MISSING DATA IN NON-PARAMETRIC TESTS OF CORRELATED DATA

Annie Green Howard

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics.

Chapel Hill
2012

Approved by:

Shrikant Bangdiwala
Lloyd J. Edwards
Gerardo Heiss
Gary Koch
Paul Stewart
Stephan R. Weinland

Abstract

ANNIE GREEN HOWARD: MISSING DATA IN NON-PARAMETRIC TESTS OF CORRELATED DATA
(Under the direction of Shrikant Bangdiwala)

Many public health studies are designed to test for differences in repeated measurements. Measurements on the same subject are not independent and therefore analysis methods must take correlation into account. A number of tests have been developed to analyze this type of data. Two prominent non-parametric methods, used often when one is not willing to make any distributional assumptions about the data, include Friedman's test and a variation on Friedman's test proposed by Koch and Sen that requires no assumptions to be made about the correlation between measurements.

While both tests require complete and balanced data, in many studies missing data can arise for a variety of reasons. Researchers have developed a number of methods to adapt Friedman's test to situations involving missing data when it can be assumed that the missing data are missing completely at random. We propose applying these same adjustments to the test statistic proposed by Koch and Sen to adapt this test to deal with data that are missing completely at random. This method involves using the sum of the reduced ranks, rather than the average rank, across all subjects to allow for meaningful comparisons across subjects. An inflation factor is used to ensure the missing data do not result in a substantial loss of power.

The assumption that the data are missing completely at random is often too strict an assumption for correlated data. Often the reason for the data to be missing is directly related to the outcome values. A new strategy is proposed for adjusting both

Friedman's test and Koch and Sen's test to informative missing data scenarios. The method put forth in this paper involves the use of single imputation to impute missing rank values along with a weighting scheme which assigns smaller weight to individuals with more missing data. Guidelines and suggestions are put forward as to when this new method would be preferred to the method currently used to address problems with missing completely at random data.

Acknowledgments

I wish to thank Dr. Shrikant Bangdiwala for his insight, encouragement and friendship throughout this process. His confidence in my skills along with his guidance and support have helped me to make it through both the ups and downs of this process. Dr. Gary Koch has also been instrumental in the development in this research and I would like to thank him for his input and ideas. I would also like to acknowledge Dr. Lloyd Edwards, Dr. Gerardo Heiss, Dr. Paul Stewart and Dr. Stephan Weiland who have offered many helpful suggestions and ideas that went into this dissertation.

I could not have completed this without the love and support of my family. In particular, Id like to acknowledge my father Dr. George Howard who offered advice and assistance at every stage of the process. His belief in me has and will continue to be instrumental in my future success. I'd also like to thank my mother, Dr. Virginia Howard, for the countless words of encouragement and support. In addition she has lived her life in such a way that her career has served to help me in the development of my personal and professional goals. My sisters, Marjorie Howard and Letitia Perdue have offered nothing but unconditional love and support which has been invaluable during this process.

I cannot thank enough the staff, faculty and students at UNC, particularly within the Biostatistics Department. This has been a difficult but rewarding process and the generosity, patience, acceptance and assistance of the friends I have made within this department have been responsible for much of my success. I'd like to particularly thank

Melissa Hobgood, Virginia Pate, Allison Deal and my fellow members both official and unofficial of The Cave. I would also like to thank specifically the CSCC for their flexibility and financial support. This work was supported in large part by Training Grant T32ES007018 from NIEHS.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction and Literature Review

1.1 Introduction

1.1.1 Motivation

Many public health studies are designed to test for a difference between repeated measurements on the same subject. These studies can be useful in a number of different contexts including, but not limited to, testing the reliability of a particular procedure using repeated measurements on the same individuals, testing for a change in average values or proportions over time, and testing for an effect of different treatments on the same subject over a follow-up period. In all of these cases, measurements on the same subject are not independent and therefore analysis methods must take into account the correlation between measurements taken on the same subject. Recently, some of the analytic approaches for this type of data analysis have been generalized to account for missing data. In large sample studies, these methods have been proven to produce accurate type I error rates in certain scenarios although the preferred analysis method differs depending on a number of factors.

Non-parametric methods are one of the approaches for which these adaptations have been developed and tested for small studies. However, these adaptations have

only been developed for scenarios where one can assume the correlation between any two measurements on the same subject is the same. This is problematic for many repeated measures studies, particularly longitudinal studies, when one can often make the assumption that the correlation between two measurements that are close together in time are likely to be more strongly correlated than measurements farther apart in time.

These non-parametric adjustments that seek to minimize bias and preserve the accuracy of type I error rates have been proven to be effective only when the probability an observation is missing does not depend on the outcome or covariate values, commonly referred to as missing completely at random (MCAR). Some researchers have investigated the power of these adjusted tests, finding the reduction in power not to be an of great concern in the case of MCAR data (Kenward and Roger, 1997; Schluchter and Elashoff, 1990; Catellier and Muller, 2000; Manor and Zucker, 2004; Kawaguchi and Koch, 2010).

In practice the reasons for missingness, if even known, can be complex and assuming MCAR can potentially lead to accuracy problems when testing hypotheses. Often in studies involving repeated measurements, missingness is informative, meaning missingness depends on the actual outcome values themselves. Current methods have not been adapted to deal with this missingness scenario and there is a strong potential for biased results when attempting to make inference when incorrectly assuming that the data are MCAR.

1.1.2 Example

The first example this research will address is a longitudinal study testing for a change in average outcome values over time. A fixed study schedule is assumed with

the same number of measurements planned to be recorded for each subject. A continuous outcome is to be measured at each of these time points. Measurements closer in time tend to be more highly correlated than measurements further apart in time and therefore equal correlation between any two measurements is highly unlikely. Specifically we will be looking at a study in which investigators were interested in testing if pain scores differed throughout a day. Participants with irritable bowel syndrome (IBS) were asked to record pain scores (on a scale from 0 to 10) at wake up, morning, midday, evening and bedtime. Measurements were collected over a large number of days and so pain scores were collapsed by averaging pain scores for each period of the day across all days. Occasionally patients forgot to record pain scores. If the pain score was missing for more than 40% of all days for a particular period the average pain score for that subject at that period of the day was set to missing. In the larger context, it is important to note that missing data occur in longitudinal studies in a number of contexts. In this case, it is reasonable to assume the chance of a subject reporting a missing value is unrelated to the outcome value.

The second example involves a situation where a number of judges were each asked to rank a number of objects. In this scenario, since the objects are naturally ranked, non-parametric methods are a natural choice of analysis. Researchers are interested in testing for a preference in objects. Therefore, one is interested in testing if the rank for each object is the same, although differences between judges are not of interest. In this scenario, some judges felt uncomfortable ranking one or more of the objects. This could happen when one object was noticeably better or worse than the remaining objects. Therefore, one would expect that these missing ranks were more likely to be higher ranks. This scenario would be similar to one in which lab measurements were collected on the same subject at one clinic visit and researchers tested for a within subject difference. We would expect ranks from the same subject, particularly as they

were taken at the same time, to be highly correlated. In the case of a particular lab outcome, we assume that the results from each aliquot, coming from the same subject, will be equally correlated with all other aliquots. Some loss of data is expected in this study as some aliquots were lost or broken during the transportation of the samples to the lab. In addition with this data set, the lab equipment cannot measure certain lab values when values fall outside of a pre-specified value, in the case of this study when the lab values are extremely high.

The final example examined in this research involves a similar situation to that described in the first example. Researchers were interested in testing to determine if there was a difference in the difference in pre and post-bowel movement pain scores throughout the day. Subjects suffering from irritable bowel syndrome (IBS) were asked to rank pain on a scale from 0 to 10 before and after every bowel movement. Based on the time stamp of these measurements, the difference in pre and post-bowel pain scores were classified as early morning, morning, afternoon or evening. This study enrolled both diarrhea predominant (IBS-D) and constipation predominant IBS (IBS-C) patients. IBS-C patients were more likely to have missing data when they were experiencing IBS symptoms. A missing value for these participants was often indicative of a higher pre-bowel pain score as they were experiencing constipation. As such, one expects missing measurements to be higher than the non-missing counterparts. A similar situation can develop in any longitudinal trial with missing data. While subjects may miss a visit randomly, some also drop out of studies due to a change in location or health status. If a patient drops out of the study due to health status, this can be due to either the participant's health improving or declining to a point that they are no longer interested or able to participate in the study. These improvements or declines can be associated with better or worse outcome values, often denotes by extremely high or extremely low outcome values.

1.2 Literature Review

1.2.1 Notation and Assumptions

We will generalize this research to a study designed with n planned measurements recorded for each of the k subjects. We denote the j^{th} measurement for the i^{th} individual as the scalar Y_{ij} and X_{ij} denotes any additional covariates collected on the i^{th} subject along with the j^{th} outcome measurement. These X_{ij} values can either vary by measurement or can be constant within a subject. For our research, we will assume all covariates are constant within a subject and therefore that $X_{ij} = X_i$ for all i.

Missing outcome values occur often in repeated measures studies. Any scenario involving missing data also involves a loss of information and therefore analysis with missing data will often differ from the analysis of the data if it was a complete data set. Missing data is a very common occurrence and one that must be accounted for in order to minimize bias and loss of efficiency both of which are associated with missing data. To classify missingness, a variable R_{ij} is specified as an indicator variable to denote if Y_{ij} is missing. R_{ij} takes the value of 1 if the Y_{ij}^{th} observation is observed and 0 otherwise. Table 1.1 below illustrates that all the information gathered can be summarized by reporting these three values (Y_{ij}, X_{ij}, and R_{ij}).

The notation $\boldsymbol{Y_i}$ will denote the full vector of missing and observed responses for the i^{th} subject (Y_{i1}, Y_{i2}, . . .Y_{in}) and the notation $\boldsymbol{R_i}$ will denote the corresponding vector for the indicators R_{ij}. Each of these vectors has n elements, corresponding to the n measurements taken on each subject.

1.2.2 Non-parametric Analysis of Complete and Balanced Data

There are a number of situations in which there are problematic and potentially influential outliers or situations in which making distributional assumptions about Y_{ij}

Table 1.1: Repeated Measures Data Set with Missing Data

	Measurement			
Subject	1	2	...	n
1	$\begin{pmatrix} Y_{11} \\ X_1 \\ R_{11} \end{pmatrix}$	$\begin{pmatrix} Y_{12} \\ X_1 \\ R_{12} \end{pmatrix}$		$\begin{pmatrix} Y_{1n} \\ X_1 \\ R_{1n} \end{pmatrix}$
2	$\begin{pmatrix} Y_{21} \\ X_2 \\ R_{21} \end{pmatrix}$	$\begin{pmatrix} Y_{22} \\ X_2 \\ R_{22} \end{pmatrix}$		$\begin{pmatrix} Y_{2n} \\ X_2 \\ R_{2n} \end{pmatrix}$
...	...	...	...	...
k	$\begin{pmatrix} Y_{k1} \\ X_k \\ R_{k1} \end{pmatrix}$	$\begin{pmatrix} Y_{k2} \\ X_k \\ R_{k2} \end{pmatrix}$		$\begin{pmatrix} Y_{kn} \\ X_k \\ R_{kn} \end{pmatrix}$

values may be inappropriate. A number of non-parametric approaches to the repeated measures analysis have been adapted to deal with such scenarios. The earliest methods focused on data in which each subject had the same number of measurements, known as balanced data, and data where there were no missing values, known as complete data (Friedman, 1937; Koch and Sen, 1968). While there are methods to test for differences in measurement within a subject if the effects are different between subjects, this research will be focusing on methods that assume the measurement effect is constant across all subjects.

Within this subset of methods, the preferred method depends on assumptions one is willing to make. One such assumption involves, what Koch and Sen refer to as, the additivity of subject effects. When this assumption is true, it is reasonable to believe that comparisons of between block rankings are meaningful. This involves assuming that the difference between two ranked measurements on the same subject is comparable to the difference in measurements of the same ranks for a different subject (Koch and

Sen, 1968; Stokes, Davis and Koch, 2000). If one assumes the additivity of subject effects, methods take this into account and adapt their methods so rank comparisons between subjects are used. The most widely accepted class of these testing methods involves the use of aligned rank tests. The premise behind all these tests is that some function of the data for the subject, usually a measure of location, is subtracted from all the Y_{ij} values. These differences are then treated as the outcome variables and are ranked within a subject (Stokes, Davis and Koch, 2000; Hodges and Lehmann, 1962; Sen, 1968; Lehmann and D'Abrera, 2006; Koch and Sen, 1968).

Friedman's Test

One of the most widely-used non-parametric methods, Friedman's test, utilizes partial rank transformation methods to test for the hypothesis of no difference in measurements within a subject while assuming no additivity of subject effects. In studies where compound symmetry can be assumed, Friedman's statistic tests the measurement effect while controlling for any subject effect. This test does not require the normality assumptions of parametric methods and also minimizes the effect of the between subject variability, allowing tests to focus on measurement effect (Friedman, 1937; Stokes, Davis and Koch, 2000).

Friedman's method assumes that all outcome variables Y_{ij} come from an n-variate continuous cumulative distribution function F_i where $F_i = G_i(y - b_i + \theta_j)$. The focus of this test is in testing for a measurement effect while controlling for subject. Therefore hypothesis testing involve testing if $\theta_j = 0$ for all j, under the constraint that $\sum \theta_j = 0$. No additional covariates are included in analysis and therefore no X_i's are involved in the test statistic.

Friedman's test replaces the original measurement values by within subject ranks. Therefore in the case of the data set described in Table 1.1, the outcome variable

Y_{ij} would be replaced with within subject rank, which will be denoted r_{ij} where $r_{ij} = 1, 2, ...n_i$. For complete and balanced data sets, we assume the total number of non-missing measurements is the same for all subjects. Thus, for any subject i, $n_i = n$. It is important to note that $\sum_{j=1}^{n} r_{ij} = \frac{n(n+1)}{2}$ for all k subjects. The data in Table 1.1 can be summarized in a new format, which allows for a display of the same data in a new format show in Table 1.2 below:

Table 1.2: Non-Parametric Test With Complete and Balanced Data

	Measurement			
Subject	1	2	...	n
1	r_{11}	r_{12}		r_{1n}
2	r_{21}	r_{22}		r_{2n}
...	...	...	...	...
k	r_{k1}	r_{k2}		r_{kn}

If there was no difference in measurements collected on the same subject, one would expect each rank to be equally likely to be located in each of the n columns. Therefore, under the null hypothesis of no measurement effect, one would expect the mean rank for each column to come from a distribution with a mean of the average rank, $\frac{n+1}{2}$. The variance of this distribution under the null hypothesis can be calculated to be $\frac{n^2-1}{12k}$, where k denotes the number of subjects. Friedman's test statistic, based off these values for the mean and variance, is:

$$\frac{12k}{n(n+1)} \sum_{j=1}^{n} \left(\bar{r}_j - \frac{1}{2}(n+1) \right)^2$$

where $\bar{r}_j$ is the average rank of the j^{th} column. In small studies, it is best to use the exact permutation distribution to test the null hypothesis of no trend across the columns, or in the case of longitudinal data no trend over time. However, when the number of blocks is sufficiently large the Friedman statistic will have a chi-squared

distribution with $n-1$ degrees of freedom.

Tied outcome variables can be dealt with by assigning the tied ranks at random to each of the tied measurements. However, the mid-rank method is a more common method that allows for the utilization of more information. This method involves giving tied measurements the average value of the ranks for which two or more observations are tied (Friedman, 1937).

This method, with the use of the mid-rank option for ties, is equivalent to combining Kruskal-Wallis tests while conditioning on subject which is equivalent to a stratified Mantel-Haenszel tests with column scores being equal to within subject ranks. Since Friedman's test can only be used in the case of complete and balanced data, in this case only the stratified Mantel-Haenszel statistic equivalent to Friedman's statistics. Both involve tests to determine if mean responses differ using within subject rank scores rather than actual measurement values (Landis, Heyman and Koch, 1978; Stokes, Davis and Koch, 2000).

Koch and Sen's Test

If it is not reasonable to assume a compound symmetric correlation structure, Koch and Sen have proposed an alternative to Friedman's test. This method also uses partial rank transformation and therefore the data structure is identical to that shown in Table 1.2. Where the two methods differ is in terms of the permutation-based distribution under the null hypothesis. In Friedman's test, when the correlation is equal between any two measurements within a subject, the distribution of the ranks under the null hypothesis is based off of the fact that each permutation of ranks is equally likely within each subject. The distribution used for the test statistic in Koch and Sen's test is based off of the premise that when compound symmetry is violated, the unique pair wise correlation between each two measurements must be taken into account. In Koch

and Sen's test, the distribution under the null hypothesis allows for only two possible permutations of ranks within a subject. The first possible permutation is the observed permutation and the second is the exact opposite permutation, specified explicitly below. For both of these permutations the correlation between any two measurements is the same, thereby preserving the correlation structure (Koch and Sen, 1968).

$$\boldsymbol{r_i} = (r_{i1}, r_{i2}, ..., r_{in})$$

$$\boldsymbol{r_i} = (n+1-r_{i1}, n+1-r_{i2}, ..., n+1-r_{in})$$

These two permutations for the i^{th} subject are the only two permutations that have the same correlation as the observed data for the i^{th} subject. Each of these permutations are assumed to be observed with equal probability under the null hypothesis that there is no difference in measurements within a subject. Koch and Sen's method tests for a measurement effect while controlling for subject, and therefore the results are combined across all subjects to get an estimated average effect across all subjects. Interest lies in tests involving $\mathbf{T}$ which is a n x 1 vector with elements $T_j = \frac{1}{k}\sum_{i=1}^{k} r_{ij}$.

Under the null hypothesis of this distribution, the expected value of T_j can be calculated based on the expected value of r_{ij} which is equal to $\frac{n+1}{2}$.

$$\begin{aligned} E[T_j] = \frac{1}{k}\sum_{i=1}^{k} E[r_{ij}] = E[r_{ij}] &= ((r_{ij})Pr(r_{ij} = r_{ij}) + (n+1-r_{ij})Pr(r_{ij} = n+1-r_{ij})) \\ &= r_{ij}\frac{1}{2} + (n+1-r_{ij})\frac{1}{2} = \frac{n+1}{2} \end{aligned}$$

Under the assumptions required for Koch and Sen's test, the covariance matrix of $\mathbf{T}$, a nxn matrix, will be denoted as $\mathbf{V}$ with each element $v_{jj'}$ calculated as:

$$\begin{aligned}
v_{jj'} = Cov(T_j, T_{j'}) &= Cov(\frac{1}{k}\sum_{i=1}^{k} r_{ij}, \frac{1}{k}\sum_{i=1}^{k} r_{ij'}) = \left(\frac{1}{k}\right)^2 \sum_{i=1}^{k} Cov(r_{ij}, r_{ij'}) \\
&= \frac{1}{k^2}\sum_{i=1}^{k} (E[(r_{ij} - E[r_{ij}])(r_{ij'} - E[r_{ij'}])]) \\
&= \frac{1}{k^2}\sum_{i=1}^{k} \left(E[r_{ij}r_{ij'}] - \left(\frac{n+1}{2}\right)^2\right) \\
&= \frac{1}{k^2}\sum_{i=1}^{k} \left((r_{ij}r_{ij'})\frac{1}{2} + (n+1-r_{ij})(n+1-r_{ij'})\frac{1}{2} - \left(\frac{n+1}{2}\right)^{\frac{1}{2}}\right) \\
&= \frac{1}{k^2}\sum_{i=1}^{k} \left(\frac{2r_{ij}r_{ij'}}{2} - \frac{r_{ij}(n+1)}{2} - \frac{r_{ij'}(n+1)}{2} + \frac{(n+1)^2}{2} - \frac{\frac{1}{2}(n+1)^2}{2}\right) \\
&= \frac{1}{k^2}\sum_{i=1}^{k} \left(r_{ij}r_{ij'} - \frac{r_{ij}(n+1)}{2} - \frac{r_{ij'}(n+1)}{2} + \left(\frac{n+1}{2}\right)^2\right) \\
&= \frac{1}{k^2}\sum_{i=1}^{k} \left(\left(r_{ij} - \frac{(n+1)}{2}\right)\left(r_{ij'} - \frac{(n+1)}{2}\right)\right)
\end{aligned}$$

Koch and Sen developed a generalized statistic which allows for the testing of any linear contrast $\mathbf{C}$ of the vector $\mathbf{T}$, which consists of all n T_j elements, for which $\mathbf{Cj} = \mathbf{0}$ where $j' = (1, ..., 1)$. The form of the generalized statistic is stated in terms of the contrast matrix as $\mathbf{T'C'}\left(\mathbf{CVC'}\right)^{-\mathbf{1}}\mathbf{CT}$. Under the null hypothesis $k^{1/2}\mathbf{T}$ is an asymptotically multivariate normal vector of rank $n-1$. Therefore, the test statistic $\mathbf{T'C'}\left(\mathbf{CVC'}\right)^{-\mathbf{1}}\mathbf{CT}$ has an asymptotically chi-squared distribution with $n-1$ degrees of freedom under the null hypothesis (Koch and Sen, 1968).

1.2.3 Missing Data Mechanisms

Missing Covariates

Although in some studies missing covariates can also be a problem, we will be assuming no missing covariates in our examples. Background on dealing with missing covariates is included for completeness since this can present substantial concerns. Various methods exist to address the problems that arise from missing covariates, but for simplicity we will deal with studies that have complete covariate data. Missing covariates can potentially be an issue in all types of studies; however, in repeated measures scenarios it can often have a greater impact on the analysis, as one missing covariate can affect multiple observations on one participant. One method used to address these concerns involves replacing the missing covariate with the mean or median value of the covariate. An alternative method involves replacing the missing value by the predicted value generated by regressing the covariate with missing values on all observed covariates. More recently, methods often used for dealing with missing outcome values have been adapted for use in dealing with missing covariates values, including maximum likelihood methods, weighted estimating equations and multiple imputations. However, these are not directly incorporated into the basic repeated measures modeling procedures in many computer-programming packages, including SAS. These can be done separately from repeated measures analysis computing procedures, although they are more computationally intensive, particularly the more complex methods, and these methods do require additional assumptions. In order to simplify analysis, the most common method of dealing with missing covariates is that all observations for a participant are deleted if one or more covariates are missing, which can lead to a much smaller sample size and to biased estimates unless the covariates are missing completely at random (Horton and Kleinman, 2007). This research will consider data with complete covariates and instead focus on scenarios involving missing outcome variables.

Missing Completely at Random (MCAR)

When outcome variables are missing completely at random (MCAR), the probability of a subject having a missing value for an observation does not depend on the subject's observed values or the covariates. MCAR data are defined explicitly to be data in which the indicator vector for missingness, $\boldsymbol{R_i}$, is independent of both $\boldsymbol{Y_i}$ and $\boldsymbol{X_i}$. This is equivalent to stating $Pr(\boldsymbol{R_i}|\boldsymbol{Y_i}, \boldsymbol{X_i}) = Pr(\boldsymbol{R_i})$ (Fitzmaurice, Laird and Waire, 2004). As a number of analyses require the assumption that missing data are MCAR, this assumption is often assumed even though it requires the strictest assumptions.

In studies with repeated measures over time, participants have a higher probability of missing later visits due to fatigue or lack of interest as the study continues. In these studies missing data are often classified at MCAR since missingness depends only on time, which is often treated as a design variable in the case on longitudinal studies. Unlike covariates, design variables are specified by the investigator and predetermined for use in the study design. If time is fixed and treated as a design variable, the missingness depends on a fixed variable and therefore not on the observed or unobserved data. Therefore time is not included in the covariate matrix and the missingness is MCAR (Fitzmaurice, Laird and Waire, 2004).

Covariate Dependent Missingness or Missing at Random (MAR)

Covariate dependent missingness, commonly referred to as missing at random (MAR), occurs when the probability of a subject having a missing value does not depend on the actual missing outcome values but could depend on a subject's covariates. Stated explicitly, covariate dependent missingness is defined to occur when $Pr(\boldsymbol{R_i}|\boldsymbol{Y_i}, \boldsymbol{X_i}) = Pr(\boldsymbol{R_i}|\boldsymbol{X_i})$. Covariate dependent missingness requires fewer assumptions than MCAR

and is still considered likely in studies involving repeated outcome measurements. Assuming MCAR requires less strict assumptions and therefore fewer methods of analysis are valid. Under the assumption of covariate dependent missingness, the distribution of the data used for analysis, the observed, is not the same as the population of interest. Therefore, the parameter estimates will be biased when using least squares methods and will only be accurate with maximum likelihood methods when the distribution of the outcome is correctly specified (Fitzmaurice, Laird and Waire, 2004).

Non-ignorable or Informative Missingness

As knowing that an observation is missing reveals no information about the actual missing values, both MCAR and covariate dependent missingness are commonly referred to as non-informative or ignorable missing data. In these cases knowing the actual missing values is not needed to conduct valid analyses. In contrast, the third type of missing data, missing not at random (MNAR), is commonly referred to as informative or non-ignorable missingness. In this situation $Pr(\boldsymbol{R_i}|\boldsymbol{Y_i}, \boldsymbol{X_i})$ cannot be simplified, meaning the probability of a subject having a missing value depends on the actual unobserved missing values. In this case, not incorporating information about the missing data will yield biased results. There is not currently a computationally simple method when dealing with this type of data for statistical estimation and testing. The most common method, which is extremely difficult to do with great accuracy, requires specifying models both for the response as well as the missing data mechanism. This requires defining $Pr(\boldsymbol{R_i}|\boldsymbol{Y_i}, \boldsymbol{X_i})$ accurately and explicitly (Fitzmaurice, Laird and Waire, 2004)

Monotonic vs. Non-Monotonic Missingness

All three of these categories can be further classified, if the order of the repeated measures has meaning, by specifying the missing data as monotonic or non-monotonic. If having a missing value forces all subsequent values for a subject to be missing as well, the missing mechanism is considered to be monotonic. This is commonly referred to as drop-out or loss to follow up in longitudinal studies. In contrast, non-monotonic missing data occur when a measurement can be observed after a missing measurement was reported for that subject. When missing data are non-monotonic, non-informative missingness is easier to assume and more likely to be valid. For example, when a subject drops out it is difficult to assume the reason for drop out is completely unrelated to the subject's missing outcome values. In contrast, if a subject has intermittent missing data, it is easier to assume the missingness is unrelated to the missing outcome values. Additionally, for subjects with intermittent missingness, the observed data which occurred after the missing data can help in making assumptions about the true missing values with greater accuracy (Fitzmaurice, Laird and Waire, 2004).

1.2.4 Missing Data in Repeated Measures Analysis

Complete-Case Analysis

There are a number of methods for dealing with missing data. The simplest approach is known as complete-case analysis, in which any subject with one or more observations missing is excluded from the analysis. Only data from subjects who have no missing data are included in the analysis. Using this method with informative missing data will result in noticeably biased results. If for example participants with high values were more likely to drop out of a study, the missingness is informative and any overall estimate of the mean value would be lower than the true population mean. The

only situation in which complete case analysis would produce unbiased analyses would be MCAR, as that is the only situation in which dropping those with missing data would be dropping a random sample of the population. However, even with MCAR data, the decrease in sample size could lead to a substantial decrease in power. With small samples, the loss of even a small number of observations can have an important effect. However, due to the ease of analysis and interpretation, complete case analysis is still considered an option of handling missing data (Fitzmaurice, Laird and Waire, 2004).

Repeated Measures ANOVA

Repeated measures analysis of variance (ANOVA) requires complete and balanced data and is therefore is often used in conjunction with complete case analysis. This method assumes the correlation between an individual's measurements are based on the individual's underlying tendencies that remain the same for all measurements. This is one of the earliest methods developed but due to ease of computation and interpretation, this method is still commonly used even though it requires making assumptions that may not always be valid. Repeated measures ANOVA assumes the individual has a latent response which is the same for all measurements thereby assuming some individuals tend to have overall higher or lower outcomes than the population. This method forces the data to have a compound symmetric correlation matrix, meaning the correlation is the same between any two time points. Compound symmetry is particularly questionable in longitudinal studies as one would expect measurements taken further apart in time to have weaker correlation than measurements closer in time (Fitzmaurice, Laird and Waire, 2004; Demidenko, 2004).

Single Imputation

An additional approach to handling missing data is imputation, in which each missing value is replaced by some estimated value. There are many approaches to imputation. The simplest case is that of single imputation, in which the missing value is replaced with one value generated as an estimate of the true unobserved value. Within this category, imputation can be broken down further into within individual imputation, where the estimates are gathered from the individual with the actual missing value, or between individual imputations, in which information from the entire sample or a portion of the sample is used to estimate the missing value for an individual. Missing values for an individual of a particular subgroup, for example females, may be imputed to be the overall mean of that particular subgroup. One of the most common within individual single imputations for monotonic missing data is last observation carried forward (LOCF), in which a subject's last known measurement is substituted for all successive missing values. However, the assumption of a stable outcome after drop out is unrealistic and the standard errors are smaller than they would be in the case of non-missing data. A number of other functions of the data, both within the individual as well as data from the overall sample, can replace the missing data. Some of the more frequently used values for single imputation include the mean value for a subject, the baseline value, or a worse or best case value, known as extreme case analysis (Fitzmaurice, Laird and Waire, 2004).

Multiple Imputation

Single imputation methods do not take into account the variation and uncertainty of predicting an unobserved value. Multiple imputation methods have been developed to address this concern. These methods involve replacing the missing value with a value based on a number of different values (Fitzmaurice, Laird and Waire, 2004). A

number of possible values for the missing value, generally somewhere between 3 and 10, are generated (Schafer, 1999). The methods of generating these imputed values can greatly affect the analyses. All methods rely on using the observed data to predict the missing data. One of the more common methods involves generating the imputations from an estimated proper prior distribution generated from the observed data. For more complex situations, Markov Chain Monte Carlo (MCMC) methods can be used. Both methods require multivariate normality; however, with minor departures, accurate inferences can still be made using these estimates (Horton and Lipsitz, 2001).

One complete data set is generated from each one of the generated estimates of the missing value. The complete data sets created from these multiple estimates of the missing values are then used to determine parameter estimates and variances. Denoting the parameter estimate or the combination of parameter estimates as $\boldsymbol{Q}$, m imputations would result in m estimates of $\boldsymbol{Q}$, denoted as $\hat{\boldsymbol{Q}}$, with each having a variance estimator $\boldsymbol{U}$. These multiple $\hat{\boldsymbol{Q}}$ estimates are then averaged to create a point estimate for $\boldsymbol{Q}$. The estimate of the variance of this point estimate incorporates both the between-imputation and within-imputation variance. In the multivariate case, where $\hat{\boldsymbol{Q}}$ is a vector of values, the within-imputation variance $\bar{\boldsymbol{U}}$ is the average of the m covariate matrices $\boldsymbol{U}$. The between imputation $\boldsymbol{B}$ is $\frac{1}{m-1}\sum_{t=1}^{m}(\hat{\boldsymbol{Q}}^{\boldsymbol{t}}-\bar{\boldsymbol{Q}})(\hat{\boldsymbol{Q}}^{\boldsymbol{t}}-\bar{\boldsymbol{Q}})^{\boldsymbol{T}}$. The total variance can then be expressed as $\boldsymbol{T}=\bar{\boldsymbol{U}}+(1+m^{-1})\boldsymbol{B}$. If we define k to be the number of elements in $\boldsymbol{Q}$, then inference can be made by comparing the statistic $\frac{(\bar{\boldsymbol{Q}}-\boldsymbol{Q_0})^T\boldsymbol{T^{-1}}(\bar{\boldsymbol{Q}}-\boldsymbol{Q_0})}{k}$ to an F distribution with k numerator degrees of freedom and $v=(m-1)\left\{(1+m^{-1})tr(\boldsymbol{B}\boldsymbol{T}^{-1})/k\right\}^{-2}$ denominator degrees of freedom. In multivariate cases, especially with only a small number of imputations, it becomes more complicated as the between-imputation covariance matrix would not be of full rank when the number of imputations is less than or equal to the number of elements in $\boldsymbol{Q}$. This can lead to an inaccurate estimate of the variance (Rubin, 1987; Schafer, 1997).

