maximizes the likelihood function as if no data were missing to obtain new parameter estimates. The new parameter estimates are substituted back into the E step and a new M step is performed. The procedure iterates through these two steps until convergence is obtained. Convergence occurs when the change of the parameter estimates from iteration to iteration becomes negligible.

Thus, the main steps involved in EM approach are (Little and Rubin 1987):

Replace missing values by estimated values.

1. Estimate parameters.
2. Reestimate the missing values assuming the new parameter estimates are correct
3. Reestimate parameters, iterating until convergence.

The advantage of the expectation maximization approach is that it has well known statistical properties and it generally performs better than methods such as listwise, pairwise data deletion, and mean substitution because it assumes incomplete cases have data missing at random rather than missing completely at random (Allison 2002; Rubin 1978).

The main disadvantage of the EM approach is that it adds no uncertainty component to the estimated data. Practically speaking, this means that while parameter estimates based upon the

EM approach are reliable, standard errors and associated test statistics are not (Allison 2002; Rubin 1978). This weakness led to the development of two newer likelihood based methods for handling missing data, the raw maximum likelihood approach (full information maximum likelihood) and multiple imputation. This method is one of several maximum likelihood (ML) approaches. In all ML strategies, observed data are used to estimate parameters, which are then used to estimate the missing scores. The EM strategy is based on a recursive process: The missing data have information that is useful in estimating various parameters, and the estimated parameter has information that is useful in finding the most likely value of the missing data (Bennett, 2001). Thus, the EM method is an iterative procedure with two steps in each iteration: In the expectation step, the process is similar to the regression-based imputation. First, starting values for the parameters (e.g., means, covariances) are obtained with available data. Regression methods are used to impute, on the basis of these initial values, the values for the missing data. When this step is completed, in the maximization step new values for the parameters are calculated with the newly imputed data along with the original observed data. Then the process starts over with the expectation step and continues until the estimates change very little from one iteration to the next (i.e., until the estimates converge; Allison, 2001).

Applications of EM-type algorithms have a long history (McLachlan & Krishnan, 1997). The popularity of the EM algorithm among statisticians is largely based on the fact that this approach allows many complex statistical problems to be reformulated as a missing data problem in a way that greatly simplifies parameter estimation (e.g., mixture models, random effects models, hierarchical linear models, unbalanced designs including repeated measures). In most applications, it is assumed that data follow a multivariate normal distribution. In addition to

taking the missing data mechanism into account, the primary advantages of the EM algorithm are simplicity and ease of computing (Dempster, *et al.*, 1977; Little & Rubin, 1987).

The EM method provides “unbiased and efficient” (Graham *et al.*, 2003) parameters and is particularly useful for procedures such as exploratory factor analysis and internal consistency calculations, which do not require hypothesis testing. Because exploratory factor analysis requires relatively large sample sizes, the ability to impute missing data that are unbiased and retain all participants is an enormous advantage and is highly recommended. The disadvantage of EM is that the standard errors and confidence intervals are not provided, so obtaining those statistics requires an additional step. For inferential analyses for which those are essential, EM may not suffice.

2.5.3 Multiple Imputation (MI)

In multiple imputation, missing values for any variable are predicted using existing values from other variables. The predicted values, called “imputes”, are substituted for the missing values, resulting in a full data set called an “imputed data set.” This process is performed multiple times, producing multiple imputed data sets (hence the term “multiple imputation”). Standard statistical analysis is carried out on each imputed data set, producing multiple analysis results. These analysis results are then combined to produce one overall analysis.

Multiple imputation accounts for missing data by restoring not only the natural variability in the missing data, but also by incorporating the uncertainty caused by estimating missing data. Maintaining the original variability of the missing data is done by creating imputed values which are based on variables correlated with the missing data and causes of missingness. Uncertainty is

accounted for by creating different versions of the missing data and observing the variability between imputed data sets.

According to Jeffery (2003), it is important to note that imputed values produced from an imputation model are not intended to be “guesses” as to what a particular missing value might be; rather, this modeling is intended to create an imputed data set which maintains the overall variability in the population while preserving relationships with other variables. Thus, in performing multiple imputation, a researcher is interested in preserving important characteristics of the data set as a whole (e.g., means, variances, regression parameters). Creating imputes is merely a mechanism to deliver an analysis which makes use of all possible information.

Multiple imputation combines the well known statistical advantages of EM and FIML with the ability of hot deck imputation to provide a raw data matrix to analyze (Allison 2000). Multiple imputation works by generating a maximum likelihood based covariance matrix and vector of means, like EM. Multiple imputation takes the process one step further by introducing statistical uncertainty into the model and using that uncertainty to emulate the natural variability among cases one encounters in a complete database. Multiple imputation then imputes actual data values to fill in the incomplete data points in the data matrix, just as hot deck imputation does (Little and Rubin 1987).

Multiple imputation has several advantages. It is fairly well understood and robust to violations of non-normality of the variables used in the analysis. Like hot deck imputation, it outputs complete raw data matrices. It is clearly superior in most cases to listwise, pairwise, and mean substitution methods of handling missing data. Disadvantages include the time intensiveness in imputing five to ten databases, testing models for each database separately, and recombining the

model results into one summary. Multiple imputation (MI) appears to be one of the most applicable methods for general purpose handling of missing data in multivariate analysis. The basic idea of multiple imputation as presented in Little and Rubin (1987) are

1. Impute missing values using an appropriate model that incorporates random variation.
2. Repeat this m times (usually 3-5 times), producing m complete data sets.
3. Perform the desired analysis on each data set using standard complete data methods.
4. Average the values of the parameter estimates across the m samples to produce a single point estimate.
5. Calculate the standard errors by (a) averaging the squared standard errors of the m estimates (b) calculating the variance of the m parameter estimates across samples, and (c) combining the quantities using a simple formula.

CHAPTER THREE

METHODOLOGY

3.1 INTRODUCTION

This chapter presents the description of the methods that were adopted in this study. Missing data reduce the representativeness of the sample and can therefore distort inferences about the population. If it is possible try to think about how to prevent data from missing before the actual data gathering takes place. For example in computer questionnaires it is often not possible to skip a question. A question has to be answered; otherwise one cannot continue to the next. So, missing values due to the participant are eliminated by this type of questionnaire. And in survey research, it is common to make multiple efforts to contact each individual in the sample, often sending letters to attempt to persuade those who have decided not to participate to change their minds (Stoop *et al.*, 2010). However, such techniques can either help or hurt in terms of reducing the negative inferential effects of missing data, because the kind of people who are willing to be persuaded to participate after initially refusing or not being home are likely to be significantly different from the kinds of people who will still refuse or remain unreachable after additional effort (Stoop *et al.*, 2010). In situations where missing data are likely to occur, the researcher is often advised to use methods of data analysis that are robust in handling missing data. An analysis is robust when we are confident that mild to moderate violations of the technique's key assumptions will produce little or no bias, or distortion in the conclusions drawn about the population.

Various methods have been proposed for treating incomplete observations in statistical analysis. This section first describes conventional ad hoc missing data method (Pairwise and then

describes the more principled methods such as the Multiple Imputation (MI), and Expectation Maximization (EM).

3.2 THE DATA

The data used in this research is a secondary data which was gotten from Institute of Agricultural Research, Zaria. The data is a replicated 2 x 3 x 4 factorial experiment in a randomized complete block design (RCBD) for the number of leaves, height and number of branches of two cowpea varieties types at 3 different rates of manure and 4 different rates of nitrogen, all in 3 blocks.

3.3 FACTORIAL EXPERIMENT

Factorial designs are very efficient for studying two or more factors. The effect of a factor can be defined as the change in response produced by a change in the level of the factor. This is referred to as the main effect. In some experiments, it may be found that the difference in the response between levels of one factor is not the same at all levels of the other factors. This is referred to as an interaction effect between factors. Collectively, main effects and interaction effects are called the factorial effects. A full factorial design can estimate all main effects and higher-order interactions.

3.4 TESTING THE ASSUMPTIONS FOR ANOVA

All the Assumptions for Anova were met.

1. The treatment groups are normally distributed.
2. The treatment groups all have the same variance.
3. The experimental units are picked at random and assigned at random to the treatment groups.

According to Sherly *et al.*, (2004), The model for a x b x c factorial design is as follows

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \varepsilon_{ijkl}$$

$$i = 1, \dots, a$$

$$j = 1, \dots, b$$

$$k = 1, \dots, c$$

$$l = 1, \dots, n$$

μ = the overall mean for the experiments of this type α_i ; the effect of the i^{th} level of factor A

β_j = the factor effect of the j^{th} level of factor B

γ_k = the effect of the k^{th} level of factor C

$(\alpha\beta)_{ij}$ = the interaction effect between the i^{th} level of factor A and the j^{th} level of factor B

$(\alpha\gamma)_{ik}$ = the interaction effect between the i^{th} level of factor A and the k^{th} level of factor C

$(\beta\gamma)_{jk}$ = the interaction effect between the j^{th} level of factor B and the k^{th} level of factor C

$(\alpha\beta\gamma)_{ijk}$ = the interaction effect among the i^{th} level of A, j^{th} level of B and the k^{th} level of factor C

ε_{ijkl} = a random effect due to sampling $\varepsilon_{ijkl} \text{ IND } (0, \sigma^2)$

3.5 ADVANTAGES OF FACTORIAL DESIGN

1. Factorial design are cost efficient
2. Factorial design may enhance external validity

- External validity: to what extent research findings can be generalize to other conditions.
- Whenever we are interested in examining treatment variations, factorial designs should be strong candidates as the designs of choice

3. Factorial designs are the only effective way to examine interaction effects

Table 3. 1: Anova Table of $a \times b \times c$ Factorial Design

SOURCE OF VARIATION	DEGREE OF FREEDOM (DF)	SUM OF SQUARES (SS)	MEAN OF SQUARE(MS)	F
BETWEEN REPLICATION	$r - 1$	SSR	$MSR = SSR / (r - 1)$	MSR / MSE
A	$a - 1$	SSA	$MSA = SSA / (a - 1)$	MSA / MSE
B	$b - 1$	SSB	$MSB = SSB / (b - 1)$	MSB / MSE
C	$c - 1$	SSC	$MSC = SSC / (c - 1)$	MSC / MSE
AB	$(a - 1)(b - 1)$	SSAB	$MSAB = SSAB / (a - 1)(b - 1)$	$MSAB / MSE$
AC	$(a - 1)(c - 1)$	SSAC	$MSAC = SSAC / (a - 1)(c - 1)$	$MSAC / MSE$
BC	$(b - 1)(c - 1)$	SSBC	$MSBC = SSBC / (b - 1)(c - 1)$	$MSBC / MSE$
ABC	$(a - 1)(b - 1)(c - 1)$	SSABC	$MSABC = SSABC / (a - 1)(b - 1)(c - 1)$	$MSABC / MSE$
ERROW	$abc(n - 1)$	SSE	$MSE = SSE / abc(n - 1)$	

3.6 BRIEF REVIEW ON SIMULATION

Monte Carlo simulation uses repeated random sampling to simulate data for a given mathematical model and evaluate the outcome. According to Paul (2012), this method was initially applied back in the 1940s, when scientists working on the atomic bomb used it to calculate the probabilities of one fissioning uranium atom causing a fission reaction in another. With uranium in short supply, there was little room for experimental trial and error. The scientists discovered that as long as they created enough simulated data, they could compute reliable probabilities and reduce the amount of uranium needed for testing.

Today, simulated data is routinely used in situations where resources are limited or gathering real data would be too expensive or impractical. By using Minitab's ability to easily create random data.

3.6.1 Importance of Simulation

1. Simulate the range of possible outcomes to aid in decision-making
2. Forecast financial results or estimate project timelines
3. Understand the variability in a process or system
4. Find problems within a process or system
5. Manage risk by understanding cost/benefit relationships

3.6.2 Description of How the Simulation of the Simulated Data Was Obtained

I used Minitab to create a random set of data that is normally distributed

Below are the procedures used in obtaining the simulated data

Select Calc >> Random Data >> Normal..

1. In the box labeled Generate ... rows of data, type in the number of rows of data that you would like to generate.
2. In the box labeled Store in Column(s):, enter the column name(s) where you want Minitab to store the data.
3. In the boxes labeled Mean: and Standard deviation:, type in the mean and standard deviation of your desired normal distribution. The default is the standard normal distribution with mean = 0 and standard deviation = 1.

3.7 DESCRIPTION OF HOW MISSING VALUES WERE OBTAINED

For selecting which method is best to be used, same amount of data points were randomly selected to be missing for both real life dataset and the simulated dataset for the different factor levels. The Real life data, the simulated data, the incomplete Real life and the incomplete simulated data are presented in the appendix.

3.8 METHODS FOR REPLACING MISSING VALUES

3.8.1 Pairwise Deletion

PD retains all available data provided by a subject. If this approach is applied to data analysis, descriptive statistics and a few inferential statistics (t -, z -, and chi-square, etc.) are computed from non- missing data on each variable (Glasser, 1964; Raymond & Robert, 1987). It is the default setting in SPSS®, SYSTAT®, and SAS® for descriptive, correlation, and regression analysis using either correlation or covariance matrices. According to Kim and Curry (1977), PD is an attractive alternative when there is a small number of missing cases on each variable relative to the total sample size, and a large number of variables are involved. PD approach utilizes information obtained from partially complete observations. Its disadvantage is that the

sample data change from variable to variable. This variability in the sample base creates practical problems, such as the determination of sample size and degrees of freedom. It is especially problematic for multivariate statistical analyses where solutions and intermediate computations are often based on the entire raw data matrix (Rubin, 1987). According to Cool (2000), it is possible to compute correlation matrices with mutually inconsistent correlations.” Because of this problem, sample correlation or covariance matrices may not be semi-positive definite Malhotra, (1987). The PD method produces biased parameter estimates and biased statistical tests unless the MCAR assumption holds. For these reasons, PD is not a satisfactory solution to the missing data problem Wilkinson (1999).