Generalized Estimating Equations

Methods have been developed that analyze all observed data without imputing the missing values or excluding subjects with missing data. One of these methods involves the use of generalized estimating equations, known also as marginal models, which extends generalized linear model theory to correlated data. As in generalized linear models, a link and a variance function are specified that connect the outcome to a linear combination of covariates. These methods do not require any distributional assumptions be made about the outcome variable but rely on quasi-likelihood methods in order to estimate parameters and test hypotheses. Instead of making assumptions about the distribution of the outcome variable, a correlation structure must specified (Liang and Zeger, 1986). When dealing with a continuous outcome, which we will be focusing on in this research, the mean model is defined as

$$\boldsymbol{\mu_i} = E\left[\boldsymbol{Y_i}|\boldsymbol{X_i}\right] = \boldsymbol{X_i}\beta$$

The specification of the variance component involves the specification of the correlation structure. The general covariance matrix of Y_i is specified as

$$\boldsymbol{V_i} = \phi \boldsymbol{A_i^{\frac{1}{2}}} \boldsymbol{W_i}(\boldsymbol{\alpha}) \boldsymbol{A_i^{\frac{1}{2}}}$$

where $\boldsymbol{A_i}$ is an n_i x n_i matrix with the elements consisting of the variance of Y_i along the diagonal. Here $\boldsymbol{W_i}$ is the working correlation matrix; an n_i x n_i matrix which is a function of the correlation parameters α. Once these are specified, the equation below can be solved in order to get parameter estimates, for both mean and covariance parameters.

$$\sum_{i=1}^{k} \frac{\boldsymbol{d\mu_i'}}{\boldsymbol{d\beta}} \boldsymbol{V_i}(\boldsymbol{Y_i} - \boldsymbol{\mu_i}(\boldsymbol{\beta})) = 0$$

In this equation $\boldsymbol{V_i}$ denotes the covariance matrix and $\boldsymbol{\mu_i}$ denotes the estimate of the mean for the i^{th} individual. Here we note that $\frac{\boldsymbol{d\mu_i'}}{\boldsymbol{d\beta}}$ is an s x s matrix where s denotes the number of mean parameters in the model as $\boldsymbol{\beta}$ is an s x 1 vector. The estimate for the covariance matrix is calculated based on the working correlation matrix. The model-based estimate of the covariance matrix is specified as:

$$\left(\sum_{i=1}^{k} \frac{\boldsymbol{d\mu_i'}}{\boldsymbol{d\beta}} \boldsymbol{V_i^{-1}} \frac{\boldsymbol{d\mu_i}}{\boldsymbol{d\beta}}\right)^{-1}$$

If the working correlation matrix is misspecified, this estimate of the covariance matrix will be incorrect and the standard errors of the estimates will be inaccurate. Even when the correlation structure is misspecified and the standard errors are invalid, the mean parameter estimates are accurate since the mean model is separate from the covariance model. However, inference about the parameter estimates will be invalid. As a solution, an empirical estimator, commonly referred to as the sandwich estimator has been derived and can be a more accurate estimator of the covariance of $\boldsymbol{Y_i}$

$$\left(\sum_{i=1}^{k} \frac{\boldsymbol{d\hat{\mu}_i'}}{\boldsymbol{d\beta}} \boldsymbol{\hat{V}_i^{-1}} \frac{\boldsymbol{d\hat{\mu}_i}}{\boldsymbol{d\beta}}\right)^{-1} \left(\sum_{i=1}^{k} \frac{\boldsymbol{d\hat{\mu}_i'}}{\boldsymbol{d\beta}} \boldsymbol{\hat{V}_i^{-1}} (\boldsymbol{Y_i} - \boldsymbol{\hat{\mu}_i})(\boldsymbol{Y_i} - \boldsymbol{\hat{\mu}_i})' \boldsymbol{\hat{V}_i^{-1}} \frac{\boldsymbol{d\hat{\mu}_i}}{\boldsymbol{d\beta}}\right) \left(\sum_{i=1}^{k} \frac{\boldsymbol{d\hat{\mu}_i'}}{\boldsymbol{d\beta}} \boldsymbol{\hat{V}_i^{-1}} \frac{\boldsymbol{d\hat{\mu}_i}}{\boldsymbol{d\beta}}\right)^{-1}$$

If the working correlation matrix is relatively accurate, the results can be more efficient than when using the model based estimator. Although there are many advantages to using generalized estimating equations, including the use of all observed data, these methods will yield unbiased estimates only in the case of MCAR (Stokes, Davis and Koch, 2000; Fitzmaurice, Laird and Waire, 2004).

Mixed Models

With recent computing advances, mixed models has become one of the most common methods that utilizes incomplete data. The linear mixed model for the i^{th} subject

is commonly expressed by the following equation where $i = 1, 2, ...k$:

$$\boldsymbol{Y_i} = \boldsymbol{X_i}\boldsymbol{\beta} + \boldsymbol{Z_i}\boldsymbol{b_i} + \boldsymbol{\epsilon_i}$$

Suppose $\boldsymbol{Y_i}$ is the n_i x 1 vector of non-missing outcomes for the i^{th} subject and $\boldsymbol{X_i}$ is the fixed effect design matrix containing the covariates of interest. In this model, $\boldsymbol{Z_i}$ is defined as a subset of the $\boldsymbol{X_i}$ matrix known as the random effects design matrix. We define $\boldsymbol{b_i}$ as the vector of unobserved random effects for the i^{th} subject and $\boldsymbol{\epsilon_i}$ as the unobserved vector of within-subject error. In this model $\boldsymbol{b_i}$ and $\boldsymbol{\epsilon_i}$ are assumed to have multivariate normal distributions and to be independent of each other. We can write this as follows:

$$\begin{pmatrix} \boldsymbol{b_i} \\ \boldsymbol{\epsilon_i} \end{pmatrix} : N \left(\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix} \quad , \quad \begin{pmatrix} \boldsymbol{\sigma_b^2} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\sigma_\epsilon^2} \end{pmatrix} \right)$$

If the Z_i matrix is a n_i x 1 column of ones, the model is simplified. Based on these assumptions, the covariance matrix for $\boldsymbol{Y_i}$ can be defined as $\sum_{\boldsymbol{i}} = \boldsymbol{\sigma_b^2}\boldsymbol{11'} + \boldsymbol{\sigma_\epsilon^2}\boldsymbol{I}$ where $\boldsymbol{1}$ is a n_i x 1 matrix and $\boldsymbol{I}$ denotes a n_i x n_i identity matrix. In the simplest case of mixed models, the components of the model are divided into between and within subject components. The between subject components are considered to be fixed effects, which are the true values of the population. The within subject effects, commonly referred to as random effects, are the random deviation of the subject from the population average. In this case the pair wise correlation for any two observations within an individual is the same and by including only a random intercept we are forcing a compound symmetric correlation structure on the data. Each individual has the same population mean and differs from this mean by a random intercept.

In more complex mixed models the overall mean, the effect of measurement, and any number of other regression coefficients are allowed to vary by subject. This is done by including additional covariates in the design matrix for the random effects

(Fitzmaurice, Laird and Waire, 2004).

Treatment of Missing Data

Mixed models and GEE have risen to the forefront as the primary methods of addressing missing data. Since both methods do not require a equal number of observations per individual, they allow for unbalanced and therefore missing data. However, mixed modeling is usually preferred over GEE in the case of small samples for a number of reasons. First, stronger assumptions about missing data are needed to use GEE, which limits GEE to situations involving covariate-dependent MCAR. Mixed modeling requires less strict assumptions and therefore is applicable for both MCAR and MAR mechanisms (Hedeker and Gibbons, 2006). Additionally the covariance matrix based on GEE estimates may not be positive definite. A number of simulation studies with missing data suggest that the likelihood methods of mixed models provide less bias and smaller mean squared errors than the quasi-likelihood methods used in GEE (Catellier and Muller, 2000).

Additionally, mixed models allow for a distinction of between and within subject variances without having to estimate a large number of covariance parameters. Thus, mixed models are preferred for longitudinal studies which have a large number of time points or a flexible timing for visits. Since correlation within a cluster, or within an individual, is incorporated in the model by the use of random effect, mixed models are ideal in the case of unbalanced data. Mixed models not only allow for a decomposition in variance, into between and within variation, but they also allow for testing of fixed effects while allowing for the variation to differ depending on the individual (Fitzmaurice, Laird and Waire, 2004). However, the tests associated with both mixed models and generalized estimating equation models rely on asymptotic properties and therefore the methodology developed works best with larger sample sizes.

Small Sample Studies

Some researchers have failed to explicitly define what they constitute to be small samples, and of those who do there, is some variation in terms of the definition. Some researchers have focused on defining small samples according to the overall number of observations while others have defined small samples by the number of subjects. Additionally, small sample sizes can be more or less of a problem depending on the number of parameters of interest, if there is an interest in the interaction terms, and what hypotheses are of interest. When dealing with mixed models, guidelines have been suggested that require a minimum of 30 sampling units and 30 repeated measurements on each unit to avoid small sample concerns. However, this is often incredibly impractical especially in the case of longitudinal studies and therefore these guidelines are often ignored (Bell et al., 2010). Some researchers have referred to studies with 30 or 40 subjects as small studies while others have defined small studies to involve as few as 10 or 12 subjects. These studies still had anywhere from 30 to 136 overall number of observations (Fouladi and Shieh, 2004; Catellier and Muller, 2000; Akritas and Brunner, 1997; Zucker, Lieberman and Manor, 2000).

There are two common approaches for dealing with small samples and missing data. The first method uses mixed modeling techniques with small sample adjustments to preserve the validity of tests. The second method involves ranking the response variable and using non-parametric methods to make inference. This allows for a relaxation in terms of assumptions about the distribution of the outcome variable and minimized the influence of outliers, which may be more influential in small studies (Friedman, 1937; Koch and Sen, 1968). Research on both of these methods has been developed in order to preserve type I error rates in small studies. However, there are still some problems with accuracy in certain scenarios depending on the combination of the covariance structure assumed, the methods of addressing missing data, the correlation within a subject and

what hypothesis is being tested. While most of this research has focused on the type I error rate, some have presented power results for certain scenarios as well (Schluchter and Elashoff, 1990; Catellier and Muller, 2000; Manor and Zucker, 2004; Mehrotra, Lu and Li, 2010).

Since current mixed modeling techniques allow for missing data, substantial research has been done using parametric methods to adjust large sample methods to be more efficient for small samples. Generally these methods have been tested in scenarios for which the mixed model assumptions, primarily that the outcome variable has a multivariate normal distribution, is true (Catellier and Muller, 2000; Fouladi and Shieh, 2004; Gao, 2007). There has been, however, some research that has attempted to examine the performance for outcomes with alternative distributions. These have resulted in relatively good results in terms of preserving type I error in particular scenarios (Manor and Zucker, 2004).

Some of the earliest small sample statistics involved adjustments to the likelihood ratio statistic. The first of these adjustments involved a general formula, using Bartlett's method of weighting the likelihood ratio statistic. With this weight, the moments of the likelihood ratio statistic are moved closer to the chi-squared distribution to which they are compared (Lawley, 1956). In small samples, the impact of nuisance parameters on the likelihood statistic can be significant. Therefore, an adjusted likelihood that involves the likelihood conditional on the nuisance parameters was developed (Cox and Reid, 1987). Bartlett's correction has been applied to the statistic based on this adjusted likelihood. When directly comparing these two methods, Bartlett's correction alone tends to produce a slightly inflated type I error rate and the adjusted likelihood proposed by Cox and Reid was overly conservative particularly for small samples. However, Bartlett's correction in combination with the adjusted likelihood statistic produces only a slightly inflated type I error rate in very small sample sizes (Manor and Zucker,

2004; Zucker, Lieberman and Manor, 2000). Using the likelihood ratio statistics, even an adjusted version, only allows for comparison between two nested models. Therefore, these methods are limited in the type of hypotheses that can be tested. Additionally, all of these studies were tested for using unbalanced data thereby suggesting that the validity of these results is limited to MCAR data (Manor and Zucker, 2004; Zucker, Lieberman and Manor, 2000).

In comparison, the Wald statistic allows for the testing of a much broader class of hypotheses and the research suggests that adjustments to the Wald statistic yield a comparable type I error rate to tests done based on the likelihood ratio statistic (Fouladi and Shieh, 2004). In the case of likelihood ratio tests, only maximum likelihood methods can be used to test fixed effects. One aspect involved with adjustments to Wald tests is the use of restricted maximum likelihood (REML) rather than maximum likelihood (ML) estimation. These estimation methods are often used in large sample methods but can also improve the small sample behavior of tests. Maximum likelihood methods generally behave well when sample size is large; however, in the case of small samples these results underestimate the variance and produce biased results. These problems with the variance estimate in ML methods arises even in the case of complete data (Manor and Zucker, 2004; Fitzmaurice, Laird and Waire, 2004). Since the precision of a test relies on accurate variance estimates, many methods of adjusting tests to small samples use restricted maximum likelihood estimates. Maximum likelihood methods use estimates of the mean model to estimate the variance without taking into account the uncertainty associated with the estimates of the mean model. The log likelihood of the mixed model maximized by ML estimates is shown below:

$$L = constant - \frac{1}{2}\sum_i ln\,|\boldsymbol{\Sigma_i}| - \frac{1}{2}\sum_i (\boldsymbol{Y_i} - \boldsymbol{X_i\beta})'\boldsymbol{\Sigma_i^{-1}}(\boldsymbol{Y_i} - \boldsymbol{X_i\beta})$$

By comparison REML methods remove the estimate of the mean model from the calculation of the estimate of the variance and thereby remove some of the bias. The log likelihood maximized by REML methods is

$$L = constant - \frac{1}{2}\sum_i ln\left|\boldsymbol{\Sigma_i}\right| - \frac{1}{2} ln \left|\sum_i \boldsymbol{X_i'\Sigma_i^{-1}X_i}\right| - \frac{1}{2}\sum_i \boldsymbol{r_i'\Sigma_i^{-1}r_i}$$

where

$$\boldsymbol{r_i = Y_i - X_i}\left(\sum_{\boldsymbol{i}} \boldsymbol{X_i'\Sigma_i^{-1}X_i}\right)^{\boldsymbol{-1}}\left(\sum_{\boldsymbol{i}} \boldsymbol{X_i'\Sigma_i^{-1}Y_i}\right)$$

When direct comparisons were made, REML statistics proved to be consistently better at preserving the type I error rate than uncorrected maximum likelihood estimates (Schluchter and Elashoff, 1990; Manor and Zucker, 2004). As an alternative to uncorrected ML methods, some researchers have suggested adjusting the actual ML statistic for small samples in order to account for bias. After allowing for a correction factor for the ML test statistic, REML and adjusted ML estimates proved to produce comparable type I error rates as well as similar power curves in certain scenarios (Schluchter and Elashoff, 1990). However, REML estimates are less likely to have inflated type I error rates in the case of non-normal outcomes (Catellier and Muller, 2000; Manor and Zucker, 2004).

A number of correction factors for Wald statistics have been proposed to improve the small sample behavior of both ML and REML tests. Most of these adjustments involve comparing the Wald statistic to critical values from a t or an F-distribution rather than a chi-squared distribution. There are numerous variations of this method that involve weighting this test statistic or using a different degrees of freedom. Changes to the degrees of freedom usually involve changing the denominator degrees of freedom when using the F distribution or the degrees of freedom for the t test. Numerous studies have

investigated which choice of weight or denominator degrees of freedom is better at preserving type I error rate in different scenarios with both MCAR and MAR data. The preferred method depends strongly on study design, covariance structure, hypothesis of interest, choice of REML or ML, and the correlation of outcome variables (Catellier and Muller, 2000; Schluchter and Elashoff, 1990).

For studies with fixed time points and missing data, researchers have suggested the best choices for the adjustments, in terms of both weights and denominator degrees of freedom, involve a function of the number of non-missing observations. Depending on a number of different factors, these adjustments can be improved upon when the number of groups, number of repeated measurements, and the number of overall subjects are also taken into account in deriving the weight or the denominator degrees of freedom (Catellier and Muller, 2000; Schluchter and Elashoff, 1990). However, even with this additional information taken into account, it has been noted that studies with fewer observations, more repeated measurements, higher correlation between measurements, and more missing data still have problems with inflated type I error rates. Specifically, one study examined sample sizes with up to 10% missingness and with higher levels of missingness the type I error rates are considerably inflated (Catellier and Muller, 2000).

For studies without fixed time points, there is no way of defining a participant with complete data so alternative methods must be used. All of these methods involve an adjustment to the degrees of freedom of the test. There are six common options that are often considered: the naïve degrees of freedom, the residual degrees of freedom, the separation of the degrees of freedom into between and within subject components, the containment method, the Satterwaite approximation and the Kenward-Rogers adjustment to the Satterwaite approximation. Determining the denominator degrees of

freedom as if the sample comes from a balanced design has been proven to be an effective method of controlling the type I error in certain MCAR small sample scenarios when testing fixed effects. In this naïve method, the denominator degrees of freedom are determined as if the tests were simply done using ANOVA with subject specific parameters specified by subject's specific linear regression. An additional option involves making the degrees of freedom the total number of observations minus number of between subject parameters in the model. This choice, known as the residual degrees of freedom, yields the same degrees of freedom for a study with many subjects with few observations per subject as for a study with few subjects but many observations per subject. To combat this issue and to account for differences in these two scenarios, an alternative option is to use the between and within degrees of freedom. This method separates the denominator degrees of freedom into two different parts that are used in different hypothesis testing scenarios. The degree of freedom for the between-subject hypotheses is the number of subjects minus the number of between-subject effects in the model (Manor and Zucker, 2004). An additional option, the containment method, allows the degrees of freedom for a fixed effect to depend on whether or not there is a corresponding random effect for that fixed effect. If there are, then the degrees of freedom is the rank contribution of the random effect to the $\begin{pmatrix} X & Z \end{pmatrix}$ matrix (SAS/STAT(R) 9.2 User's Guide, Second Edition). Otherwise the degrees of freedom are the residual degrees of freedom mention above, which is the total number of observations minus the rank of $\begin{pmatrix} X & Z \end{pmatrix}$. One of the most effective methods at preserving type I error in small samples is the Satterthwaite approximation. This method approximates the degrees of freedom to be $\frac{2S_i^4}{Appr\left(Var\left(S_i^2\right)\right)}$ where $S_i^2 = Var(\hat{\beta}_i)$ (Manor and Zucker, 2004). Kenward and Roger adjusted the Satterwaite approximation so the uncertainty about the estimate of the covariance matrix was taken into account. The same Satterthwaite approximation for the degrees of freedom is used but before this

is calculated, the covariance matrix is inflated (Kenward and Roger, 1997). This adjustment lowers the bias and improves the type I error rate, although again in specific situations the error rate remains inflated (Fouladi and Shieh, 2004; Kenward and Roger, 1997). Only this method and the Satterthwaite approximation are a function of the observed data. A small sample simulation study, which did not test the Kenward-Roger method, found the Satterwaite and the naïve REML method performed the best in the particular small sample MCAR scenarios. However, even these methods had problems with inflated type I error rates in certain scenarios (Manor and Zucker, 2004).

1.2.5 Missing Data in Non-parametric Analysis

Friedman's test was developed to analyze data with one observation per cell. This applies only to situations with balanced data with no missingness and exactly one observation per subject for each measurement. Methods have been developed to adapt Friedman's test to more general scenarios. These adaptations generally focus on one of two situations although some do incorporate both. The first of these involves data with more than one observation per cell and the second involves missing data while dealing with at most one observation per cell. The focus of this research will be on applications of the second type.

Durbin has been recognized as one of the first researchers to investigate alternative to Friedman's test with missing data. However, his methods do require an equal number of observations per subject (Durbin, 1951). As it is more common in the case of missing data to have an uneven number of observations per subject, research has further developed these methods to allow for incomplete and unbalanced data. The majority of these methods have focused on inflating or weighting the contribution of each subject to the statistic by some function of the number of observations the individual contributes.

Bernard and van Elteren adapted Durbin's model for scenarios with an arbitrary number of observations for any subject at any time. This statistic completes the k x n table used in Friedman's test by forcing the rank of any missing observation to be zero and then ranking all other observations from 1 to n_i. The ranks for the i^{th} subject are then weighted by a factor of $\frac{n_i^3 - \sum_{\gamma} \gamma^3 t_{i\gamma}}{12 n_i (n_i - 1)}$ where $t_{i\gamma}$ is the number of ties of size γ for subject i. Based on the calculated mean and variance of the distribution of these ranks, a statistic is generated which for large samples under the null hypothesis is compared to a chi-squared statistic with n-1 degrees of freedom (Bernard and Elteren, 1953). The complexity of these calculations has led many researchers to attempt to find simpler methods for dealing with these scenarios (Prentice, 1979; Mack and Skillings, 1980; Skillings and Mack, 1981; Rai, 1987; Wittkowski, 1988).

Van Elteren developed a more usable statistic to deal with scenarios with only two measurements. Friedman's test involves combining data across subjects to determine if the average rank for measurements are different. Van Elteren's test, which is two measurement form of the Friedman's test, involves testing a linear combination of within subject Wilcoxin rank sum tests. A general test statistic was proposed with no specific linear combination specified. A locally most powerful test was derived which involved a linear combination that inflated the test statistic for each subject based on some function of the number of measurements collected for that subject. The inflation factor of $(n_i + 1)^{-1}$ yielded the most powerful test for the hypothesis of no difference between two populations, or in the case of longitudinal studies, two time points (Elteren, 1960). When there is a true constant effect across subjects, van Elteren's method was developed to preserve type I error rate and with the intention to have better power than the alternatives (Mehrotra, Lu and Li, 2010). However, if within strata samples sizes are small, Van Elteren's statistic has been shown to have low power (Kawaguchi and Koch, 2010).

Just as a method involving a combination of Wilcoxin rank sum statistics conditioning on subject has been developed, methods have been developed that are combinations of Kruskal-Wallis statistics conditioning on subject. These are stratified Mantel-Haenszel tests, which as mentioned previously are equivalent to Friedman's test in the case of complete data. However, unlike Friedman's test these can deal with missing data in the case of MCAR data (Landis, Heyman and Koch, 1978; Stokes, Davis and Koch, 2000). In a case such as one we are focusing on, with a continuous outcome and one in which only one observation per time point is possible, Van Elteren type adjustments to these tests have been developed. These adjustments, of applying an inflation factor of $(n_i + 1)^{-1}$ to the ranks of each subject, have been tested in scenarios involving incomplete and unbalanced study designs with great success. These values, just as those provided for the two measurement case by Van Elteren, improve the power by inflating the contribution of subjects with fewer observations (Prentice, 1979). Using this inflation factor in the statistical calculations is equivalent to selecting different scores in stratified Mantel-Haenzel methods. This Van Elteren inflation factor, combined with the ranks, is commonly referred to as the modified ridit scores and the use of these methods has become widely used as SAS and other statistical packages have made it part of standard software (*SAS/STAT(R) 9.2 User's Guide, Second Edition*, N.d.).

A number of other researchers have proposed alternative inflation factors to deal specifically with Friedman-type statistics, although most were established to ease computation. Overall the most effective methods work well with MCAR data as they assign a subject a weight inversely proportional to their sample size which allows for subjects with smaller number of observations to contribute more than in unadjusted tests (Mack and Skillings, 1980; Skillings and Mack, 1981; Rai, 1987; Wittkowski, 1988)

1.3 Proposed Research

1.3.1 Background and Motivation for the Research Problem

In the case of complete and balanced data, non-parametric methods are commonly used to test for a within subject difference while controlling for the effect of subject. These methods minimize the influence of extreme outliers and require no distributional assumption regarding the outcome variable. Methods have been developed to address missing data in these scenarios however these methods are limited to MCAR data and compound symmetric correlation structures. Often one or both of these assumptions are not realistic in studies with repeated measurements.

Longitudinal studies, with measurements recorded over a number of time points, are one of the most widely used repeated measures designs. In this case, the assumption of compound symmetry is often unrealistic as measurements recorded closer in time are likely to be more strongly correlated than measurements further apart. While some researchers still use Friedman type tests for this type of data, it is important to note that the estimate of the variance associated with Friedman's test is incorrect and is actually an inflated estimate if the true correlation structure if actual correlation is not compound symmetric. Other methods of analysis for repeated measure studies have been proven to yield biased results in scenarios in which a compound symmetry correlation structure is incorrectly assumed (Gurka, Edwards and Muller, 2011). It is therefore important to adapt methods of handling MCAR data to test statistics where the assumption of compound symmetry is not required.

For situations where both compound symmetry can and cannot be assumed, the assumption that any missing data are MCAR can be questionable. Particularly problematic in longitudinal studies, non-MCAR missing data often occur as patients who drop out of the study do so as an either a direct or indirect result of their outcomes. However,

even with other types of studies informative missing data arise due to any number of other reasons, one such reason being detection limits, above or below which outcomes cannot be measured. As such, informatively missing data are common in repeated measures studies. Research methods have not dealt with adapting non-parametric methods to address the issue of informative missing data. This research seeks to adjust current methods for handling missing outcome data in small repeated measures studies when the missingness is not MCAR. We will be focusing on non-parametric methods; looking at non-parametric methods with only one covariate of interest, subject, in order to account for any potential subject effect. We will adapt methods used for these scenarios in the case of equal correlation between any two measurements and for scenarios where this equal correlation cannot be assumed.

1.3.2 Proposed Method

Current methods of addressing missing data in these tests often involve some form of an inflation factor for each subject that is some function of the number of non-missing measurements collected on that subject. This is generally done to ensure acceptable statistical power by inflating the ranks of participants with more missing values so they contribute as much to an overall test statistic as participants with complete data. This has become a generally accepted method developed of dealing with missing data for MCAR data (Stokes, Davis and Koch, 2000; Landis, Heyman and Koch, 1978). As the statistical justification of Friedman's test is similar to that of the test proposed by Koch and Sen, we propose applying this adjustment to Koch and Sen's test statistic.

When using these adjustments, one is making an assumption regarding the missing data, which in the case of informative missing data is not valid. In a non-parametric setting, we propose imputing values for the missing ranks. These values would be used to complete each block and eliminate all missing values, therefore creating blocks

of equal size and creating a scenario for which Friedman's test and Koch and Sen's test would be appropriate. The information about the reason for missingness should be used in determining the value for the imputed rank. This method would utilize Wittkowski's imputation methods and then use these imputed ranks to produce a test statistic from the generated data (Wittkowski, 1988). In addition, we propose using weights to account for the uncertainty associated with imputation, assigning smaller weights to individuals with fewer observations.

The statistical power for these tests is extremely sensitive to sample size and for even moderate sample sizes the power can be extremely low. Therefore, the effect of missing data on an already low powered statistical test can be more extreme. Therefore, the focus of this research will be on developing methods that preserve type I error rates while minimizing the loss of power due to missing data. The performance of these tests will be done using simulated data. In these simulation studies, the number of observations per individual, the number of individuals and the degree of correlation between measurements will be allowed to vary. In addition, when testing the methods developed to work with informatively missing data, we will vary the percent of the missing data that are informative. By varying all these factors, we aim to develop some guidelines as to when the methods proposed will be most useful. We will also apply these proposed tests to data sets in which the guidelines developed suggest they would be an improvement over current methods.

Chapter 2

MCAR: Without Assuming Compound Symmetry

2.1 Introduction

2.1.1 Introduction and Motivation

This research was motivated by a longitudinal study in which measurements were collected on the same subject over a period of time. This study is interested in determining if the outcome changed over time. Measurements on the same subject are correlated and therefore any subject effect must be accounted for even though differences between subjects are not of interest.

When testing non-parametrically if there is a difference in measurements over time while controlling for the effect of subject, one option for analyzing complete data sets is Koch and Sen's test. Koch and Sen's test relies on methodology similar to Friedman's but addresses scenarios in which equal correlation between any two measurements on the same subject cannot be assumed. This test evaluates whether there are differences in outcome with preservation of the actual correlation structure, without the test being affected by variation between subjects. This is done by using within subject ranks so as to evaluate their equality. Diagonal symmetry is assumed, meaning if the outcome

vector for the i^{th} subject is denoted $\mathbf{Y_i}$, the distribution is the same for $\mathbf{Y_i} - \mathbf{E}[\mathbf{Y_i}]$ and $\mathbf{E}[\mathbf{Y_i}] - \mathbf{Y_i}$. Under the null hypothesis only two permutations of ranks are possible for each subject: the permutation observed and the exact opposite permutation. The correlation between two measurements on the same subject is preserved in the case of both of these possible permutations (Koch and Sen, 1968).

Koch and Sen's method, like Friedman's test, requires complete and balanced data. No research has been done on specifically adapting Koch and Sen's test to scenarios involving missing data; however, numerous researchers have developed adaptations to Friedman's test that allow for MCAR data (Prentice, 1979; Mack and Skillings, 1980; Skillings and Mack, 1981; Rai, 1987; Wittkowski, 1988). The most widely accepted of these methods involve ranking all non-missing observations within a subject, thereby allowing each subject to have a different number of ranked observations. Missing observations within a subject are excluded from the calculation of the test statistic. To address the loss of information associated with the missing observations, the ranks for subjects with missing observations is inflated so as to ensure relatively equal contribution from each subject in the calculation of the test statistic (Prentice, 1979). This method, which is equivalent to the stratified Mantel-Haenszel test using modified ridit scores, has become a widely accepted method of testing for differences within blocks (Landis, Heyman and Koch, 1978; Stokes, Davis and Koch, 2000). Using this test when the correlation structure is not compound symmetric could lead to an over- or an under-estimation of the covariance structure for Friedman's test statistic. Due to the similarities between Friedman's and Koch and Sen's tests, we propose using these methods for Koch and Sen's test as a method to preserve type I error rate while maximizing power with MCAR data.

2.1.2 Notation, Assumptions and Terminology

We assume a longitudinal study design within the context notation specified in Section 1.2.2 where the j^{th} measurement represents the measurement at the j^{th} time point. Koch and Sen's test focuses on the effect of time when no other covariates are of interest. Therefore, no $\boldsymbol{X_i}$'s are involved in the test statistic. Missing data can be intermittent throughout a study but in longitudinal studies, missing data can also occur as a result of a subject dropping out of the study. With loss to follow up, the missingness is less likely to be MCAR as subjects who drop out are likely doing so due to either good or bad outcomes. Therefore, to address the issue of MCAR data, we will assume a non-monotonic missing data pattern. The set up for this test, as seen in table Table 1.2, involves the ranking of the Y_{ij} values within each subject with r_{ij} denoting the within subject rank of the i^{th} individual at the j^{th} time point. In the case of complete data, it is important to note that $\sum_{i=1}^{n} = \frac{n(n+1)}{2}$.

The aim of this test is to determine if the outcomes are different within a subject without assuming a compound symmetric correlation structure. As mentioned in Section 1.2.2, this test preserves the correlation structure of the observed data as only two possibilities of ranks are possible under the null hypothesis of no difference in measurements.

2.2 Reduced Rank Adjustment

Koch and Sen's test, like Friedman's, was developed for the case of complete and balanced data. In the case of unbalanced and incomplete data, basing a test statistic on the average rank can be misleading, as subjects with missing observations would not have a complete rank vector of numbers from 1 to n. In addition the distributional assumption under the null hypothesis, of two equally likely permutations, is violated.

In the case of missing data, each subject has a different expected rank based on the number of non-missing measurements. Therefore, testing if the average rank of the j^{th} measurement is equal to $\frac{n+1}{2}$ would not be an appropriate test. In the case of Friedman's test, the use the reduced ranks, rather than average rank, has been proposed as a way to adapt this test to scenarios involving missing data. First proposed by Bernard and Van Elteren, this method involves calculating a reduced rank by subtracting the expected rank for the subject from the observed rank (Bernard and Elteren, 1953). This makes the value comparable across subjects with different numbers of observed outcome values.

We propose using the "reduced rank" method that has been widely used in cases of unbalanced and incomplete data in Friedman's test. The reduced rank method takes into account different subjects having a different number of observations and therefore different expected ranks. In the case of unbalanced and incomplete data, using the sum of the reduced ranks is generally accepted as a preferred over the average rank as a method of controlling for subject effect while testing for a difference in measurements.

If the null hypothesis that all measurements are equal is true it is expected the sum of the reduced ranks for each of the j^{th} measurements will be very close to the value of zero. If a particular measurement tends to be higher or lower than the other measurements then the value for the sum of the reduced rank for those measurements will be further away from zero (Bernard and Elteren, 1953).

The generalized form of Koch and Sen's test statistic involves a test of contrasts involving the elements of the vector $\mathbf{T}$. For complete and balanced data, if a contrast matrix with diagonal elements $k - \frac{k}{n}$ and off diagonal elements $-\frac{k}{n}$ is specified this creates a test statistic consisting of the sum of the reduced ranks. The j^{th} element of

the (n-1) vector **CT** can be calculated as follows:

$$kT_j - \frac{k}{n}\sum_{i=1}^{n} T_j = k\frac{1}{k}\sum_{i=1}^{k} r_{ij} - \frac{k}{n}\sum_{j=1}^{n}\frac{1}{k}\sum_{i=1}^{k} r_{ij} = \sum_{i=1}^{k} r_{ij} - \frac{k}{n}\left(\frac{n(n+1)}{2}\right) = \sum_{i=1}^{k}\left(r_{ij} - \frac{n+1}{2}\right)$$

Koch and Sen's test statistic, based on this vector **CT** relies on this value as well as the $Var(\mathbf{CT}) = \mathbf{CVC}'$ (Koch and Sen, 1968). The calculation of this test statistic for this specific contrast relies on the fact that $\sum_{j=1}^{n} T_j = \frac{n(n+1)}{2}$, which is not the case with incomplete data. Therefore, while we must note that our revised test statistic cannot be proposed as a simple contrast matrix, it is important to note it is of similar format.

We propose a test statistic that is based on the vector $\boldsymbol{\mu_K}$ which has elements μ_{Kj} which are shown below:

$$\mu_{Kj} = \sum_{i=1}^{k}\left(r_{ij} - \frac{n_i+1}{2}\right)$$

As proven in Section 1.2.2, $E[r_{ij}] = \frac{n_i+1}{2}$. Therefore,

$$E[\mu_{Kj}] = \sum_{i=1}^{k}\left(E[r_{ij}] - \frac{n_i+1}{2}\right) = 0$$

In the case of the null hypothesis being true, we can determine the asymptotic behavior of this vector in the situation where the number of the measurements remains bounded and only the number of subjects goes to infinity. While the sums of the reduced ranks for all n measurements are not linearly independent, suppose we select the $n-1$ vector where one μ_{Kj} is removed arbitrarily. This results in a $\boldsymbol{\mu_K}$ vector which is composed of $n-1$ linearly independent sums. By the central limit theorem and Lyapunov's condition, the $n-1$ vector μ_K has an asymptotically normal distribution of dimension $n-1$ with a covariance matrix equal to the covariance of the sum of the

reduced ranks. We propose a test statistic $K = \boldsymbol{\mu}_{\boldsymbol{K}}' \boldsymbol{V}_{\boldsymbol{K}}^{-1} \boldsymbol{\mu}_{\boldsymbol{K}}$ where $\boldsymbol{V}_{\boldsymbol{K}}$ denotes the covariance matrix of the $\boldsymbol{\mu}_{\boldsymbol{K}}$ vector. If this is the case, then the test statistic K has a chi-squared distribution with $n-1$ degrees of freedom under the null hypothesis (Koch and Sen, 1968; Sen and Puri, 1967).

2.3 Inflation Factor for Ranks

After the method of reduced ranks was proposed by Bernard and Van Elteren to address problems in Friedman-type statistics, a number of researchers expressed concern about the decrease in power that would result from the loss of information due to the missing data. For this reason, a number of inflation factors were proposed which inflate the ranks of those subjects with more missing data, thereby allowing their contribution to be substantial in comparison with the contribution of subjects with more complete data (Prentice, 1979; Rai, 1987; Wittkowski, 1988)

We propose updating the test statistic based on the vector μ_K with a similar vector calculated based on the inflated ranks, specifically the inflation factor proposed by Prentice. Prentice's weight, $\frac{1}{n_i+1}$, both simplifies variance calculations and allows for each subject to have a more equal contribution to the test statistic. The weighted test statistic will be of similar format to the test statistic K proposed above but will be based on the inflated reduced rank vector $\boldsymbol{\mu}_{\boldsymbol{U}}$ and the variance matrix for this weighted vector. Just as with the $\boldsymbol{\mu}_{\boldsymbol{K}}$ vector, the $\boldsymbol{\mu}_{\boldsymbol{U}}$ vector consists of $n-1$ elements with one μ_{Uj} will be omitted from the vector in order to preserve linear independence. The μ_{Uj} element of this vector is calculated below:

$$\mu_{Uj} = \sum_{\substack{i=1 \\ n_{ij}>0}}^{k} \left(\frac{1}{n_i + 1} r_{ij} - \frac{1}{2} \right)$$

By the same methodology as shown in Section 2.2, we note that the vector μ_U has an asymptotically normal distribution with a covariance matrix $\boldsymbol{V_U}$. We will define $v_{jj'}$ as the element in the j^{th} row and the j'^{th} column of the $\boldsymbol{V_U}$ matrix.

$$v_{jj} = Var(\mu_{Uj}) \quad = \quad \sum_{\substack{i=1 \\ n_{ij}>0}}^{k} \left(\frac{1}{n_i + 1} r_{ij} - \frac{1}{2} \right)^2$$

$$v_{jj'} = Cov(\mu_{Uj}, \mu_{Uj'}) \quad = \quad \sum_{\substack{i=1 \\ n_{ij}>0, n_{ij'}>0}}^{k} \left(\left(\frac{1}{n_i + 1} r_{ij} - \frac{1}{2} \right) \left(\frac{1}{n_i + 1} r_{ij'} - \frac{(n+1)}{2} \right) \right)$$

Calculations for both the covariance and variance elements can be found in Appendix A. The test statistic, U, that we will calculate as $\boldsymbol{\mu_U' V_U^{-1} \mu_U}$ will be asymptotically distributed as chi-squared with $n - 1$ degrees of freedom.

2.4 Simulations

The goal of our simulation study was to maintain generalizability while still varying the simulations enough to be able specify guidelines in which this test statistic would be appropriate. Data were generated from a multivariate normal distribution. In each scenario 10% of the overall observations were set to missing. As we were primarily focused situations where compound symmetry cannot be assumed, commonly the case in longitudinal studies, we assumed an autoregressive correlation structure which corresponds more to longitudinal studies. The correlation between two consecutive measurements was varied, including 0.1, 0.3 and 0.5, to allow for some variation in terms of the degree of correlation. The number of measurements on each subject as well as the number of subjects to vary. We selected all possible combinations of 10, 50 and 100 subjects and 5 and 10 measurements. With this test, as with most non-parametric tests, low

power could potentially be a problem. In order to combat this issue, 0.1 was the lowest correlation tested. For each scenario, 10,000 simulated data sets were generated. The selected combinations can be seen in Table 2.1 below.

In each scenario, 10% of the overall observations were set to missing. The missing

Table 2.1: Data Sets Generated

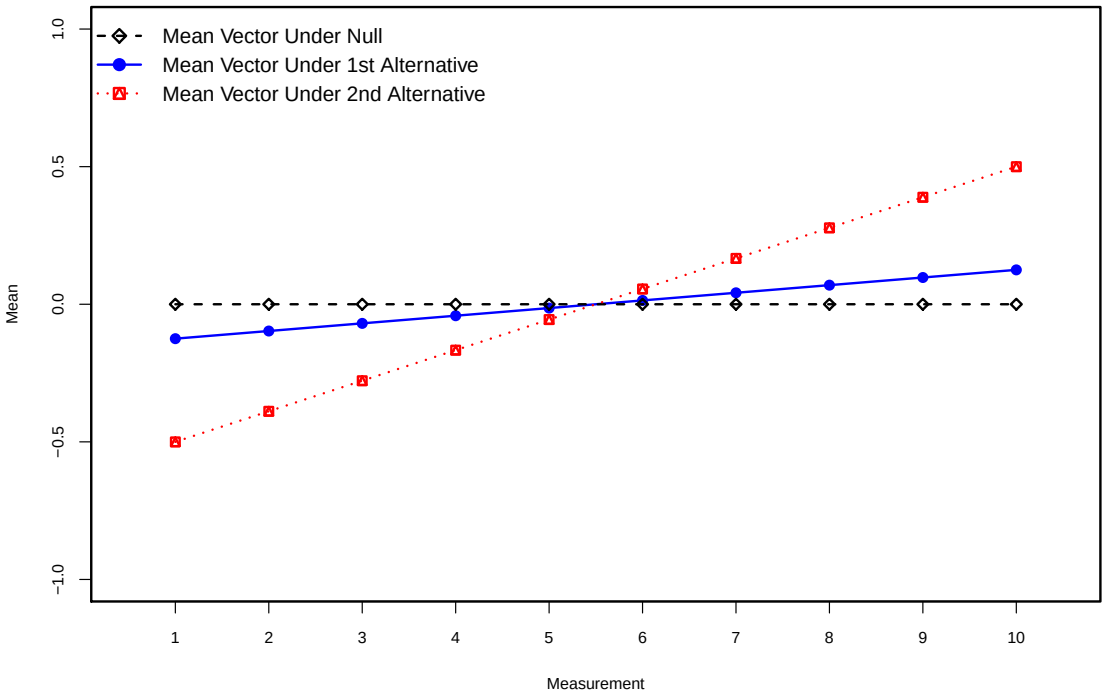
Number of Subjects	Number of Observations Per Subject	Correlation
10	5	0.1
10	5	0.01
10	10	0.1
10	10	0.01
50	5	0.1
50	5	0.01
50	10	0.1
50	10	0.01
100	5	0.1
100	5	0.01
100	10	0.1
100	10	0.01

data were assumed to be missing completely at random. Although it is acknowledged that monotonic missing data does occur in longitudinal studies, only non-monotonic missing data were generated, although this does not lessen the generalizability of our results.

Three different mean vectors were chosen in order to allow for the examination of both the type I error rate and power. The overall mean of all three mean matrices was zero although one of the mean vectors had a linear increase in the mean as a function of observation number. The three mean vectors for the 10 observations per subject can be seen in Figure 2.1. These mean vector with the linear increase in outcome was used to examine power. As some scenarios included a larger number of observations, the

linear increase per observation was minimized as a quarter of the variance of the data, 0.25 and equivalent to the variance of the data, 1.

Figure 2.1: Mean Vectors for Null and Alternative Hypotheses



2.5 Results

2.5.1 Type I Error Rates

For the simulation study, we compared the type I error rates for the updated method, using the generated missing data, to that of the type I error rates of Koch and Sen's test using the original complete data set. The results for the simulations can be seen in Figure 2.2. The type I error rates for our method, which can handle missing data, are comparable to the type I error rates calculated based on the original complete data set.

Therefore, our method does not appear to noticeable change the power of the complete test. The sample size, both the number of subjects and the number of observations for each subject, appear to have the some influence on type I error rates. As expected, having fewer subjects leads to larger problems with type I error rates. In situations with only 10 subjects, the type I error rates appear to be so extremely stringent that the effectiveness of the test is highly questionable. Type I error rates appear to be overly stringent for studies with a larger number of measurements collected on each subject. When 10 measurements were collected on each subject, the type I error rates were noticeably more stringent, although the difference between 5 and 10 measurements is less pronounced when data are collected on more subjects. As the number of subjects increase, the stringent nature of the type I error rate does not appear to be problematic for the test involving the complete data set nor for the test proposed in this paper.

While there were some differences in the type I error rates depending on the pair-

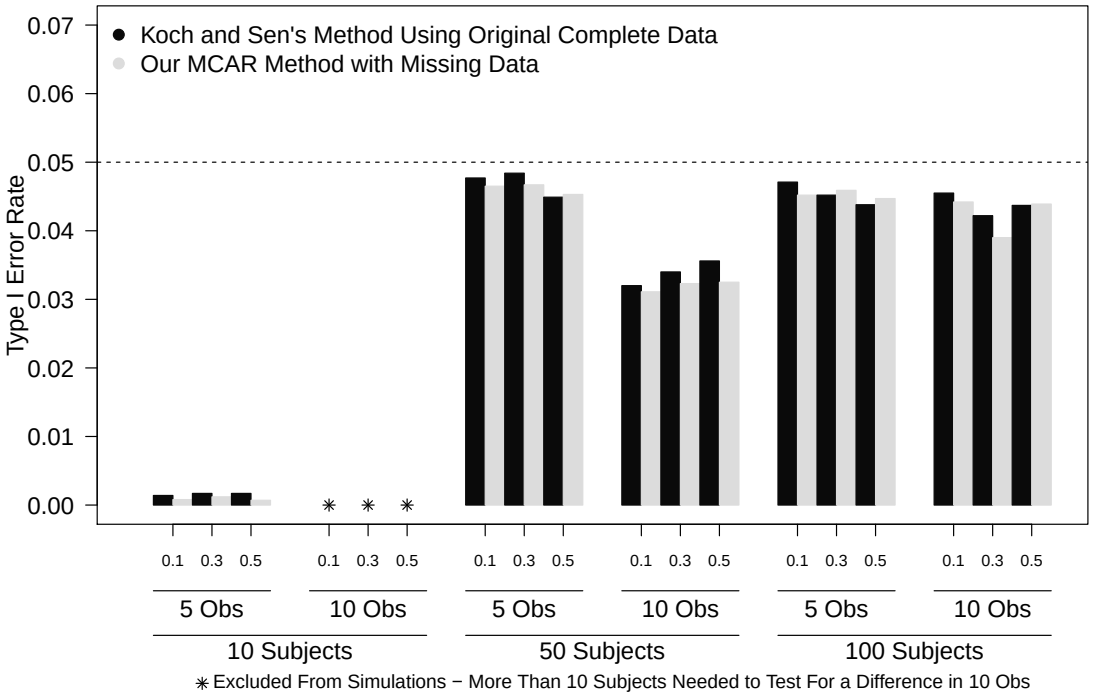


Figure 2.2: Type I Error Rates

wise correlation specified between two sequential measurements, these differences were not as clearly defined nor as noteworthy as the differences due to the number of subjects and number of measurements.

2.5.2 Power

As seen in Figure 2.3 and Figure 2.4 below, the power of the test proposed in this paper is relatively similar to the power of the test even when using the complete data set. Although, in the case of only 10 subjects in the study, both this test and the test using the complete data, have extremely low statistical power with which to detect a difference in outcome measurements.

As expected, an increase in the number of study subjects as well as a more

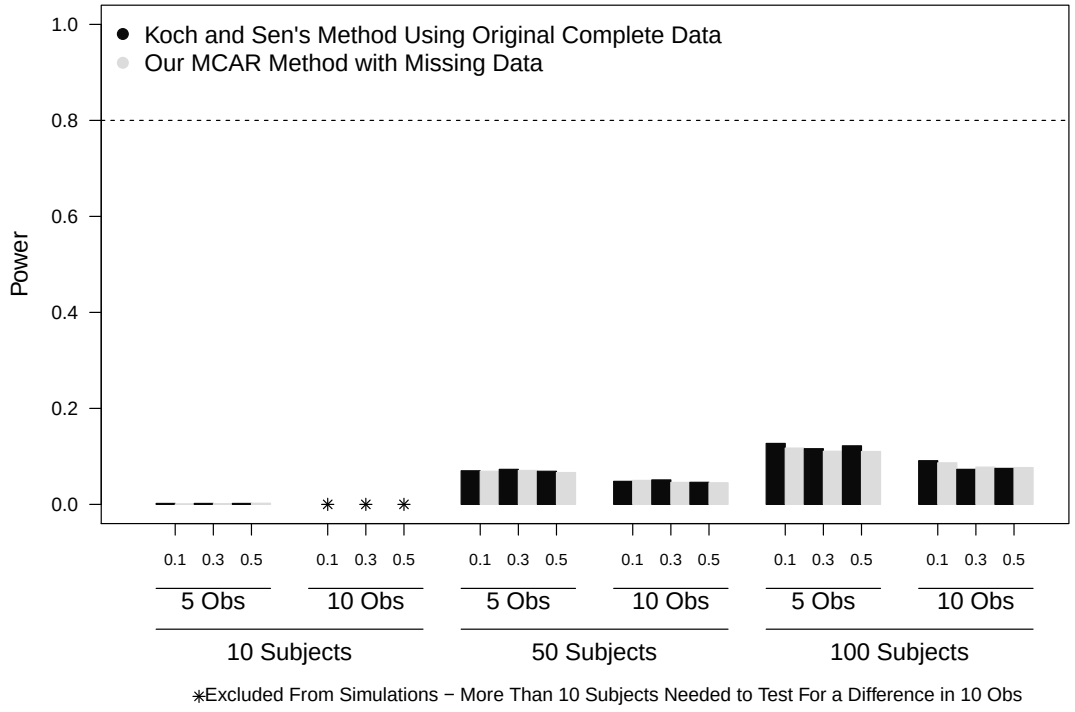


Figure 2.3: Power (Under Linear Increase of 0.25)

extreme alternative hypothesis are associated with an increase in statistical power.

With more measurements per subject, the power appears to decrease. As with the type I error rates, the differences in power between differing strengths of the correlation were minimal in the case of the less extreme alternative shown in Figure 2.3 and in the when the number of subjects was substantially greater than the number of observations collected on each subject in Figure 2.4.

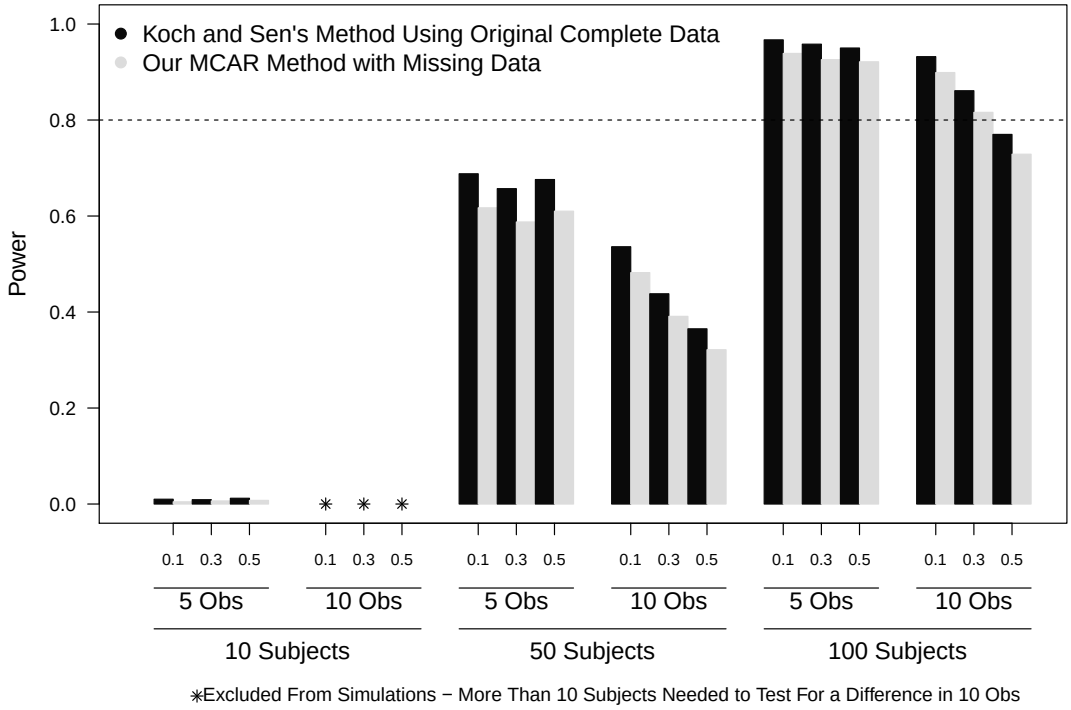


Figure 2.4: Power (Under a Linear Increase of 1)

2.5.3 Asymptotic Behavior

The distribution of our test statistic, based on 5,000 simulations, was examined in order to evaluate the asymptotic behavior of our revised test statistic under the null hypothesis. As mentioned in Section 2.3, under the null our distribution is asymptotically distributed chi-squared distribution with n-1 degrees of freedom. We considered a

scenario with 500 subjects and 5 measurements collected on each subject. The distribution of these test statistics along with the probability density function for a chi-squared distribution with 4 degrees of freedom is shown in Figure 2.5 below.

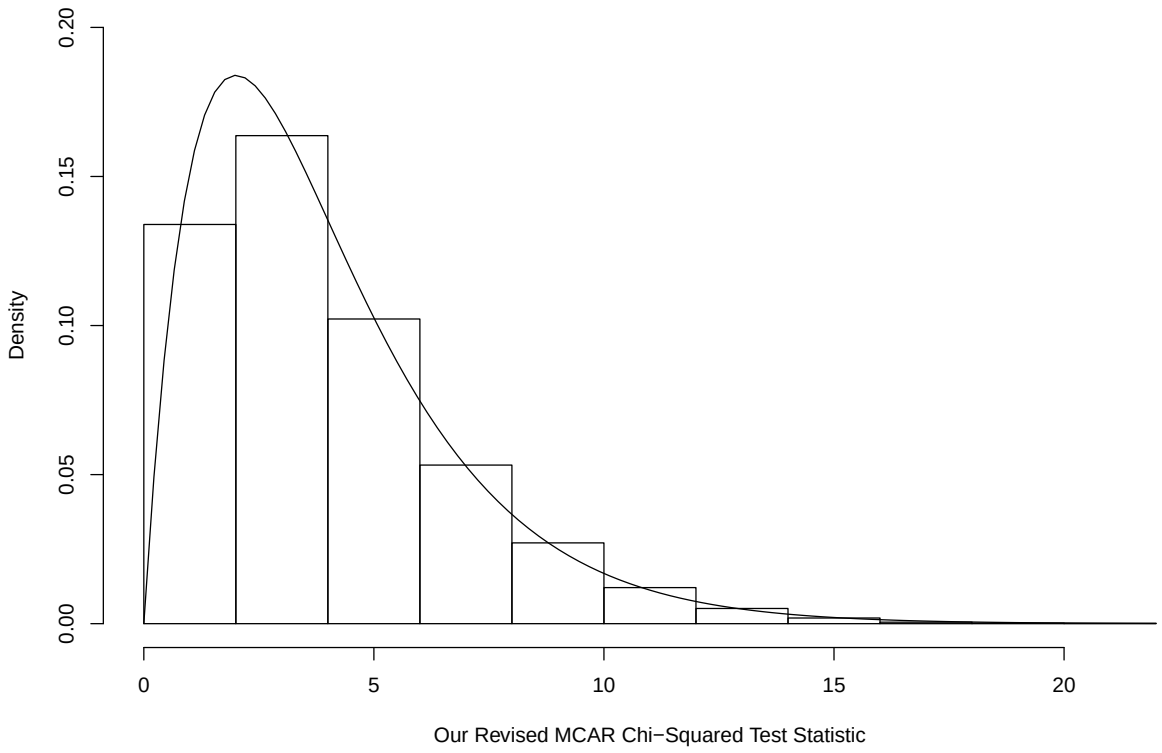


Figure 2.5: Asymptotic Behavior of Our Revised Test Statistic

2.6 Data Example

We will use this method to analyze data from forty-seven individuals who suffer from irritable bowel syndrome (IBS). In this study, participants were asked to report pain, on a scale from 0 to 10, at five times throughout the day: wake up, morning, midday, evening and bedtime. Researchers were interested in determining if levels of pain were different throughout the day. Participants were reminded by alarms to record

their pain at these time points.

Data were collected over a number of visits. At each visit, data was requested for multiple days. With so much data, it is of interest to limit the focus of the study to data from the first visit as well as combine data across all days. The average pain score across all days for wake up, morning, midday, evening and bedtime were calculated for all participants. If a subject was missing pain scores for more than forty percent of the days the average pain score for that period of the day was set to missing for an individual. Using this criteria, there were 14 missing average pain scores, resulting in 5.96% of the data missing. As seen in Figure 2.6 below, the distribution of the average pain score in this study does not appear to be normally distributed. Therefore, the nonparametric test for difference in average pain scores would be ideal, as it requires no distributional assumptions to be made about the average pain score. The average pain score for each period of time can be found in Table A.4 in Appendix A.

The distribution of the average pain score across all periods of the data for both the data set including subjects with missing values and for the data set once those subjects are excluded is shown in Figure 2.7 below. Using the macro given in Appendix A, the revised Koch and Sen's test statistic was calculated to be 5.16 which, with 4 degrees of freedom, yields a p-value of 0.27. Therefore, we fail to reject the null hypothesis. This data set suggests that the average pain score does not significantly differ by period of the day. The only method of using the original Koch and Sen's test statistic would be to remove all subjects with any missing observations. In this case, rather than data on 47 participant, only 38 participants are included in this analysis. The test statistic using only these participants is 6.27 with a p-value of 0.18. Both tests in this case, fail to reject the null hypothesis which the data shown in Figure 2.7 suggest.

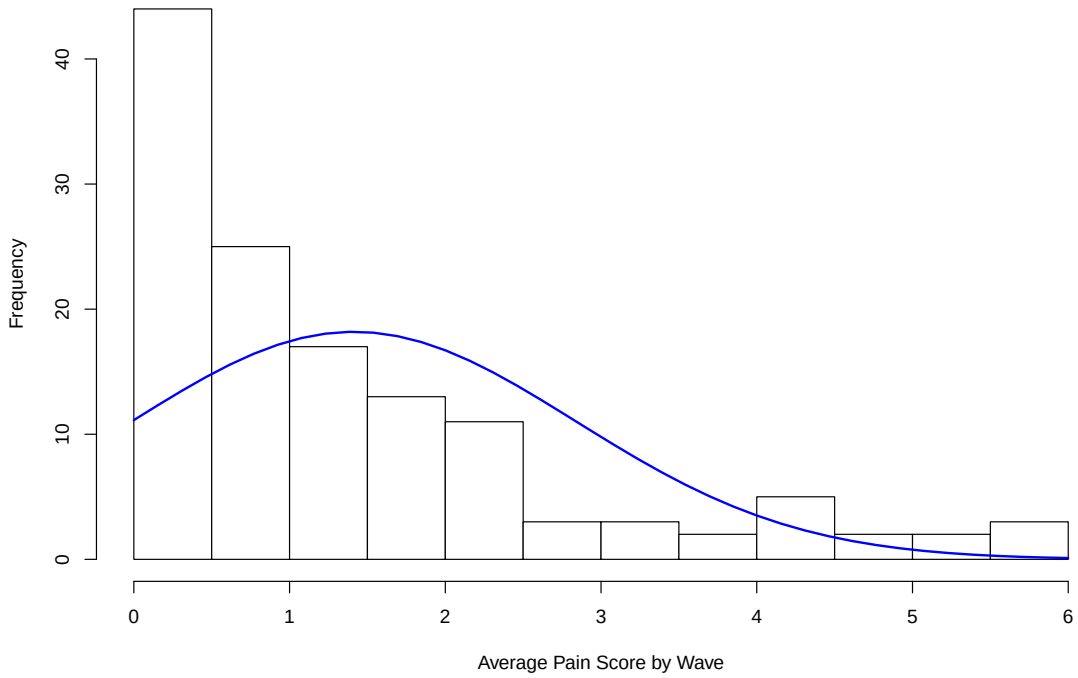


Figure 2.6: Histogram of Average Pain Score by Wave with Normal Curve

2.7 Discussion

Koch and Sen's test is an effective method of testing for a difference in measurements when one does not want to make any assumptions regarding the distribution of the outcome measurement. However, often in the case of studies involving repeated measurements, missing data occurs and Koch and Sen's test can only analyze complete and balanced data. The method proposed in this paper serves as an effective adaptation of Koch and Sen's test to cases with MCAR data. The power and the type I error rates for this method are comparable to analysis done using the complete data. With the smaller correlations examined in this paper, particularly most autoregressive correlation structures, we acknowledge that the power of both the revised test and the complete test provide lower statistical power. For smaller sample sizes, of 10 subjects

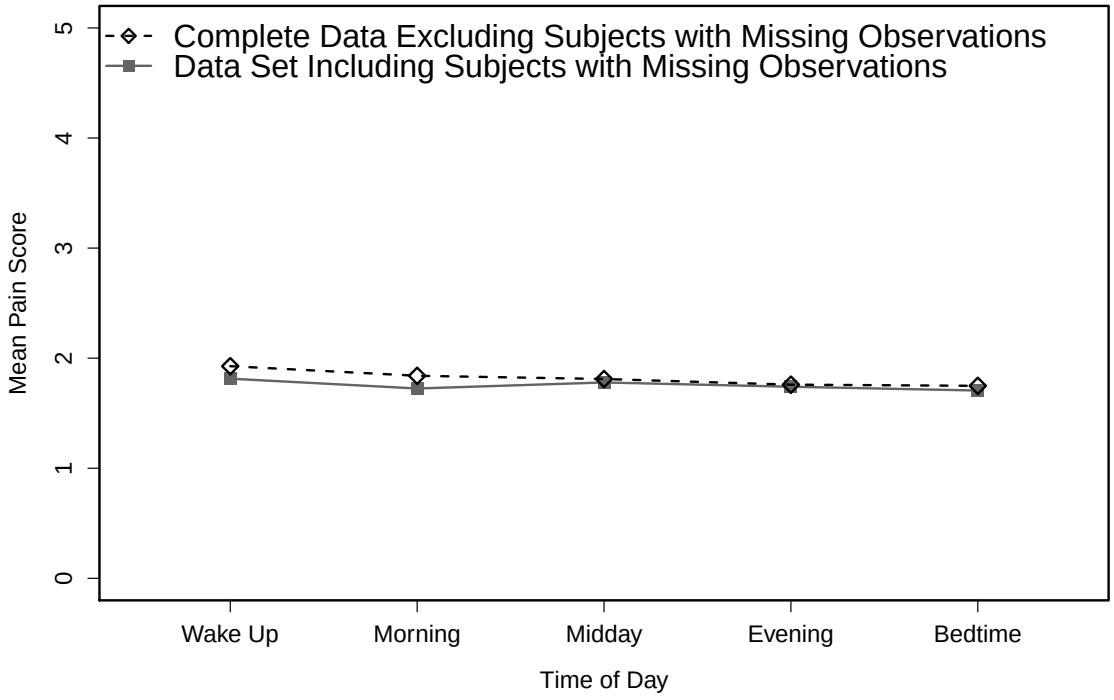


Figure 2.7: Average Pain Score by Period of Day

or less, the power under very minor alternative hypotheses is very small and essentially makes it very difficult to reject the null hypothesis of no difference in measurements when a difference does exist. However, this method, as well as the complete method, provide adequate statistical power when the sample size is greater than 10 subjects. We recommend using this method for testing for differences between measurements on the same subject when there is missing data, the number of subjects is greater than 10 and no distributional assumptions can be made about the outcome variables.

The power of this test presents more problematic issue with this method. Even with relatively large sample sizes (more than 100 subjects), this test only has around 10% power in the case of less extreme alternative hypotheses. For smaller sample sizes, the power is even smaller and essentially makes it very difficult to reject the null hypothesis of no difference in measurements even when a difference does exist. This seems to be

a function of the correlation and therefore with very small correlations, including most autoregressive correlation structures, we recommend using this method with great caution.

Koch and Sen's test was developed to work in situations where the only assumption was that the correlation was not compound symmetric, thereby allowing for a variety of correlation structures. Further research should investigate alternative correlation structures that result in higher correlation between measurements. These may yield higher statistical power for this test. It was felt that an autoregressive structure would be one of the more common correlation structures, as a logical choice for longitudinal studies. However, we acknowledge that this did limit the scope of this research and the investigation into the performance of this test. As Koch and Sen's test allows for any correlation structure, it would be of interest to examine the performance when compound symmetry does hold and compare these results to their Friedman counterparts.

With regards to the covariance calculations used in this paper, it is important to note that the covariance estimates are composed only from data for those subjects with the measurements observed for both the j^{th} and the j'^{th} measurement. It is noted that we could include more information in these calculations by breaking the covariance into the correlation and variance components and allowing for all subjects with a missing j^{th} measurement to contribute to the variance calculations of the j'^{th} measurement for the correlation estimate and vice-versa. For the purpose of this paper, we felt it was important to not use this method due to the increase in the amount of computations that would be required. Using the covariance estimates proposed in this paper, calculating test statistics requires only minor computational adjustments to most major statistical software. In addition by using the covariance estimates proposed in this paper, we allow for situations in which the variance of those with the j^{th} but not the j'^{th} measurement, or vice versa, differed from the rest of the collected data. However, it

would be of interest to compare the performance of the statistic proposed in this paper and the similar statistic using the alternative covariance estimates.

In addition, the method proposed in this paper makes a number of assumptions regarding the missing data, one of the stricter being that the missing data are missing completely at random. When the missing data are not missing completely at random, which is often the case in longitudinal studies, the performance of this test may be called into question and revised methods may need to be developed.

Chapter 3

Informative Missing: Assuming Compound Symmetry

3.1 Introduction

3.1.1 Introduction and Motivation

This research was motivated by study to test for a preference among k objects. All n subjects were asked to rank all k objects. As the outcomes provided in this data set were naturally ranked within each subject, this research question lends itself naturally to non-parametric tests. If one assumes that some subjects felt uncomfortable ranking one or more of the k objects, missing data are present in this scenario. Assuming the unease of the subject was due to those objects being particularly better than the remaining objects, one would expect these missing ranks to be more likely to be assigned higher ranks. As all k objects are independent of each other, we can easily assume a compound symmetric correlation structure.

One of the most common tests for a difference in measurements while controlling for the effect of subject, in the case of equal correlation between any two measurements for the same subject, is Friedman's test. By using within subject ranks to determine if each rank is equally likely for each measurement, the results are not dependent on

the variation between subjects. Under the null hypothesis, each permutation of ranks is equally likely for all subjects.

Friedman's test requires complete and balanced data, which does not often occur in practice. A number of researchers have since developed methods to adjust Friedman's test to address missing data. These methods have been developed under the assumption of MCAR data and have been tested and yield accurate type I error rates in these scenarios involving this type of missingness (Prentice, 1979; Mack and Skillings, 1980; Rai, 1987; Skillings and Mack, 1981; Wittkowski, 1988). This would be applicable if the missingness was due to vials being dropped or lost during the shipping process as these can be assumed to be missing completely at random. However, in scenarios where the measurement is reported as missing when values fall outside of prespecified detection limits, the missingness is informative. If the missing data truly are informative, analyzing the results under this invalid assumption could lead to biased results.

We will examine the performance of these MCAR adaptations to Friedman's test, specifically examining the inflation factor proposed by Prentice, in the case of informative missing data. The focus will be primarily on evaluating the type I error rates and power. In this paper a new method is proposed that aims to be less biased, preserve the type I error rate and improve power in the case of strictly informative missing data. This method will use single imputation to impute the ranks for missing outcome variables rather than removing missing observations from the calculation of the test statistic completely. The information known about the reason for missingness will be used to generate imputed values. For example if it is known higher values cannot be read by the lab equipment, then the imputed value will be the average of the highest possible ranks for that subject. The ranks for each subject will then be weighted by a function of the number of missing observations per subject so that subjects with more missing data will be given less weight to account for the uncertainty associated with

imputation.

3.1.2 Notation, Assumptions and Terminology

We will generalize this research to a study designed with a planned number of measurements, n, to be recorded for each k subjects. The within subject ranks are calculated and in the case of complete data the data can be summarized in the manner specified in Table 1.1. It is assumed that there were n preplanned measurements for each individual. We also assume any missing data are non-monotonic as the actual number of the measurement, j, has no inherent meaning in the calculation of this statistic. Therefore, if a subject has a missing j^{th} measurement than any subsequent measurement $j^{'}$ for $j^{'} > j$ may still be observed.

3.1.3 MCAR Data Using Friedman Methodology

Reduced Rank Adjustment

When analyzing data with a different number of observations collected on each subject, the expected rank for each subject does not remain the same. Therefore Friedman's test if the average within subject rank for each measurement is equal to $\frac{n+1}{2}$ would be inappropriate in this scenario. For situations involving missing data, most methods developed using Friedman-type methods have proposed using Bernard and Van Elteren's method of reduced ranks to resolve this issue (Bernard and Elteren, 1953; Wittkowski, 1988; Mack and Skillings, 1980; Skillings and Mack, 1981; Rai, 1987; Prentice, 1979). By calculating a reduced rank, subtracting the subject specific expected rank from the observed rank, differences between subjects in the number of observed measurement are taken into account. This makes the value meaningful for subjects with different numbers of non-missing outcome values. When the reduced ranks are used it is important to note that any subject with missing data for the j^{th} measurement is not included

in the calculations. Thereby under the null hypothesis any of the permutations from 1 to n_i are equally likely for all the observed measurements.

Rather than averaging the ranks across all subjects, as was done in Friedman's test, the reduced rank method calculates a statistic by summing the reduced ranks over all subjects. A test statistic is calculated based on these values. If the null hypothesis is true and all permutations of ranks, from 1 to n_i are equally likely, the sum of the reduced ranks for each of the j^{th} measurements will be very close to zero. If a particular measurement tends to be higher or lower than other measurements, meaning each rank is not equally likely at each measurement, then the value for the sum of the reduced rank for those measurements will be further away from zero (Bernard and Elteren, 1953).

Inflation Factor

This loss of information due to the missing data can result in a substantial decrease in power. To address this issue, a number of researchers proposed inflating the ranks, and therefore the expected values of the ranks, of individuals with missing observations. This would ensure a greater equality across subjects in terms of a subject's contribution to the calculation of the test statistic. By multiplying the rank by some function of the number of non-missing observations per subject, this method increases the value of a subject's non-missing observations in the calculations in an effort to account for the loss of information due to the missing data. The use of inflation factors results in substantially higher power than Bernard and Van Elteren's original test statistic (Prentice, 1979; Mack and Skillings, 1980; Skillings and Mack, 1981; Rai, 1987; Wittkowski, 1988).