3.8.2 Expectation Maximization

The EM algorithm is aimed at estimating a parameter θ , such that $P(X/\theta)$ is maximum. The variable X can be defined as a random vector from a parameterized family. To estimate the parameter adequately, a log likelihood function is introduced and the likelihood of θ , denoted as $L(\theta)$ is defined by Dempster *et al.*, (1977)

$$L(\theta) = \ln P(X/\theta),$$

Due to the \ln function, $L(\theta)$ is an ever increasing function such that the θ that will maximize $L(\theta)$ will also maximize $P(X/\theta)$. As mentioned earlier, the EM algorithm is an iteration procedure aimed at maximizing $L(\theta)$. The iteration is performed such that the most recent value of $L(\theta)$ is better than the ones in previous iterations. As a result,

$$L(\theta_n) > L(\theta_{n-1})$$

3.2

Where n represents the n th iteration. The algorithm is also aimed at reaching to the maximum value much faster. This can be achieved by maximizing the difference between two consecutive values of $L(\theta)$. The EM algorithm provides a natural framework that enables missing data to be reestimated. The missing data may be viewed as some hidden variable that, if available, would make the maximization process easier. By letting Z denote a hidden variable.

Dempster *et al.*, (1977) and Allison (2002) defined maximum probability as

$$P(X/\theta) = \sum_z P(X/Z, \theta) P(Z/\theta) \tag{3.3}$$

The difference, $L(\theta_n) - L(\theta_{n-1}) = \ln \left[\sum_z P(X/Z, \theta) P(Z/\theta) - \ln P(X/\theta_{n-1}) \right]$ 3.4

More formally, the E-step is aimed at determining the conditional expectation $E_{Z/X, \theta_n} [P(X, Z/\theta)]$ whereas the maximization step maximizes this expression with respect to θ .

3.8.2.1 Execution of EM Using IBM SPSS

Many statistical packages can now implement expectation maximization. To execute this technique with SPSS

1. Choose Missing Value Analysis from the Analyze menu.
2. Transfer all numerical variables that are related to the study or issue into the box labelled Quantitative Variables. Exclude irrelevant variables
3. Transfer all categorical variables that are related to the study or issue into the box labeled Categorical Variables

4. Select the EM option
5. Press the EM button, and select Save completed data.
6. Choose Write a new data file. Press File and type a filename.
7. Open this new file-which should include the data together with some of the missing data completed.

3.8.3 Multiple Imputation

MI is a principled missing data method that provides valid statistical inferences under the MAR condition, Little and Rubin (2002). MI was proposed to impute missing data while acknowledging the uncertainty associated with the imputed values (Little and Rubin 2002). Specifically, MI acknowledges the uncertainty by generating a set of m plausible values for each unobserved data point, resulting in m complete data sets, each with one unique estimate of the missing values. The m complete data sets are then analyzed individually using standard statistical procedures, resulting in m slightly different estimates for each parameter. At the final stage of MI, m estimates are pooled together to yield a single estimate of the parameter and its corresponding standard error (SE). The pooled SE of the parameter estimate incorporates the uncertainty due to the missing data treatment (the between imputation uncertainty) into the uncertainty inherent in any estimation method (the within imputation uncertainty). Consequently, the pooled SE is larger than the SE derived from a single imputation method (e.g., mean substitution) that does not consider the between imputation uncertainty. Thus, MI minimizes the bias in the SE of a parameter estimate derived from a single imputation method.

In sum, MI handles missing data in three steps: (1) imputes missing data m times to produce m complete data sets; (2) analyzes each data set using a standard statistical procedure; and (3)

combines the m results into one using formulae from Rubin (1987) or Schafer (1997). Below we discuss each step in greater details.

3.8.3.1 Step 1: Imputation

The imputation step in MI is the most complicated step among the three steps. The aim of the imputation step is to fill in missing values multiple times using the information contained in the observed data. Many imputation methods are available to serve this purpose. The preferred method is the one that matches the missing data pattern. Given a univariate or monotone missing data pattern, one can impute missing values using the regression method (Rubin 1987), or the predictive mean matching method if the missing variable is continuous (Heitjan and Little 1991 ; Schenker and Taylor 1996). When data are missing arbitrarily, one can use the Markov Chain Monte Carlo (MCMC) method (Schafer (1997), or the fully conditional specification (also referred to as chained equations) if the missing variable is categorical or non-normal (Raghunathan *et al.* 2001 ; van Buuren 2007 ; van Buuren *et al.*, 1999 ; van Buuren *et al.*, 2006).

3.8.3.2 Step 2: Statistical Analysis

The second step of MI analyzes the m sets of data separately using a statistical procedure of a researcher's choice. At the end of the second step, m sets of parameter estimates are obtained from separate analyses of m data sets.

3.8.3.3 Step 3: Combining Results

The third step of MI combines the m estimates into one. Rubin (1987) provided formulae for combining m point estimates.

Let \hat{Q}_i and \hat{U}_i be the point and variance estimates from the i th imputed data set, $i = 1, 2, \dots, m$.

Then the point estimate for Q from multiple imputations is the average of the m complete data estimates.

$$\bar{Q} = \frac{1}{m} \sum_{i=1}^m \hat{Q}_i \quad 3.6$$

Let \bar{U} be the within imputation variance, which is the average of the m complete data estimates

$$\bar{U} = \frac{1}{m} \sum_{i=1}^m \hat{U}_i \quad 3.7$$

And B be the between imputation variance

$$B = \frac{1}{m-1} \sum_{i=1}^m (\hat{Q}_i - \bar{Q})^2 \quad 3.8$$

Then the variance estimate associated with \bar{Q} is the total variance

$$T = \bar{U} + \left(1 + \frac{1}{m}\right) B \quad 3.9$$

The Statistic $(\hat{Q} - \bar{Q}) T^{-1/2}$ is approximately distributed as a t-distribution with

V_m degrees of freedom (Rubin, 1987)

$$V_m = (m-1) \left[1 + \frac{\bar{U}}{(m-1)B} \right]^2 \quad 3.10$$

When the complete data degrees of freedom V_0 is small and there is only a modest proportion of missing data the computed degrees of freedom V_m^*

$$V_m^* = \left[\frac{1}{V_m} + \frac{1}{V_{obs\hat{}}} \right]^{-1} \quad 3.11$$

Where

$$V_{obs\hat{}} = \frac{V_0 + 1}{V_0 + 3} V_0 (1 - \gamma) \quad 3.12$$

$$\gamma = \frac{\text{tr}(m^{-1} \hat{B})}{T} \quad 3.13$$

Similar to the univariate inferences, multivariate inferences based on wald's tests can also be derived from the m imputed data sets.

CHAPTER FOUR

DATA ANALYSIS AND INTERPRETAION

4.1 INTRODUCTION

This chapter seeks to analyze and interpret the results obtained from the different methods for estimating missing values in a randomized complete block design. These methods include the pairwise deletion (PD) method, expectation maximization (EM) method and multiple imputation (MI) method. For selecting which method is best to be recommended for use we examined their mean estimates, standard error estimates and P-values of the three methods with the real-life data set (RLDS) estimates and the simulated data set (SIMDS) estimates for assessing the power of each of the methods. The data used for the study is a replicated 2 x 3 x 4 factorial experiment in a randomized complete block design (RCBD) for the number of leaves, height and number of branches of two cowpea varieties types at 3 different rates of manure and 4 different rates of nitrogen, all in 3 blocks. Data points were randomly selected to be missing for both the real-life data and the simulated data for the different factors levels. The real-life data and the simulated data are presented in the appendix.

4.2 PRESENTATION AND INTERPRETATION OF RESULTS

Below are the results obtained from the three different methods which were discussed in the methodology for estimating missing data.

4.2.1 Presentation and Interpretation of Results for the Real-Life Data and Simulated Data Sets

In this section we applied all the three methods under study on the real data and simulated data set. The means, standard error and P-values obtained from the different methods were compared for height, number of leaves and branches. Below are the results.

Table 4. 1: Estimated Means and Standard Errors for Plant Height Based On the Rlds and the Three Methods (PD, MI and EM).

MEANS					STANDARD ERROR			
REPS	RLDS	PD	MI	EM	RLDS	PD	MI	EM
1	33.750	34.337	33.542	33.978	0.727	0.854	0.645	0.748
2	34.958	35.223	34.592	34.868	0.727	0.918	0.645	0.748
3	32.50	32.656	32.867	32.519	0.727	0.797	0.645	0.748
V								
1	32.944	33.371	33.344	33.263	0.594	0.684	0.527	0.610
2	34.528	33.774	33.989	34.313	0.594	0.684	0.570	0.610
N								
0	30.333	30.986	31.044	30.915	0.840	1.002	0.745	0.863
1	33.556	33.672	34.056	33.863	0.840	0.933	0.745	0.863
2	35.556	35.374	34.411	34.814	0.840	0.998	0.745	0.863
3	35.500	36.257	35.156	35.562	0.840	0.933	0.745	0.863
M								
0	30.750	30.588	31.450	30.972	0.727	0.857	0.645	0.748
1	33.792	34.188	34.258	33.615	0.727	0.814	0.645	0.748
2	36.667	37.441	35.292	36.778	0.727	0.844	0.645	0.748
V*N	RLDS	PD	MI	EM	RLDS	PD	MI	EM

0 1	29.778	30.736	31.156	30.850	1.188	1.413	1.053	1.221
0 2	30.889	31.236	30.933	30.980	1.188	1.413	1.053	1.221
1 1	33.222	33.456	34.400	33.836	1.188	1.413	1.053	1.221
1 2	33.809	33.889	33.711	33.889	1.188	1.220	1.053	1.221
2 1	35.111	34.933	34.844	35.035	1.188	1.323	1.053	1.221
2 2	36	35.814	33.978	34.592	1.188	1.509	1.053	1.221
3 1	33.667	34.359	32.978	33.332	1.188	1.323	1.053	1.221
3 2	37.333	38.155	37.333	37.792	1.188	1.323	1.053	1.221
V*M								
0 1	30.083	30.002	30.967	30.538	1.029	1.247	0.912	1.057
0 2	31.417	31.175	31.933	31.405	1.029	1.184	0.912	1.057
1 1	32.083	32.531	33.933	32.45	1.029	1.124	0.912	1.057
1 2	35.500	35.844	34.583	34.781	1.029	1.184	0.912	1.057
2 1	36.667	37.579	35.133	36.802	1.029	1.195	0.912	1.057
2 2	36.667	37.302	35.45	36.754	1.029	1.184	0.912	1.057
N*M								
0 0	25.167	25.604	27.100	26.638	1.455	1.837	1.290	1.495
0 1	32.500	31.954	32.967	32.687	1.455	1.68	1.290	1.495
0 2	34.500	33.962	33.500	33.728	1.455	1.839	1.290	1.495
0 3	30.833	30.833	32.233	30.833	1.455	1.494	1.290	1.495
1 0	34.333	35.066	34.3	33.834	1.455	1.680	1.290	1.495
1 1	32.667	33.563	34.533	33.40	1.455	1.679	1.290	1.495
0 2	33.833	33.788	34.367	32.893	1.455	1.680	1.290	1.495
0 3	34.333	34.333	33.833	34.333	1.455	1.494	1.290	1.495
2 0	31.500	32.288	31.733	32.272	1.455	1.680	1.290	1.495
2 1	35.500	35.500	34.667	35.500	1.455	1.494	1.290	1.495
2 2	38.333	38.371	35.367	37.819	1.455	1.68	1.290	1.495

2 3	41.333	43.604	39.400	41.519	1.455	1.837	1.290	1.495
V*N*M								
0 0 1	26.333	27.632	28.600	28.005	2.057	2.612	1.824	2.115
0 0 2	24.000	23.576	25.600	25.270	2.057	2.614	1.824	2.115
0 1 1	30.667	29.576	31.867	31.041	2.057	2.614	1.824	2.115
0 1 2	34.333	34.333	34.067	34.333	2.057	2.113	1.824	2.115
0 2 1	34.667	34.132	34.067	34.438	2.057	2.612	1.824	2.115
0 2 2	34.333	33.792	32.933	33.018	2.057	2.610	1.824	2.115
0 3 1	28.667	28.667	29.333	28.667	2.057	2.113	1.824	2.115
0 3 2	33.000	33.000	35.133	33.000	2.057	2.113	1.824	2.115
1 0 1	32.000	32.000	33.533	32.000	2.057	2.113	1.824	2.115
1 0 2	36.667	38.132	35.067	35.669	2.057	2.612	1.824	2.115
1 1 1	32.000	33.792	35.733	33.467	2.057	2.610	1.824	2.115
1 1 2	33.333	33.333	33.333	33.333	2.057	2.113	1.824	2.115
1 2 1	30.000	30.000	32.333	30.000	2.057	2.113	1.824	2.115
1 2 2	37.667	37.576	36.400	35.787	2.057	2.614	1.824	2.115
1 3 1	34.333	34.333	34.133	34.333	2.057	2.113	1.824	2.115
1 3 2	34.333	34.333	33.533	34.333	2.057	2.113	1.824	2.115
2 0 1	31.000	32.576	31.333	32.545	2.057	2.614	1.824	2.115
2 0 2	32.000	32.000	32.133	32.000	2.057	2.113	1.824	2.115
2 1 1	37.000	37.000	35.600	37.000	2.057	2.113	1.824	2.115
2 1 2	34.000	34.000	33.733	34.000	2.057	2.113	1.824	2.115
2 2 1	40.667	40.667	38.133	40.667	2.057	2.113	1.824	2.115
2 2 2	36.000	36.076	32.600	34.972	2.057	2.614	1.824	2.115
2 3 1	38.000	40.076	35.467	36.996	2.057	2.614	1.824	2.115
2 3 2	44.667	47.132	43.333	46.042	2.057	2.612	1.824	2.115

Table 4.1 above present the means and standard errors of the three factors levels on plant heights for the real-life data set before data points are randomly selected as missing which is compared with means and standard errors obtained after data points have been randomly selected as missing and have been estimated by PD, MI and EM. It was observed by inspection that the means obtained by the three methods for each of the levels of the three factors compared with the real-life data are similar with slight departure. Generally, the EM estimates are more consistent than the PD and MI with PD less consistent. Observing the standard errors obtained from the three methods, the EM proved to be more superior over the PD and MI since its standard error values are approximately similar to that of the real life-data set while the PD and MI standard error values are very high or very low compared to the real-life data set standard error values. For instance the mean standard error for V*N*M for the real-life data set is approximately 2.06 while that of PD is 2.36, MI is 1.82 and EM is 2.12 this is an evidence of the EM to be the best method for estimating missing data point(s) from a data set since its value of 2.12 is approximately closer to the value of the real-life data value of 2.06 while PD is greater and MI value is less. Likewise, same evidence was observed for the levels of the main factors (V, N and M) and their interactions, the EM claimed superiority.