Test Statistic

One of the more common choices for this inflation factor, proposed by Prentice, involves multiplying the reduced rank by $(n_i+1)^{-1}$. We will denote the vector of reduced ranks, when using this inflation factor, as a vector $\boldsymbol{\mu_P}$ consisting of $n-1$ elements. The j^{th} element of this vector, corresponding to the j^{th} measurement, will be denoted as μ_{Pj} and can be specified as follows:

$$\mu_{Pj} = \sum_{i=1}^{k} \left((n_i+1)^{-1} r_{ij} - \frac{1}{2} \right)$$

One μ_{Pj} is arbitrarily omitted so the vector will be of full rank. It is important to note that any subject with a missing j^{th} measurement does not contribute to the calculation of the μ_{Pj} element of this vector.

We denote $\boldsymbol{V_P}$ as the (n x 1) x (n x 1) covariance matrix associated with $\boldsymbol{\mu_P}$. The $v_{jj'}$ element of this matrix denotes the covariance between the j^{th} and $j'th$ measurements. Assuming measurements from different subjects are independent, we can calculate the elements of the covariance matrix as:

$$v_{jj} = Var\left(\mu_{Pj}\right) = \sum_{\substack{i=1 \\ n_{ij}>0}}^{k} \frac{n_i - 1}{12(n_i+1)}$$

$$v_{jj'} = Cov\left(\mu_{Pj}, \mu_{Pj'}\right) = -\sum_{\substack{i=1 \\ n_{ij}>0, n_{ij'}>0}}^{k} \frac{1}{12(n_i+1)}$$

It is important to note that these summations only include subjects with $n_{ij} > 0$ and $n_{ij'} > 0$. Therefore the summation, while specifically from 1 to k does not include k elements unless there are no subjects with missing j^{th} or j'^{th} measurements. The Friedman-type test statistic, $P = \boldsymbol{\mu_p'} \boldsymbol{V_P^{-1}} \boldsymbol{\mu_p}$, is assumed to have a chi-squared distribution with n-1 degrees of freedom under the null hypothesis of no difference in

measurements within a subject (Prentice, 1979).

Application to Informative Missing Data

For complete data, under the null hypothesis, all outcome values are equally likely to be any of the n ranks. When missing data exists, all of the non-missing measurements are equally likely to be any of n_i ranks. The missing measurements are equally likely to be at any of the n time points if the null hypothesis is true. Friedman's test, and Prentice's test statistic based on Friedman's test, tests the hypothesis that the outcomes at each measurement are not significantly different. In the case of informatively missing data, as the number of subjects goes to infinity, the higher missing values will be equally spread out across all n measurements. Therefore, the type I error rate for this the test for a difference in measurements should not be drastically affected by the informative missing data.

Bias associated with $\boldsymbol{\mu_P}$ as an estimator of the vector of reduced ranks, μ, can be explicitly calculated in the case of strictly informative missing data. The expected value of each element of the vector $\boldsymbol{\mu_P}$ is:

$$E[\mu_{Pj}] = E\left[\sum_{\substack{i=1 \\ n_{ij}>0}}^{k}\left(\frac{1}{n_i+1}r_{ij} - \frac{1}{2}\right)\right] = \sum_{\substack{i=1 \\ n_{ij}>0}}^{k}\left(\frac{1}{n_i+1}E[r_{ij}] - \frac{1}{2}\right)$$

For subjects with a non-missing j^{th} measurement, $Pr\left(r_{ij} = a\right) = 0$ for any value where $a > n_i$ and under the null hypothesis all ranks less than or equal to n_i are equally likely for each measurement. Therefore the expected value of r_{ij} is

$$\begin{aligned} E[r_{ij}] &= 1\left(Pr\left(r_{ij}=1\right)\right) + ... + n_i\left(Pr\left(r_{ij}=n_i\right)\right) + ... + n\left(Pr\left(r_{ij}=n\right)\right) \\ &= 1\left(\frac{1}{n_i}\right) + ... + n_i\left(\frac{1}{n_i}\right) + (n_i+1)(0) + ... + (n)(0) = \frac{(n_i+1)}{2} \end{aligned}$$

Subjects with a missing j^{th} measurement are not included in the calculation of the estimate of μ_{Pj}. Based on these calculations of the expected values, we can calculate the bias of Prentice's test statistic based on the true value of μ_j, the reduced rank, which is defined as $\sum_{i=1}^{k}\left[r_{ij}-\frac{n+1}{2}\right]$.

$$E[\mu_{Pj}]-\mu_j=\sum_{\substack{i=1\\ n_{ij}>0}}^{k}\left(\frac{1}{n_i+1}\left(\frac{n_i+1}{2}\right)-\frac{1}{2}\right)-\sum_{i=1}^{k}\left(r_{ij}-\frac{n+1}{2}\right)=-\sum_{i=1}^{k}\left(r_{ij}-\frac{n+1}{2}\right)$$

As k approaches infinity, under the null hypothesis the bias will be zero. Therefore, the test statistic, even in the case of informative missing data, yields an asymptotically unbiased estimator of the reduced rank vector $\boldsymbol{\mu}$. Prentice's method therefore yields an unbiased estimator and a test statistic for which the type I error rate of the test statistic is not greatly impacted by informatively missing data.

While in terms of bias and type I error rate, Prentice's test statistic performs well other substantial problems do arise. If the null hypothesis is in fact not true and one, or multiple, measurements are significantly different from others, then informative missing data can lead to substantial problems with statistical power. In situations with few subjects or only a few measurements reported for each subject, the power for Friedman's test is already low and therefore any decrease in power could potentially be problematic (Friedman, 1937; Stokes, Davis and Koch, 2000). In the scenario proposed in this paper, higher values are more likely to be missing. If many of these higher values are missing this could result in the estimates of the reduced rank to be much lower than they are in reality thereby making the reduced ranks for the measurements closer in range than they truly are.

3.2 Method

The sum of the reduced ranks, when combined with an inflation factor, is a proven and effective non-parametric method of estimating the differences between measurements in the case of MCAR data. However, in the case of informative missing data, by not taking into account the information known about the reason for missingness, Prentice's method results in an unnecessary loss of power. This paper proposes a new method that utilizes reduced ranks, weights and imputation to adapt these Friedman-type tests to informative missing data. It is of interest to compare both the type I error rate and power of Prentice's method with the new method.

3.2.1 Imputation

Prentice's method ranks only non-missing measurements, thereby excluding all information about missing measurements from the calculations. In situations where the probability that a measurement is missing is related directly to the value of the measurement, excluding these values can result in a substantial loss of information. We propose a method which imputes missing ranks based on information known about the reason for missingness. For simplicity, we assume strictly informative missingness data, although it is acknowledged that this is rarely the case in practice. We propose using single imputation methods to address the scenario where higher measurements are more likely to be missing. With minor adjustments, the methods proposed here can also be applied to situations where lower measurements have a higher probability of being missing.

In scenarios where higher measurements are more likely to have missing values, imputation is based on the assumption that non-zero probabilities can be assigned only to the highest $n - n_i$ ranks. Often in these scenarios there is no way to distinguish multiple missing values from each other. Therefore the probabilities of each of the higher

ranks are assumed to be equal for each missing measurement on the same subject. This method assumes the probabilities assigned to each of the highest ranks are $\frac{1}{n-n_i}$. Therefore, the expected value for the missing observations can be calculated as below.

$$\begin{aligned} E[r_{ij}] &= 1\left(Pr(r_{ij}=1)\right) + ... + (n_i+1)\left(Pr(r_{ij}=n_i+1)\right) + n\left(Pr(r_{ij}=n)\right) \\ &= \frac{1}{n-n_i}(n_i+1) + \frac{1}{n-n_i}(n_i+2) + ... + \frac{1}{n-n_i}(n) = \frac{n+n_i+1}{2} \end{aligned}$$

Imputation using this expected value rather than the missing value is done for all missing measurements. For all non-missing observations, the ranks remain the within subject rankings from 1 to n_i. After imputation, it is important to note that the average rank for each subject is now the same as the average rank would be for each subject in the case of no missing data.

3.2.2 Subject-specific Weight for Ranks

After imputation, each subject has a complete set of n ranks. The revised test statistic μ_{Rj} will therefore be based on k elements, one from each subject. Subjects that were missing the j^{th} measurement and therefore were excluded from the calculations of μ_{Pj}, will be included and given the imputed value for calculations of μ_{Rj}. To account for the uncertainty associated with imputation, particularly as this imputation assumes the only reason for missingness is the actual outcome value, a weight is proposed that will assign less weight to subjects with more missing data. These weighted reduced ranks will then be used to calculate the estimate for $\boldsymbol{\mu}$ as well as the statistic for testing if there is a significant difference in measurements within a subject.

The method of the reduced ranks involves subtracting the expected value of the weighted ranks from the actual value of the weighted ranks. After imputation, the

expected value for r_{ij} can be calculated as follows:

$$\begin{aligned} E[r_{ij}] &= \left[1\left(Pr\left(r_{ij}=1\right)\right)+...+n_i\left(Pr\left(r_{ij}=n_i\right)\right)+\frac{n+n_i+1}{2}\left(Pr\left(r_{ij}=\frac{n+n_i+1}{2}\right)\right)\right] \\ &= \left[1\left(\frac{1}{n}\right)+...+n_i\left(\frac{1}{n}\right)+\frac{n+n_i+1}{2}\left(\frac{n-n_i}{n}\right)\right]=\frac{n+1}{2} \end{aligned}$$

If the weight assigned to the ranks of the i^{th} subject is denoted as w_i then $\boldsymbol{\mu_R}$, the revised estimate of $\boldsymbol{\mu}$, can be expressed in general terms with elements, μ_{Rj}, defined as the weighted rank minus the expected value of the weighted rank under the null hypothesis. Just as with Prentice's statistic, one μ_{Rj} is omitted to insure the vector is of full rank.

$$\mu_{Rj}=\sum_{i=1}^{k} w_i\left(r_{ij}-\frac{n+1}{2}\right)$$

It is important that non-missing values are given higher weights than missing values, as there is a degree of uncertainty in imputation. Therefore we will assign a subject specific weight which assigns less weight to individuals with more missing data. For subjects with many missing measurements, imputation adds very little information and therefore we want to ensure the weight for these subjects is much smaller. We propose a weight $\frac{1}{n-n_i+1}$ which gives subjects with complete data the same weight they would be given in the case of complete data. However, subjects with very few non-missing observations are assigned a very small weight to account for the high level of uncertainty. Based on this weight, our estimate of the elements of the $\boldsymbol{\mu_R}$ vector becomes the following:

$$\mu_{Rj}=\sum_{i=1}^{k}\frac{1}{n-n_i+1}\left(r_{ij}-\frac{n+1}{2}\right)$$

3.2.3 Test Statistic

The hypothesis test proposed by Prentice involves testing the null hypothesis that the inflated ranks for the j^{th} measurement are on average close to the expected value of the inflated ranks, which is equivalent to testing if $\boldsymbol{\mu_P} = \mathbf{0}$. In a similar fashion, if the weighted ranks of the method proposed in this paper are close to their expected value, then $\boldsymbol{\mu_R} = \mathbf{0}$. Therefore, this revised estimate of the reduced ranks can be used to test the same hypothesis as Prentice's statistic.

We denote $\boldsymbol{V_R}$ as the $(n-1)$ x $(n-1)$ covariance matrix associated with $\boldsymbol{\mu_R}$ with the element in the j^{th} column and j'^{th} row denoted as $v_{jj'}$. We can calculate this covariance matrix, after first calculating the variance and covariance of the actual ranks.

$$\begin{aligned} Var[r_{ij}] &= \frac{n_i^3 + 3n^2 n_i - 3n n_i^2 - n_i}{12n} \\ &= \frac{(n_i - n))^3 + (n^3 - n_i)}{12n} \end{aligned}$$

and

$$Cov[r_{ij}] = E[r_{ij}, r_{ij'}] - E[r_{ij}]^2 = -\frac{n_i^3 - 3n n_i^2 + 3n^2 n_i - n_i}{12n(n-1)}$$

Based on these calculations, since ranks from different subjects are independent,

$$\begin{aligned} Var\left(\mu_{Rj}\right) &= \sum_{i=1}^{k} \left(\frac{1}{n - n_i + 1}\right)^2 \left(\frac{(n_i - n)^3 + (n^3 - n_i)}{12n}\right) \\ &= \sum_{i=1}^{k} \left(\frac{1}{n - n_i + 1}\right)^2 \left(\frac{n_i^3 + 3n^2 n_i - 3n n_i^2 - n_i}{12n}\right) \\ Cov\left(\mu_{Rj}, \mu_{Rj'}\right) &= -\sum_{i=1}^{k} \left(\frac{1}{n - n_i + 1}\right)^2 \left(\frac{n_i^3 - 3n n_i^2 + 3n^2 n_i - n_i}{12n(n-1)}\right) \end{aligned}$$

The explicit calculation for these values can be found in Appendix B. By the central limit theorem and Lyapunov's condition, the $n-1$ vector μ_R has an asymptotically normal distribution of dimension $n-1$ with a covariance matrix composed of the elements specified above. Therefore, the revised Friedman-type test statistic, $R = \boldsymbol{\mu}_{\boldsymbol{R}}' \boldsymbol{V}_{\boldsymbol{R}}^{-1} \boldsymbol{\mu}_{\boldsymbol{R}}$, has a chi-squared distribution with $n-1$ degrees of freedom under the null hypothesis (Koch and Sen, 1968; Sen and Puri, 1967).

3.2.4 Calculation of Bias

Bias associated with $\boldsymbol{\mu}_{\boldsymbol{R}}$ as an estimator of the vector of reduced ranks, $\boldsymbol{\mu}$, can be explicitly calculated in the case of strictly informative missing data. The expected value of each element of the vector $\boldsymbol{\mu}_{\boldsymbol{R}}$ is:

$$E[\mu_{Rj}] = E\left[\sum_{i=1}^{k}\left(\frac{1}{n-n_i+1}\right)\left(r_{ij}-\frac{n+1}{2}\right)\right] = \sum_{i=1}^{k}\left(\frac{1}{n-n_i+1}\right)\left(E[r_{ij}]-\frac{n+1}{2}\right)$$

Under the null hypothesis, as shown in Section 3.2.2, when using this revised test statistic, the expected value of $r_{ij} = \frac{n+1}{2}$. Therefore, under the null hypothesis $E[\mu_{Rj}] = 0$. Therefore the bias of the test statistic is easily calculated below:

$$E[\mu_{Rj}] - \mu_j = 0 - \sum_{i=1}^{k}\left(r_{ij}-\frac{n+1}{2}\right)$$

Just as with Prentice's statistic, this value approaches 0 as k goes to infinity so both Prentice's method as well as the revised method proposed in this paper produce asymptotically unbiased estimates of the reduced rank vector. Therefore, other methods must be used to compare the two test statistics.

3.2.5 Comparison of Type I Error Rate

In the case of informative missing data, it is important to note that $Pr(r_{ij}$ is missing) $= Pr(r_{ij} > n_i) = \frac{1}{n-n_i}$ for all n measurements under the null hypothesis. Therefore, regardless of the value of j, the missing values are excluded with equal probability from all n measurements. By the same justification, the inclusion of the imputed value is equally likely to be at any value at any of the n measurements, again with probability $\frac{1}{n-n_i}$ for all n measurements. Therefore when testing if one or more measurements has a significantly higher reduced rank than the others, is not affected if the informatively high values are excluded with equal likelihood from each measurement. Therefore, whether the missing value is excluded from the test statistic, or an imputed value is substituted, under the null hypothesis this does not change the probability that one would reject the null hypothesis.

3.2.6 Comparison of Power

In unbalanced designs, the Pitman efficiency can be largely dependent on the alternative hypothesis (Prentice, 1979). Therefore, in an effort to generalize our results, we will address the comparison of power between our revised method and Prentice's method using less explicit guidelines. The general alternative hypothesis of our statistical test is that measurements within a subject are statistically different. As mentioned previously, we are only considering situations where the differences within a subject are constant across all subjects.

In these scenarios, the alternative hypothesis requires the reduced ranks of one or more measurements to be significantly greater than the reduced rank of others. In this situation, it is important to note that $Pr(r_{ij}$ is missing) is not the same for all n measurements. This probability, under the alternative hypothesis, is dependent on the measurement number j. Without losing generalizability we cannot specifically give an

explicit function of this probability.

However, we can note that when using Prentice's method, all observations with a missing value are excluded from the test statistic. In the case of strictly informative missing data, this would exclude the highest reduced ranks from the value of the reduced rank for the j^{th} measurement thereby making this reduced rank closer to the reduced rank of the other measurements. In doing so, this makes it harder to reject the null hypothesis when in fact the alternative is true.

In contrast, although the revised method does not impute the highest values, it does minimize the difference in the reduced rank for that measurement. Therefore using this statistic allows for the clearer delineation between the reduced ranks and thereby increases the power as compared to Prentice's test statistic.

3.3 Simulations

3.3.1 Generation of Data sets

To narrow the scope of our research, only a few factors were allowed to vary so specific guidelines could be established. For each combination of factors 10,000 complete data sets were generated. Of primary concern was how well the method would work in scenarios with a blend of informative missing data and missing completely at random data. We allowed the percentage of these two types of missingness to vary while maintaining the same amount of overall missing data. All data were generated from a multivariate normal distribution. In each scenario 10% of the overall observations were set to missing and then a certain percentage out of this 10% was set to be informatively missing data and the remaining was set to be MCAR. Therefore each data set generated had the same number of missing observations, although some of those had a higher percentage of the missingness generated by informative missingness. As our derivations

of the revised test statistic assumed 100% of the missing data were informative this was the first option selected. However, we also chose to examine scenarios where 80%, 50% and 20% out of the total 10% missing data were forced to be informative missing data.

While the percent of the missing data that were informative was of primary interest, we also were interested in varying the number of subjects and the number of observations within each subject. It was of interest to examine the performance of our test in a relatively small sample scenario as well as a more moderate sample size. Therefore, the number of subjects considered included 10 and 50 and the number of measurements per subject for each of those options were 5 and 10 respectively. Table 3.1 below illustrates all of the variations generated in our scenarios.

Table 3.1: Data Sets Generated

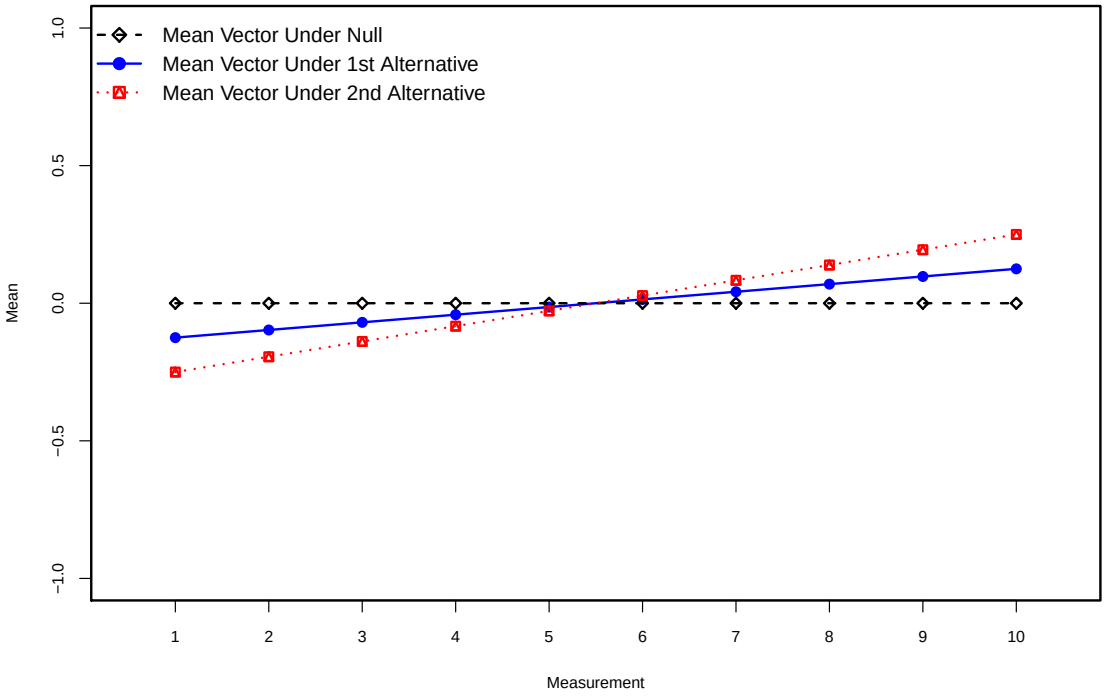
Number of Subjects	Number of Observations Per Subject	% Of Missing Data that is Informative
10	5	100%
10	5	80%
10	5	50%
10	5	20%
10	10	100%
10	10	80%
10	10	50%
10	10	20%
50	5	100%
50	5	80%
50	5	50%
50	5	20%
50	10	100%
50	10	80%
50	10	50%
50	10	20%

Continuous outcome variables were generated from the same multivariate normal distribution, with the covariance matrix calculated based on the specification of the variance and correlation. For generalizability we standardized our outcome variable with the mean of each measurement to be zero and variance of each measurement to be one. As measurements taken on the same individual at the same time point are likely to be highly correlated the correlation matrix was compound symmetric with $\rho = 0.9$. Based off of the mean and covariance matrix, multivariate normal data were generated using the RANDNORMAL function in PROC IML.

Once the complete data set was generated, the data were sorted by outcome and the highest observations, up to the prespecified number of informative missing observations, were set to missing. While we acknowledge that informative missing data realistically involve increasing the probability of a subject being missing, we felt that our method allowed for optimal control over the percent that is truly informatively missing. The remaining amount of missing data, if any, was then eligible to be set to missing according to MCAR patterns. Using PROC SURVEYSELECT, a simple random sample of the remaining non-missing observations were randomly selected. Those selected were then set to missing.

Three different mean vectors were chosen in order to allow for the examination of both the type I and power. All three mean vectors, in the case of 10 observations collected on each subject, can be seen in Figure 3.1 below. The overall mean of all three mean matrices was zero although two of the three mean vectors had a linear increase in the mean as a function of observation number. These two mean vectors with the linear increases in outcome were used to examine power. As some scenarios included a larger number of observations, the linear increase per observation was minimized to half the variance of the data and a quarter of the variance of the data, 0.5 and 0.25 respectively.

Figure 3.1: Mean Vectors for Null and Alternative Hypotheses



3.3.2 Imputation

The single imputation method first ranked the observations within a block from one to the total number of non-missing observations within that block. The remaining ranks that were not assigned, which consisted of the highest ranks for that subject, were averaged and this average rank was imputed as the value for all missing observations within that block. As mentioned in Section 3.2.1, this was equivalent to assigning all missing values the value $\frac{n+n_i+1}{2}$.

3.3.3 Calculation and Comparison of Type I Error Rate

Prentice's weighting of the reduced ranks is an option included in many statistical packages. These adjusted test statistics were calculated using PROC FREQ in SAS 9.2

with the additional specification of CMH2 and SCORES=MODRIDIT.

The revised method proposed in this paper can be calculated using Friedman type test statistics using the weighted ranks after imputation rather than the ranks. Using the weighted rank is equivalent to weighting the reduced ranks. The outcomes must first be ranked and weighted and then test statistics and corresponding p-values can be calculated using PROC FREQ with the CMH2 option and SCORES=TABLE.

For both these methods, type I error rates were calculated. Using a standard rejection level of 0.05, if the p-value for the test statistic fell below 0.05, the test rejected the null hypothesis. The type I error of these methods were calculated as the total percentage of data sets for which the null hypothesis was rejected when there was in fact no linear change in the outcome variable over time. Power was calculated in a similar fashion, under the two alternative hypotheses.

3.4 Results

3.4.1 Type I Error Rate

The type I error rates for both Prentice's method as well as the method presented in the paper are shown in Figures 3.2 and 3.3. For reference, the type I error rates for the original complete data set are shown as well. Overall, the type I error rates did not appear to be very different between the two methods that can handle missing data. There appears to be a slight inflation of the type I error rate when using the revised method proposed in this paper in the case of 50 subjects and 5 observations with 100% of the missingness being informative. Also in the case of 50 subjects and 10 observations with 80% of the missingness being informative, there is a slight inflation of the type I error rate. Other than those two cases, for all three tests, the type I error rate was well preserved and under 0.05. The exact type I error rates are presented in

tabular format in Appendix B. Type I error rates were not calculated for the scenario of 10 subjects and 10 observations per subject as data on more than 10 subjects is needed to test for a difference in 10 measurements.

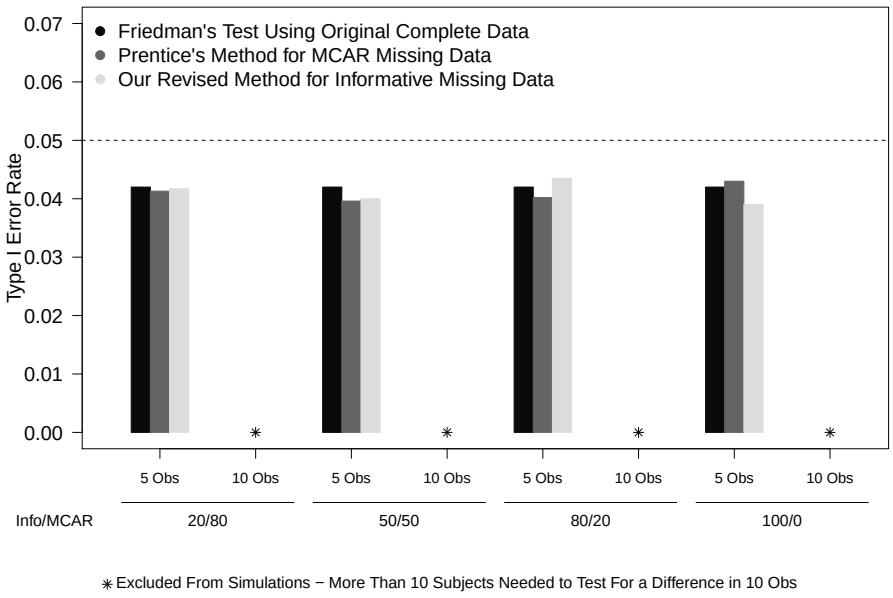


Figure 3.2: Type I Error Rates by % Informative Missing - 10 Subjects

3.4.2 Power

The power calculations for both the proposed method and Prentice's under an alternative of a linear increase of 0.25 across all n measurements, are shown in Figures 3.4 and 3.5. Under this less extreme alternative, in the case of 10 subjects, the power of our revised test is very similar to that of Prentice's test. As expected, for smaller sample sizes, the power for Prentice's test statistic, our revised test statistic and the original complete data set, is relatively low. For 50 subjects, our method shows a slight improvement in power over that of Prentice's method when 100% of the missing data are informatively missing and in the case of 5 observations per subject, our revised method yields a relatively similar power to that of Prentice's test when at least 50% of

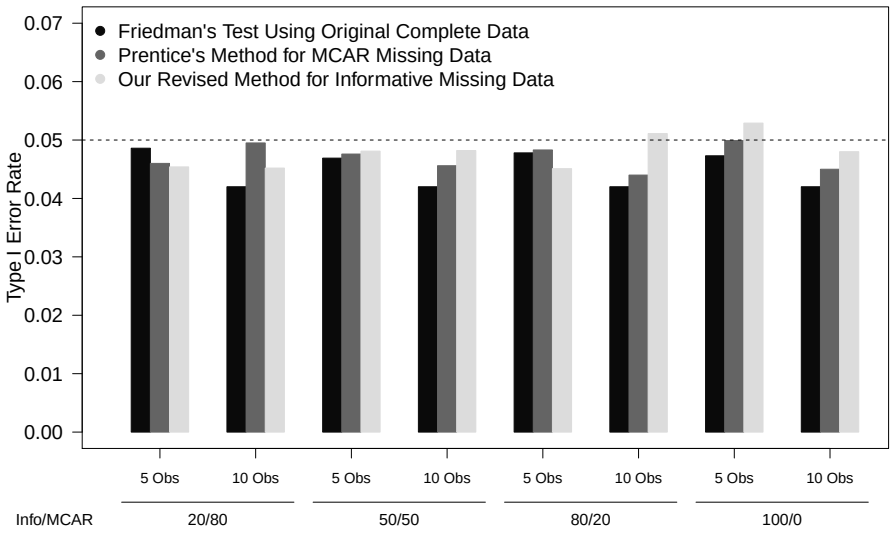


Figure 3.3: Type I Error Rates by % Informative Missing - 50 Subjects

the missingness is informatively missing.

Power for scenarios under an alternative of a linear increase of 0.50 are shown in Figures 3.6 and 3.7. In the case of this more extreme alternative hypothesis, with 10 subjects, our revised method shows an improvement in power over Prentice's test when 100% of the data are informatively missing and comparable power to Prentice's method when at least half of the missing data is informatively missing. For a sample size of 50, this extreme alternative results in extremely high power with almost equivalent power for all three tests.

3.4.3 Asymptotic Behavior

The distribution of our test statistic, based on 5,000 simulations, was examined in order to evaluate the asymptotic behavior of our revised test statistic under the null hypothesis. As mentioned in Section 3.2.3, under the null our distribution is

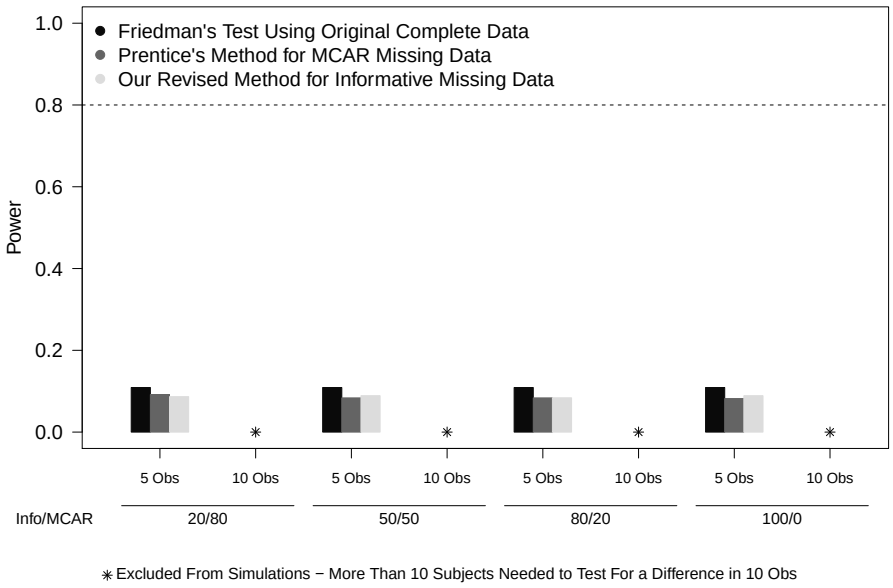


Figure 3.4: Power by % Informative Missing (Increase of 0.25) - 10 Subjects

asymptotically distributed chi-squared distribution with n-1 degrees of freedom. We considered a scenario with 250 subjects and 5 measurements collected on each subject. The distribution of these test statistics along with the probability density function for a chi-squared distribution with 4 degrees of freedom is shown in Figure 3.8 below.

3.5 Data Example

A data set composed of the rankings of k objects by n subjects, as proposed in Section 3.1.1 was used. This data set set is of an identical structure to that used for Kendall's test involving n objects being ranked by k judges (Kendall and Smith, 1939). The data set proposed by Kendall was used to test for agreement and concordance, although the same type of data set could be used to test for a difference in ranks while controlling for subject.

We have data collected from a Kendall-type scenario where 20 individuals each

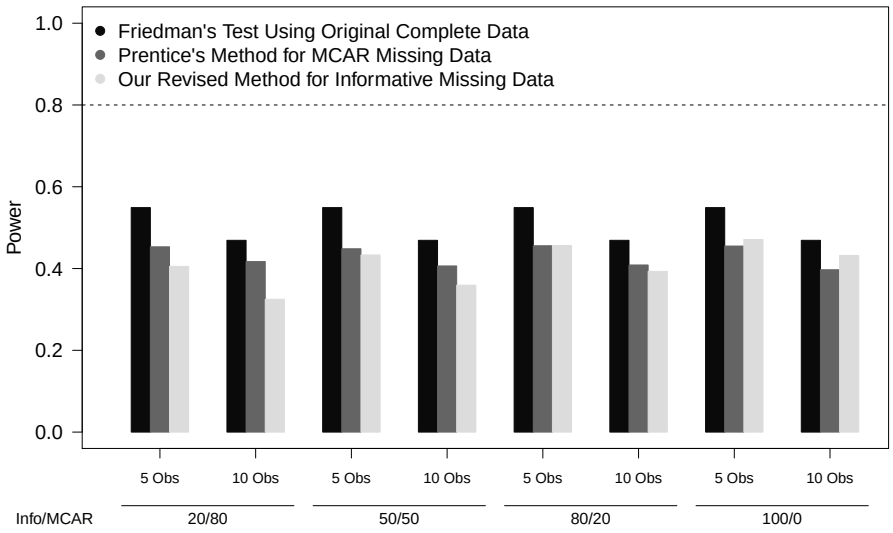


Figure 3.5: Power by % Informative Missing (Increase of 0.25) - 50 Subjects

ranked 4 objects. This study sought to gather data to test the hypothesis that each rank is equally likely for each object, combining data across all subjects. Since there are four distinct objects we expect the true correlation between rankings of the same subject to be equal between any two objects. This allows for the assumption of compound symmetry to be made. The within-subject ranks are shown in Table B.4 in Appendix B. As the data are already ranked within each subject, non-parametric within subject tests are a preferred analysis method. When dealing with the complete data set, Friedman's test would be an effective method of testing for a difference in ranks while controlling for any potential subject effect.

In truth the rank each subject gave each object is known. However, suppose not all subjects felt comfortable ranking all objects. In particular, suppose the second object was of a slightly higher quality, and therefore some subjects felt uncomfortable ranking this object. We can therefore suppose that the missing data in this study are, at least in some substantial part, informatively missing as the missing values are more likely to

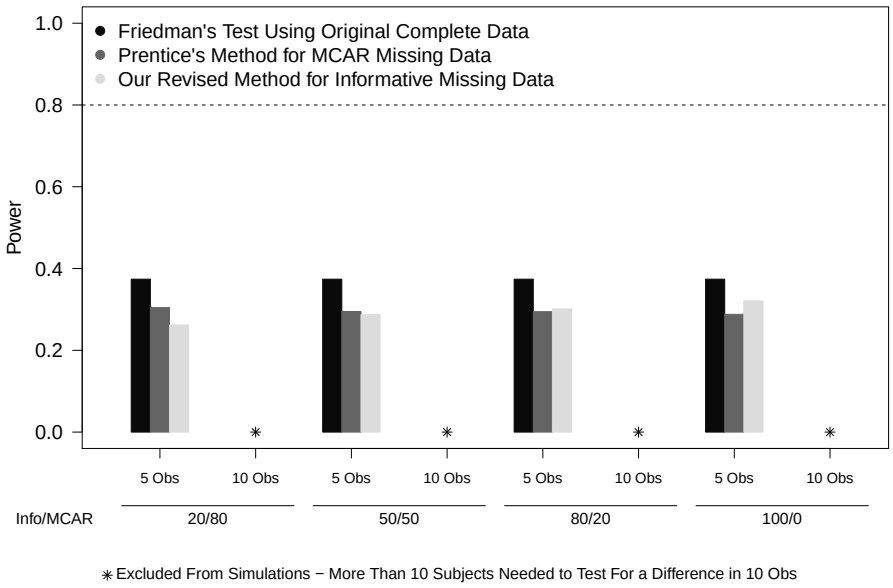


Figure 3.6: Power by % Informative Missing (Increase of 0.50) - 10 Subjects

be higher values. The data set with missing data is shown in Table B.5 in Appendix B as well.

As all ranks were actually collected on all subjects, we can compare both Prentice's method as well as the method proposed in this paper to the results from Friedman's test done with the complete data set. As seen in Figure 3.9 below, the original complete data set appears to show slightly more differences between objects than the data set with simulated missing data. When Friedman's test is performed on the complete data set, the test statistic is 5.34, which with 3 degrees of freedom yields a p-value of 0.1485. Prentice's test statistic, intended for MCAR missing data, provides a test statistic of 1.98, which with the same degrees of freedom, yields a p-value of 0.5756. The revised test statistic proposed in this paper, gives a value closer to the true value calculated using the complete data set. The test statistic is 2.47 with a corresponding p-value of 0.4801. Here, unlike in many studies, we can compare ranks between the missing and the observed data. The missing data appear to be informative, as 62.5% of the missing

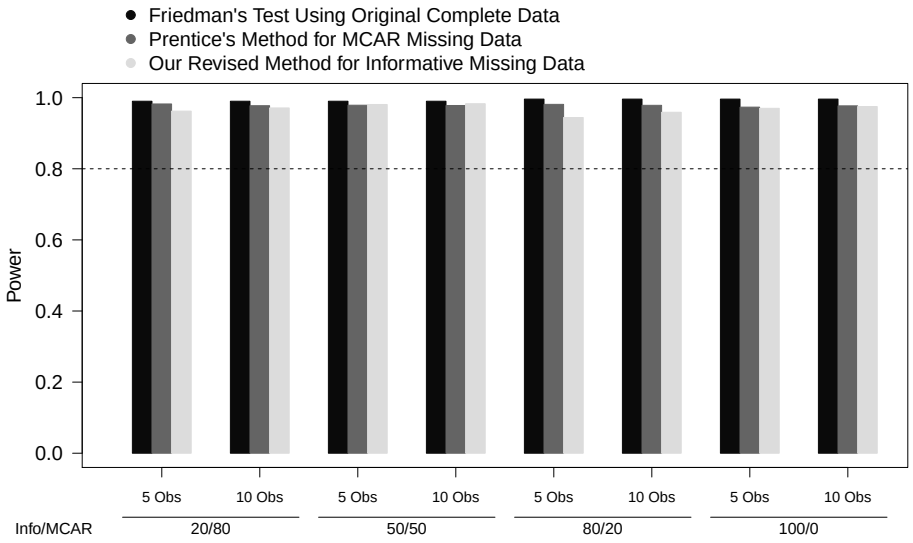


Figure 3.7: Power by % Informative Missing (Increase of 0.50) - 50 Subjects

data has a true rank of four, compared to the observed data where only 20.8% was given a rank of four. Therefore, as missing values are more likely to be of higher ranks, the revised method produces a test statistic closer to the actual value and therefore in this case this test statistic would be preferred over Prentice's.

3.6 Discussion

Our method offers an improvement in power in certain scenarios, which in the case of a within subject test for differences, can be an important improvement to current methods. When the number of subjects is substantially larger than the number of measurements collected on each subject, our method proves power that is at least as high, if not higher, as Prentice's test when at least half of missing data are informatively missing. The most marked improvements appear to be when the power is not extremely high nor extremely low, in this case in the situations of the less extreme alternative and

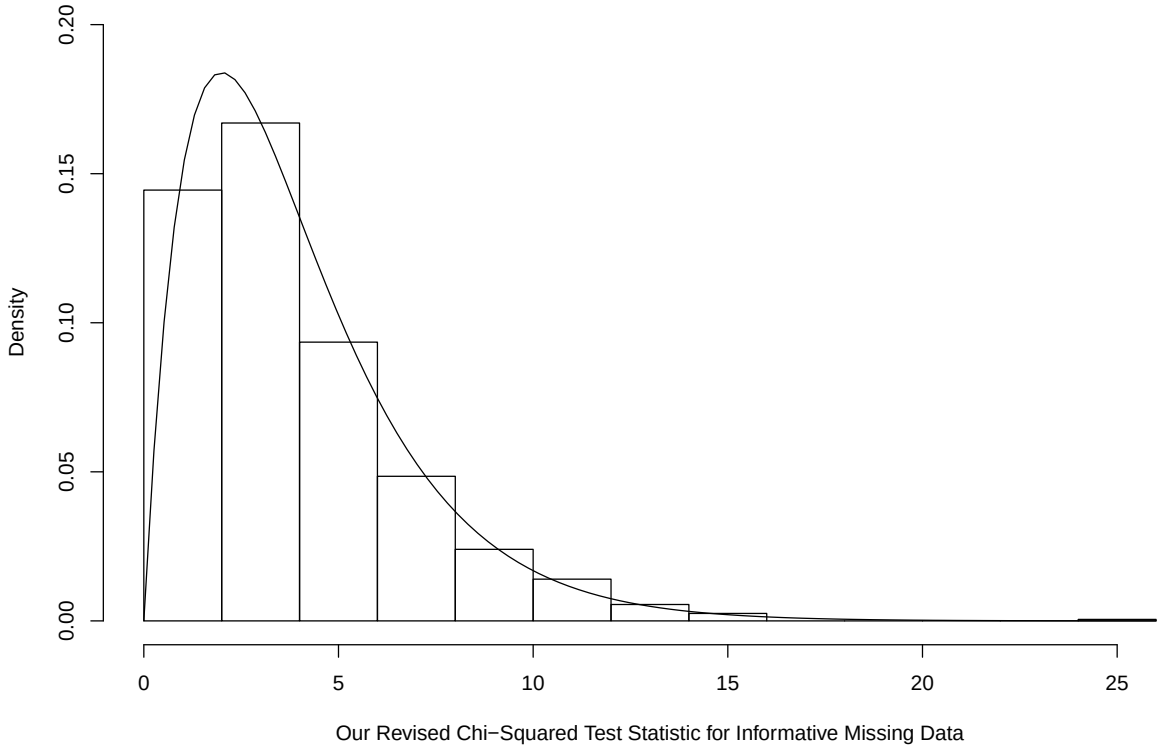
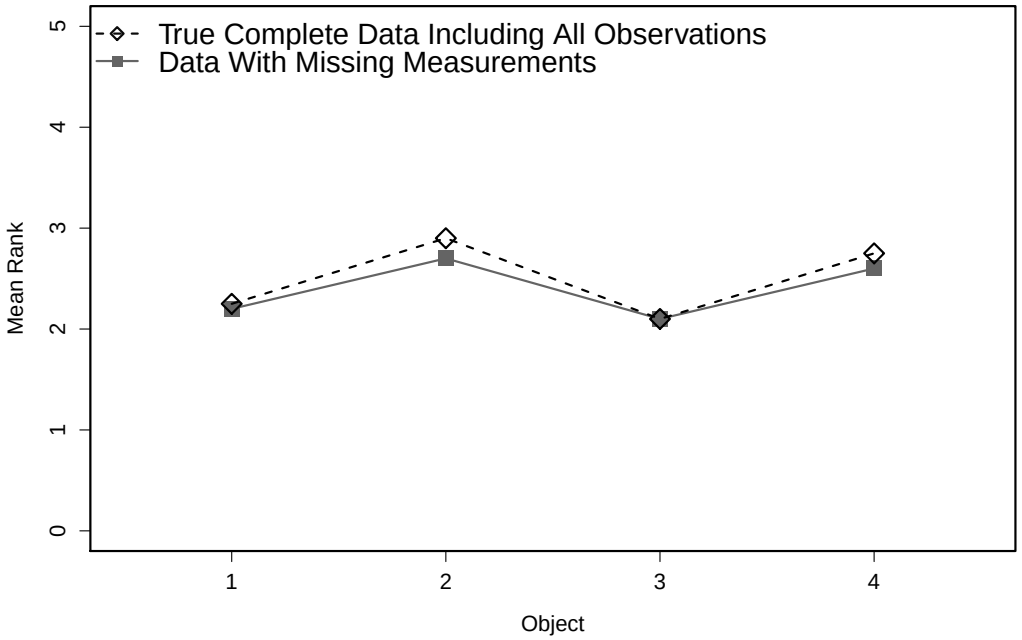


Figure 3.8: Asymptotic Behavior of Our Revised Test Statistic

higher sample sizes and in the case of the more extreme alternative and smaller sample sizes. However, it is important to note that the revised method proposed in this paper can result is slightly inflated type I error rates when the sample size is moderately large.

It is important to note that making assumptions about the percent of the missing data that are truly informative missing is difficult to do and, as such, this presents itself as one limitation of the guidelines proposed in this paper. Additionally, we must acknowledge that due to the large number of possibilities, we could not fully evaluate the performance of our revised test statistic under all alternative hypotheses. The alternative hypotheses for our tests were meant to evaluate a smaller departure from the null hypothesis and a relatively drastic departure. In addition both of our alternative hypotheses tested the performance in the case of a linear increase in outcome. Further

Figure 3.9: Mean Rank for Each Object



research should be done to test the performance of this test under a wider variety of alternative hypotheses including both different levels of linear increases as well as non-linear alternative hypotheses.

Chapter 4

Informative Missing: Without Assuming Compound Symmetry

4.1 Introduction

4.1.1 Introduction and Motivation

This research is motivated by longitudinal studies in which repeated measurements are collected on the same subject over a period of time. The study is interested in determining if the difference in pre and post-bowel pain scores differ throughout the day. Based on the time stamp of the pre-bowel movement pain score, the difference in pain score was categorized into early morning, morning, afternoon or evening. Researchers were interested in testing if the difference was the same throughout the day. Measurements on the same subject are expected to be correlated and therefore any subject effect must be accounted for; however, differences between subjects are not of interest.

In scenarios where compound symmetry cannot be assumed, one method of testing this hypothesis has been proposed by Koch and Sen (Koch and Sen, 1968). Using similar methods and assumptions to those proposed by Friedman, this method tests for a difference in measurements while controlling for any potential subject effect. However,

the test requires complete data, which can be rare when dealing with longitudinal data. Missing data in these studies can be MCAR, with missingness due to patients missing observations for reasons unrelated to their outcomes. In Chapter 2 of this paper we proposed a method of adjusting Koch and Sen's test statistic to analyze data sets with MCAR missing data. One common form of missing data in longitudinal studies is loss to follow up. This type of missing data are likely to be informatively missing as drop out can be a result of subjects dropping out of the study due to either adverse health outcomes or improved health outcomes. In the case of the example above, a subject may stop reporting pain scores if they are feeling large amounts of pain. In these cases, assuming the data are MCAR, when in reality the data are informative, can lead to substantial problems with bias and inaccuracy. The performance of the revised test proposed in Chapter 2 when the MCAR assumption is violated has not been examined.

In this chapter we propose using a method similar to that proposed in Chapter 3 to address the informative missing data. This method will use single imputation to impute missing ranks and a weighting scheme to account for the uncertainty associated with the imputation. The new method proposed seeks to have higher power and comparable type I error rate to the test proposed in Chapter 2, when the missing data truly are informatively missing.

4.1.2 Notation, Assumptions and Terminology

A longitudinal study design, as specified in Section 1.2.2, is assumed. The j^{th} measurement represents the measurement at the j^{th} time point for the i^{th} subject. Koch and Sen's test focuses on the effect of time when no other covariates are of interest. Missing data can be intermittent throughout a study but in longitudinal studies, missing data can also occur as a result of a subject dropping out of the study. Often when a subject drops out of a study it is related to either particularly negative

or positive results. Therefore, it is often assumed that missing data due to loss to follow up is informative missing. Since the goal of this paper is to adapt the methods of Chapter 2 to informative missing data, we will assume the informative missing data are monotonic. Therefore if any subject is missing the j^{th} measurement, the subject will subsequently have a missing j'^{th} measurement for all $j' > j$. The general set up for this test, as seen in table Table 1.2, involves the ranking of the Y_{ij} values within each subject. Here we allow r_{ij} to denote the within subject rank of the i^{th} individual at the j^{th} time point. In the case of complete data, it is important to note that $\sum_{i=1}^{n} r_{ij} = \frac{n(n+1)}{2}$.

The aim of this test is to determine if the outcomes are different within a subject without making any assumptions about the correlation structure. As mentioned in Section 1.2.2, Koch and Sen's test, which uses complete and balanced data, preserves the correlation structure of the observed data as only two possible permutations of ranks are possible for each subject under the null hypothesis of no difference in measurements.

4.1.3 MCAR Data Using Koch and Sen's Methodology

Reduced Rank Adjustment

Koch and Sen's test statistic was developed for situations of complete and balanced data. However, this is often not practical for the analysis of longitudinal studies. Researchers have proposed the reduced rank method as a way to address the problem of incomplete and unbalanced data sets for tests involving within subject ranks (Bernard and Elteren, 1953). This method, which is explicitly described in Chapter 2 of this paper, ranks all non-missing observations and calculates a reduced rank by subtracting the expected value of the rank from the observed rank. Testing to see if the sum of these reduced ranks are different from each other is often seen as an alternative to testing if the average ranks across all subjects are equal. In these cases, subjects with a missing j^{th} observation are not included in the summation and therefore not in the

calculation of the test statistic.

Inflation Factor

Due to the loss of information that results from missing data, the power from the test is lower than desired. In an effort to rectify this problem, a number of researchers proposed different inflation factors that inflate the values of the reduced rank for subjects with more missing data, thereby minimizing the impact of the loss of information from subjects with fewer observations (Prentice, 1979; Mack and Skillings, 1980; Skillings and Mack, 1981; Rai, 1987; Wittkowski, 1988). One of the most widely used inflation factors is $\frac{1}{n_i+1}$, which has been proposed by Prentice. His inflation factor has been proven to yield acceptable power in the case of MCAR data in scenarios involving Friedman's test and has been applied to Koch and Sen's test with some success as shown in Chapter 2 of this document (Prentice, 1979; Stokes, Davis and Koch, 2000).

Test Statistic

The test statistic proposed in Chapter 2 is based on the inflated reduced rank vector $\boldsymbol{\mu_U}$ and the corresponding covariance matrix. The $\boldsymbol{\mu_U}$ matrix consists of $n-1$ elements with one μ_{Uj} omitted from the vector in order to preserve linear independence. The μ_{Uj} element of this vector is shown below:

$$\mu_{Uj} = \sum_{\substack{i=1 \\ n_{ij}>0}}^{k} \left(\frac{1}{n_i+1} r_{ij} - \frac{1}{2} \right)$$

The vector $\boldsymbol{\mu_U}$ has an asymptotically normal distribution with a covariance matrix $\boldsymbol{V_U}$ where $v_{jj'}$ is used to denote the element in the j^{th} row and the j'^{th} column of the

$\boldsymbol{V_U}$ matrix.

$$v_{jj} = Var(\mu_{Uj}) = \sum_{\substack{i=1 \\ n_{ij}>0}}^{k} \left(\frac{1}{n_i + 1} r_{ij} - \frac{1}{2} \right)^2$$

$$v_{jj'} = Cov(\mu_{Uj}, \mu_{Uj'}) = \sum_{\substack{i=1 \\ n_{ij}>0, n_{ij'}>0}}^{k} \left(\left(\frac{1}{n_i + 1} r_{ij} - \frac{1}{2} \right) \left(\frac{1}{n_i + 1} r_{ij'} - \frac{(n+1)}{2} \right) \right)$$

Calculations for both the covariance and variance elements can be found in Appendix A. The test statistic, U, that we will calculate as $\boldsymbol{\mu_U' V_U^{-1} \mu_U}$ will be asymptotically distributed as chi-squared with $n - 1$ degrees of freedom.

Application to Informative Missing Data

Under the null hypothesis there is no difference between measurements at different time points. This test statistic is testing the same hypothesis as Friedman's test but relying on the fact that under this null hypothesis only two permutations are possible for each subject. Assuming no difference in measurements, as the number of subjects goes to infinity, the missing measurements are spread out evenly across all n time points. Therefore, the impact of the informatively missing data will be similar for all measurements and the type I error rate of this test will not be greatly impacted by informative missing data.

For the scenario proposed in this paper, in which higher values are more likely to be missing, all subjects with a non-missing j^{th} measurement have the following expected

rank:

$$\begin{aligned} E[r_{ij}] &= r_{ij}\left(Pr(r_{ij} = r_{ij})\right) + (n_i - r_{ij} + 1)\left(Pr(r_{ij} = n_i - r_{ij} + 1)\right) \\ &= r_{ij}\left(\frac{1}{2}\right) + (n_i - r_{ij} + 1)\left(\frac{1}{2}\right) = \frac{n_i + 1}{2} \end{aligned}$$

And therefore,

$$E[\mu_{Uj}] = \sum_{\substack{i=1 \\ n_{ij}>0}}^{k} \left(\frac{1}{n_i + 1} E[r_{ij}] - \frac{1}{2}\right) = \sum_{\substack{i=1 \\ n_{ij}>0}}^{k} \left(\frac{1}{n_i + 1}\frac{n_i + 1}{2} - \frac{1}{2}\right) = 0$$

Based on the calculations above, the bias of the $\boldsymbol{\mu_U}$ vector in estimating the true vector of reduced ranks, $\boldsymbol{\mu}$, can be calculated explicitly in the case of informatively missing data.

$$E[\mu_{Uj}] - \mu_j = 0 - \sum_{i=1}^{k} \left(r_{ij} - \frac{n+1}{2}\right) = -\sum_{i=1}^{k} \left(r_{ij} - \frac{n+1}{2}\right)$$

As k approaches infinity, under the null hypothesis the bias will be zero. Therefore, the test statistic, even in the case of informative missing data, yields an asymptotically unbiased estimator of the reduced rank vector $\boldsymbol{\mu}$. Therefore the test statistic proposed in Chapter 2 yields an unbiased estimator and a test statistic for which the type I error rate is not greatly impacted by informatively missing data.

While in terms of bias and type I error rate, this test statistic performs well other issues do arise. If the null hypothesis is in fact not true and one, or multiple, measurements are significantly different from others, then informative missing data can lead to substantial problems with statistical power. In situations with few subjects or only a few measurements reported for each subject, the power for this test has been shown to be low and therefore any decrease in power could potentially be problematic. In the scenario proposed in this paper, higher values are more likely to be missing. If many

of these higher values are missing this could result in the estimates of the reduced rank for one measurement to be much lower than they are in reality thereby making the reduced ranks for the measurements closer in range than they truly are.

4.2 Method

The sum of the reduced ranks, when combined with an inflation factor, is a proven and effective non-parametric method of estimating the differences between measurements in the case of MCAR data. However, in the case of informative missing data, by not taking into account the information known about the reason for missingness, this adjustment results in an unnecessary loss of power. This paper proposes a new method that utilizes reduced ranks, weights and imputation to adapt Koch and Sen's test to scenarios involving informatively missing data. It is of interest to compare both the type I error rate and power of the new method to that proposed in Chapter 2 to handle MCAR data.

4.2.1 Imputation

The method for MCAR data ranks only non-missing measurements, thereby excluding all information about missing measurements. In situations where the probability that a measurement is missing is related directly to the value of the measurement, excluding these values can result in a substantial loss of information. We propose a method which imputes missing ranks based on information known about the reason for missingness. For simplicity, we assume strictly informative missingness data, although it is acknowledged that this is rarely the case in practice. We propose using single imputation methods to address the scenario where higher measurements are more likely to be missing. With minor adjustments, the methods proposed here can also be applied to situations where lower measurements have a higher probability of being missing.

In scenarios where higher measurements are more likely to have missing values, imputation is based on the assumption that non-zero probabilities should be assigned to the highest $n - n_i$ ranks. Often in these scenarios there is no way to distinguish multiple missing values from each other. Therefore the probabilities associated with each of the higher ranks are assumed to be equal for each missing measurement for the same subject. This method assumes the probabilities assigned to each of the highest ranks are $\frac{1}{n-n_i}$. Therefore, the expected value for the missing observations can be calculated as below.

$$\begin{aligned} E[r_{ij}] &= 1\left(Pr(r_{ij}=1)\right) + ... + (n_i+1)\left(Pr(r_{ij}=n_i+1)\right) + n\left(Pr(r_{ij}=n)\right) \\ &= \frac{1}{n-n_i}(n_i+1) + \frac{1}{n-n_i}(n_i+2) + ... + \frac{1}{n-n_i}(n) = \frac{n+n_i+1}{2} \end{aligned}$$

Imputation using this expected value rather than the missing value is done for all missing measurements. For all non-missing observations, the ranks remain the within subject rankings from 1 to n_i. After imputation, it is important to note that the average rank for each subject is now the same as the average rank would be for each subject in the case of no missing data.

4.2.2 Subject-specific Weight for Ranks

, After imputation, each subject has a complete set of n ranks. Now, regardless of whether or not r_{ij} is an imputed value or an observed value, we have one observed permutation of ranks from 1 to n. Therefore, there are two possible possibilities, the observed permutation after imputation or the permutation that is the exact opposite. The revised test statistic will therefore be based on k elements, one from each subject. Subjects that were missing the j^{th} measurement, and therefore were excluded from the calculations of μ_{Uj}, will be included and given the imputed value for calculations

of the revised estimate. To account for the uncertainty associated with imputation, particularly as this imputation assumes the only reason for missingness is the actual outcome value, a weight is proposed that will assign less weight to subjects with more missing data. These weighted reduced ranks will then be used to calculate the estimate for $\boldsymbol{\mu}$ as well as the statistic for testing if there is a significant difference in measurements within a subject.

The method of the reduced ranks involves subtracting the expected value of the weighted ranks from the actual value of the weighted ranks. After imputation, the expected value for r_{ij}, regardless of whether r_{ij} is an imputed value or not, can be calculated as follows:

$$\begin{aligned} E[r_{ij}] &= r_{ij}\left(Pr(r_{ij} = r_{ij})\right) + (n - r_{ij} + 1)\left(Pr(r_{ij} = n - r_{ij} + 1)\right) \\ &= \frac{n+1}{2} \end{aligned}$$

If the weight assigned to the ranks of the i^{th} subject is denoted as w_i then $\boldsymbol{\mu_F}$, the revised estimate of $\boldsymbol{\mu}$, can be expressed in general terms with the j^{th} element, μ_{Fj}, defined as the weighted rank minus the expected value of the weighted rank under the null hypothesis. Just as with Prentice's statistic, one μ_{Fj} is omitted to insure the vector is of full rank.

$$\mu_{Fj} = \sum_{i=1}^{k} w_i \left(r_{ij} - \frac{n+1}{2} \right)$$

It is important that non-missing values are given higher weights than missing values, as there is a degree of uncertainty in imputation. Therefore we will assign a subject specific weight which assigns less weight to individuals with more missing data. For subjects with many missing measurements, imputation adds very little information and therefore we want to ensure the weight for these subjects is much smaller. We propose a weight $\frac{1}{n-n_i+1}$ which gives subjects with complete data the same weight they

would be given in the case of complete data. However, subjects with very few non-missing observations are assigned a very small weight to account for the high level of uncertainty. Based on this weight, our estimate of the elements of the $\boldsymbol{\mu_F}$ vector becomes the following:

$$\mu_{Fj} = \sum_{i=1}^{k} \frac{1}{n - n_i + 1} \left(r_{ij} - \frac{n+1}{2} \right)$$

4.2.3 Test Statistic

The hypothesis test proposed in Chapter 2 involves testing the null hypothesis that the inflated ranks for the j^{th} measurement are on average close to the expected value of the inflated ranks, which is equivalent to testing if $\boldsymbol{\mu_U} = \mathbf{0}$. In a similar fashion, if the weighted ranks of the method proposed in this paper are close to their expected value, then $\boldsymbol{\mu_F} = \mathbf{0}$. Therefore, this revised estimate of the reduced ranks can be used to test the same hypothesis as Prentice's statistic.

We denote $\boldsymbol{V_F}$ as the $(n-1)$ x $(n-1)$ covariance matrix associated with $\boldsymbol{\mu_F}$ with the element in the j^{th} column and j'^{th} row denoted as $v_{jj'}$. The variance and covariance of r_{ij} after imputation are the same as they would be in the case of complete data. Therefore the elements of the covariance matrix are very similar to those elements shown in Chapter 2. A more explicit illustration of the calculation of these elements of the covariance matrix can be found in Appendix C. Since ranks from different subjects are independent,

$$\begin{aligned} Var\left(\mu_{Fj}\right) &= \sum_{i=1}^{k} \left(\frac{1}{n - n_i + 1} \right)^2 \left(r_{ij} - \frac{(n+1)}{2} \right)^2 \\ Cov\left(\mu_{Fj}, \mu_{Fj'}\right) &= \sum_{i=1}^{k} \left(\frac{1}{n - n_i + 1} \right)^2 \left(r_{ij} - \frac{n+1}{2} \right) \left(r_{ij'} - \frac{n+1}{2} \right) \end{aligned}$$

By the central limit theorem and Lyapunov's condition, the $n-1$ vector $\boldsymbol{\mu_F}$ has an asymptotically normal distribution of dimension $n-1$ with a covariance matrix composed of the elements specified above. Therefore, the revised test statistic, $F = \boldsymbol{\mu_F'} \boldsymbol{V_F^{-1}} \boldsymbol{\mu_F}$, has a chi-squared distribution with $n-1$ degrees of freedom under the null hypothesis (Koch and Sen, 1968; Sen and Puri, 1967).

4.2.4 Calculation of Bias

Bias associated with $\boldsymbol{\mu_F}$ as an estimator of the vector of reduced ranks, $\boldsymbol{\mu}$, can be explicitly calculated in the case of strictly informative missing data. The expected value of each element of the vector $\boldsymbol{\mu_F}$ is:

$$E[\mu_{Fj}] = E\left[\sum_{i=1}^{k}\left(\frac{1}{n-n_i+1}\right)\left(r_{ij}-\frac{n+1}{2}\right)\right] = \sum_{i=1}^{k}\left(\frac{1}{n-n_i+1}\right)\left(E[r_{ij}]-\frac{n+1}{2}\right)$$

Under the null hypothesis, as shown in Section 4.2.2, the expected value of $r_{ij} = \frac{n+1}{2}$. Therefore, under the null hypothesis $E[\mu_{Fj}] = 0$. Therefore the bias of the test statistic is easily calculated below:

$$E[\mu_{Fj}] - \mu_j = 0 - \sum_{i=1}^{k}\left(r_{ij}-\frac{n+1}{2}\right)$$

Just as in the case of the test statistic proposed in Chapter 2 to handle MCAR data, this value approaches 0 as k goes to infinity. Therefore both the method for MCAR data and the revised method proposed in this paper produce asymptotically unbiased estimates of the reduced rank vector. Therefore, other methods must be used to compare the two test statistics.

4.2.5 Comparison of Type I Error Rate

In the case of strictly informative missing data, it is important to note that $Pr(r_{ij}$ is missing) does not depend on j under the null hypothesis. Therefore, regardless of the value of j, informatively missing values are excluded with equal probability from all n measurements. By the same justification, the inclusion of the imputed value is equally likely to be at any value at any of the n measurements. Therefore when testing if one or more measurements has a significantly higher reduced rank than the others, is not affected by the exclusion of informatively high values. Therefore, whether the missing value is excluded from the test statistic, or an imputed value is substituted, under the null hypothesis this does not noticeably change the probability that one would reject the null hypothesis.

4.2.6 Comparison of Power

In an unbalanced study design, the Pitman efficiency can be largely dependent on the alternative hypothesis. Therefore, in an effort to generalize our results, we will address the comparison of power between our revised method and the method proposed for MCAR data using less explicit guidelines. The general alternative hypothesis of our statistical test is that measurements within a subject are statistically different. As mentioned previously, we are only considering situations where the differences within a subject are constant across all subjects.

In these scenarios, the alternative hypothesis requires the reduced ranks of one or more measurements to be significantly greater than the reduced rank of others. In this situation, it is important to note that $Pr(r_{ij}$ is missing) is not the same for all n measurements. This probability, under the alternative hypothesis, is dependent on the measurement number j. Without a loss of generalizability, we cannot specifically give an explicit function of this probability.

However, we can note that when using the MCAR method, all observations with a missing value are excluded from the test statistic. In the case of strictly informative missing data, this would exclude the highest reduced ranks from the value of the reduced rank for the j^{th} measurement thereby making this reduced rank closer to the reduced rank of the other measurements. In doing so, this makes it harder to reject the null hypothesis when it should be rejected.

In contrast, although the revised method does not impute the exact high value, it does minimize the difference in the reduced rank for that measurement. Therefore using this statistic allows for the clearer delineation between the reduced ranks and thereby increases the power as compared to the test statistic proposed in Chapter 2.

4.3 Simulations

4.3.1 Generation of Data sets

To narrow the scope of our research, only a few factors were allowed to vary so specific guidelines could be established. For each combination of factors, 10,000 complete data sets were generated in order to calculate the type I error rate and power. Of primary concern was how well the method would work in scenarios with a blend of informative missing data and missing completely at random data. We allowed the percentage of these two types of missingness to vary while maintaining the same amount of overall missing data. All data were generated from a multivariate normal distribution. In each scenario 10% of the overall observations were set to missing and a certain percentage out of this 10% were set to be informatively missing data and the remaining were MCAR. Therefore each data set generated had the same number of missing observations, although some of those had a higher percentage of the missingness generated by informative missingness. As our derivations of the revised test statistic assumed

100% of the missing data were informative this was the first option selected. However, we also chose to examine scenarios where 80%, 50% and 20% out of the total 10% missing data were forced to be informative missing data.

While the percent of the missing data that were informative was of primary interest, we also were interested in varying the number of subjects and the number of observations within each subject. It was of interest to examine the performance of our test in a relatively small sample even though as shown in Chapter 2 this can present substantial problems in terms of statistical power. Therefore, we selected a much larger sample size as well, testing scenarios involving 10, 50 and 100 subjects with the number of measurements per subject being 5 and 10. Table 4.1 below illustrates all of the variations generated in our scenarios.

Continuous outcome variables were generated from the same multivariate normal distribution, with the covariance matrix calculated based on the specification of the variance and correlation. To narrow the scope of our research, we focused primarily on longitudinal studies. For generalizability we standardized our outcome variable, making the mean zero and the variance of each measurement one. If measurements are taken on the same individual across a wide range of time, the correlation is usually relatively small between measurements. To maintain a reasonably small correlation as expected in longitudinal studies, while still generating a range of correlations, we chose to narrow the simulations to situations where the correlation between any two consecutive measurements is 0.1, 0.3 and 0.5. An autoregressive correlation structure was assumed. Using the mean and covariance matrix, multivariate normal data were generated using the RANDNORMAL function in PROC IML.

Once the complete data set was generated, the data were sorted by outcome and the highest observations, up to the prespecified number of informative missing observations, were set to missing. While we acknowledge that informative missing data

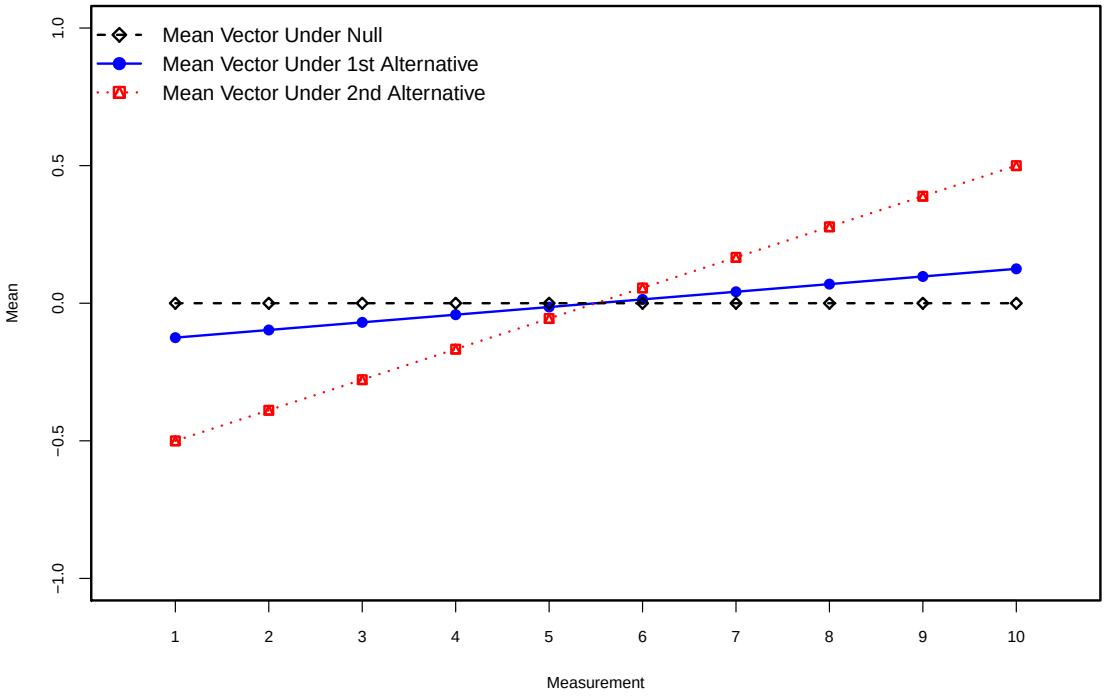
realistically involve increasing the probability of a subject being missing, we felt that our method allowed for optimal control over the percent that is truly informatively missing. The remaining missing data, if any, was then eligible to be set to missing according to MCAR patterns. Using PROC SURVEYSELECT, a simple random sample of the remaining non-missing observations were randomly selected. Those selected were then set to missing.