Table 4. 2:Estimated Means and Standard Errors for Plant Height Based on the SIMDS and the Three Methods (PD, MI AND EM).

MEANS					STANDARD ERROR S			
REPS	SIMDS	PD	MI	EM	SIMDS	PD	MI	EM
1	32.375	32.223	32.375	32.386	0.495	0.564	0.477	0.455
2	33.542	33.518	33.408	33.454	0.495	0.607	0.477	0.455
3	33.417	33.506	33.45	33.505	0.495	0.51	0.477	0.455
V								
1	33.278	33.122	33.206	33.209	0.405	0.446	0.390	0.371
2	32.944	33.042	32.95	33.021	0.405	0.452	0.390	0.371
N								
0	32.556	32.652	32.789	32.851	0.572	0.662	0.551	0.525
1	33.000	32.87	32.967	32.898	0.572	0.595	0.551	0.525
2	32.722	32.786	32.711	32.851	0.572	0.659	0.551	0.525
3	34.167	34.020	33.844	33.86	0.572	0.617	0.551	0.525
M								
0	32.208	31.974	32.133	32.155	0.495	0.566	0.477	0.455
1	33.875	33.953	34.125	34.068	0.495	0.524	0.477	0.455
2	33.25	33.32	32.975	33.121	0.495	0.557	0.477	0.455
V*N								
0 1	33.000	33.374	33.311	33.448	0.809	0.934	0.779	0.743
0 2	32.111	31.929	32.267	32.254	0.809	0.934	0.779	0.743
1 1	33.667	33.406	33.6	33.462	0.809	0.874	0.779	0.743
1 2	32.333	32.333	32.333	32.333	0.809	0.806	0.779	0.743
2 1	32.556	31.857	32.289	32.268	0.809	0.874	0.779	0.743
2 2	32.889	33.716	33.133	33.434	0.809	0.997	0.779	0.743
3 1	33.889	33.85	33.622	33.657	0.809	0.874	0.779	0.743
3 2	34.444	34.19	34.067	34.062	0.809	0.874	0.779	0.743
V*M								
0 1	32.500	32.090	32.450	32.383	0.701	0.824	0.675	0.643

0 2	31.917	31.857	31.817	31.928	0.701	0.783	0.675	0.643
1 1	34.333	34.333	34.333	34.333	0.701	0.698	0.675	0.643
1 2	33.417	33.572	33.917	33.804	0.701	0.782	0.675	0.643
2 1	33.000	32.942	32.833	32.911	0.701	0.790	0.675	0.643
2 2	33.500	33.697	33.117	33.331	0.701	0.782	0.675	0.643
N*M								
0 0	33.000	33.144	33.067	33.173	0.991	1.214	0.954	0.910
0 1	31.833	31.442	31.733	31.527	0.991	1.110	0.954	0.910
0 2	32.333	31.641	32.067	32.255	0.991	1.215	0.954	0.910
0 3	31.667	31.667	31.667	31.667	0.991	0.987	0.954	0.910
1 0	33.000	33.202	33.567	33.545	0.991	1.110	0.954	0.910
1 1	34.500	34.500	34.500	34.500	0.991	0.987	0.954	0.910
0 2	32.000	32.109	32.433	32.229	0.991	1.110	0.954	0.910
0 3	36.000	36.000	36.000	36.000	0.991	0.987	0.954	0.910
2 0	31.667	31.609	31.733	31.836	0.991	1.110	0.954	0.910
2 1	32.667	32.667	32.667	32.667	0.991	0.987	0.954	0.910
2 2	33.833	34.609	33.633	34.069	0.991	1.110	0.954	0.910
2 3	34.833	34.394	33.867	33.912	0.991	1.214	0.954	0.910
V*N*M								
0 0 1	33.333	34.570	34.133	34.339	1.401	1.726	1.349	1.287
0 0 2	32.667	31.718	32.000	32.007	1.401	1.727	1.349	1.287
0 1 1	33.000	32.218	32.800	32.387	1.401	1.727	1.349	1.287
0 1 2	30.667	30.667	30.667	30.667	1.401	1.396	1.349	1.287
0 2 1	32.667	30.570	31.867	31.804	1.401	1.726	1.349	1.287
0 2 2	32.000	32.712	32.267	32.706	1.401	1.724	1.349	1.287
0 3 1	31.000	31.000	31.000	31.000	1.401	1.396	1.349	1.287
0 3 2	32.333	32.333	32.333	32.333	1.401	1.396	1.349	1.287
1 0 1	34.333	34.333	34.333	34.333	1.401	1.396	1.349	1.287
1 0 2	31.667	32.070	32.8	32.756	1.401	1.726	1.349	1.287
1 1 1	34.667	34.667	34.667	34.667	1.401	1.396	1.349	1.287

V*N*M	SIMDS	PD	MI	EM	SIMDS	PD	MI	EM
1 1 2	34.333	34.333	34.333	34.333	1.401	1.396	1.349	1.287
1 2 1	32.000	32.000	32.000	32.000	1.401	1.396	1.349	1.287
1 2 2	32.000	32.218	32.867	32.458	1.401	1.727	1.349	1.287
1 3 1	36.333	36.333	36.333	36.333	1.401	1.396	1.349	1.287
1 3 2	35.667	35.667	35.667	35.667	1.401	1.396	1.349	1.287
2 0 1	31.333	31.218	31.467	31.671	1.401	1.727	1.349	1.287
2 0 2	32.000	32.000	32.000	32.000	1.401	1.396	1.349	1.287
2 1 1	33.333	33.33333	33.333	33.333	1.401	1.396	1.349	1.287
2 1 2	32.000	32.000	32.000	32.000	1.401	1.396	1.349	1.287
2 2 1	33.000	33.000	33.000	33.000	1.401	1.396	1.349	1.287
2 2 2	34.667	36.21778	34.267	35.137	1.401	1.727	1.349	1.287
2 3 1	34.333	34.21778	33.533	33.639	1.401	1.727	1.349	1.287
2 3 2	35.333	34.57037	34.2	34.185	1.401	1.726	1.349	1.287

Table 4.2 above present the means and standard errors of the three factors levels on plant heights for the simulated data set before data points are randomly selected as missing which is compared with means and standard errors obtained after data points have been randomly selected as missing and have been estimated by PD, MI and EM. It was observed by inspection that the means obtained by the three methods for each of the levels of the three factors compared with the simulated data set are similar with slight departure from the means of the simulated data but generally, the EM estimates are more consistent than the PD and MI because from the overall means of the levels of each factors and their interaction we observed that the EM obtained overall means (33.11) that are equal to the overall means of the simulated data (33.11) while PD and MI obtained means (33.08 respectively) that are slightly less. Observing the standard errors obtained by the three methods we observed that the MI displayed superiority over the PD and

EM since its standard error values are approximately similar to that of the simulated data set while the PD standard error values are higher than the simulated data set and EM standard error values are slightly less than the standard error values of the simulated data set. For instance the mean standard error for V*N*M for the simulated data set is approximately 1.40 while that of PD is 1.55, MI is 1.35 and EM is 1.29, this is an evidence of the MI to be the best method for estimating missing data point(s) from a data set since its value is approximately closer to the simulated data value of 1.40 while PD is greater and EM value is less. Likewise, same result was observed for the levels of the main factors (V, N and M) and their interactions, the MI claimed superiority. The results obtained showed that the MI and EM are best for estimating missing values based on their means and standard error obtained. However, the results from the real-life data set revealed that the EM is best for estimating missing data, this could be that MI might be affected by outliers or slight departure from normality though the real-life data is assumed normal but the simulated is normally distributed with $N(0,1)$ which is specific.

Table 4. 3:Estimated Means and Standard Errors for Plant Leave Based on the RLDS and the Three Methods (PD, MI AND EM).

MEANS OBTAINEDS					STANDARD ERRORS			
REPS	RLDS	PD	MI	EM	RLDS	PD	MI	EM
1	39.500	39.613	37.358	38.761	1.639	1.791	1.482	1.612
2	33.458	33.222	34.367	32.545	1.639	1.727	1.482	1.612
3	31.083	29.916	30.183	29.76	1.639	1.86	1.482	1.612
V								
1	37.000	36.667	36.400	36.047	1.338	1.425	1.210	1.316
2	32.361	31.834	31.539	31.331	1.338	1.411	1.210	1.316
N								
0	23.611	23.780	25.178	23.577	1.892	1.961	1.712	1.861
1	29.833	31.225	33.267	30.497	1.892	1.962	1.712	1.861
2	40.056	38.697	38.289	38.59	1.892	2.040	1.712	1.861
3	45.222	43.301	39.144	42.091	1.892	2.089	1.712	1.861
M								
0	26.708	27.898	29.067	25.433	1.639	1.778	1.482	1.612
1	36.750	35.812	36.567	36.015	1.639	1.740	1.482	1.612
2	40.583	39.042	36.275	39.619	1.639	1.691	1.482	1.612
V*N								
0 1	29.556	29.389	30.622	28.953	2.676	2.675	2.421	2.633
0 2	17.667	18.171	19.733	18.2	2.676	2.858	2.421	2.633
1 1	37.000	38.833	38.644	37.15	2.676	2.858	2.421	2.633
1 2	22.667	23.616	27.889	23.844	2.676	2.675	2.421	2.633
2 1	42.333	41.894	39.778	41.386	2.676	2.858	2.421	2.633
2 2	37.778	35.500	36.800	35.795	2.676	2.892	2.421	2.633
3 1	39.111	36.551	36.556	36.697	2.676	3.050	2.421	2.633

3 2	51.333	50.051	41.733	47.485	2.676	2.858	2.421	2.633
V*M	RLDS	PD	MI	EM	RLDS	PD	MI	EM
0 1	28.333	30.500	31.283	27.698	2.318	2.395	2.096	2.280
0 2	25.083	25.295	26.85	23.167	2.318	2.622	2.096	2.280
1 1	39.333	38.458	39.917	38.359	2.318	2.504	2.096	2.280
1 2	34.167	33.166	33.217	33.671	2.318	2.418	2.096	2.280
2 1	43.333	41.042	38.000	42.083	2.318	2.504	2.096	2.280
2 2	37.833	37.042	34.550	37.155	2.318	2.273	2.096	2.280
N*M								
0 0	15.500	17.174	18.400	14.567	3.278	3.397	2.965	3.224
0 1	23.833	28.931	30.067	25.356	3.278	3.773	2.965	3.224
0 2	30.333	28.500	33.767	28.716	3.278	3.398	2.965	3.224
0 3	37.167	36.986	34.033	33.091	3.278	3.769	2.965	3.224
1 0	27.500	26.333	29.167	28.329	3.278	3.776	2.965	3.224
1 1	33.000	33.000	35.633	33.000	3.278	3.020	2.965	3.224
0 2	39.667	38.007	41.267	38.669	3.278	3.397	2.965	3.224
0 3	46.833	45.909	40.200	44.062	3.278	3.717	2.965	3.224
2 0	27.833	27.833	27.967	27.833	3.278	3.02	2.965	3.224
2 1	32.667	31.743	34.100	33.135	3.278	3.396	2.965	3.224
2 2	50.167	49.583	39.833	48.386	3.278	3.776	2.965	3.224
2 3	51.667	47.007	43.200	49.121	3.278	3.397	2.965	3.224
V*N*M								
0 0 1	18.667	18.667	22.867	18.667	4.636	4.271	4.193	4.560
0 0 2	12.333	15.681	13.933	10.468	4.636	5.283	4.193	4.560
0 1 1	30.333	37.681	32.533	29.846	4.636	5.283	4.193	4.560
0 1 2	17.333	20.181	27.600	20.865	4.636	5.283	4.193	4.560
0 2 1	34.667	34.667	36.267	34.667	4.636	4.271	4.193	4.56