Three different mean vectors were chosen in order to allow for the examination of both the type I and power. The three mean vectors for the 10 observations per subject can be seen in Figure 4.1. The overall mean of all three mean matrices was zero although two of the three mean vectors had a linear increase in the mean as a function of observation number. These two mean vectors with the linear increases in outcome were used to examine power. As some scenarios included a larger number of observations, the linear increase per observation was varied. The less extreme alternative hypothesis was a linear increase of a quarter of the variance of the data, 0.25, and the more extreme alternative was a linear increase of 1.

4.3.2 Imputation

The single imputation method first ranked the observations within a block from one to the total number of non-missing observations within that block. The remaining ranks that were not assigned, which consisted of the highest ranks for that subject, were averaged and this average rank was imputed as the value for all missing observations within that block. As mentioned in Section 4.2.1, this was equivalent to assigning all missing values the value $\frac{n+n_i+1}{2}$.

Figure 4.1: Mean Vectors for Null and Alternative Hypotheses



4.3.3 Calculation and Comparison of Type I Error Rate

SAS currently does not offer packages that can easily compute Koch and Sen's test statistic. For this reason, a simple SAS macro using the IML language was developed which calculated the elements of the $\boldsymbol{\mu_F}$ vector and the corresponding $\boldsymbol{V_F}$ matrix along with the test statistic and p-value. The macro developed can be seen in Section C.3 in Appendix C.

For both the MCAR method, calculated from the macro in Appendix A, and the revised method put forth in this chapter, the type I error rates were calculated. Using a standard rejection level of 0.05, if the p-value for the test statistic fell below 0.05, the test rejected the null hypothesis. The type I error rates were calculated as the total percentage of data sets for which the null hypothesis was rejected when there was in

fact no linear change in the outcome variable over time. Power was calculated in a similar fashion, under the two alternative hypotheses.

4.4 Results

4.4.1 Type I Error Rate

Type I error rate for Koch and Sen's test using the original complete data was calculated for comparison the MCAR method and our revised method for informative missing data in this paper. The error rates for 10, 50 and 100 subjects can be seen respectively in Figures 4.2, 4.3 and 4.4. The type I error rates for all three tests, in the case of 10 subjects, are extremely stringent. For the cases of 50 and 100 subjects, the type I error rates for the most part are all under the 0.05 level, with the exception of our revised method which has a slightly inflated type I error rate in some cases with 100 subjects. In the case of 50 subjects, the scenario with more observations collected on each subject has more stringent type I error rates than the fewer observations scenario. Overall, the type I error rates for the MCAR method and the informative method proposed in this paper do not appear to differ substantially when the number of subjects is noticeably larger than the number of observations collected on each subject. In cases where 10 observations are collected on each subject the type I error rate of our revised method is more stringent than the MCAR alternative in some situations. There does not appear to be any noticeable difference between type I error rates and correlation. Exact values can be found in Appendix C. Simulations were not run in the case of 10 subjects and 10 observations per subject as more than 10 subjects are needed to evaluate a difference in 10 measurements.

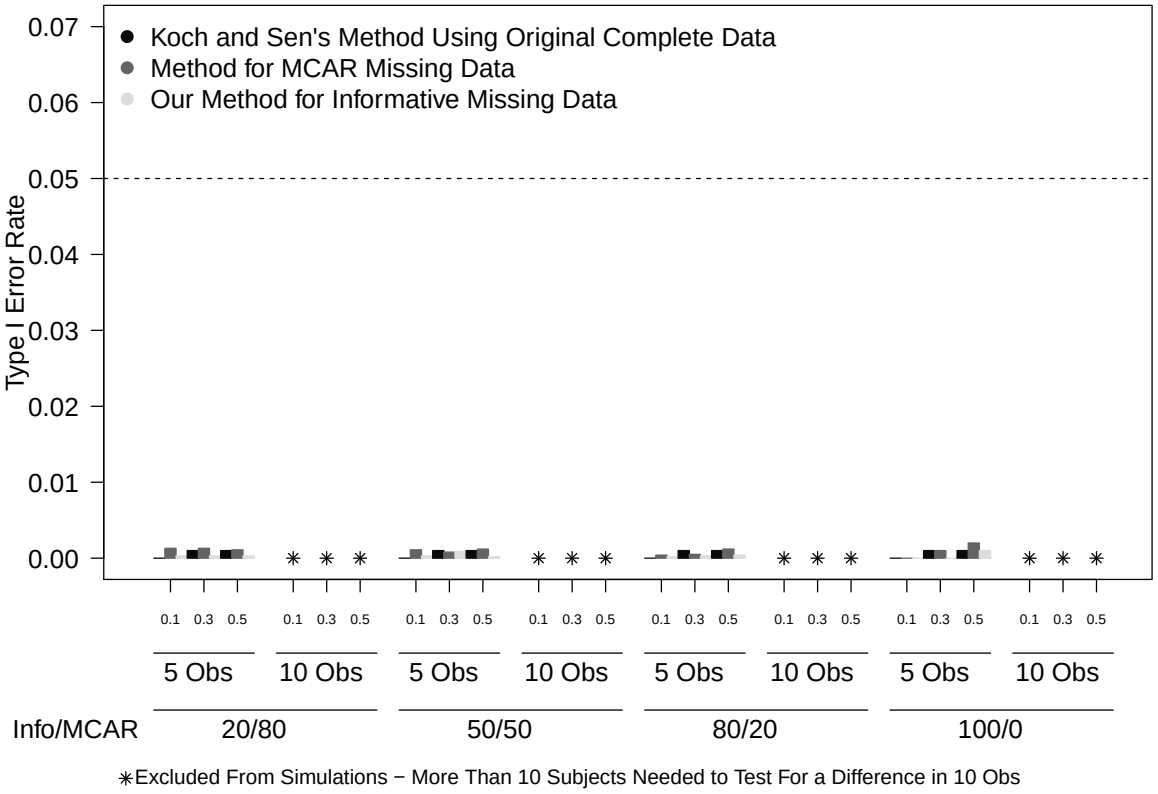


Figure 4.2: Type I Error Rate by % Informative Missing - 10 Subjects

4.4.2 Power

Graphs visually displaying the power under the less extreme alternative, a linear increase of 0.25, for 10, 50 and 100 subjects are shown in Figures 4.5, 4.6, and 4.7 respectively. The exact values are given in Appendix C. All three tests show almost no power to detect any difference in measurements under this alternative when data is collected on only 10 subjects. Our revised method for informative missing data has lower power than the MCAR alternative, for 50 subjects, when 10 observations are collected on each subject and almost equivalent power to the MCAR test when 5 observations are collected. When the sample size is 100, our revised test shows at least equivalent, and at times, an improvement in power over the MCAR method for

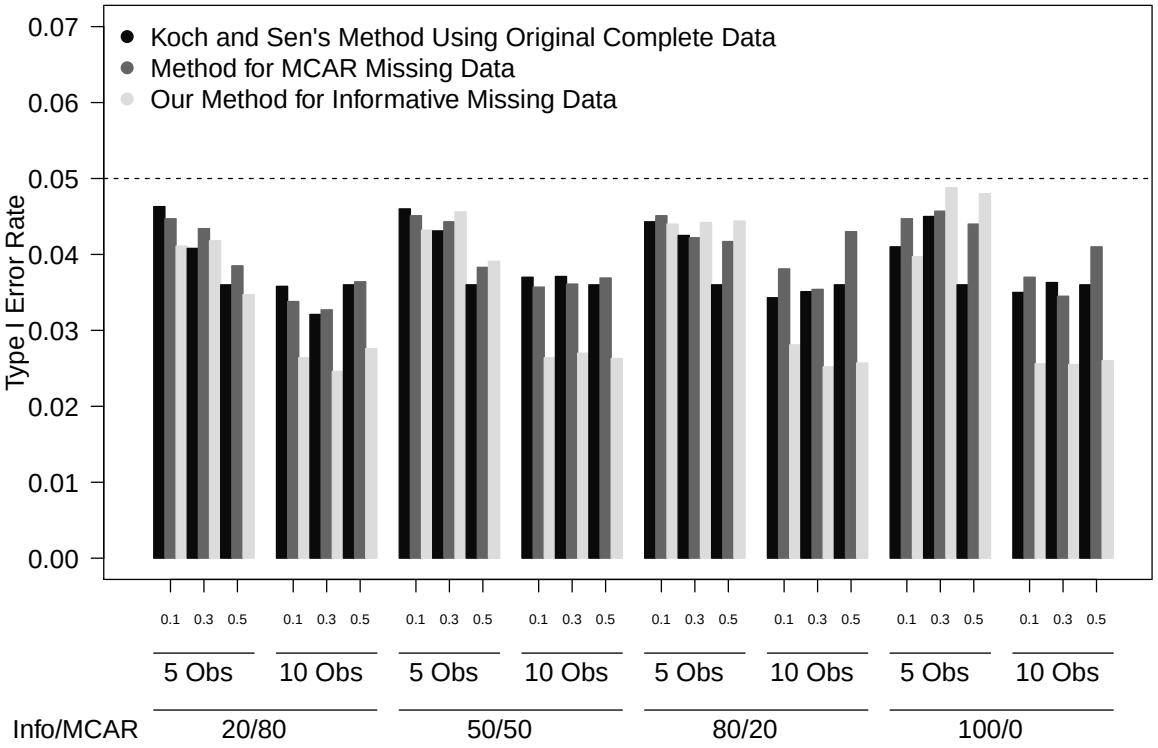


Figure 4.3: Type I Error Rate by % Informative Missing - 50 Subjects

5 observations when at least 50% of the missing data are informatively missing. The revised method also appears to to offer a slight improvement in power over the MCAR alternative when all of the missing data is informative and 10 observations are collected on each subject.

Statistical power under the more extreme alternative of a linear trend of 1 across all measurements, is shown for cases of 10, 50 and 100 subjects in Figures 4.8, 4.9, and 4.10. The power for the revised method is compared to the power of the MCAR test and the power of Koch and Sen's test using the original complete data. In the case of 10 subjects, shown in Figure 4.8, the power is extremely low and all three methods have essentially no power to detect a difference in measurements. For situations involving 50 or 100 subjects, there are more noticeable differences in power between the original

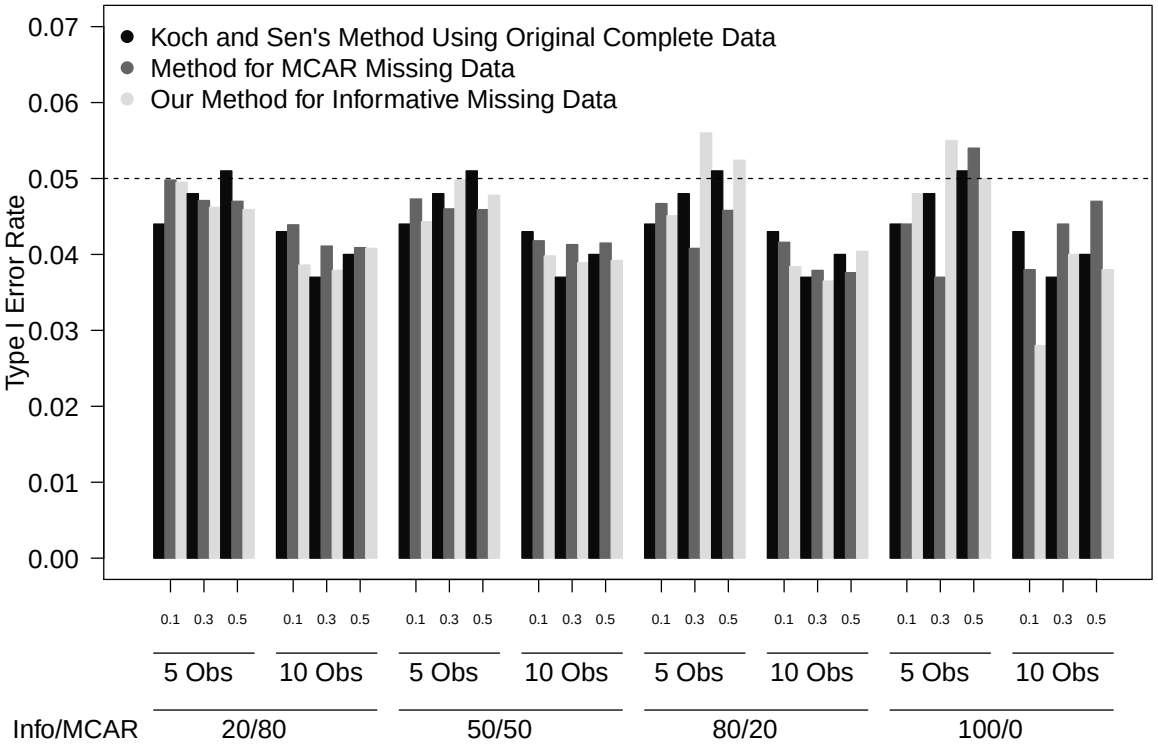


Figure 4.4: Type I Error Rate by % Informative Missing- 100 Subjects

complete data and the methods proposed to handle missing data. With regards to comparisons between the two methods that allow for missing data, has better power than the MCAR alternatives when only 5 measurements are reported on each subject and over 50% of the missing data are informatively missing. The revised method proposed in this paper has similar power to the MCAR method when exactly 50% of the missing data are informatively missing and there are only 5 measurements collected on each subject. Our method shows a slight improvement in power over the MCAR test when all of the missing data are informatively missing and the correlation is lower.

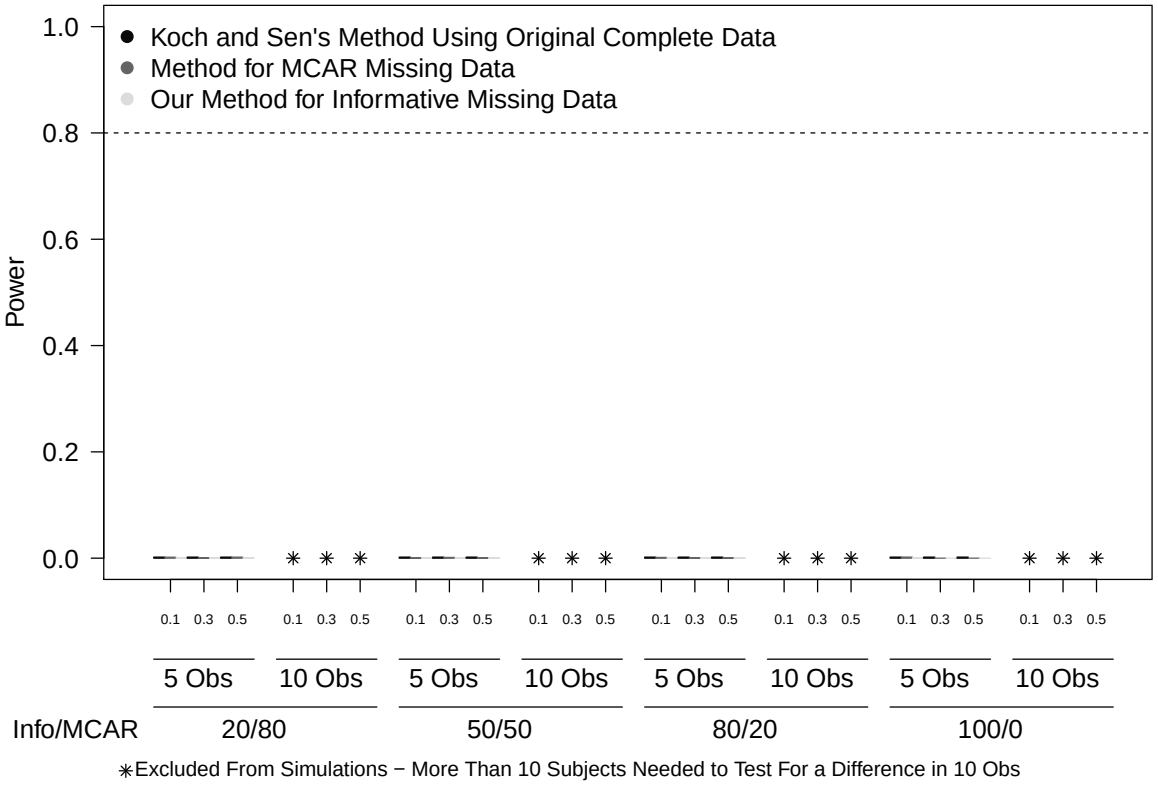


Figure 4.5: Power by % Informative Missing (Increase of 0.25) - 10 Subjects

4.4.3 Asymptotic Behavior

The distribution of our test statistic, based on 5,000 simulations, was examined in order to evaluate the asymptotic behavior of our revised test statistic under the null hypothesis. As mentioned in Section 4.2.3, under the null our distribution is asymptotically distributed chi-squared distribution with n-1 degrees of freedom. We considered a scenario with 500 subjects and 10 measurements collected on each subject. The distribution of these test statistics along with the probability density function for a chi-squared distribution with 9 degrees of freedom is shown in Figure 4.11 below.

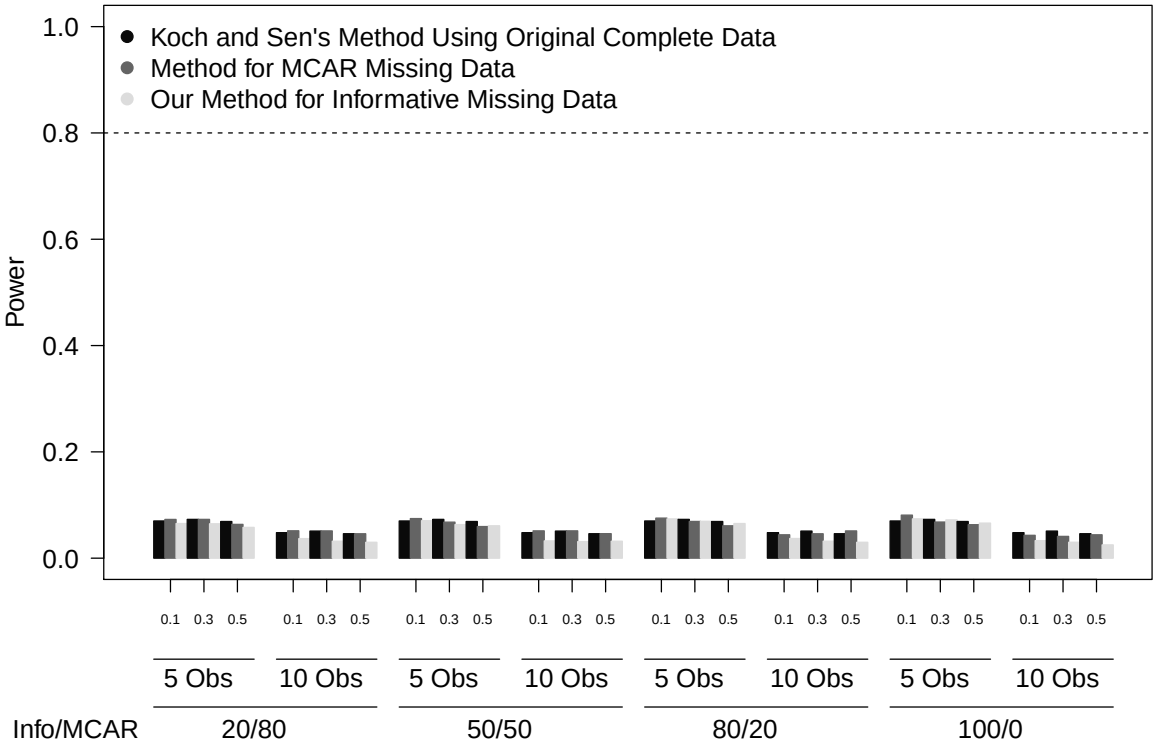


Figure 4.6: Power by % Informative Missing (Increase of 0.25) - 50 Subjects

4.5 Data Example

We will use this method to analyze data collected from individuals who suffer from irritable bowel syndrome (IBS). This data set is intended to be a specific case of the longitudinal study design described in Section 4.1.1. In this study, participants were asked to report pain, on a scale from 0 to 10, before and after any bowel movement. The timing of these pain measurements were classified into one of four categories based on the time stamp. Midnight to six a.m. was considered as early morning, from six a.m. to noon was considered morning, from noon to six p.m. was considered afternoon and from six p.m. to midnight was considered evening. Researchers were interested in determining if the difference in pre and post bowel movement differed throughout the

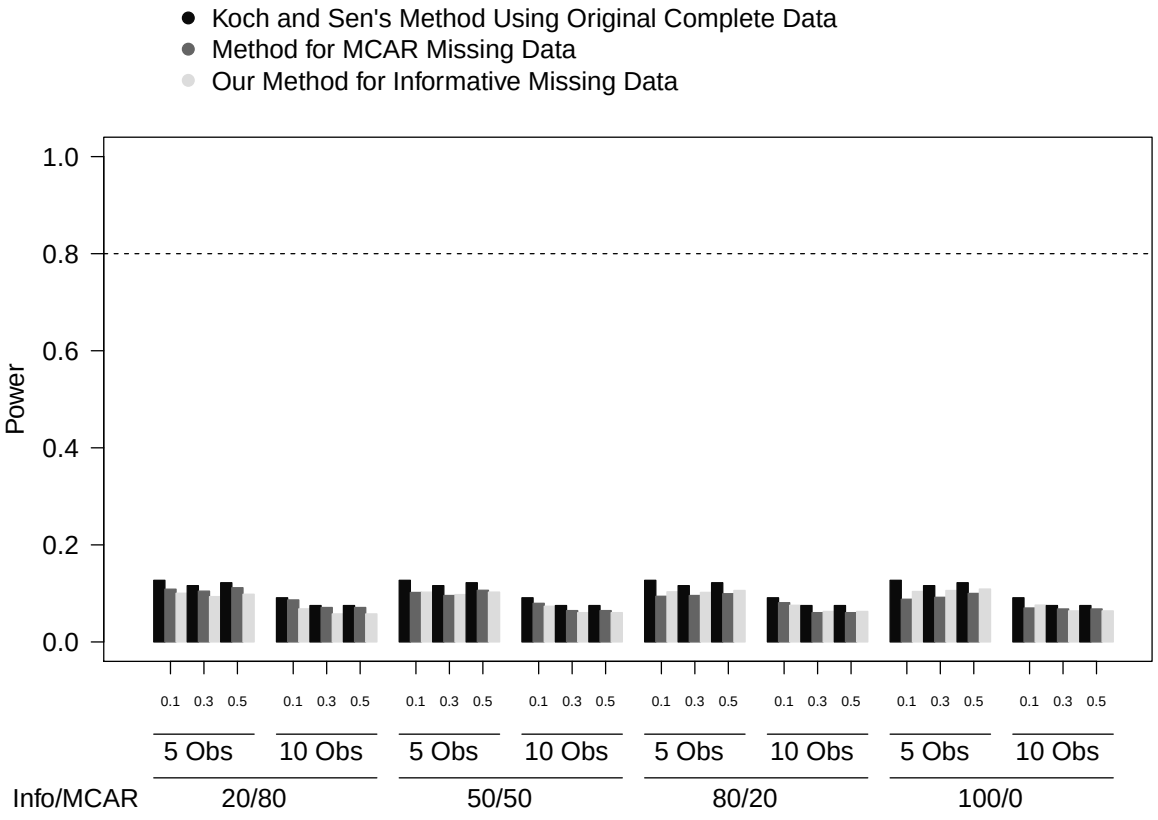


Figure 4.7: Power by % Informative Missing (Increase of 0.25) - 100 Subjects

day. Therefore, both pre and post measurements were required for a measurement to be included in this analysis. Data had been collected over a number of different days, and due to irregularities in certain days, the average difference in pre and post bowel measurements for each time period across all days was used as the outcome of interest.

Both participants with diarrhea predominant IBS and constipation predominant IBS were included in the study. Both types of participants were missing some observations at some time points. However, those patients with constipation predominant IBS reported less pain measurements overall as they had fewer bowel movements. During bouts of constipation, was usually when these participants were in the most pain and when these individuals were most likely to not be reporting pain measurements. Therefore, we have reason to believe that a substantial amount of the missing observations were

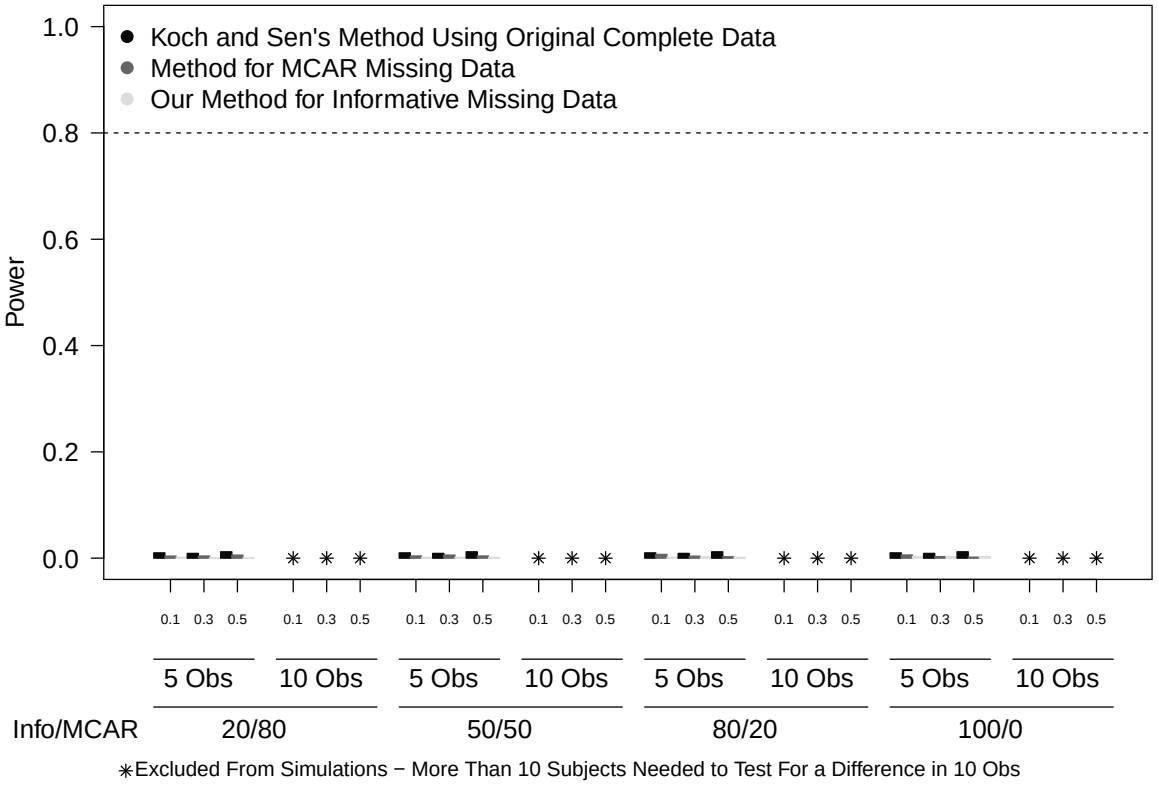


Figure 4.8: Power by % Informative Missing (Increase of 1) - 10 Subjects

likely to be higher pain scores. We could therefore likely assume that at over half the data should be informatively missing data.

Data were collected on 37 participants, 18 with constipation prominent IBS and 19 with diarrhea prominent IBS. Almost 34% of the measurements were missing, 62% of which were measurements of the constipation prominent patients. Just over 50% of the missing measurements were from the early morning time period of midnight to 6 a.m. Using the nonparametric test for difference in pre and post bowel movement pain scores would be ideal, as it requires no distributional assumptions to be made about the difference in pain scores. The average difference in pre and post pain scores for each period of time along with the within subject rankings of the average pain scores are found in Table C.10 and Table C.11 in Appendix C.

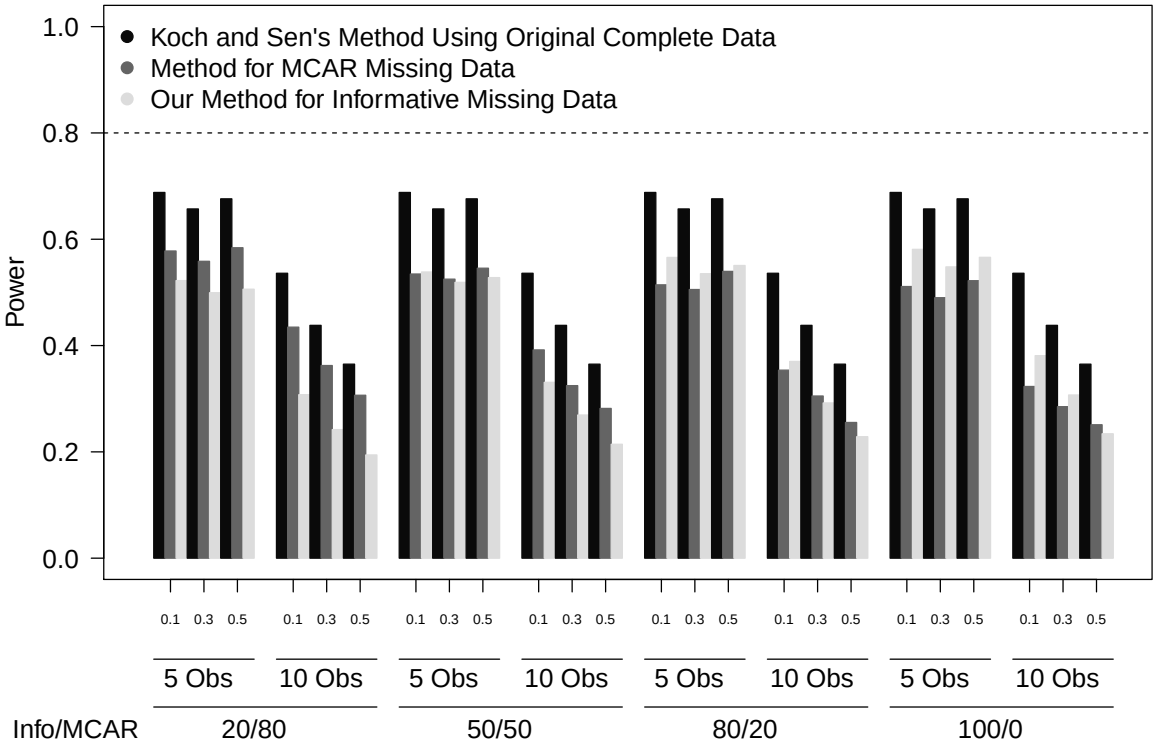


Figure 4.9: Power by % Informative Missing (Increase of 1) - 50 Subjects

Using the macro also found in Appendix C, the revised test statistic for this data set is 7.30 which with 3 degrees of freedom yields a p-value of 0.063. With this p-value, we fail to reject the null hypothesis that the difference in pre and post bowel movement pain score are different throughout the day. For comparison, the MCAR test statistic, proposed in Chapter 2, yielded a test statistic of 3.08 with a p-value of 0.3802. Also for comparison, when all subjects with any missing observations were removed, leaving only 9 subjects, Koch and Sen's test statistic was 4.85 with a corresponding p-value of 0.1828. However, it is important to note that there is a large amount of missing data here, and most commonly this missing data is found in the early morning time period. Therefore, due to the large number of missing observations at that time period, these results should be interpreted with caution. Figure 4.12 illustrates the mean difference in

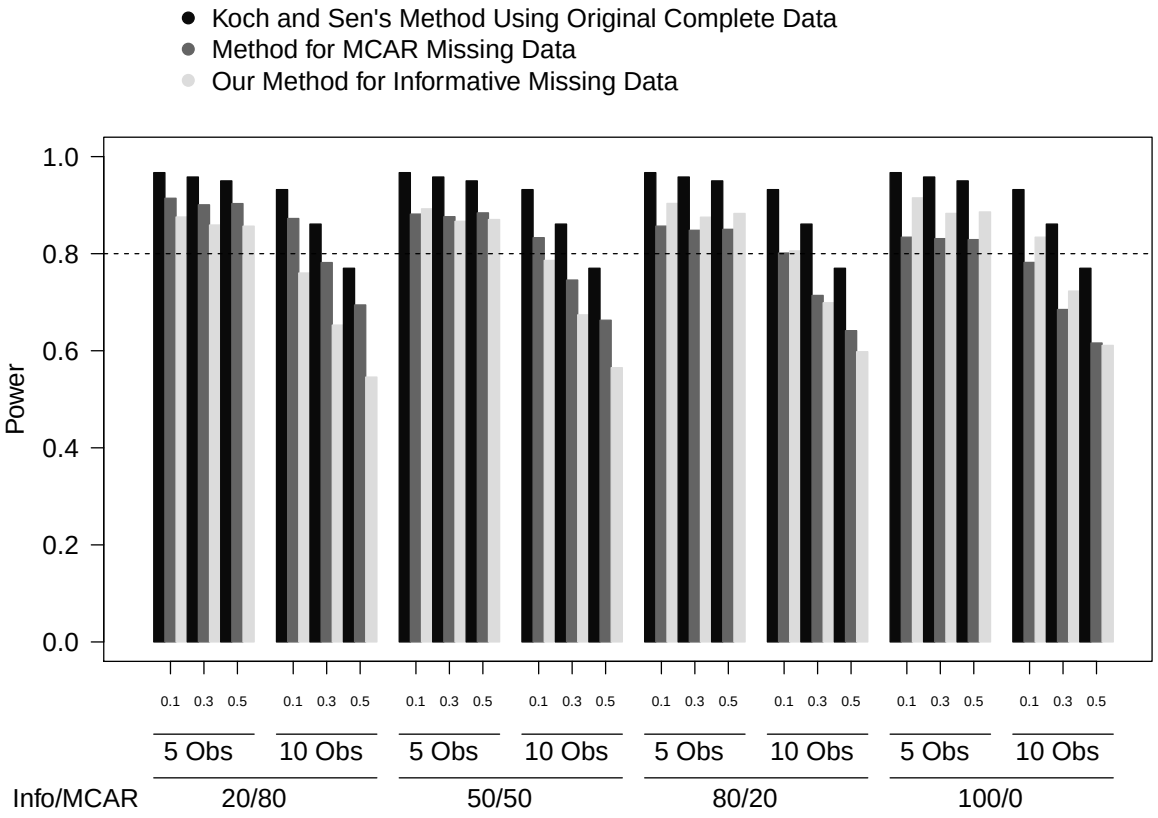


Figure 4.10: Power by % Informative Missing (Increase of 1) - 100 Subjects

pre and post-bowel movement pain scores by time of day. The dotted line, representing the averages for the 9 subjects with all 4 measurements, does appear to differ from the solid line which includes data from all 37 participants.

4.6 Discussion

Koch and Sen's test is an effective method of testing for a difference in measurements when one does not want to make any assumptions regarding the distribution of the outcome measurement. However, often in the case of studies involving repeated measurements, missing data occur and Koch and Sen's test can only analyze complete and balanced data. In Chapter 2, we proposed a test to handle MCAR data. However,

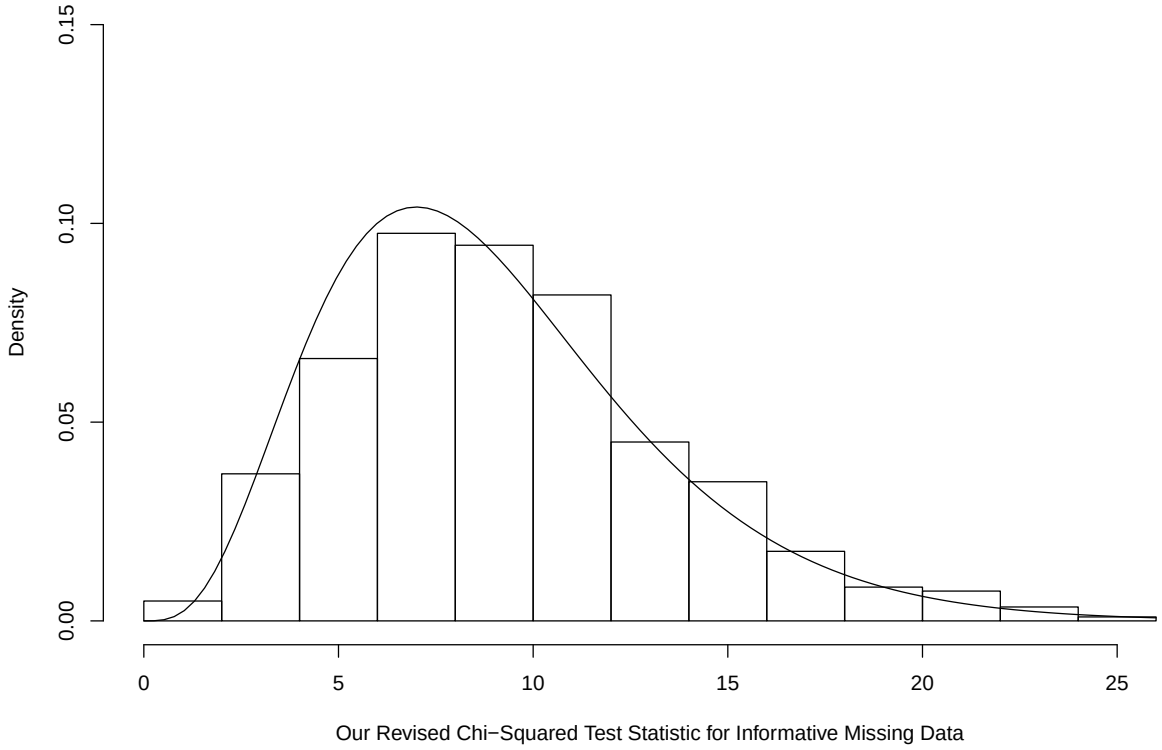
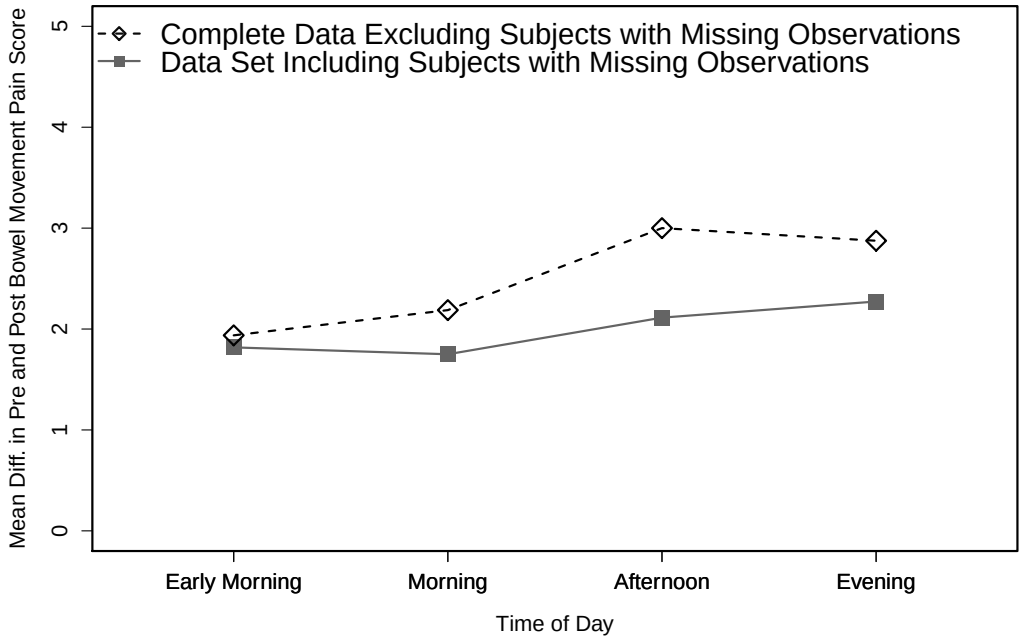


Figure 4.11: Asymptotic Behavior of Our Revised Test Statistic

in this paper, the revised method proposed offers a slight improvement in power over the MCAR method in certain scenarios. With the smaller correlations examined in this paper, particularly the autoregressive correlation structure, we acknowledge that the power the revised test as well as the MCAR test and the test using the original complete data set yield relatively low statistical power. Therefore, the improvement in power provided by this revised test can be very crucial.

For smaller sample sizes, of 10 subjects or less, the type I error rate and power under the less extreme alternative hypotheses is very small and essentially makes it very difficult to reject the null hypothesis altogether. The revised test proposed in this chapter provides at least as much power, if not more, as the MCAR test when at least half of the missing data can be assumed to be informatively missing and the number of

Figure 4.12: Mean Difference in Pre and Post BM Pain Score by Time of Day



subjects is substantially larger than the number of observations collected on each subject. Therefore, in these situations we would recommend using the method proposed in this paper, since it performs at the least very similar to the MCAR method and can actually led to an improvement in statistical power. One should however be aware, in the case of large sample sizes, there may be a slight inflation of type I error rate.

Koch and Sen's test was developed to work in situations where the only assumption was that the correlation was not compound symmetric, thereby allowing for a variety of correlation structures. Further research should investigate alternative correlation structures that result in higher correlation between measurements. These may yield higher statistical power for this test. It was felt that an autoregressive structure would be one of the more common correlation structures, as a logical choice for longitudinal studies. However, we acknowledge that this did limit the scope of this research and the

investigation into the performance of this test. As Koch and Sen's test allows for any correlation structure, it would be of interest to examine the performance when compound symmetry does hold and compare these results to their Friedman counterparts.

With regards to the covariance calculations used in this paper, it is important to note that the covariance estimates are composed only from data for those subjects with the measurements observed for both the j^{th} and the j'^{th} measurement. It is noted that we could include more information in these calculations by breaking the covariance into the correlation and variance components and allowing for all subjects with a missing j^{th} measurement to contribute to the variance calculations of the j'^{th} measurement for the correlation estimate and vice-versa. For the purpose of this paper, we felt it was important to not use this method due to the increase in the amount of computations that would be required. Using the covariance estimates proposed in this paper, calculating test statistics requires only minor computational adjustments to most major statistical software. In addition by using the covariance estimates proposed in this paper, we allow for situations in which the variance of those with the j^{th} but not the j'^{th} measurement, or vice versa, differed from the rest of the collected data. However, it would be of interest to compare the performance of the statistic proposed in this paper and the similar statistic using the alternative covariance estimates.

Table 4.1: Data Sets Generated

Number of Subjects	Number of Observations Per Subject	% Of Missing Data that is Informative
10	5	100%
10	5	80%
10	5	50%
10	5	20%
10	10	100%
10	10	80%
10	10	50%
10	10	20%
50	5	100%
50	5	80%
50	5	50%
50	5	20%
50	10	100%
50	10	80%
50	10	50%
50	10	20%
100	5	100%
100	5	80%
100	5	50%
100	5	20%
100	10	100%
100	10	80%
100	10	50%
100	10	20%

Chapter 5

Proposed Guidelines and Future Research

5.1 Summary and Guidelines

When analyzing small sample sizes, or in any situation where one is not willing to make assumptions regarding the outcome variable in the case of repeated measurements, often non-parametric methods should be used to test for a difference in measurements on the same subject. There are a variety of ways to handle missing data in all these scenarios. One could remove all measurements collected on subjects who were missing one or more measurements. Alternatively one could use methods which account for missing data and utilize the known non-missing measurements for hypothesis testing. As a final alternative, one could impute the missing values and complete the data set, then use complete data methods to analyze the data.

In the case of MCAR data, the first two of these methods would both be appropriate to handle missing data. However, using the first method of excluding all measurements for subjects missing one or more measurements can result in a much smaller sample size. This can present problems particularly when the sample size for a study is already relatively small. When compound symmetry cannot be assumed and the missing data

can be assumed to be MCAR, we propose using the method proposed in this paper. Our method allows for the use of all available data, as compared to the only alternative, which involves removing all observations from the analysis from any subject missing even one measurement. Based on our simulation study, which compared Koch and Sen's test using the original complete data to our method, our method shows similar type I error rates and power to that of the original complete data set. Therefore, when contrasted with the alternative of excluding all subjects with missing data, our method yields similar type I error rates and power while including the maximum amount of information.

When the missing data are informatively missing, we proposed a method which used imputation and complete data analysis methods. While our method does not provide much difference in terms of type I error rate, our method does shows improvement over MCAR methods with regards to statistical power. In scenarios where the number of subjects is substantially larger than the number of measurements collected on each subject, our method has proven to provide either equivalent power or provide improvements in statistical power over the MCAR alternatives. While, often the exact proportion of missing data that are informatively missing are not known, we propose using our method as it has the potential to improve power when one is willing to assume at least 50% of the missing data are informatively missing and the number of subjects is substantially larger than the number of measurements collected on each subject.

5.2 Future Research

5.2.1 Imputation Assumptions and Limitations

Although in this research we focused on informative missing data where higher outcomes are more likely to be missing, as mentioned earlier, these methods can easily

be adapted to scenarios where lower outcomes are more likely to be missing. The ranks for those missing observations are imputed under these assumptions that the highest ranks, or similarly the lowest ranks, are those that are likely to be missing. The idea that the extreme measurements are those most likely to be missing is often a reasonable assumption in many cases. As mentioned earlier, when subjects drop out of a study or cannot make it to a study visit, this may often be a result of extremely bad or extremely good health which would have resulted in an extreme outcome measurement if the missing measurement were recorded. For instance, one can likely hypothesize that if a subject did not report pain for a number of time points in a row, that subject may be suffering from extreme pain and imputing the highest ranks seems to be a reasonable assumption. However, imputing the most extreme values is not always appropriate. Most notably, in the event of death, when a subject has missing values, imputing a value would be inappropriate. Adaptations to our imputation method and the performance of our method against these will be investigated in order to allow for informative missing data in the event of death.

5.2.2 Performance in Alternative Scenarios

The guidelines above were developed based on the simulation studies done for this research. There are a number of factors that we did not vary and some variations which we did not consider. One much consideration, is in terms of the true distribution of the outcome variables used in these simulations. The outcome variable in our simulation studies were generated from a multivariate normal distribution. Non-parametric methods are used more often in scenarios where the outcome variables have a more distinct distribution. Future research will investigate the performance of the tests proposed in this paper, when the outcome variables are generated from distributions further away from the multivariate normal distribution. The guidelines proposed in this paper will

be evaluated for these scenarios. One specific distribution of interest is the proposed is a multivariate t-distribution with only a few degrees of freedom. This distribution has heavier tales than the multivariate normal and, as such, it would be of interest to examine the performance of ours tests in these scenarios.

In addition, our research was limited in terms of the correlation scenarios examined. Our research was focused only on scenarios of compound symmetry and autoregressive correlation structures. Compound symmetry is realistic for scenarios involving litters of animals or in cases where the block group is not subject but rather some collection of subjects at some location. One would expect any two measurements, in this case, collected from the same block, to have equal correlation. To generate scenarios where one could not assume compound symmetry, we chose an autoregressive scenario as a way of selecting a simple correlation structure where measurements further apart are less strongly correlated than measurements closer together. However, it should be noted, particularly when many measurements are collected on the same subject, the correlation between measurements that are farther apart is essentially zero under the assumption of autoregressive correlation. In reality, measurements on the same subject, even if they were extremely far apart, would still be correlated. As such, we will test our methods in a scenario where some minimal level of correlation between measurements collected on the same subject is assumed. For example, one option we plan to pursue is to simulate data with a correlation structure of ρ,ρ,$^{1.25}\rho$,$^{1.75}$, ect. As an alternative we also plan to examine scenarios where a simple a floor correlation between any two measurements from the same subject is assumed.

Our scenarios were limited in the number of possible observations that could be collected on any subject. As such, we are currently working on evaluating future research which will investigate the performance of our test in cases where even fewer observations are collected on each subject, specifically examining situations with only

three observations collected on each subject.

In terms of the covariance calculations used in this paper, it is important to note that the covariance estimates are composed only from data for those subjects with the measurements observed for both the j^{th} and the j'^{th} measurement. It is noted that we could include more information in these calculations by breaking the covariance into the correlation and variance components and allowing for all subjects with a missing j^{th} measurement to contribute to the variance calculations of the j'^{th} measurement for the correlation estimate and vice-versa. For the purpose of this paper, we felt it was important to not use this method due to the increase in the amount of computations that would be required. By using the covariance estimates proposed in this paper, calculating test statistics requires only minor computational adjustments to most major statistical software. In addition by using the covariance estimates proposed in this paper, we allow for situations in which the variance of those with the j^{th} but not the j'^{th} measurement, or vice versa, differed from the rest of the collected data. However, we plan to compare the performance of the statistic proposed in this paper and the similar statistic using the alternative covariance estimates.

5.2.3 Alternative Tests

The within subject test proposed by Koch and Sen, involves tests for a overall difference in measurement effect. Often, in the case when compound symmetry cannot be assumed, interest lies in testing for a trend across measurements. We will investigate a non-parametric counterpart to Koch and Sen's test which will be able to test for a trend over time rather than just testing for a general difference in measurements. The focus of this test at this time is in the calculation of subject specific Spearman correlation coefficients. Under the assumptions put forward by Koch and Sen, this test will be based on the assumption that the observed correlation coefficient and the

inverse of the observed correlation coefficient are equally likely with probability $\frac{1}{2}$.

This research has been based on developing methods to test for a difference in within-subject measurements, thereby focusing on situations in which we assume between rank comparisons are not equivalent between different subjects. It is of interest to expand these methods to aligned rank scenarios in which one takes into account between subject comparisons after accounting for differences in subjects (Stokes, Davis and Koch, 2000; Hodges and Lehmann, 1962; Sen, 1968; Lehmann and D'Abrera, 2006; Koch and Sen, 1968). Our methods could be applied to a variety of aligned rank tests, specifically those where a compound symmetric correlation structure can be assumed and one where it cannot. We propose to develop methods for aligned rank tests, which involve ranking all measurements after subtracting some subject specific measure of location from the measurements for each subject. The guidelines for applying this method to aligned ranks tests may vary depending on the measure of location for each subject that is subtracted from the outcome values. For some scenarios proposed in our simulation study, these aligned rank tests could provide more statistical power than the within subject methods proposed in this research. By using tests involving overall comparison, between subject comparisons can be used to detect differences in measurements **??**. These tests do require the assumption that observations from difference subjects are comparable after a subject specific measure is subtracted from all observations. One is often willing to make this assumption, particularly in scenarios where randomization is used and one is willing to assume a homogenous variance across all subjects.

Appendix A

Chapter 2

A.1 Variance

$$
\begin{aligned}
v_{jj} = Var(\mu_{Kj}) &= \sum_{\substack{i=1 \\ n_{ij}>0}}^{k} Var\left(\frac{1}{n_i+1}r_{ij} - \frac{1}{2}\right) = \sum_{\substack{i=1 \\ n_{ij}>0}}^{k} \left(\frac{1}{n_i+1}\right)^2 Var(r_{ij}) \\
&= \sum_{\substack{i=1 \\ n_{ij}>0}}^{k} \left(\frac{1}{n_i+1}\right)^2 \left(E[r_{ij}^2] - E[r_{ij}]^2\right) \\
&= \sum_{\substack{i=1 \\ n_{ij}>0}}^{k} \left(\frac{1}{n_i+1}\right)^2 \left((r_{ij}r_{ij})\frac{1}{2} + (n_i+1-r_{ij})^2\frac{1}{2} - \left(\frac{n_i+1}{2}\right)^2\right) \\
&= \sum_{\substack{i=1 \\ n_{ij}>0}}^{k} \left(\frac{1}{n_i+1}\right)^2 \left(r_{ij}^2 - r_{ij}(n_i+1) + \frac{(n_i+1)^2}{2} - \frac{\frac{1}{2}(n_i+1)^2}{2}\right) \\
&= \sum_{\substack{i=1 \\ n_{ij}>0}}^{k} \left(\left(\frac{r_{ij}}{(n_i+1)}\right)^2 - \frac{r_{ij}(n_i+1)}{(n_i+1)^2} + \left(\frac{1}{2}\right)^2\right) \\
&= \sum_{\substack{i=1 \\ n_{ij}>0}}^{k} \left(\frac{1}{n_i+1}r_{ij} - \frac{1}{2}\right)^2
\end{aligned}
$$

A.2 Covariance

$$
\begin{aligned}
v_{jj'} &= Cov(\mu_{Kj}, \mu_{Kj'}) \\
&= Cov\left(\sum_{\substack{i=1 \\ n_{ij}>0}}^{k}\left(\frac{1}{n_i+1}r_{ij}-\frac{1}{2}\right), \sum_{i=1}^{k}\left(\frac{1}{n_i+1}r_{ij'}-\frac{1}{2}\right)\right) \\
&= \sum_{\substack{i=1 \\ n_{ij}>0}}^{k} Cov\left(\left(\frac{1}{n_i+1}r_{ij}-\frac{1}{2}\right), \left(\frac{1}{n_i+1}r_{ij'}-\frac{1}{2}\right)\right) \\
&= \sum_{\substack{i=1 \\ n_{ij}>0}}^{k}\left(\frac{1}{n_i+1}\right)^2\left(E\left[r_{ij}r_{ij'}-r_{ij}\frac{n_i+1}{2}-r_{ij'}\frac{n_i+1}{2}+\left(\frac{n_i+1}{2}\right)^2\right]-0\right) \\
&= \sum_{\substack{i=1 \\ n_{ij}>0}}^{k}\left(\frac{1}{n_i+1}\right)^2\left(E[r_{ij}r_{ij'}]-\left(\frac{n_i+1}{2}\right)^2\right) \\
&= \sum_{\substack{i=1 \\ n_{ij}>0}}^{k}\left(\frac{1}{n_i+1}\right)^2\left((r_{ij}r_{ij'})\frac{1}{2}+(n_i+1-r_{ij})(n_i+1-r_{ij'})\frac{1}{2}-\left(\frac{n_i+1}{2}\right)^2\right) \\
&= \sum_{\substack{i=1 \\ n_{ij}>0}}^{k}\left(\frac{2r_{ij}r_{ij'}}{2(n_i+1)^2}-\frac{r_{ij}(n_i+1)}{2(n_i+1)^2}-\frac{r_{ij'}(n_i+1)}{2(n_i+1)^2}+\frac{(n_i+1)^2}{2(n_i+1)^2}-\frac{\frac{1}{2}(n_i+1)^2}{2(n_i+1)^2}\right) \\
&= \sum_{\substack{i=1 \\ n_{ij}>0}}^{k}\left(\frac{r_{ij}r_{ij'}}{(n_i+1)^2}-\frac{r_{ij}}{2(n_i+1)}-\frac{r_{ij'}}{2(n_i+1)}+\left(\frac{1}{2}\right)^2\right) \\
&= \sum_{\substack{i=1 \\ n_{ij}>0}}^{k}\left(\left(\frac{1}{n_i+1}r_{ij}-\frac{1}{2}\right)\left(\frac{1}{n_i+1}r_{ij'}-\frac{1}{2}\right)\right)
\end{aligned}
$$

A.3 SAS Macro for Statistic with MCAR data

This macro requires the input of three variables. The variable numTime denotes the maximum number of measurements collected on any subject and n denotes the number of subjects. Data set, is the name of the SAS data set. The outcome variables should be ranked prior to using this macro. There should be one record for each subject and each subject should have multiple variables which denote the rank at the j^{th} time point. The only variables that should be included in the data set are these within subject ranks. If these macro variables are specified correctly this macro will output a data set k which provides the Koch and Sen test statistic, the degrees of freedom and the corresponding p-value.

```
%macro koch(data set,numTime,n);
proc iml;
    numTime = &numTime;
    n = &n;
    dimension = numTime*n;
/*Reading data set into IML*/
    use &data set;
    read all var _NUM_ into X[colname=varNames];
/*Creating Indicator Variable for Nonmissing*/
    NOTMISS=j(n,numTime,1);
    DO g=1 TO n;
        DO d=1 TO numTime;
            IF X[g, d]=. THEN NOTMISS[g, d]=0;
        END;
    END;
/*Creating inflation factor vector*/
    NI=1/(NOTMISS[,+]+1);
/*Calculation of inflated mu vector*/
    INF_X=j(n,numTime,.);
    DO e=1 TO n;
        DO f=1 to numTime;
            INF_X[e,f]=X[e,f]*NI[e,1];
        END;
    END;
/*Creating Expected Value Matrix*/
    Y=((1/2))*j(n,numTime);
```

```
    DIFF=INF_X-Y;
/*Creating Mu Matrix*/
    MU=DIFF[+,];
/*Creating Variance Matrix*/
    VNTIME=t(1:numTime);
    SUBJ =j(n,1,1);
    IND=SUBJ@VNTIME;
/*Creating elements for summation*/
    Z = j(numTime*n,numTime,1);
    DO a=1 to dimension;
        DO b=1 to numTime;
            Z[a,b]=DIFF[ceil(a/numTime),mod(a-1,numTime)+1]*DIFF[ceil(a/numTime),b];
            IF Z[a,b]=. then Z[a,b]=0;
        END;
    END;
    NEW_Z=IND||Z;
    create z from NEW_Z;
    append from NEW_Z;
    create MU from MU;
    append from MU;
run;
quit;
/*Creating variance matrix as summation of variance components*/
data z; set z; rename COL1=row; run;
%macro combine_var(inds,c,outds);
proc sql; create table &outds as
    select distinct row,
    %DO s=2 %TO &c; sum(col&s) as fcol&s %IF &s¡&c %THEN,;
    %END;
    from &inds
    group by row
    order by row;
quit;
%mend;

    %combine_var(z,&numTime+1,p);
data p (drop=row); set p; run;

    proc iml;
    use p;
    read all var _NUM_ into VAR;
    use mu;
    read all var _NUM_ into MU;
```

```
    numTime = &numTime;
    n = &n;
    dimension = numTime*n;
/*Making Mu Vector and Covariance Matrix Singular*/
    MU_U =MU[2:numTime];
    VAR_U=VAR[2:numTime,2:numTime];
/*Inverting Covariance Matrix*/
    INV_VAR=inv(VAR_U);
    KOCH=t(MU_U)*INV_VAR*MU_U;
    DATASET = KOCH ||numTime;
    create k_&data set from DATASET;
    append from DATASET;
run;
quit;
/*Creation of final data set*/
    data k (drop=numTimes);
    set k_&data set;
    dof=col2-1;
    p_value=1-probchi(col1,col2-1);
    rename col1=Koch;
    rename col2=numTimes;
    run;
%mend Koch;
```

A.4 Tables

Table A.1: Type I Error Rates

Number of Subjects	Number of Obs Per Subject	Correlation	Revised	Complete
10	5	0.1	0.0008	0.0014
10	5	0.3	0.0012	0.0017
10	5	0.5	0.0007	0.0017
10	10	0.1	-	-
10	10	0.3	-	-
10	10	0.5	-	-
50	5	0.1	0.0465	0.0477
50	5	0.3	0.0467	0.0484
50	5	0.5	0.0453	0.0449
50	10	0.1	0.0311	0.032
50	10	0.3	0.0323	0.034
50	10	0.5	0.0325	0.0356
100	5	0.1	0.0452	0.0471
100	5	0.3	0.0459	0.0452
100	5	0.5	0.0447	0.0438
100	10	0.1	0.0442	0.0455
100	10	0.3	0.039	0.0422
100	10	0.5	0.0439	0.0437

Table A.2: Power Under a Linear Increase of 0.25

Number of Subjects	Number of Obs Per Subject	Correlation	Revised	Complete
10	5	0.1	0.0008	0.002
10	5	0.3	0.0012	0.002
10	5	0.5	0.0024	0.002
10	10	0.1	-	-
10	10	0.3	-	-
10	10	0.5	-	-
50	5	0.1	0.069	0.07
50	5	0.3	0.0707	0.073
50	5	0.5	0.0663	0.069
50	10	0.1	0.05	0.048
50	10	0.3	0.0458	0.051
50	10	0.5	0.045	0.046
100	5	0.1	0.1173	0.127
100	5	0.3	0.1107	0.116
100	5	0.5	0.1102	0.122
100	10	0.1	0.0866	0.091
100	10	0.3	0.0777	0.073
100	10	0.5	0.0767	0.075

Table A.3: Power Under a Linear Increase of 1

Number of Subjects	Number of Obs Per Subject	Correlation	Revised	Complete
10	5	0.1	0.0046	0.01
10	5	0.3	0.0064	0.009
10	5	0.5	0.0076	0.012
10	10	0.1	-	-
10	10	0.3	-	-
10	10	0.5	-	-
50	5	0.1	0.6171	0.688
50	5	0.3	0.5876	0.657
50	5	0.5	0.6101	0.676
50	10	0.1	0.4821	0.536
50	10	0.3	0.391	0.438
50	10	0.5	0.3214	0.365
100	5	0.1	0.9387	0.967
100	5	0.3	0.9257	0.958
100	5	0.5	0.9213	0.95
100	10	0.1	0.8989	0.932
100	10	0.3	0.816	0.861
100	10	0.5	0.7288	0.77

Table A.4: Average Pain Scores By Period of the Day for IBS Study

Subject	Wake Up	Morning	Midday	Evening	Bedtime
D01	4.20	3.75	3.00	2.77	3.77
D02	0.86	0.79	0.23	0.31	0.45
D03	2.25	3.15	4.29	4.62	3.00
D05	1.25	1.00	2.14	1.08	1.18
D06	5.73	6.00	5.88	5.50	4.88
D07	0.50	0.71	0.50	0.60	0.46
D09	0.27	0.27	.	.	0.08
D10	0.64	0.87	0.35	0.23	0.38
D11	0.92	0.08	0.17	0.17	0.22
D12	0.73	1.31	0.93	0.92	0.50
D13	0.00	0.31	0.20	0.38	0.23
D14	0.64	0.43	0.64	1.07	0.62
D15	1.00	0.75	0.64	1.38	2.25
D16	2.50	2.27	2.08	2.15	2.38
D17	5.21	4.07	4.08	4.15	2.09
D18	1.86	1.43	1.50	1.71	1.15
D19	0.21	0.13	0.62	0.15	0.17
D20	0.00	0.00	0.00	0.17	0.00
D21	0.29	0.54	1.92	3.50	.
D23	0.33	0.50	0.00	0.33	.
D24	2.41	1.65	1.63	1.31	1.88
D25	1.15	1.86	2.07	0.92	0.70
D26	1.93	1.58	1.62	1.50	1.11
D27	0.95	1.50	1.79	1.70	.
D28	3.14	1.21	1.13	0.86	.
D29	0.47	0.47	0.00	0.00	0.13
M02	4.63	4.05	3.37	3.38	3.80
M03	3.00	2.93	4.14	4.00	3.63
M04	2.29	2.07	2.64	2.06	2.08
M05	1.67	1.29	0.69	0.87	1.42
M06	0.87	0.22	0.30	0.31	0.07
M07	1.43	1.43	0.86	0.71	1.31
M08	0.92	0.77	1.27	0.64	0.75
M09	1.93	2.36	2.00	1.69	2.23
M11	1.47	1.38	0.93	0.93	0.47

Table A.4 Continued: Average Pain Scores By Period of the Day for IBS Study

Subject	Wake Up	Morning	Midday	Evening	Bedtime
M12	0.00	0.00	0.22	0.25	.
M13	4.64	4.14	3.43	4.15	5.45
M15	1.86	2.00	3.29	1.73	.
M16	4.31	4.85	4.75	4.75	5.00
M17	4.40	5.00	4.85	5.21	5.00
M18	1.00	1.40	0.87	0.79	1.27
M20	1.46	0.50	0.25	0.27	0.25
M21	1.85	2.07	3.23	2.82	3.17
M22	0.00	0.09	0.00	0.00	0.00
M23	2.93	2.79	2.67	3.15	3.00
M24	3.36	3.38	2.92	3.10	.