V*N*M	RLDS	PD	MI	EM	RLDS	PD	MI	EM
0 2 2	26.000	22.333	31.267	22.766	4.636	5.286	4.193	4.56
0 3 1	29.667	30.986	33.467	27.612	4.636	5.281	4.193	4.560
0 3 2	44.667	42.986	34.600	38.571	4.636	5.281	4.193	4.560
1 0 1	37.333	36.833	36.533	35.527	4.636	5.286	4.193	4.560
1 0 2	17.667	15.833	21.8	21.132	4.636	5.286	4.193	4.560
1 1 1	38.333	38.333	42.467	38.333	4.636	4.271	4.193	4.560
1 1 2	27.667	27.667	28.8	27.667	4.636	4.271	4.193	4.560
1 2 1	37.000	33.681	37.867	35.004	4.636	5.283	4.193	4.560
1 2 2	42.333	42.333	44.667	42.333	4.636	4.271	4.193	4.560
1 3 1	44.667	44.986	42.800	44.572	4.636	5.281	4.193	4.560
1 3 2	49.000	46.833	37.600	43.552	4.636	5.286	4.193	4.560
2 0 1	32.667	32.667	32.467	32.667	4.636	4.271	4.193	4.560
2 0 2	23.000	23.000	23.467	23.000	4.636	4.271	4.193	4.560
2 1 1	42.333	40.486	40.933	43.271	4.636	5.281	4.193	4.560
2 1 2	23.000	23.000	27.267	23.000	4.636	4.271	4.193	4.560
2 2 1	55.333	57.333	45.200	54.486	4.636	5.286	4.193	4.560
2 2 2	45.000	41.833	34.467	42.285	4.636	5.286	4.193	4.560
2 3 1	43.000	33.681	33.400	37.909	4.636	5.283	4.193	4.560
2 3 2	60.333	60.333	53.000	60.333	4.636	4.271	4.193	4.560

Table 4.3 above present the means and standard errors of the three factors levels on plant leaves for the real-life data set before data points are randomly selected as missing which is compared with means and standard errors obtained after data points have been randomly selected as missing and have been estimated by PD, MI and EM. It was observed by inspection that the means obtained by the three methods for each of the levels of the three factors compared with the

means of the real-life data are similar with slight departure. Generally, the EM estimates are more consistent than the PD and MI. Observing the standard errors obtained from the three methods, the EM proved to be more superior over the PD and MI since its standard error values are approximately similar to that of the real life-data set while the PD and MI standard error values are very high or very low compared to the real-life data set standard error values. For instance the mean standard error for V*N*M for the real-life data set is approximately 4.64 while that of PD is 4.90, MI is 4.19 and EM is 4.56 this is an evidence of the EM to be the best method for estimating missing data point(s) from a data set since its value of 4.56 is approximately close to the real-life data value of 4.64 while PD value is greater and MI value is less. Likewise, same evidence was observed for the levels of the main factors (V, N and M) and their interactions, the EM claimed superiority.

Table 4. 4:Estimated Means and Standard Errors For Plant Leave Based on the SIMDS and the Three Methods (PD, MI AND EM).

MEANS					STANDARD ERROR			
REPS	SIMDS	PD	MI	EM	SIMDS	PD	MI	EM
1	32.5	33.12	33.742	33.986	1.302	1.768	1.355	1.331
2	37.417	37.066	36.842	36.615	1.302	1.704	1.355	1.331
3	34.75	34.219	34.125	33.914	1.302	1.836	1.355	1.331
V								
1	33.944	33.983	34.139	34.005	1.063	1.406	1.106	1.087
2	35.833	35.621	35.667	35.672	1.063	1.392	1.106	1.087
N								
0	33.778	32.985	33.633	33.253	1.503	1.935	1.564	1.537
1	34.222	34.853	34.822	34.796	1.503	1.936	1.564	1.537
2	38.222	38.464	37.644	37.876	1.503	2.013	1.564	1.537
3	33.333	32.905	33.511	33.428	1.503	2.062	1.564	1.537
M								
0	35.333	35.806	35.283	35.429	1.302	1.755	1.355	1.331
1	34.208	33.219	34.15	33.991	1.302	1.717	1.355	1.331
2	35.125	35.38	35.275	35.095	1.302	1.668	1.355	1.331
V*N								
0 1	31.222	30.736	31.444	31.006	2.126	2.64	2.212	2.174
0 2	36.333	35.234	35.822	35.5	2.126	2.82	2.212	2.174
1 1	34.111	34.153	34.356	34.035	2.126	2.821	2.212	2.174
1 2	34.333	35.553	35.289	35.557	2.126	2.639	2.212	2.174
2 1	37.556	37.900	37.111	37.509	2.126	2.82	2.212	2.174
2 2	38.889	39.028	38.178	38.242	2.126	2.854	2.212	2.174
3 1	32.889	33.141	33.644	33.469	2.126	3.010	2.212	2.174

3 2	33.778	32.669	33.378	33.387	2.126	2.82	2.212	2.174
V*M	SIMDS	PD	MI	EM	SIMDS	PD	MI	EM
0 1	33.583	34.198	34.367	34.218	1.841	2.363	1.916	1.883
0 2	37.083	37.415	36.2	36.64	1.841	2.587	1.916	1.883
1 1	33.333	32.417	33.25	33.139	1.841	2.471	1.916	1.883
1 2	35.083	34.021	35.05	34.842	1.841	2.386	1.916	1.883
2 1	34.917	35.333	34.8	34.657	1.841	2.471	1.916	1.883
2 2	35.333	35.427	35.75	35.533	1.841	2.243	1.916	1.883
N*M								
0 0	33.500	32.913	32.833	32.914	2.603	3.352	2.710	2.662
0 1	31.833	34.159	34.300	34.686	2.603	3.723	2.710	2.662
0 2	41.000	41.021	39.100	39.63	2.603	3.353	2.710	2.662
0 3	35.000	35.132	34.900	34.486	2.603	3.72	2.710	2.662
1 0	35.000	33.209	35.233	34.012	2.603	3.726	2.710	2.662
1 1	34.833	34.833	34.833	34.833	2.603	2.98	2.710	2.662
0 2	32.667	32.663	33.7	34.164	2.603	3.352	2.710	2.662
0 3	34.333	32.17	32.833	32.954	2.603	3.668	2.710	2.662
2 0	32.833	32.833	32.833	32.833	2.603	2.98	2.710	2.662
2 1	36.000	35.566	35.333	34.87	2.603	3.351	2.710	2.662
2 2	41.000	41.709	40.133	39.833	2.603	3.726	2.710	2.662
2 3	30.667	31.413	32.800	32.845	2.603	3.352	2.710	2.662
V*N*M								
0 0 1	29.667	29.667	29.667	29.667	3.682	4.214	3.832	3.765
0 0 2	37.333	36.159	36.000	36.161	3.682	5.213	3.832	3.765
0 1 1	31.667	32.659	33.733	33.701	3.682	5.213	3.832	3.765
0 1 2	32.000	35.659	34.867	35.671	3.682	5.213	3.832	3.765
0 2 1	38.333	38.333	38.333	38.333	3.682	4.214	3.832	3.765

V*N*M								
0 2 2	43.667	43.709	39.867	40.928	3.682	5.216	3.832	3.765
0 3 1	34.667	36.132	35.733	35.172	3.682	5.211	3.832	3.765
0 3 2	35.333	34.132	34.067	33.8	3.682	5.211	3.832	3.765
1 0 1	34.667	33.209	35.333	34.018	3.682	5.216	3.832	3.765
1 0 2	35.333	33.209	35.133	34.006	3.682	5.216	3.832	3.765
1 1 1	33.667	33.667	33.667	33.667	3.682	4.214	3.832	3.765
1 1 2	36.000	36.000	36.000	36.000	3.682	4.214	3.832	3.765
1 2 1	27.667	27.659	29.733	30.661	3.682	5.213	3.832	3.765
1 2 2	37.667	37.667	37.667	37.667	3.682	4.214	3.832	3.765
1 3 1	37.333	35.132	34.267	34.212	3.682	5.211	3.832	3.765
1 3 2	31.333	29.209	31.400	31.696	3.682	5.216	3.832	3.765
2 0 1	29.333	29.333	29.333	29.333	3.682	4.214	3.832	3.765
2 0 2	36.333	36.333	36.333	36.333	3.682	4.214	3.832	3.765
2 1 1	37	36.132	35.667	34.739	3.682	5.211	3.832	3.765
2 1 2	35.000	35.000	35.000	35.000	3.682	4.214	3.832	3.765
2 2 1	46.667	47.709	43.267	43.533	3.682	5.216	3.832	3.765
2 2 2	35.333	35.709	37.000	36.133	3.682	5.216	3.832	3.765
2 3 1	26.667	28.159	30.933	31.024	3.682	5.213	3.832	3.765
2 3 2	34.667	34.667	34.667	34.667	3.682	4.214	3.832	3.765

Table 4.4 above present the means and standard errors of the three factors levels on plant leaves for the simulated data set before data points are randomly selected as missing which is compared with means and standard errors obtained after data points have been randomly selected as missing and have been estimated by PD, MI and EM. It was observed by inspection that the means obtained by the three methods for each of the levels of the three factors compared with the

means of the simulated data are similar with slight departure. Generally, the EM estimates are more consistent than the PD and MI. Observing the standard errors obtained from the three methods, the EM proved to be more superior over the PD and MI since its standard error values are approximately similar to that of the simulated data set while the PD and MI standard error values are very high or very low compared to the simulated data set standard error values. For instance the mean standard error for V*N*M for the simulated data set is approximately 3.682 while that of PD is 4.839, MI is 3.832 and EM is 3.765 this is an evidence of the EM to be the best method for estimating missing data point(s) from a data set since its value of 3.765 is approximately close to the simulated data value of 3.682 while PD and MI values are greater. Likewise, same evidence was observed for the levels of the main factors (V, N and M) and their interactions, the EM claimed superiority.

Table 4. 5: Estimated Means and Standard Errors for Plant Branches Based on the RLDS and the Three Methods (PD, MI AND EM).

MEANS					STANDARD ERRORS			
REPS	RLDS	PD	MI	EM	RLDS	PD	MI	EM
1	9.708	9.573	9.275	9.74	0.432	0.613	0.396	0.424
2	8.583	8.843	8.933	8.680	0.432	0.598	0.396	0.424
3	6.750	6.694	7.425	7.185	0.432	0.672	0.396	0.424
V								
1	9.139	8.888	8.872	9.000	0.352	0.485	0.323	0.346
2	7.556	7.852	8.217	8.070	0.352	0.520	0.323	0.346
N								
0	7.111	7.411	7.133	7.163	0.498	0.697	0.457	0.490
1	7.722	7.743	8.311	7.888	0.498	0.792	0.457	0.490
2	8.778	8.793	9.233	9.047	0.498	0.637	0.457	0.490
3	9.778	9.532	9.500	10.042	0.498	0.717	0.457	0.490
M								
0	7.083	7.116	7.633	7.149	0.432	0.521	0.396	0.424
1	8.875	8.841	9.000	8.895	0.432	0.615	0.396	0.424
2	9.083	9.153	9.000	9.560	0.432	0.699	0.396	0.424
V*N								
0 1	7.889	8.113	7.600	7.812	0.705	0.928	0.647	0.693
0 2	6.333	6.71	6.667	6.513	0.705	1.048	0.647	0.693
1 1	8.667	7.687	9.000	8.494	0.705	1.142	0.647	0.693
1 2	6.778	7.799	7.622	7.281	0.705	1.087	0.647	0.693
2 1	10.000	10.11	9.956	10.114	0.705	0.869	0.647	0.693
2 2	7.556	7.477	8.511	7.981	0.705	0.928	0.647	0.693
3 1	10.000	9.643	8.933	9.580	0.705	0.928	0.647	0.693

3 2	9.556	9.421	10.067	10.504	0.705	1.087	0.647	0.693
V*M	RLDS	PD	MI	EM	RLDS	PD	MI	EM
0 1	7.250	6.999	7.717	7.250	0.610	0.738	0.560	0.600
0 2	6.917	7.234	7.550	7.048	0.610	0.738	0.560	0.600
1 1	9.833	9.556	9.733	9.627	0.610	0.965	0.560	0.600
1 2	7.917	8.126	8.267	8.164	0.610	0.778	0.560	0.600
2 1	10.333	10.110	9.167	10.123	0.610	0.821	0.560	0.600
2 2	7.833	8.196	8.833	8.997	0.610	1.160	0.560	0.600
N*M								
0 0	4.333	4.333	5.067	4.333	0.863	0.980	0.792	0.849
0 1	7.500	7.632	8.167	7.762	0.863	1.207	0.792	0.849
0 2	7.667	7.667	8.533	7.667	0.863	0.98	0.792	0.849
0 3	8.833	8.833	8.767	8.833	0.863	0.98	0.792	0.849
1 0	8.667	8.634	8.2	8.478	0.863	1.104	0.792	0.849
1 1	7.667	6.898	8.167	7.297	0.863	1.412	0.792	0.849
0 2	9.167	9.382	10.167	9.417	0.863	1.207	0.792	0.849
0 3	10.000	10.449	9.467	10.39	0.863	1.208	0.792	0.849
2 0	8.333	9.267	8.133	8.676	0.863	1.51	0.792	0.849
2 1	8.000	8.699	8.600	8.604	0.863	1.476	0.792	0.849
2 2	9.500	9.331	9.000	10.058	0.863	1.104	0.792	0.849
2 3	10.500	9.314	10.267	10.904	0.863	1.508	0.792	0.849
V*N*M								
0 0 1	5.000	5.000	5.467	5.000	1.221	1.386	1.120	1.200
0 0 2	3.667	3.667	4.667	3.667	1.221	1.386	1.120	1.200
0 1 1	7.667	6.662	8.333	7.665	1.221	1.718	1.120	1.200
0 1 2	7.333	8.602	8.000	7.859	1.221	1.719	1.120	1.200
0 2 1	8.667	8.667	9.067	8.667	1.221	1.386	1.120	1.200

V*N*M	RLDS	PD	MI	EM	RLDS	PD	MI	EM
0 2 2	6.667	6.667	8.000	6.667	1.221	1.386	1.120	1.200
0 3 1	7.667	7.667	8.000	7.667	1.221	1.386	1.120	1.200
0 3 2	10.000	10.000	9.533	10.000	1.221	1.386	1.120	1.200
1 0 1	9.667	9.602	9.133	9.289	1.221	1.719	1.120	1.200
1 0 2	7.667	7.667	7.267	7.667	1.221	1.386	1.120	1.200
1 1 1	9.333	7.797	9.733	8.594	1.221	2.46	1.120	1.200
1 1 2	6.000	6.000	6.600	6.000	1.221	1.386	1.120	1.200
1 2 1	9.333	9.662	10.200	9.674	1.221	1.718	1.120	1.200
1 2 2	9.000	9.102	10.133	9.161	1.221	1.719	1.120	1.200
1 3 1	11.000	11.162	9.867	10.952	1.221	1.718	1.120	1.200
1 3 2	9.000	9.736	9.067	9.827	1.221	1.716	1.120	1.200
2 0 1	9.000	9.736	8.200	9.147	1.221	1.716	1.120	1.200
2 0 2	7.667	8.797	8.067	8.205	1.221	2.460	1.120	1.200
2 1 1	9.000	8.602	8.933	9.224	1.221	1.719	1.120	1.200
2 1 2	7.000	8.797	8.267	7.984	1.221	2.460	1.120	1.200
2 2 1	12	12	10.600	12.000	1.221	1.386	1.120	1.200
2 2 2	7	6.662	7.40	8.115	1.221	1.718	1.120	1.200
2 3 1	11.333	10.102	8.933	10.122	1.221	1.719	1.12	1.2
2 3 2	9.667	8.527	11.6	11.685	1.221	2.451	1.12	1.2

Table 4.5 above present the means and standard errors of the three factors levels on plant branches for the real-life data set before data points are randomly selected as missing which is compared with means and standard errors obtained after data points have been randomly selected as missing and have been estimated by PD, MI and EM. It was observed by inspection that the means obtained by the three methods for each of the levels of the three factors compared with the

means of the real-life data are similar with slight departure. Generally, the EM estimates are more consistent than the PD and MI. Observing the standard errors obtained from the three methods, the EM proved to be more superior over the PD and MI since its standard error values are approximately similar to that of the real life-data set while the PD and MI standard error values are very high or very low compared to the real-life data set standard error values. For instance the mean standard error for V*N*M for the real-life data set is approximately 1.221 while that of PD is 1.717, MI is 1.12 and EM is 1.2 this is an evidence of the EM to be the best method for estimating missing data point(s) from a data set since its value of 1.2 is approximately similar to the real-life data value of 1.221 while PD and MI values are greater than and less than the real-life data value of 1.221 respectively. Likewise, same evidence was observed for the levels of the main factors (V, N and M) and their interactions, the EM claimed superiority.

Table 4. 6: Estimated Means and Standard Errors for Plant Branches Based on the SIMDS and the Three Methods (PD, MI AND EM).

MEANS OBTAINED					STANDARD ERROR OBTAINED			
REPS	SIMDS	PD	MI	EM	SIMDS	PD	MI	EM
1	8.125	8.219	8.183	8.196	0.247	0.287	0.203	0.199
2	8.25	8.284	8.2	8.159	0.247	0.28	0.203	0.199
3	8.375	8.222	8.067	8.081	0.247	0.315	0.203	0.199
V								
1	8.028	8.109	8.117	8.06	0.202	0.227	0.166	0.163
2	8.472	8.374	8.183	8.231	0.202	0.243	0.166	0.163
N								
0	7.889	7.978	7.844	7.868	0.285	0.326	0.234	0.230
1	8.444	8.308	8.289	8.217	0.285	0.371	0.234	0.230
2	8.389	8.467	8.389	8.414	0.285	0.298	0.234	0.230
3	8.278	8.215	8.078	8.083	0.285	0.336	0.234	0.230
M								
0	8.125	7.956	7.942	7.97	0.247	0.244	0.203	0.199
1	8.375	8.396	8.317	8.248	0.247	0.288	0.203	0.199
2	8.25	8.374	8.192	8.219	0.247	0.327	0.203	0.199
V*N								
0 1	7.778	7.948	7.978	7.919	0.403	0.435	0.331	0.325
0 2	8.000	8.007	7.711	7.818	0.403	0.491	0.331	0.325
1 1	8.444	8.334	8.378	8.164	0.403	0.535	0.331	0.325
1 2	8.444	8.282	8.2	8.27	0.403	0.509	0.331	0.325
2 1	8.667	8.886	8.644	8.687	0.403	0.407	0.331	0.325
2 2	8.111	8.048	8.133	8.141	0.403	0.434	0.331	0.325
3 1	7.222	7.271	7.467	7.471	0.403	0.434	0.331	0.325

3 2	9.333	9.16	8.689	8.695	0.403	0.509	0.331	0.325
V*M	SIMDS	PD	MI	EM	SIMDS	PD	MI	EM
0 1	7.75	7.456	7.517	7.509	0.349	0.346	0.287	0.281
0 2	8.5	8.456	8.367	8.43	0.349	0.346	0.287	0.281
1 1	8.5	8.623	8.5	8.39	0.349	0.452	0.287	0.281
1 2	8.25	8.169	8.133	8.105	0.349	0.364	0.287	0.281
2 1	7.833	8.25	8.333	8.281	0.349	0.384	0.287	0.281
2 2	8.667	8.498	8.05	8.157	0.349	0.543	0.287	0.281
N*M								
0 0	7.5	7.5	7.5	7.5	0.494	0.459	0.406	0.398
0 1	8.667	7.989	7.933	8.046	0.494	0.565	0.406	0.398
0 2	8.167	8.167	8.167	8.167	0.494	0.459	0.406	0.398
0 3	8.167	8.167	8.167	8.167	0.494	0.459	0.406	0.398
1 0	7.5	7.411	7.533	7.526	0.494	0.517	0.406	0.398
1 1	8.5	8.678	8.6	8.375	0.494	0.661	0.406	0.398
0 2	8.667	9.239	8.833	8.89	0.494	0.565	0.406	0.398
0 3	8.833	8.256	8.3	8.2	0.494	0.566	0.406	0.398
2 0	8.667	9.022	8.5	8.579	0.494	0.707	0.406	0.398
2 1	8.167	8.256	8.333	8.229	0.494	0.691	0.406	0.398
2 2	8.333	7.995	8.167	8.186	0.494	0.517	0.406	0.398
2 3	7.833	8.223	7.767	7.881	0.494	0.706	0.406	0.398
V*N*M								
0 0 1	7.333	7.333	7.333	7.333	0.698	0.649	0.574	0.563
0 0 2	7.667	7.667	7.667	7.667	0.698	0.649	0.574	0.563
0 1 1	8.667	7.49	7.733	7.705	0.698	0.804	0.574	0.563
0 1 2	8.667	8.489	8.133	8.387	0.698	0.805	0.574	0.563
0 2 1	7.667	7.667	7.667	7.667	0.698	0.649	0.574	0.563

0 2 2	8.667	8.667	8.667	8.667	0.698	0.649	0.574	0.563
V*N*M	SIMDS	PD	MI	EM	SIMDS	PD	MI	EM
0 3 1	7.333	7.333	7.333	7.333	0.698	0.649	0.574	0.563
0 3 2	9	9	9	9	0.698	0.649	0.574	0.563
1 0 1	7.667	7.489	7.733	7.718	0.698	0.805	0.574	0.563
1 0 2	7.333	7.333	7.333	7.333	0.698	0.649	0.574	0.563
1 1 1	8.667	9.022	8.867	8.417	0.698	1.152	0.574	0.563
1 1 2	8.333	8.333	8.333	8.333	0.698	0.649	0.574	0.563
1 2 1	9.333	9.99	9.267	9.395	0.698	0.804	0.574	0.563
1 2 2	8	8.489	8.4	8.385	0.698	0.805	0.574	0.563
1 3 1	8.333	7.99	8.133	8.031	0.698	0.804	0.574	0.563
1 3 2	9.333	8.521	8.467	8.369	0.698	0.803	0.574	0.563
2 0 1	8.333	9.021	8.867	8.706	0.698	0.803	0.574	0.563
2 0 2	9	9.022	8.133	8.453	0.698	1.152	0.574	0.563
2 1 1	8	8.489	8.533	8.37	0.698	0.805	0.574	0.563
2 1 2	8.333	8.022	8.133	8.088	0.698	1.152	0.574	0.563
2 2 1	9	9	9	9	0.698	0.649	0.574	0.563
2 2 2	7.667	6.99	7.333	7.372	0.698	0.804	0.574	0.563
2 3 1	6	6.489	6.933	7.047	0.698	0.805	0.574	0.563
2 3 2	9.667	9.958	8.6	8.715	0.698	1.148	0.574	0.563

Table 4.6 above present the means and standard errors of the three factors levels on plant branches for the simulated data set before data points are randomly selected as missing which is compared with means and standard errors obtained after data points have been randomly selected as missing and have been estimated by PD, MI and EM. It was observed by inspection that the means obtained by the three methods for each of the levels of the three factors compared with the

means of the simulated data are similar with slight departure. Generally, the EM estimates are more consistent than the PD and MI. Observing the standard errors obtained from the three methods, the MI and EM proved to be more superior over the PD since its standard error values are approximately similar to that of the real simulated data set while the PD standard error values very high compared to the simulated data set standard error values. For instance the mean standard error for V*N*M for the simulated data set is approximately 0.698 while that of PD is 0.804, MI is 0.574 and EM is 0.563 this is an evidence that MI and EM are best methods for estimating missing data point(s) from a data set since their values of 0.574 and 0.563 are approximately close to the simulated data value of 0.698 while PD value is greater than the simulated data value of 0.698 respectively. Likewise, same evidence was observed for the levels of the main factors (V, N and M) and their interactions, the MI and EM claimed superiority where the MI values are consistently more approximately closer to that of the simulated data values.

Table 4. 7:P-Values Obtained with the RLDS, Simds, Pw, MI And EM For Plant Heights.

$\alpha = 0.05$

SV	RLDS	PD	MI	EM	SIMDS	PD	MI	EM
REPS	0.068	0.11	0.174	0.092	0.199	0.193	0.207	0.157
V	0.066	0.155	0.391	0.23	0.563	0.901	0.645	0.722
N	0.0001	0.003	0.002	0.002	0.199	0.394	0.449	0.451
M	0.0001	0.0001	0.0001	0.0001	0.066	0.045	0.018	0.018
V*N	0.56	0.545	0.051	0.169	0.588	0.279	0.439	0.309
V*M	0.257	0.317	0.939	0.53	0.576	0.609	0.778	0.716
N*M	0.004	0.009	0.025	0.018	0.101	0.127	0.182	0.072
V*N*M	0.038	0.058	0.024	0.031	0.902	0.89	0.943	0.93

Table 4.7 above present the P-values obtained from the analysis of variance (ANOVA) for plant heights which is presented in the appendix for the RLDS and SIMDS compared with the P-values obtained with PD, MI and EM estimates for the missing values which were randomly selected and declared as missing for the RLDS and SIMDS. We observed that at five percent significance level, the P-values for REPS, V, V*N and V*M for the RLDS were not significant and similarly for PD, MI and EM. However, N, M, N*M and V*N*M were significant for the RLDS at five percent significance level and similarly for MI and EM but for PD its P-values of 0.003, 0.0001 and 0.009 were significant for N, M and N*M respectively though not significant for V*N*M with a P-value of 0.058. This is a further evidence of weakness of the PD method for missing data estimation. Hence, for the simulated data P-values we found that none of the factors and their combinations were significant at five percent significance level but, for the PD, MI and EM estimates we found that factor M was significant with their P-values of 0.045, 0.018 and 0.018 respectively.