Appendix B

Chapter 3

B.1 Variance

$$
\begin{aligned}
E[r_{ij}^2] &= (1^2)\frac{1}{n} + (2^2)\frac{1}{n} + ... + (n_i^2)\frac{1}{n} + \left(\frac{n+n_i+1}{2}\right)^2 \frac{n-n_i}{n} \\
&= \frac{1}{n}\left(\sum_{s=1}^{n_i} s^2 + \left(\frac{n^2+n_i^2+2nn_i+2n+2n_i+1}{4}\right)(n-n_i)\right) \\
&= \frac{1}{n}\left(\left(\frac{n_i(n_i+1)(2n_i+1)}{6}\right) + \left(\frac{n^3-n_i^3+n^2n_i-nn_i^2+2n^2-2n_i^2+n-n_i}{4}\right)\right) \\
&= \frac{(4n_i^3+6n_i^2+2n_i)+(3n^3-3n_i^3+3n^2n_i-3nn_i^2+6n^2-6n_i^2+3n-3n_i)}{12n} \\
&= \frac{3n^3+n_i^3+3n^2n_i-3nn_i^2+6n^2+3n-n_i}{12n} \\
E[r_{ij}]^2 &= \frac{n^2+2n+1}{4} \\
&= \frac{3n^3+6n^2+3n}{12n} \\
Var[r_{ij}] &= \frac{n_i^3+3n^2n_i-3nn_i^2-n_i}{12n} \\
&= \frac{(n_i-n))^3+(n^3-n_i)}{12n}
\end{aligned}
$$

Therefore

$$
\begin{aligned}
Var\left(\mu_{Rj}\right) &= \sum_{i=1}^{k} w_i^2 \left(\frac{(n_i-n)^3+(n^3-n_i)}{12n}\right) \\
&= \sum_{i=1}^{k} w_i^2 \left(\frac{n_i^3+3n^2n_i-3nn_i^2-n_i}{12n}\right)
\end{aligned}
$$

B.2 Covariance

Note:

$$2\sum_{k=1}^{n} k(k-1) = \frac{k(k+1)(k-1)(3k+2)}{12}$$

Therefore:

$$\begin{aligned}
E[r_{ij}r_{ij'}] &= \frac{1}{\binom{n}{2}}(1)(2) + ... + \frac{1}{\binom{n}{2}}(n_i-1)n_i + \frac{\binom{n-n_i}{1}}{\binom{n}{2}}(1)(\frac{n+n_i+1}{2}) + ... + \\
&\quad \frac{\binom{n-n_i}{1}}{\binom{n}{2}}(n_i)\left(\frac{n+n_i+1}{2}\right) + \frac{\binom{n-n_i}{2}}{\binom{n}{2}}\left(\frac{n+n_i+1}{2}\right)\left(\frac{n+n_i+1}{2}\right) \\
&= \frac{1}{n(n-1)}\left(\frac{n_i(n_i+1)(n_i-1)(3n_i+2)}{12}\right) + \\
&\quad \frac{2(n-n_i)}{n(n-1)}\left(\frac{n_i(n_i+1)}{2}\right)\left(\frac{n+n_i+1}{2}\right) + \\
&\quad \frac{(n-n_i)(n-n_i-1)}{n(n-1)}\left(\frac{n+n_i+1}{2}\right)^2 \\
&= \frac{3n_i^4+2n_i^3-3n_i^2-2n_i}{12n(n-1)} + \frac{n_i^2n^2+nn_i^2+n^2n_i+nn_i-n_i^4-2n_i^3-n_i^2}{2n(n-1)} + \\
&\quad \frac{n^4+n_i^4+n^3+3n_i^3-n^2+3n_i^2-2n^2n_i^2-3n^2n_i-nn_i^2-2nn_i-n+n_i}{4n(n-1)} \\
&= \frac{3n^4+3n^3-n_i^3-3n^2-3n^2n_i+3nn_i^2-3n+n_i}{12n(n-1)} \\
E[r_{ij}]^2 &= \left(\frac{n^2+2n+1}{4}\right) \\
&= \frac{3n^4+3n^3-3n^2-3n}{12n(n-1)} \\
Cov[r_{ij}] &= E[r_{ij}, r_{ij'}] - E[r_{ij}]^2 = -\frac{n_i^3-3nn_i^2+3n^2n_i-n_i}{12n(n-1)}
\end{aligned}$$

Therefore

$$Cov\left(\mu_{Rj}, \mu_{Rj'}\right) = -\sum_{i=1}^{k} w_i^2\left(\frac{n_i^3-3nn_i^2+3n^2n_i-n_i}{12n(n-1)}\right)$$

B.3 Tables

Table B.1: Type I Error Rates

Number of Subjects	Number of Obs Per Subject	% Informative Missing	Revised	Prentice	Complete
10	5	20	0.0417	0.0413	0.042
10	5	50	0.04	0.0396	0.042
10	5	80	0.0435	0.0402	0.042
10	5	100	0.039	0.043	0.042
10	10	20	-	-	-
10	10	50	-	-	-
10	10	80	-	-	-
10	10	100	-	-	-
50	5	20	0.0454	0.046	0.0469
50	5	50	0.0481	0.0476	0.0469
50	5	80	0.0451	0.0483	0.0469
50	5	100	0.0529	0.0499	0.0473
50	10	20	0.0452	0.0495	0.042
50	10	50	0.0482	0.0456	0.042
50	10	80	0.0511	0.044	0.042
50	10	100	0.048	0.045	0.042

Table B.2: Power Under a Linear Increase of 0.25

Number of Subjects	Number of Obs Per Subject	% Informative Missing	Revised	Prentice	Complete
10	5	20	0.0866	0.0919	0.109
10	5	50	0.089	0.0836	0.109
10	5	80	0.0838	0.0836	0.109
10	5	100	0.089	0.082	0.109
10	10	20	-	-	-
10	10	50	-	-	-
10	10	80	-	-	-
10	10	100	-	-	-
50	5	20	0.4052	0.4531	0.549
50	5	50	0.4332	0.4483	0.549
50	5	80	0.4564	0.4557	0.549
50	5	100	0.471	0.455	0.549
50	10	20	0.3248	0.417	0.469
50	10	50	0.3593	0.4063	0.469
50	10	80	0.393	0.4087	0.469
50	10	100	0.432	0.397	0.469

Table B.3: Power Under the Alternative of a Linear Increase of 0.5

Number of Subjects	Number of Obs Per Subject	% Informative Missing	Revised	Prentice	Complete
10	5	20	0.262	0.3044	0.374
10	5	50	0.287	0.2949	0.374
10	5	80	0.3012	0.2945	0.374
10	5	100	0.321	0.288	0.374
10	10	20	-	-	-
10	10	50	-	-	-
10	10	80	-	-	-
10	10	100	-	-	-
50	5	20	0.9622	0.9823	0.99
50	5	50	0.9711	0.9774	0.99
50	5	80	0.9808	0.9788	0.99
50	5	100	0.983	0.978	0.99
50	10	20	0.9442	0.9812	0.996
50	10	50	0.959	0.9785	0.996
50	10	80	0.9701	0.9734	0.996
50	10	100	0.975	0.977	0.996

Table B.4: Complete Ranking of 4 Objects by 20 Subjects

	Object			
Subject	1	2	3	4
1	2	3	1	4
2	3	2	1	4
3	3	4	1	2
4	1	2	3	4
5	1	3	4	2
6	1	3	4	2
7	3	2	1	4
8	3	1	4	2
9	4	1	2	3
10	4	3	1	2
11	4	3	2	1
12	1	3	2	4
13	3	4	2	1
14	1	3	2	4
15	1	4	2	3
16	3	4	1	2
17	1	4	3	2
18	4	2	1	3
19	1	3	2	4
20	1	4	3	2

Table B.5: Ranking of 4 Objects by 20 Subjects with Missing Data

	Object			
Subject	1	2	3	4
1	2	3	1	4
2	3	2	1	4
3	3	4	1	2
4	1	2	3	4
5	1	3	4	2
6	1	3	4	2
7	3	2	1	4
8	3	1	4	2
9	4	1	2	3
10	4	3	1	2
11	4	3	2	1
12	1	3	2	4
13	3	.	2	1
14	1	.	2	3
15	1	.	2	3
16	3	.	1	2
17	1	.	3	2
18	3	.	1	2
19	1	.	2	3
20	1	.	3	2

Appendix C

Chapter 4

C.1 Variance

$$
\begin{aligned}
v_{jj} = Var(\mu_{Kj}) &= \sum_{i=1}^{k} Var\left(\frac{1}{n-n_i+1}\left(r_{ij}-\frac{n+1}{2}\right)\right) = \sum_{i=1}^{k}\left(\frac{1}{n-n_i+1}\right)^2 Var(r_{ij}) \\
&= \sum_{i=1}^{k}\left(\frac{1}{n-n_i+1}\right)^2 \left(E[r_{ij}^2]-E[r_{ij}]^2\right) \\
&= \sum_{i=1}^{k}\left(\frac{1}{n-n_i+1}\right)^2 \left((r_{ij}r_{ij})\frac{1}{2}+(n+1-r_{ij})(n+1-r_{ij})\frac{1}{2}-\left(\frac{n+1}{2}\right)^2\right) \\
&= \sum_{i=1}^{k}\left(\frac{1}{n-n_i+1}\right)^2 \left(r_{ij}^2-r_{ij}(n+1)+\frac{(n+1)^2}{2}-\frac{\frac{1}{2}(n+1)^2}{2}\right) \\
&= \sum_{i=1}^{k}\left(\frac{1}{n-n_i+1}\right)^2 \left(r_{ij}-\frac{n+1}{2}\right)^2
\end{aligned}
$$

C.2 Covariance

$$
\begin{aligned}
v_{jj'} &= Cov(\mu_{Kj}, \mu_{Kj'}) \\
&= Cov\left(\sum_{i=1}^{k}\left(\frac{1}{n-n_i+1}\left(r_{ij}-\frac{n+1}{2}\right)\right), \sum_{i=1}^{k}\left(\frac{1}{n-n_i+1}\left(r_{ij'}-\frac{n+1}{2}\right)\right)\right) \\
&= \sum_{i=1}^{k} Cov\left(\left(\frac{1}{n-n_i+1}\left(r_{ij}-\frac{n+1}{2}\right)\right), \left(\frac{1}{n-n_i+1}\left(r_{ij'}-\frac{n+1}{2}\right)\right)\right) \\
&= \sum_{i=1}^{k}\left(\frac{1}{n-n_i+1}\right)^2\left(E\left[r_{ij}r_{ij'}-r_{ij}\frac{n+1}{2}-r_{ij'}\frac{n+1}{2}+\left(\frac{n+1}{2}\right)^2\right]-0\right) \\
&= \sum_{i=1}^{k}\left(\frac{1}{n-n_i+1}\right)^2\left(E[r_{ij}r_{ij'}]-\left(\frac{n+1}{2}\right)^2\right) \\
&= \sum_{i=1}^{k}\left(\frac{1}{n-n_i+1}\right)^2\left((r_{ij}r_{ij'})\frac{1}{2}+(n+1-r_{ij})(n+1-r_{ij'})\frac{1}{2}-\left(\frac{n+1}{2}\right)^2\right) \\
&= \sum_{i=1}^{k}\left(\frac{1}{n-n_i+1}\right)^2\left(r_{ij}r_{ij'}-\frac{r_{ij}(n+1)}{2}-\frac{r_{ij'}(n+1)}{2}+\frac{(n+1)^2}{2}-\frac{\frac{1}{2}(n+1)^2}{2}\right) \\
&= \sum_{i=1}^{k}\left(\frac{1}{n-n_i+1}\right)^2\left(r_{ij}r_{ij'}-\frac{r_{ij}(n+1)}{2}-\frac{r_{ij'}(n+1)}{2}+\left(\frac{n+1}{2}\right)^2\right) \\
&= \sum_{i=1}^{k}\left(\frac{1}{n-n_i+1}\right)^2\left(r_{ij}-\frac{n+1}{2}\right)\left(r_{ij'}-\frac{n+1}{2}\right)
\end{aligned}
$$

C.3 SAS Macro for Statistic with Informative Missing data

This macro requires the input of three variables. The variable numTime denotes the maximum number of measurements collected on any subject and n denotes the number of subjects. Data set, is the name of the SAS data set. The outcome variables should be ranked prior to using this macro. There should be one record for each subject and each subject should have multiple variables which denote the rank at the j^{th} time point. The only variables that should be included in the data set are these within subject ranks. If these macro variables are specified correctly this macro will output a data set k_info which provides the Koch and Sen test statistic for informative missing data, the degrees of freedom and the corresponding p-value.

```
%macro koch_inform(data set,numTime,n);
proc iml;
    numTime = &numTime;
    n = &n;
    dimension = numTime*n;
/*Reading data set into IML*/
    use &data set;
    read all var _NUM_ into X[colname=varNames];
/*Creating Indicator Variable for Nonmissing*/
    NOTMISS=j(n,numTime,1);
    DO g=1 TO n;
        DO d=1 TO numTime;
            IF X[g, d]=. THEN NOTMISS[g, d]=0;
        END;
    END;
/*Creating inflation factor vector*/
    NI=1/(NOTMISS[,+]+1);
/*Imputed Value*/
    IM=((NOTMISS[,+])+numTime+1)/2;
/*Imputed Matrix*/
    NEWX=j(n,numTime,1);
    DO g=1 TO n;
        DO d=1 TO numTime;
            IF X[g,d]=. THEN NEWX[g,d]=IM[g,];
            ELSE NEWX[g,d]=X[g,d];
        END;
```

```
    END;
/*Creating Expected Value Matrix*/
    Y=j(n,numTime,(numTime+1)/2);
    U_DIFF=NEWX-Y;
/*Creating Mu Matrix*/
    DIFF=j(n,numTime);
    DO e=1 TO n;
        DO f=1 to numTime;
            DIFF[e,f]=U_DIFF[e,f]*NI[e,1];
        END;
    END;
/*Creating Mu Matrix*/
    MU=DIFF[+,];
/*Creating Variance Matrix*/
    VNTIME=t(1:numTime);
    SUBJ =j(n,1,1);
    IND=SUBJ@VNTIME;
/*Creating elements for summation*/
    WEIGHT=NI@j(numTime,1,1);
    Z = j(numTime*n,numTime,1);
    DO a=1 to dimension;
        DO b=1 to numTime;
            Z[a,b]=(U_DIFF[ceil(a/numTime),mod(a-1,numTime)+1]
                *U_DIFF[ceil(a/numTime),b])*(WEIGHT[a,1]**2);
            IF Z[a,b]=. then Z[a,b]=0;
        END;
    END;
    NEW_Z=IND||Z;
    create z from NEW_Z;
    append from NEW_Z;
    create MU from MU;
    append from MU;
run;
quit;
/*Creating variance matrix as summation of variance components*/
data z; set z; rename COL1=row; run;
%macro combine_var(inds,c,outds);
proc sql; create table &outds as
    select distinct row,
    %DO s=2 %TO &c; sum(col&s) as fcol&s %IF &s¡&c %THEN,;
    %END;
    from &inds
    group by row
```

```
    order by row;
quit;
%mend;

    %combine_var(z,&numTime+1,p);
data p (drop=row); set p; run;

    proc iml;
    use p;
    read all var _NUM_ into VAR;
    use mu;
    read all var _NUM_ into MU;
    numTime = &numTime;
    n = &n;
    dimension = numTime*n;
/*Making Mu Vector and Covariance Matrix Singular*/
    MU_U =MU[2:numTime];
    VAR_U=VAR[2:numTime,2:numTime];
/*Inverting Covariance Matrix*/
    INV_VAR=inv(VAR_U);
    KOCH=t(MU_U)*INV_VAR*MU_U;
    DATASET = KOCH ||numTime;
    create k_&data set from DATASET;
    append from DATASET;
run;
quit;
/*Creation of final data set*/
    data k_info (drop=numTimes);
    set k_&data set;
    dof=col2-1;
    p_value=1-probchi(col1,col2-1);
    rename col1=Koch;
    rename col2=numTimes;
    run;
%mend koch_inform;
```

C.4 Tables

Table C.1: Type I Error Rates - 10 Subjects

Number of Subjects	Number of Obs Per Subject	% Informative Missing	Correlation	Revised	Prentice	Complete
10	5	20	0.1	0.0003	0.0013	0
10	5	20	0.3	0.0003	0.0013	0.001
10	5	20	0.5	0.0003	0.0011	0.001
10	5	50	0.1	0.0003	0.0011	0
10	5	50	0.3	0.0009	0.0008	0.001
10	5	50	0.5	0.0002	0.0012	0.001
10	5	80	0.1	0.0002	0.0004	0
10	5	80	0.3	0.0003	0.0005	0.001
10	5	80	0.5	0.0004	0.0012	0.001
10	5	100	0.1	0	0	0
10	5	100	0.3	0	0.001	0.001
10	5	100	0.5	0.001	0.002	0.001

Table C.2: Type I Error Rates - 50 Subjects

Number of Subjects	Number of Obs Per Subject	% Informative Missing	Correlation	Revised	Prentice	Complete
50	5	20	0.1	0.0411	0.0447	0.0463
50	5	20	0.3	0.0418	0.0434	0.0408
50	5	20	0.5	0.0347	0.0385	0.036
50	10	20	0.1	0.0264	0.0338	0.0358
50	10	20	0.3	0.0246	0.0327	0.0321
50	10	20	0.5	0.0276	0.0364	0.036
50	5	50	0.1	0.0432	0.0451	0.046
50	5	50	0.3	0.0456	0.0443	0.0431
50	5	50	0.5	0.0391	0.0383	0.036
50	10	50	0.1	0.0264	0.0357	0.037
50	10	50	0.3	0.027	0.0361	0.0371
50	10	50	0.5	0.0263	0.0369	0.036
50	5	80	0.1	0.044	0.0451	0.0443
50	5	80	0.3	0.0442	0.0422	0.0425
50	5	80	0.5	0.0444	0.0417	0.036
50	10	80	0.1	0.0281	0.0381	0.0343
50	10	80	0.3	0.0252	0.0354	0.0351
50	10	80	0.5	0.0257	0.043	0.036
50	5	100	0.1	0.0397	0.0447	0.041
50	5	100	0.3	0.0488	0.0457	0.045
50	5	100	0.5	0.048	0.044	0.036
50	10	100	0.1	0.0256	0.037	0.035
50	10	100	0.3	0.0255	0.0345	0.0363
50	10	100	0.5	0.026	0.041	0.036

Table C.3: Type I Error Rates - 100 Subjects

Number of Subjects	Number of Obs Per Subject	% Informative Missing	Correlation	Revised	Prentice	Complete
100	5	20	0.1	0.0495	0.0498	0.044
100	5	20	0.3	0.0462	0.0471	0.048
100	5	20	0.5	0.0459	0.047	0.051
100	10	20	0.1	0.0386	0.0439	0.043
100	10	20	0.3	0.0379	0.0411	0.037
100	10	20	0.5	0.0408	0.0409	0.04
100	5	50	0.1	0.0443	0.0473	0.044
100	5	50	0.3	0.0498	0.046	0.048
100	5	50	0.5	0.0478	0.0459	0.051
100	10	50	0.1	0.0398	0.0418	0.043
100	10	50	0.3	0.0389	0.0413	0.037
100	10	50	0.5	0.0392	0.0415	0.04
100	5	80	0.1	0.0451	0.0467	0.044
100	5	80	0.3	0.056	0.0408	0.048
100	5	80	0.5	0.0524	0.0458	0.051
100	10	80	0.1	0.0384	0.0416	0.043
100	10	80	0.3	0.0365	0.0379	0.037
100	10	80	0.5	0.0404	0.0376	0.04
100	5	100	0.1	0.048	0.044	0.044
100	5	100	0.3	0.055	0.037	0.048
100	5	100	0.5	0.05	0.054	0.051
100	10	100	0.1	0.028	0.038	0.043
100	10	100	0.3	0.04	0.044	0.037
100	10	100	0.5	0.038	0.047	0.04

Table C.4: Power Under a Linear Increase of 0.25 - 10 Subjects

Number of Subjects	Number of Obs Per Subject	% Informative Missing	Correlation	Revised	Prentice	Complete
10	5	20	0.1	0.0008	0.0018	0.002
10	5	20	0.3	0.0006	0.0004	0.002
10	5	20	0.5	0.0006	0.0019	0.002
10	10	20	0.1	-	-	-
10	10	20	0.3	-	-	-
10	10	20	0.5	-	-	-
10	5	50	0.1	0.0006	0.0004	0.002
10	5	50	0.3	0.0005	0.0014	0.002
10	5	50	0.5	0.0003	0.0009	0.002
10	10	50	0.1	-	-	-
10	10	50	0.3	-	-	-
10	10	50	0.5	-	-	-
10	5	80	0.1	0.0007	0.0014	0.002
10	5	80	0.3	0.0001	0.0005	0.002
10	5	80	0.5	0.0004	0.0004	0.002
10	10	80	0.1	-	-	-
10	10	80	0.3	-	-	-
10	10	80	0.5	-	-	-
10	5	100	0.1	0.001	0.002	0.002
10	5	100	0.3	0	0	0.002
10	5	100	0.5	0	0	0.002
10	10	100	0.1	-	-	-
10	10	100	0.3	-	-	-
10	10	100	0.5	-	-	-

Table C.5: Power Under a Linear Increase of 0.25 - 50 Subjects

Number of Subjects	Number of Obs Per Subject	% Informative Missing	Correlation	Revised	Prentice	Complete
50	5	20	0.1	0.0652	0.0729	0.07
50	5	20	0.3	0.0647	0.073	0.073
50	5	20	0.5	0.0579	0.0636	0.069
50	10	20	0.1	0.0368	0.0515	0.048
50	10	20	0.3	0.0319	0.0513	0.051
50	10	20	0.5	0.0299	0.0459	0.046
50	5	50	0.1	0.0705	0.0745	0.07
50	5	50	0.3	0.0632	0.0678	0.073
50	5	50	0.5	0.0609	0.0595	0.069
50	10	50	0.1	0.0325	0.0514	0.048
50	10	50	0.3	0.0312	0.0514	0.051
50	10	50	0.5	0.0319	0.0461	0.046
50	5	80	0.1	0.0739	0.0754	0.07
50	5	80	0.3	0.0694	0.0691	0.073
50	5	80	0.5	0.0652	0.0609	0.069
50	10	80	0.1	0.037	0.044	0.048
50	10	80	0.3	0.032	0.0459	0.051
50	10	80	0.5	0.0299	0.0513	0.046
50	5	100	0.1	0.074	0.081	0.07
50	5	100	0.3	0.072	0.068	0.073
50	5	100	0.5	0.066	0.063	0.069
50	10	100	0.1	0.033	0.043	0.048
50	10	100	0.3	0.03	0.041	0.051
50	10	100	0.5	0.025	0.044	0.046

Table C.6: Power Under a Linear Increase of 0.25 - 100 Subjects

Number of Subjects	Number of Obs Per Subject	% Informative Missing	Correlation	Revised	Prentice	Complete
100	5	20	0.1	0.1004	0.1087	0.127
100	5	20	0.3	0.0934	0.1047	0.116
100	5	20	0.5	0.0983	0.1117	0.122
100	10	20	0.1	0.0682	0.0866	0.091
100	10	20	0.5	0.0578	0.071	0.075
100	10	20	0.5	0.0578	0.071	0.075
100	5	50	0.1	0.1026	0.102	0.127
100	5	50	0.3	0.0976	0.0957	0.116
100	5	50	0.5	0.1027	0.1065	0.122
100	10	50	0.1	0.0736	0.0796	0.091
100	10	50	0.5	0.0604	0.0646	0.075
100	10	50	0.5	0.0604	0.0646	0.075
100	5	80	0.1	0.1033	0.0941	0.127
100	5	80	0.3	0.1021	0.0957	0.116
100	5	80	0.5	0.1061	0.0994	0.122
100	10	80	0.1	0.0756	0.0809	0.091
100	10	80	0.5	0.0629	0.0604	0.075
100	10	80	0.5	0.0629	0.0604	0.075
100	5	100	0.1	0.104	0.088	0.127
100	5	100	0.3	0.106	0.092	0.116
100	5	100	0.5	0.109	0.1	0.122
100	10	100	0.1	0.076	0.07	0.091
100	10	100	0.5	0.064	0.068	0.075
100	10	100	0.5	0.064	0.068	0.075

Table C.7: Power Under a Linear Increase of 1 - 10 Subjects

Number of Subjects	Number of Obs Per Subject	% Informative Missing	Correlation	Revised	Prentice	Complete
10	5	20	0.1	0.0015	0.0039	0.01
10	5	20	0.3	0.0011	0.0041	0.009
10	5	20	0.5	0.0006	0.0058	0.012
10	10	20	0.1	-	-	-
10	10	20	0.3	-	-	-
10	10	20	0.5	-	-	-
10	5	50	0.1	0.0015	0.0042	0.01
10	5	50	0.3	0.0017	0.0059	0.009
10	5	50	0.5	0.0014	0.0041	0.012
10	10	50	0.1	-	-	-
10	10	50	0.3	-	-	-
10	10	50	0.5	-	-	-
10	5	80	0.1	0.002	0.0071	0.01
10	5	80	0.3	0.0023	0.0038	0.009
10	5	80	0.5	0.0016	0.0027	0.012
10	10	80	0.1	-	-	-
10	10	80	0.3	-	-	-
10	10	80	0.5	-	-	-
10	5	100	0.1	0.003	0.006	0.01
10	5	100	0.3	0.003	0.003	0.009
10	5	100	0.5	0.003	0.002	0.012
10	10	100	0.1	-	-	-
10	10	100	0.3	-	-	-
10	10	100	0.5	-	-	-

Table C.8: Power Under a Linear Increase of 1 - 50 Subjects

Number of Subjects	Number of Obs Per Subject	% Informative Missing	Correlation	Revised	Prentice	Complete
50	5	20	0.1	0.5222	0.5779	0.688
50	5	20	0.3	0.4994	0.5585	0.657
50	5	20	0.5	0.506	0.584	0.676
50	10	20	0.1	0.3076	0.4346	0.536
50	10	20	0.3	0.2419	0.3624	0.438
50	10	20	0.5	0.1942	0.3065	0.365
50	5	50	0.1	0.5384	0.5346	0.688
50	5	50	0.3	0.5189	0.5249	0.657
50	5	50	0.5	0.5278	0.5455	0.676
50	10	50	0.1	0.3308	0.3918	0.536
50	10	50	0.3	0.2693	0.3246	0.438
50	10	50	0.5	0.2143	0.2817	0.365
50	5	80	0.1	0.5658	0.5142	0.688
50	5	80	0.3	0.5351	0.5054	0.657
50	5	80	0.5	0.5506	0.5395	0.676
50	10	80	0.1	0.3701	0.3537	0.536
50	10	80	0.3	0.292	0.3051	0.438
50	10	80	0.5	0.2285	0.2553	0.365
50	5	100	0.1	0.581	0.511	0.688
50	5	100	0.3	0.548	0.49	0.657
50	5	100	0.5	0.566	0.522	0.676
50	10	100	0.1	0.381	0.323	0.536
50	10	100	0.3	0.307	0.285	0.438
50	10	100	0.5	0.234	0.251	0.365

Table C.9: Power Under a Linear Increase of 1 - 100 Subjects

Number of Subjects	Number of Obs Per Subject	% Informative Missing	Correlation	Revised	Prentice	Complete
100	5	20	0.1	0.8756	0.9142	0.967
100	5	20	0.3	0.8591	0.9007	0.958
100	5	20	0.5	0.8567	0.9032	0.95
100	10	20	0.1	0.7602	0.8725	0.932
100	10	20	0.3	0.6529	0.7817	0.861
100	10	20	0.5	0.5458	0.6944	0.77
100	5	50	0.1	0.8924	0.8816	0.967
100	5	50	0.3	0.8667	0.8762	0.958
100	5	50	0.5	0.8704	0.8841	0.95
100	10	50	0.1	0.7862	0.833	0.932
100	10	50	0.3	0.674	0.7457	0.861
100	10	50	0.5	0.5651	0.6629	0.77
100	5	80	0.1	0.9035	0.8568	0.967
100	5	80	0.3	0.8752	0.8482	0.958
100	5	80	0.5	0.8829	0.8502	0.95
100	10	80	0.1	0.8057	0.8014	0.932
100	10	80	0.3	0.6989	0.714	0.861
100	10	80	0.5	0.598	0.6414	0.77
100	5	100	0.1	0.915	0.834	0.967
100	5	100	0.3	0.883	0.831	0.958
100	5	100	0.5	0.886	0.829	0.95
100	10	100	0.1	0.834	0.782	0.932
100	10	100	0.3	0.723	0.685	0.861
100	10	100	0.5	0.611	0.616	0.77

Table C.10: Avg. Difference in BM Pain Scores By Period of Day

Subject	Early Morning	Morning	Afternoon	Night
C01	.	-1	1	2
C03	.	5	2	2
C04	.	1	1	.
C05	.	0	-1	.
C06	.	2	-2	.
C08	.	.	0	.
C09	0	-1	0.5	-0.67
C11	.	1	.	1
C12	.	0.5	2.67	0.5
C13	.	0	.	.
C14	.	0	2	0.67
C15	1	-0.5	0	0.38
C16	.	-1.5	.	.
C18	.	0	0	.
C19	.	0.17	0.8	0
C20	3	2	.	.
C21	.	0	.	0.5
C22	.	3	.	.
D01	.	0.67	1	2
D02	.	0	0.4	.
D03	.	0	-0.06	-0.21
D04	1	.	1	.
D05	1	3.2	1.33	1.5
D08	.	-1.6	0.13	0
D09	.	0	2.33	.
D10	0	0	2.08	2.13
D11	.	0	0	.
D12	.	0.25	0	0.5
D14	4	5	4.33	6.2
D15	.	7	4	.
D16	.	0.83	0	.
D17	.	3.5	0.8	-0.25
D18	-1	.	0	-0.33
D19	2	2.33	3.11	4
D20	0.5	0	1	-1.5
D21	.	4	-0.33	4
D22	1.5	1.93	3	1.67

Table C.11: Ranked Difference in BM Pain Scores By Period of Day

Subject	Early Morning	Morning	Afternoon	Night
C01	.	1	2	3
C03	.	3	1.5	1.5
C04	.	1.5	1.5	.
C05	.	2	1	.
C06	.	2	1	.
C08	.	.	1	.
C09	3	1	4	2
C11	.	1.5	.	1.5
C12	.	1.5	3	1.5
C13	.	1	.	.
C14	.	1	3	2
C15	4	1	2	3
C16	.	1	.	.
C18	.	1.5	1.5	.
C19	.	2	3	1
C20	2	1	.	.
C21	.	1	.	2
C22	.	1	.	.
D01	.	1	2	3
D02	.	1	2	.
D03	.	3	2	1
D04	1.5	.	1.5	.
D05	1	4	2	3
D08	.	1	3	2
D09	.	1	2	.
D10	1.5	1.5	3	4
D11	.	1.5	1.5	.
D12	.	2	1	3
D14	1	3	2	4
D15	.	2	1	.
D16	.	2	1	.
D17	.	3	2	1
D18	1	.	3	2
D19	1	2	3	4
D20	3	2	4	1
D21	.	2.5	1	2.5
D22	1	3	4	2

Bibliography

Akritas, Michael G. and Edgar Brunner. 1997. "A unified approach to rank tests for mixed models." *Journal of Statistical and Planning and Inference* 61:249–277.

Bell, Bethany A., Grant B. Morgan, Jason A. Schoeneberger and Brandon L. Loudermilk. 2010. Dancing the Sample Size Limbo with Mixed Models: How Low Can You Go? In *Proceedings of the SAS Global Forum 2010 Conference*. SAS Global Forum.

Bernard, A. and PH. Van Elteren. 1953. "A generalization of the method of m rankings." *Indagationes Mathematicae* 15:358–369.

Catellier, Diane and Keith Muller. 2000. "Tests for Gaussian repeated measures with missing data in small samples." *Statistics in Medicine* 19:1101–1114.

Cox, D.R. and N. Reid. 1987. "Parameter Orthogonality and Approximate Conditional Inference." *Jouranl of the Royal Statistical Society. Series B (Methodological)* 49(1):1–39.

Demidenko, Eugene. 2004. *Mixed Models: Theory and Applications*. Wiley-Interscience.

Durbin, J. 1951. "Incomplete Blocks in Ranking Experiments." *British Journal of Mathematical and Statistical Psychology* 4(2):85–90.

Elteren, PH. Van. 1960. "On the Combination of Independent Two Sample Tests of Wilcoxin." *Bulletin of the International Statistical Institute* 37:1–13.

Fitzmaurice, Garrett, Nan Laird and James Waire. 2004. *Applied Longitudinal Analysis*. Wiley-Interscience.

Fouladi, Rachel T. and Yann-Yann Shieh. 2004. "A Comparison of Two General Approaches to Mixed Model Longitudinal Analyses Under Small Sample Conditions." *Communications in Statistics - Simulation and Computation* 33(3):807–824.

Friedman, Milton. 1937. "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance." *Journal of the American Statistical Association* 32(200):675–701.

Gao, Xin. 2007. "A Nonparametric Procedure for the Two-Factor Mixed Model with Missing Data." *Biometrical Journal* 49(5):774–788.

Gurka, Matthew J., Lloyd J. Edwards and Keith E. Muller. 2011. "Avoiding bias in mixed model inference for fixed effects." *Statistics in Medicine* 30(22):2696–2707.

Hedeker, Donald R. and Robert D. Gibbons. 2006. *Longitudinal Data Analysis*. John Wiley and Sons.

Hodges, JL and EL Lehmann. 1962. "Rank Methods for Combination of Independent Experiments in Analysis of Variance." *The Annals of Mathematical Statistics* 33(2):482–497.

Horton, Nicholas and Ken Kleinman. 2007. "Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models." *Journal of the American Statistical Association* 61(1):79–90.

Horton, Nicholas and Stuart Lipsitz. 2001. "Multiple Imputation in Practice: Comparison of Software Packages for Regression Models WIth Missing Variables." *The American Statistician* 55(3):244–254.

Kawaguchi, Atsushi and Gary Koch. 2010. "Multivariate Mann-Whitney Estimators for the Comparison of Two Treatments in a Three-Period Crossover Study with Randomly Missing Data." *Journal of Biopharmaceutical Statistics* 20(4):720–744.

Kendall, M.G. and B. Babington Smith. 1939. "The Problem of m Rankings." *The Annals of Mathematical Statistics* 10(3):275–287.

Kenward, Michael and James Roger. 1997. "Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood." *Biometrics* 53(3):983–997.

Koch, Gary G. and Pranab Kumar Sen. 1968. "Some aspects of the statistical analysis of the 'mixed model'." *Biometrics* pp. 27–48.

Landis, J Richard, Eugene R Heyman and Gary Koch. 1978. "Average Partial Association in Three-Way Contingency Tables: A Review and Discussion of Alternative Tests." *International Statistical Review* 46(3):237–254.

Lawley, D.N. 1956. "A General Method for Approximating to the Distribution of Likelihood Ratio Criteria." *Biometrika* 43(3/4):295–303.

Lehmann, Erich L. and H.J.M. D'Abrera. 2006. *Nonparametrics: Statistical Methods Based on Ranks*. Springer.

Liang, Kung-Yee and Scott L. Zeger. 1986. "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika* 73(1):13–22.

Mack, Gregory and John Skillings. 1980. "A Friedman-Type Rank Test for Main Effects in a Two-Factor ANOVA." *Journal of the American Statistical Association* 75(372):947–951.

Manor, Orly and David Zucker. 2004. "Small sample inference for the fixed effects in the mixed linear model." *Computational Statistics and Data Analysis* 46:801–817.

Mehrotra, Devan V, Xiaomin Lu and Xiaoming Li. 2010. "Rank-Based Analyses of Stratified Experiments: Alternatives to the van Elteren Test." *American Statistical Association* 62(2):121–130.

Prentice, M.J. 1979. "On the problem of m incomplete rankings." *Biometrika* 66(1):167–70.

Rai, S.C. 1987. "Rank Analysis of Block Designs Having Different Cell Frequencies." *Biometrical Journal* 29(3):293–298.

Rubin, Donald. 1987. *Multiple Imputation for Nonresponse in Surveys*. Wiley-Interscience.

SAS/STAT(R) 9.2 User's Guide, Second Edition. N.d.

Schafer, JL. 1997. *Analysis of Incomplete Multivariate Data*. Chapman and Hall/CRC.

Schafer, Joseph. 1999. "Multiple Imputation: A Primer." *Statistical Methods in Medical Research* 8(1):3–15.

Schluchter, Mark and Janet Elashoff. 1990. "Small-sample adjustments to tests with unbalanced repeated measures assuming several covariance structures." *Journal of Statistical Computation and Simulation* 37(1):69–87.

Sen, Pranab Kumar. 1968. "On a Class of Aligned Rank Order Tests in Two-way Layouts." *The Annals of Mathematical Statistics* 39(4):1115–1124.

Sen, Pranab Kumar and Madan Lal Puri. 1967. "On the theory of rank order tests for location in the multivariate one sample problem." *The Annals of Mathematical Statistics* pp. 1216–1228.

Skillings, John H. and Gregory A. Mack. 1981. "On the Use of a Friedman-Type Statistic in Balanced and Unbalanced Block Designs." *Technometrics* 23(2):171–177.

Stokes, Maura, Charles Davis and Gary Koch. 2000. *Categorical Data Analysis Using the SAS System*. SAS Institute, Inc.

Wittkowski, Knut M. 1988. "Friedman-Type Statistics and Consistent Multiple Comparisons for Unbalanced Designs with Missing Data." *Journal of the American Statistical Association* 83(404):1163–1170.

Zucker, David M., Offer Lieberman and Orly Manor. 2000. "Improved small sample inference in the mixed linear model:Bartlett correction and adjusted likelihood." *Journal of the Royal Statistical Society: Series B* 62(4):827–838.