**Table 4. 8:P-Values Obtained with the RLDS, SIMDS, PW, MI and EM for Plant Leaves
 $\alpha = 0.05$**

SV	RLDS	PD	MI	EM	SIMDS	PD	MI	EM
REPS	0.002	0.003	0.005	0.001	0.036	0.274	0.222	0.027
V	0.018	0.022	0.007	0.015	0.215	0.415	0.334	0.284
N	0.0001	0.0001	0.0001	0.0001	0.096	0.189	0.225	0.137
M	0.0001	0.0001	0.001	0.0001	0.81	0.53	0.794	0.728
V*N	0.0001	0.0001	0.004	0.0001	0.661	0.843	0.75	0.739
V*M	0.872	0.956	0.731	0.996	0.705	0.813	0.966	0.919
N*M	0.617	0.316	0.92	0.523	0.162	0.749	0.706	0.848
V*N*M	0.224	0.024	0.031	0.367	0.053	0.305	0.497	0.055

Table 4.8 above present the P-values obtained from the analysis of variance (ANOVA) for plant leaves which is presented in the appendix for the RLDS and SIMDS compared with the P-values obtained with PD, MI and EM estimates for the missing values which were randomly selected and declared as missing for the RLDS and SIMDS. We observed that at five percent significance level, the P-values for REPS, V, N, M and V*N for the RLDS were significant and similarly for PD, MI and EM. However, V*M, N*M and V*N*M were not significant for the RLDS at five percent significance level. Similarly, the P-values for EM is also not significant for V*M, N*M and V*N*M but for the PD and MI their P-values were not significant for V*M and N*M but significant for V*N*M. this is a further evidence of weakness of the PD and MI method for missing data estimation. Hence, for the simulated data P-values we found that none of the factors and their combinations were significant at five percent significance level but, for the REPS it was significant with P-value of 0.036 and similarly for EM. While PD and MI were not significant for all the factors and their combinations. This is an evidence for the EM to be the best method for missing data estimates.

**Table 4. 9:P-Values Obtained With the RLDS, SIMDS, PW, MI and EM for Plant Branches
 $\alpha = 0.05$**

SV	RLDS	PD	MI	EM	SIMDS	PD	MI	EM
REPS	0.0001	0.011	0.004	0.0001	0.775	0.983	0.88	0.917
V	0.003	0.155	0.158	0.064	0.126	0.431	0.777	0.461
N	0.002	0.167	0.003	0.001	0.515	0.74	0.375	0.402
M	0.003	0.041	0.026	0.001	0.775	0.43	0.42	0.559
V*N	0.548	0.492	0.168	0.166	0.011	0.039	0.051	0.023
V*M	0.196	0.41	0.456	0.557	0.236	0.172	0.071	0.078
N*M	0.313	0.369	0.383	0.25	0.396	0.401	0.412	0.388
V*N*M	0.448	0.766	0.502	0.611	0.497	0.475	0.735	0.717

Table 4.9 above present the P-values obtained from the analysis of variance (ANOVA) for plant branches which is presented in the appendix for the RLDS and SIMDS compared with the P-values obtained with PD, MI and EM estimates for the missing values which were randomly selected and declared as missing for the RLDS and SIMDS. We observed that at five percent significance level, the P-value for REPS, V, N and M for the RLDS were significant. Similarly, the P-values obtained with PD, MI and EM for REPS and M were significant. While their P-values of 0.155, 0.158 and 0.064 respectively for V were not significant at five percent significance level and it was observed that PD P-values of 0.167 were not significant but MI and EM P-values of 0.003 and 0.001 respectively were significant at five percent significance level which correspond to the same conclusion as for the RLDS. This shows PD weakness in estimating missing values. Hence for the simulated data set P-values, we found that the REPS and all the factors and three out of the four factors combinations were not significant at five percent significance level.

CHAPTER FIVE

SUMMARY, CONCLUSION AND RECOMMENDATION

5.1 INTRODUCTION

This chapter presents the summary, conclusion and recommendations based on the results obtained in chapter four.

5.2 SUMMARY AND CONCLUSION

During the last decade, the missing data treatments reported in Yiran and Chao-Ying (2013) have shown much improvement in terms of decreased use of ad hoc methods (e.g,LD and PD) and increased use of principled methods (e.g., FIML, EM, and MI). Yet several research practices still persisted including, not explicitly acknowledging the presence of missing data, not describing the approach used in dealing with missing data, not testing assumptions assumed. The three methods used were illustrated with a replicated factorial design data set using IBM SPSS and MINITAB for the simulation. The performance of the three missing data methods was compared with that of the complete data set and simulated data set in terms of bias, standard error and P-values.

The main objective of this research work is to determine the most efficient method (Pairwise, Multiple Imputation and Expectation Maximization) of estimating missing data using their estimated means, standard error and Anova results for both the real life data set and simulated data set. From the results of the analysis in chapter four of this research work we found out and generally concluded that the expectation maximization (EM) is the best method for estimating missing values compared to pairwise deletion (PD) and multiple imputation (MI). This conclusion is backed by rigorous inspection of their estimated means, standard errors and P-values from the Anova results. The EM estimated consistent and more precise means for each

factor levels and at overall means similar to that of the real-life data sets and simulated data sets. Likewise, its standard error values were approximately closer and in some cases as we observed they were similar and equal to that of the real-life data sets and simulated data sets.

Results showed that PD method is the worst of all the three methods. Its estimated mean and standard error values are inconsistent and far away (greater than) from those of the real-life data sets and simulated data sets (that is to say it overestimates compared to the MI and EM methods). The P-values for the three methods were also considered and EM method was found to be more superior over the PD and MI methods and again it was observed that the PD method performed poorly because in cases where a particular factor is not significant from the real data or simulated data sets it claimed significance or vice versa. The same was observed for MI and in some cases their P-values obtained were larger than that obtained from the real-life data and simulated data sets but the EM produced P-values that were in line with the conclusion from the real-life data and simulated data sets. These evidences lead to the general conclusion that the expectation maximization (EM) method for estimating missing data in this is more appropriate in terms of consistency and precision.

5.3 RECOMMENDATION

Based on the researcher observation, the following are therefor recommended for selecting the best method for estimating missing values for a data set.

Quality of research will be enhanced if

1. researchers explicitly acknowledge missing data problems and the conditions under which they occurred
2. Principled methods are employed to handle missing

3. The use of expectation maximization (EM) as the best method for estimating missing values, is recommended, since from the analysis with close inspection of its results we found that it's the best due to its consistency and approximate similar values as that from the real-life data and simulated data sets.

5.4 CONTRIBUTION TO KNOWLEDGE

The contributions to knowledge from this research work are as follows;

This research demonstrates three principled missing data methods that are applicable for a variety of research contexts in experimental design. Before applying any of the principled methods, one should make every effort to prevent missing data from occurring. Toward this end, the missing data rate should be kept at minimum by designing and implementing data collection carefully. When missing data are inevitable, one needs to closely examine the missing data mechanism, missing rate, missing pattern, and the data distribution before deciding on a suitable missing data method. When implementing a missing data method, a researcher should be mindful of issues related to its proper implementation, such as, statistical assumptions, the specification of the imputation/estimation model, a suitable number of imputations, and criteria of convergence.

5.5 AREA FOR FURTHER RESEARCH

1. Application of Artificial neural network method in estimating missing values in replicated factorial design.
2. Estimating missing values in replicated factorial design using full information maximum likelihood method.

3. Comparing the efficient of Artificial neural network method, full information maximum likelihood and Expectation Maximization in estimating missing values in replicated factorial design.

REFERENCES

- Adam, D. and Jyoti, T. S. (2010). *Statistical Power Analysis with Missing Data*. Routledge Academic; 1st edition.
- Alan, A. O. (2005). Working With Missing Values. *Journal of Marriage and Family* 67: 1012–1028.
- Allison, P. D. (2002). Missing data. Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-136. Thousand Oaks, CA: Sage.
- Anderson, A. B., Basilevsky, A., and Hum, D. P. J. (1983). Missing Data: A Review of the Literature, in *Handbook of Survey Research*, eds. P. H. Rossi, J. D. Wright, and A. Anderson, New York: Academic Press, pp. 415-492.
- Arbuckle, J. L. (1995). *Amos for Windows. Analysis of moment structures. Version 3.5*. Chicago: SmallWaters Corp. <http://www.rpajournal.com/dev/wp-content/uploads/2012/05/A1.pdf>.
- Baraldi, A.N., and Enders, C.K. (2010). An introduction to modern missing data analyses. *Journal of School of Psychology*, 48:5 – 37.
- Becker, W. E., and Powers, J. (2001). Student performance, attrition, and class size given missing student data. Manuscript submitted for publication. <http://www.indiana.edu/~leeehman/mdpaper.pdf>.
- Becker, W. E., & Walstad, W. B. (1990). Data loss from pretest to posttest as a sample selection problem. *The Review of Economics and Statistics*, 72 (1):184-188.
- Bennett, D. A. (2001). How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health*, 25:464–469.
- Bernaards, C. A. , and Sijtsma, K. (1999). Factor analysis of multidimensional polytomous item response data from ignorable item nonresponse. *Multivariate Behavioral Research* 34: 277-314.
- Buu, A. (1999). Analysis of longitudinal data with missing values: A methodological comparison. Unpublished doctoral dissertation, Indiana University. <http://www.indiana.edu/~leeehman/mdpaper.pdf>.
- Buuren .V. S. (2007). Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification. *Statistical Methods In Medical Research* , 16(3): 219-242.
- Buuren, V. S., Boshuizen .H.C, Knook, D.L. (2006). Multiple Imputation of Missing Blood Pressure Covariates In Survival Analysis. *Statistics In Medicine*, 8: 681-694.

- Buruuen. V. S., Boshuizen , C. H, and Knook, D. L . (1999). Multiple Imputation Of Missing Blood Pressure Covariates in Survival Analysis. *Statistics in Medicine*, 18 (6):681–694.
- Cohen, J, and Cohen, P. (1983). Applied multiple regression/ correlation analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Cool, A. L. (2000). A review of methods for dealing with missing data. Paper presented at the *Annual Meeting of the Southwest Educational Research Association*, Dallas, TX. (ERIC Document Reproduction Service No. ED 438 31.
- Craig, K. E. and Deborah, L. B. (2001). The Relative Performance of Full Information Maximum Likelihood Estimation for Missing Data in Structural Equation Models. *Educational Psychology Papers and Publications*, 8:430-457.
- David, A. P. (2007). Estimating missing values from the general social survey: An application of multiple imputation. *Social Science Quarterly*, 88(2): 573-584.
- David, C. H. (2008). The analysis of missing data. In Outhwaite, W. and Turner, S. *Handbook of Social Science Methodology*. London: Sage.
- David, R. J. and Rebekah, Y. (2011). Toward Best Practices in Analysing Datasets With Missing Data: Comparisons and Recommendations. *Journal of Marriage and Family*, 73 (5):926-945.
- De Boeck, P. and Wilson, M. (Eds.) 2004. Explanatory item response models. A generalized linear and Nonlinear approach. New York: Springer.
- Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–39.
- Ender, C. K. (2004). The Impact of missing Data on Sample Reliability Estimates. Implications for Reliability Reporting ractices. *Educational and Psychological Measurement*, 64:419–436.
- Enders, C. K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling*, 8:128-141.
- Enders, C. K. and Bandalos, D. L. (2001), “The Relative Performance of Full Information Maximum Likelihood Estimation for Missing Data in Structural Equation Models,” *Structural Equation Modeling*, 8:430–457.
- Finch, W. H. (2008). Imputation Methods for Missing Categorical Questionnaire. *Journal of Data Science*, 8:361-378.

- Glasser, M. (1964). Linear regression with missing observations among the independent variables. *Journal of the American Statistical Association*, 59:834-844.
- Graham, J. W, Hofer .S. M, Donaldson .S. I, MacKinnon .D. P, Schafer .J.L. (1997). Analysis with missing data in prevention research. In *The Science of Prevention: Methodological Advances from Alcohol and Substance Abuse Research*, ed. K Bryant, M Windle, S West,
- Graham, J. W, Cumsille, P. E., and Elek-Fisk, E. (2003). Methods for handling missing data. In *Research Methods in Psychology*, ed. JA Schinka, WF Velicer, pp. 87–114. Volume 2 of *Handbook of Psychology*, ed. IB Weiner. New York: Wiley
- Graham, J. W. and Schafer, J. L. (1999). On the performance of multiple imputation for multivariate data with small sample size. In *Statistical Strategies for Small Sample Research*, ed. R Hoyle, 1:1–29. Thousand Oaks, CA: Sage
- Graham, J. W., & Hofer, S. M. (2000). Multiple imputation in multivariate research. In T.D. Little, K.U. Schnabel, and J. Baumert (Eds.), *Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples*. pp. 201-218. Mahwah, NJ: Erlbaum.
- Heitjan, D. F., and Little, R. J. A. (1991). Multiple Imputation in the Fatal Accident Reporting System. *Applied Statistics*, 40: 13-29.
- Holt, D. (1997). Missing data and nonresponse. In J. P. Keeve (Ed.) *Educational research, methodology, and measurement: An international handbook* (2ed.). New York: Elsevier Science Ltd.
- Jeffrey, C. W (2003). Multiple Imputation For Missing Data. What is it and how can I use? *Annual meeting of the American Educational Research Association*, Chicago, IL. pp. 2-16.
- John, L. J. (1966). An Alternate Approach to Missing Value Estimation. *The American Statistician*, 20(5): 27-29.
- Kim, J. O. and Curry, J. (1977). The treatment of missing data in multivariate analysis. *Sociological Methods & Research*, 6 (2):215-240.
- Little, R. J. and Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.
- Little, R. J, Roderick J. A. and Rubin, B. D. (2002). *Statistical Analysis with Missing Data*. Second edition, New York: John Wiley and Sons, Inc.
- Little, R. J. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of American Statistical Association*, 90:1112–1121.

- McLachlan, G. J., and Krishnan, T. (1997). *The EM algorithm and extensions*. New York: Wiley.
- Muthén, B., Kaplan, D., and Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, 52:431–462.
- Nancy, L. G. (2004). *Advances in Clinical trial Biostatistics*. National Institute of health. Bethesda, Maryland, U.S.A.
http://www.statsinfoindia.weebly.com/uploads/7/3/9/1/7391294/biostatisti_cs.pdf.
- Neale, M. C. (1994). *Statistical modeling* (2nd Edition). Department of Psychiatry: Medical College of Virginia.
- Newsom, R. B. (2010). Frequentist q-values for multiple test procedures. *The Stata Journal* 10(4): 568-584.
- Peugh, J. L., and Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, 74(4):525 – 556.
- Pigott, T. D. (2001). A review of methods for missing data. *Educational Research and Evaluation*, 7:353–383.
- Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J. and Solenberger, P . (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology*, 27(1): 85-95.
- Raghunathan, T. E. (2004). What Do We With Missing Data? Some Options for Analysis of Incomplete Data.” *Annual Review of Public Health*, 25:99– 117.
- Ravindra S. L, Erandathie .L and Keith .P (2006). Comparison of missing value imputation methods for crop yield data. *Environmetrics*, 17(4): 339-349.
- Raymond, M. R., and Roberts, D. M. (1987). A comparison of methods for incomplete data in selection research. *Educational and Psychological Measurement*, 47:13-26.
- Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology*, 47:537–570.
- Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, 63:581–592.
- Rubin, D. B. (1978), Multiple imputations in sample surveys A phenomenological Bayesian approach to nonresponse. In Proceedings of the Survey Research Methods Section, Alexandria, VA, *American Statistical Association*, pp. 20–34.

- Rubin, D. B. (1996), "Multiple Imputation After 18+Years". *Journal of the American Statistical Association*, 91:473–489.
- Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley and Sons, Inc.
- Rufus, L. C. (2006). Solutions for Missing Data in Structural Equation Modeling. *Research and Practice in Assessment*. 1(1):1-6.
- Schafer, J. L. and Olsen, M. K. (1998) Multiple imputation for multivariate missing data problems: A data analyst's perspective. *Multivariate Behavior Research*, 33:545-571.
- Schafer, J. L. and Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2):147 – 177.
- Schafer, J. L., and Maren K. O. (1998). "Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective" *Multivariate Behavioral Research*, 33:545-571.
- Schafer, J. L. (1997) *Analysis of incomplete multivariate data*. Chapman and Hall, London. Book No. 72, Chapman and Hall series Monographs on Statistics and Applied Probability
- Schenker, N., Taylor, J. M. (1996) . Partially Parametric Techniques for Multiple Imputations. *Computational Statistics and Data Analysis*, 22(4):425–446.
- Shireley .D, Stanley, W. and Daniel .C. (2004). *Statistics for Research*. (3rd Ed)John Wiley & Sons, Inc. Publication.
- Shirley, D. Stanley, W. and Daniel, C. (2003). *Statistics for Research*. Wiley Series in Probability and Statistics.
- Stoop, I., Billient,J.,Koch, A., and Fitzgerald, R. (2010) *Improving Survey Response: Lessons Learned from the European Social Survey*. Wiley
- Stubbendick, A. L. and Ibrahim, J. G. (2003). Maximum Likelihood methods for nonignorable missing responses and covariates in random effects models. *Biometrics*, 59:1140 – 1150.
- Stubbendick, A. L. and Ibrahim, J. G. (2006). Likelihood-based inference with nonignorable missing responses and covariates in models for discrete longitudinal data. *Statistica Sinica*, 16:1143 – 1167.
- Tabachnick, B. G., and Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Needham Heights, MA: Allyn and Bacon.
- Thijs, H., Molenberghs, G., Michiels, B., Verbeke, G., and Curran, D. (2002). Strategies to fit pattern-mixture models. *Biostatistics*, 3(2):245-265.

- Trivellore E. R, James M. L, John, V. H. and Peter, S. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. Statistics Canada, Catalogue No. 12001
- Vriens, M, and Melton, E. (2002). "Managing Missing Data." *Marketing Research*. 14:12-17.
- Wilkinson, L., and Task Force on Statistical Inference APA Board of Scientific Affairs.(1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54:594-604.
- Wothke, W. (1999). Longitudinal and multi- group modeling with missing data. In T.D. Little, K.U. Schnabel, and J. Baumert (Eds.) *Modeling longitudinal and multiple group data:Practical issues, applied approaches and specific examples*. Mahwah, NJ: Lawrence Erlbaum.
- Wothke, W. (2000). Longitudinal and multi-group modeling with missing data. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), *Modeling longitudinal and multiple group data: Practical issues, applied approaches and specific examples* (pp.219–240). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Wu, K., and Wu, L. (2007). Generalized linear mixed models with informative dropouts and missing covariates. *Metrika*, 66: 1 – 18.
- Yilmaz, K. Abdullay, Y. and Alamali, N. (2010). An application of expectation and Maximization, Multiple imputation and Neural Network methods for Missing Values. *World Applied Sciences Journal*. 28(9): 1281-1285.
- Yoon, Y. and Peterson, L. L. (1990). Artificial neural networks: a new technique. *In Proceedings of the conference on Trends and directions in expert systems*, pp. 417-422.

APPENDICES

APPENDIX I: REAL LIFE DATA SET

Rap	Var	Nit	Man	Hei	Lea	Bra
1	1	0	0	24	14	4
1	1	0	1	35	44	11
1	1	0	2	35	36	10
1	1	1	0	28	21	5
1	1	1	1	35	49	9
1	1	1	2	41	56	11
1	1	2	0	36	41	9
1	1	2	1	35	49	12
1	1	2	2	38	61	13
1	1	3	0	34	39	11
1	1	3	1	36	53	13
1	1	3	2	45	67	15
2	1	0	0	30	23	7
2	1	0	1	31	34	10
2	1	0	2	29	33	8
2	1	1	0	34	39	10
2	1	1	1	34	43	10
2	1	1	2	42	45	11
2	1	2	0	39	36	9
2	1	2	1	29	29	9
2	1	2	2	43	58	13
2	1	3	0	26	26	6
2	1	3	1	37	43	11
2	1	3	2	35	23	10
3	1	0	0	25	19	4
3	1	0	1	30	34	8
3	1	0	2	29	29	9
3	1	1	0	30	31	8
3	1	1	1	27	23	9
3	1	1	2	28	26	5
3	1	2	0	29	27	8
3	1	2	1	26	33	7
3	1	2	2	41	47	10
3	1	3	0	26	24	6
3	1	3	1	30	38	9
3	1	3	2	34	39	9
1	2	0	0	21	11	2
1	2	0	1	34	18	10
1	2	0	2	32	20	10
1	2	1	0	30	17	6

1	2	1	1	31	29	6
1	2	1	2	30	29	10
1	2	2	0	35	20	9
1	2	2	1	35	49	10
1	2	2	2	34	49	11
1	2	3	0	32	47	13
1	2	3	1	34	49	11
1	2	3	2	40	80	12
2	2	0	0	26	16	6
2	2	0	1	38	18	9
2	2	0	2	32	28	7
2	2	1	0	35	17	9
2	2	1	1	34	27	10
2	2	1	2	36	20	8
2	2	2	0	34	29	7
2	2	2	1	39	30	10
2	2	2	2	37	39	4
2	2	3	0	37	47	5
2	2	3	1	35	49	8
2	2	3	2	47	51	9
3	2	0	0	25	10	3
3	2	0	1	38	17	4
3	2	0	2	32	21	6
3	2	1	0	38	18	7
3	2	1	1	35	27	2
3	2	1	2	36	20	3
3	2	2	0	34	29	4
3	2	2	1	39	48	7
3	2	2	2	37	47	6
3	2	3	0	30	40	12
3	2	3	1	34	49	8
3	2	3	2	47	50	8

APPENDIX II: REAL LIFE DATA SET MISSING AT RANDOM

REP	VAR	NIT	MAN	HEI	LEA	BRA
1	1	0	0	.	14	4
1	1	0	1	35	44	.
1	1	0	2	35	36	10
1	1	1	0	28	.	5
1	1	1	1	35	49	9
1	1	1	2	41	56	.
1	1	2	0	.	41	9
1	1	2	1	35	.	12
1	1	2	2	38	61	13
1	1	3	0	34	39	11
1	1	3	1	36	53	13
1	1	3	2	45	.	.
2	1	0	0	30	23	7
2	1	0	1	31	34	10
2	1	0	2	.	33	.
2	1	1	0	.	39	10
2	1	1	1	34	43	.
2	1	1	2	42	.	11
2	1	2	0	39	36	9
2	1	2	1	29	29	9
2	1	2	2	43	58	13
2	1	3	0	26	.	6
2	1	3	1	37	.	11
2	1	3	2	.	23	10
3	1	0	0	25	19	4
3	1	0	1	30	.	8
3	1	0	2	29	29	9
3	1	1	0	30	31	.
3	1	1	1	.	23	.
3	1	1	2	28	26	5
3	1	2	0	29	27	8
3	1	2	1	26	33	.
3	1	2	2	41	.	10
3	1	3	0	26	24	6
3	1	3	1	30	38	.
3	1	3	2	34	39	9
1	2	0	0	21	.	2
1	2	0	1	.	18	10
1	2	0	2	32	20	10
1	2	1	0	30	.	.
1	2	1	1	31	29	6
1	2	1	2	30	29	10
1	2	2	0	35	20	9
1	2	2	1	35	49	.

1	2	2	2	34	49	11
1	2	3	0	32	47	13
1	2	3	1	34	49	11
1	2	3	2	.	80	.
2	2	0	0	.	16	6
2	2	0	1	38	18	9
2	2	0	2	32	28	.
2	2	1	0	35	17	9
2	2	1	1	34	27	10
2	2	1	2	36	20	.
2	2	2	0	34	29	7
2	2	2	1	.	30	10
2	2	2	2	.	39	4
2	2	3	0	37	.	5
2	2	3	1	35	49	.
2	2	3	2	47	51	9
3	2	0	0	25	10	3
3	2	0	1	38	.	4
3	2	0	2	32	21	.
3	2	1	0	38	18	7
3	2	1	1	35	27	2
3	2	1	2	36	20	.
3	2	2	0	.	.	4
3	2	2	1	39	48	7
3	2	2	2	37	.	.
3	2	3	0	30	40	12
3	2	3	1	34	.	8
3	2	3	2	47	50	

APPENDIX III: COMPLETE SIMULATED DATA SET

REP	VAR	NIT	MAN	HEI	LEA	BRA
1	1	0	0	30	32	8
1	1	0	1	31	34	8
1	1	0	2	32	28	8
1	1	1	0	32	28	7
1	1	1	1	32	31	9
1	1	1	2	31	30	7
1	1	2	0	36	42	7
1	1	2	1	31	26	10
1	1	2	2	34	55	9
1	1	3	0	30	32	6
1	1	3	1	39	36	9
1	1	3	2	34	22	5
2	1	0	0	34	30	8
2	1	0	1	35	33	9
2	1	0	2	32	32	7
2	1	1	0	35	36	8
2	1	1	1	34	44	8
2	1	1	2	31	41	8
2	1	2	0	34	40	8
2	1	2	1	34	20	10
2	1	2	2	34	41	9
2	1	3	0	33	34	10
2	1	3	1	34	44	7
2	1	3	2	35	26	5
3	1	0	0	36	27	6
3	1	0	1	37	37	6
3	1	0	2	30	28	10
3	1	1	0	32	31	11
3	1	1	1	38	26	9
3	1	1	2	38	40	9
3	1	2	0	28	33	8
3	1	2	1	31	37	8
3	1	2	2	31	44	9
3	1	3	0	30	38	6
3	1	3	1	36	32	9
3	1	3	2	34	32	8
1	2	0	0	32	38	8
1	2	0	1	30	38	7
1	2	0	2	30	33	9
1	2	1	0	30	23	9
1	2	1	1	34	24	9
1	2	1	2	30	38	8
1	2	2	0	34	41	9

1	2	2	1	32	35	7
1	2	2	2	33	30	7
1	2	3	0	32	40	9
1	2	3	1	32	25	9
1	2	3	2	36	19	11
2	2	0	0	35	31	7
2	2	0	1	31	29	7
2	2	0	2	35	43	8
2	2	1	0	28	36	9
2	2	1	1	34	49	7
2	2	1	2	35	36	9
2	2	2	0	31	47	9
2	2	2	1	32	42	8
2	2	2	2	32	42	7
2	2	3	0	34	40	9
2	2	3	1	37	34	11
2	2	3	2	36	48	10
3	2	0	0	31	43	8
3	2	0	1	34	39	8
3	2	0	2	31	33	10
3	2	1	0	34	37	8
3	2	1	1	35	35	9
3	2	1	2	31	31	8
3	2	2	0	31	43	8
3	2	2	1	32	36	9
3	2	2	2	39	34	9
3	2	3	0	31	26	9
3	2	3	1	38	35	8
3	2	3	2	34	37	8

APPENDIX IV:INCOMPLETE SIMULATED DATA

REP	VAR	NIT	MAN	HEI	LEA	BRA
1	1	0	0	.	32	8
1	1	0	1	31	34	.
1	1	0	2	32	28	8
1	1	1	0	32	.	7
1	1	1	1	32	31	9
1	1	1	2	31	30	.
1	1	2	0	.	42	7
1	1	2	1	31	.	10
1	1	2	2	34	55	9
1	1	3	0	30	32	6
1	1	3	1	39	36	9
1	1	3	2	34	.	.
2	1	0	0	34	30	8
2	1	0	1	35	33	9
2	1	0	2	.	32	.
2	1	1	0	.	36	8
2	1	1	1	34	44	.
2	1	1	2	31		8
2	1	2	0	34	40	8
2	1	2	1	34	20	10
2	1	2	2	34	41	9
2	1	3	0	33	.	10
2	1	3	1	34	.	7
2	1	3	2	.	26	5
3	1	0	0	36	27	6
3	1	0	1	37	.	6
3	1	0	2	30	28	10
3	1	1	0	32	31	.
3	1	1	1	38	26	.
3	1	1	2	38	40	9
3	1	2	0	28	33	8
3	1	2	1	31	37	.
3	1	2	2	31	.	9
3	1	3	0	30	38	6
3	1	3	1	36	32	.
3	1	3	2	34	32	8
1	2	0	0	32	.	8
1	2	0	1	.	38	7
1	2	0	2	30	33	9
1	2	1	0	30	.	.
1	2	1	1	34	24	9
1	2	1	2	30	38	8
1	2	2	0	34	41	9

1	2	2	1	32	35	.
1	2	2	2	33	30	7
1	2	3	0	32	40	9
1	2	3	1	32	25	9
1	2	3	2	.	19	.
2	2	0	0	.	31	7
2	2	0	1	31	29	7
2	2	0	2	35	43	.
2	2	1	0	28	36	9
2	2	1	1	34	49	7
2	2	1	2	35	36	.
2	2	2	0	31	47	9
2	2	2	1	.	42	8
2	2	2	2	.	42	7
2	2	3	0	34	.	9
2	2	3	1	37	34	.
2	2	3	2	36	48	10
3	2	0	0	31	43	8
3	2	0	1	34	.	8
3	2	0	2	31	33	.
3	2	1	0	34	37	8
3	2	1	1	35	35	9
3	2	1	2	31	31	.
3	2	2	0	.	.	8
3	2	2	1	32	36	9
3	2	2	2	39	.	.
3	2	3	0	31	26	9
3	2	3	1	38	.	8
3	2	3	2	34	37	.

APPENDIX V: REAL LIFE DATA SET

Between-Subjects Factors

		N
REPLICATE	1.00	24
	2.00	24
	3.00	24
VARIETY	1.00	36
	2.00	36
NITROGEN	.00	18
	1.00	18
	2.00	18
	3.00	18
MANURE	.00	24
	1.00	24
	2.00	24

INCOMPLETE DATA SET FOR PLANT HEIGHT

Between-Subjects Factors

		N
REPLICATE	1	20
	2	18
	3	22
VARIETY	1	30
	2	30
NITROGEN	0	14
	1	16
	2	14
	3	16
MANURE	0	19
	1	21
	2	20

INCOMPLETE DATA SET FOR PLANT LEAVE

Between-Subjects Factors

		N
REPLICATE	1	19
	2	20
	3	18
VARIETY	1	28
	2	29
NITROGEN	0	15
	1	15
	2	14
	3	13
MANURE	0	18
	1	19
	2	20

INCOMPLETE DATA SET FOR PLANT BRANCHES

Between-Subjects Factors

		N
REPLICATE	1	18
	2	19
	3	16
VARIETY	1	27
	2	26
NITROGEN	0	14
	1	11
	2	15
MANURE	3	13
	0	22
	1	17
	2	14

APPENDIX VI:P-VALUES OBTAINED WITH THE RLDS, PW, MI AND EM FOR PLANT HEIGHT

Tests of Between-Subjects Effects

Dependent Variable:HEIGHT

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Corrected Model	1403.847 ^a	25	56.154	4.422	.000
Intercept	81945.014	1	81945.014	6453.038	.000
REPLICATE	72.528	2	36.264	2.856	.068
VARIETY	45.125	1	45.125	3.554	.066
NITROGEN	324.597	3	108.199	8.521	.000
MANURE	420.194	2	210.097	16.545	.000
VARIETY * NITROGEN	26.486	3	8.829	.695	.560
VARIETY * MANURE	35.583	2	17.792	1.401	.257
NITROGEN * MANURE	292.028	6	48.671	3.833	.004
VARIETY * NITROGEN * MANURE	187.306	6	31.218	2.458	.038
Error	584.139	46	12.699		
Total	83933.000	72			
Corrected Total	1987.986	71			

Tests of Between-Subjects Effects

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Corrected Model	1224.137 ^a	25	48.965	3.657	.000
Intercept	66127.218	1	66127.218	4938.517	.000
REPLICATE	63.237	2	31.618	2.361	.110
VARIETY	28.336	1	28.336	2.116	.155
NITROGEN	226.960	3	75.653	5.650	.003
MANURE	437.349	2	218.675	16.331	.000
VARIETY * NITROGEN	29.043	3	9.681	.723	.545
VARIETY * MANURE	31.847	2	15.923	1.189	.317
NITROGEN * MANURE	274.826	6	45.804	3.421	.009
VARIETY * NITROGEN * MANURE	184.132	6	30.689	2.292	.058
Error	455.263	34	13.390		
Total	70632.000	60			
Corrected Total	1679.400	59			

Tests of Between-Subjects Effects

Dependent Variable:HEIGHT

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Corrected Model	819.043 ^a	25	32.762	3.282	.000
Intercept	81608.000	1	81608.000	8175.077	.000
REPLICATE	36.270	2	18.135	1.817	.174
VARIETY	7.476	1	7.476	.749	.391
NITROGEN	176.369	3	58.790	5.889	.002
MANURE	189.703	2	94.852	9.502	.000
VARIETY * NITROGEN	83.631	3	27.877	2.793	.051
VARIETY * MANURE	1.268	2	.634	.063	.939
NITROGEN * MANURE	161.101	6	26.850	2.690	.025
VARIETY * NITROGEN * MANURE	163.226	6	27.204	2.725	.024
Error	459.197	46	9.983		
Total	82886.240	72			
Corrected Total	1278.240	71			

Tests of Between-Subjects Effects

Dependent Variable:HEIGHT

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Corrected Model	1243.425 ^a	25	49.737	3.707	.000
Intercept	82198.480	1	82198.480	6126.362	.000
REPLICATE	67.507	2	33.754	2.516	.092
VARIETY	19.841	1	19.841	1.479	.230
NITROGEN	224.274	3	74.758	5.572	.002
MANURE	405.610	2	202.805	15.115	.000
VARIETY * NITROGEN	70.641	3	23.547	1.755	.169
VARIETY * MANURE	17.272	2	8.636	.644	.530
NITROGEN * MANURE	231.571	6	38.595	2.877	.018
VARIETY * NITROGEN * MANURE	206.708	6	34.451	2.568	.031
Error	617.190	46	13.417		
Total	84059.094	72			
Corrected Total	1860.615	71			

P-VALUES OBTAINED WITH THE RLDS, PW, MI AND EM FOR PLANT LEAVE

Tests of Between-Subjects Effects

Dependent Variable:LEAVE

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Corrected Model	11702.181 ^a	25	468.087	7.261	.000
Intercept	86597.347	1	86597.347	1343.286	.000
REPLICATE	903.861	2	451.931	7.010	.002
VARIETY	387.347	1	387.347	6.008	.018
NITROGEN	5148.819	3	1716.273	26.623	.000
MANURE	2464.361	2	1232.181	19.113	.000
VARIETY * NITROGEN	1938.819	3	646.273	10.025	.000
VARIETY * MANURE	17.694	2	8.847	.137	.872
NITROGEN * MANURE	287.972	6	47.995	.744	.617
VARIETY * NITROGEN * MANURE	553.306	6	92.218	1.430	.224
Error	2965.472	46	64.467		
Total	101265.000	72			
Corrected Total	14667.653	71			

Tests of Between-Subjects Effects

Dependent Variable:LEAVE

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Corrected Model	9389.941 ^a	25	375.598	6.864	.000
Intercept	64167.321	1	64167.321	1172.659	.000
REPLICATE	775.362	2	387.681	7.085	.003
VARIETY	316.163	1	316.163	5.778	.022
NITROGEN	2982.672	3	994.224	18.169	.000
MANURE	1170.769	2	585.385	10.698	.000
VARIETY * NITROGEN	1590.916	3	530.305	9.691	.000
VARIETY * MANURE	4.887	2	2.444	.045	.956
NITROGEN * MANURE	405.205	6	67.534	1.234	.316
VARIETY * NITROGEN * MANURE	941.902	6	156.984	2.869	.024
Error	1696.304	31	54.719		
Total	77660.000	57			
Corrected Total	11086.246	56			

Tests of Between-Subjects Effects

Dependent Variable:LEAVE

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Corrected Model	5872.861 ^a	25	234.914	4.454	.000
Intercept	83082.467	1	83082.467	1575.301	.000
REPLICATE	623.448	2	311.724	5.910	.005
VARIETY	425.347	1	425.347	8.065	.007
NITROGEN	2218.059	3	739.353	14.019	.000
MANURE	866.361	2	433.181	8.213	.001
VARIETY * NITROGEN	789.322	3	263.107	4.989	.004
VARIETY * MANURE	33.334	2	16.667	.316	.731
NITROGEN * MANURE	102.946	6	17.158	.325	.920
VARIETY * NITROGEN * MANURE	814.043	6	135.674	2.572	.031
Error	2426.072	46	52.741		
Total	91381.400	72			
Corrected Total	8298.933	71			

Tests of Between-Subjects Effects

Dependent Variable:LEAVE

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Corrected Model	10594.530 ^a	25	423.781	6.794	.000
Intercept	81715.454	1	81715.454	1310.144	.000
REPLICATE	1019.369	2	509.685	8.172	.001
VARIETY	400.257	1	400.257	6.417	.015
NITROGEN	3727.287	3	1242.429	19.920	.000
MANURE	2609.804	2	1304.902	20.921	.000
VARIETY * NITROGEN	1581.204	3	527.068	8.450	.000
VARIETY * MANURE	.482	2	.241	.004	.996
NITROGEN * MANURE	325.953	6	54.325	.871	.523
VARIETY * NITROGEN * MANURE	930.175	6	155.029	2.486	.036
Error	2869.083	46	62.371		
Total	95179.067	72			
Corrected Total	13463.613	71			

P-VALUES OBTAINED WITH THE RLDS, PW, MI AND EM FOR PLANT BRANCHES

Tests of Between-Subjects Effects

Dependent Variable:BRANCHES

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Corrected Model	368.681 ^a	25	14.747	3.299	.000
Intercept	5016.681	1	5016.681	1122.197	.000
REPLICATE	107.028	2	53.514	11.971	.000
VARIETY	45.125	1	45.125	10.094	.003
NITROGEN	74.708	3	24.903	5.571	.002
MANURE	58.028	2	29.014	6.490	.003
VARIETY * NITROGEN	9.597	3	3.199	.716	.548
VARIETY * MANURE	15.083	2	7.542	1.687	.196
NITROGEN * MANURE	32.750	6	5.458	1.221	.313
VARIETY * NITROGEN * MANURE	26.361	6	4.394	.983	.448
Error	205.639	46	4.470		
Total	5591.000	72			
Corrected Total	574.319	71			

Tests of Between-Subjects Effects

Dependent Variable:BRANCHES

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Corrected Model	291.132 ^a	25	11.645	2.020	.038
Intercept	3163.588	1	3163.588	548.872	.000
REPLICATE	61.211	2	30.605	5.310	.011
VARIETY	12.358	1	12.358	2.144	.155
NITROGEN	31.535	3	10.512	1.824	.167
MANURE	41.671	2	20.836	3.615	.041
VARIETY * NITROGEN	14.259	3	4.753	.825	.492
VARIETY * MANURE	10.629	2	5.314	.922	.410
NITROGEN * MANURE	39.274	6	6.546	1.136	.369
VARIETY * NITROGEN * MANURE	19.022	6	3.170	.550	.766
Error	155.623	27	5.764		
Total	4083.000	53			
Corrected Total	446.755	52			

Tests of Between-Subjects Effects

Dependent Variable:BRANCHES

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Corrected Model	216.772 ^a	25	8.671	2.304	.007
Intercept	5256.542	1	5256.542	1397.002	.000
REPLICATE	46.514	2	23.257	6.181	.004
VARIETY	7.736	1	7.736	2.056	.158
NITROGEN	61.800	3	20.600	5.475	.003
MANURE	29.884	2	14.942	3.971	.026
VARIETY * NITROGEN	19.896	3	6.632	1.763	.168
VARIETY * MANURE	6.004	2	3.002	.798	.456
NITROGEN * MANURE	24.573	6	4.096	1.088	.383
VARIETY * NITROGEN * MANURE	20.364	6	3.394	.902	.502
Error	173.086	46	3.763		
Total	5646.400	72			
Corrected Total	389.858	71			

Tests of Between-Subjects Effects

Dependent Variable:BRANCHES

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Corrected Model	339.098 ^a	25	13.564	3.140	.000
Intercept	5244.834	1	5244.834	1213.982	.000
REPLICATE	79.118	2	39.559	9.156	.000
VARIETY	15.579	1	15.579	3.606	.064
NITROGEN	87.065	3	29.022	6.717	.001
MANURE	74.463	2	37.231	8.618	.001
VARIETY * NITROGEN	22.950	3	7.650	1.771	.166
VARIETY * MANURE	5.122	2	2.561	.593	.557
NITROGEN * MANURE	35.292	6	5.882	1.361	.250
VARIETY * NITROGEN * MANURE	19.509	6	3.251	.753	.611
Error	198.736	46	4.320		
Total	5782.669	72			
Corrected Total	537.834	71			