Sample Size Considerations for Clustered and Correlated Outcomes in Clinical Research

XIAOHAN XU
Southern Methodist University, daisyxxh@gmail.com

Follow this and additional works at: https://scholar.smu.edu/hum_sci_statisticalscience_etds

Recommended Citation
https://scholar.smu.edu/hum_sci_statisticalscience_etds/10

This Dissertation is brought to you for free and open access by the Statistical Science at SMU Scholar. It has been accepted for inclusion in Statistical Science Theses and Dissertations by an authorized administrator of SMU Scholar. For more information, please visit http://digitalrepository.smu.edu.
SAMPLE SIZE CONSIDERATIONS FOR CLUSTERED AND CORRELATED OUTCOMES IN CLINICAL RESEARCH

Approved by:

Dr. Hong Zhu
Associate Professor in Department of Population & Data Sciences, UTSW

Dr. Chul Ahn
Professor in Department of Population & Data Sciences, UTSW

Dr. Daniel F. Heitjan
Professor in Department of Statistical Science, SMU & Population & Data Sciences, UTSW

Dr. Hon Keung Tony Ng
Professor in Department of Statistical Science, SMU
SAMPLE SIZE CONSIDERATIONS FOR CLUSTERED AND CORRELATED OUTCOMES IN CLINICAL RESEARCH

A Dissertation Presented to the Graduate Faculty of the
Dedman College
Southern Methodist University
in
Partial Fulfillment of the Requirements
for the degree of
Doctor of Philosophy
with a
Major in Biostatistics
by
Xiaohan Xu

B.A., Statistics, Southwestern University of Finance and Economics
M.S., Biostatistics, Georgetown University

December 21, 2019
ACKNOWLEDGMENTS

First of all, I would like to express my special appreciation and thanks to my advisor Dr. Hong Zhu for her endless support of my Ph.D study and related research. It has been an honor to be her first Ph.D. student. Her guidance and immense knowledge helped me in all the time of research and writing of this thesis. I appreciate all her contributions to make me become a biostatistician and could not have imagined having a better advisor and mentor for my Ph.D study.

Secondly, I would like to express my sincere gratitude to our director Dr. Daniel Heitjan for his continuous encouragement and support of my Ph.D study for the past four years. I also greatly appreciate working with Dr. Chul Ahn on my dissertation. He has provided very important guidance and many insightful suggestions. I am grateful for Dr. Hon Keung Ng, for his time, interest, and helpful comments to improve my dissertation, and also for his excellent teaching during my PhD study.

In addition, I would like to thank all of my friends. Thank you for listening, offering me advice and supporting me throughout the journey.

Last but not least, I would like to thank my family for all their love and encouragement. Thank my parents a lot for supporting me accomplish my dreams. And special thanks to my husband for his love and accompany. All of you made my life wonderful in Dallas.
Sample Size Considerations for Clustered and Correlated Outcomes in Clinical Research

Advisor: Dr. Hong Zhu

Doctor of Philosophy degree conferred December 21, 2019
Dissertation completed September 16, 2019

In this dissertation, we investigate sample size calculations for three different study designs: stratified cluster randomization trials (CRTs), paired experimental designs and paired cluster experimental designs.

Stratified CRTs have been frequently employed in clinical and healthcare research. Comparing with simple randomized CRTs, stratified CRTs reduce the imbalance of baseline prognostic factors among different intervention groups. Clusters are often naturally formed with random sizes in CRTs. With varying cluster size, commonly used ad hoc approaches ignore the variability in cluster size, which may underestimate (overestimate) the required number of clusters for each group per stratum and lead to underpowered (overpowered) clinical trials. In Chapter 2, we propose a closed-form sample size formula for estimating the required total number of subjects and for estimating the number of clusters for each group per stratum, based on Cochran-Mantel-Haenszel statistic for stratified cluster randomization design with binary outcomes, accounting for both clustering and varying cluster size. We investigate the impact of various design parameters on the relative change in number of clusters due to varying cluster size. Simulation studies are conducted to evaluate the finite-sample performance of the proposed sample size formula. A real application example of a pragmatic stratified CRT of a triad of chronic kidney disease (CKD), diabetes and hypertension is presented for illustration.
In paired experimental design, each study unit contributes a pair of observations. Investigators often encounter incomplete observations of paired outcomes in the data collected. Some study units contribute complete pairs of observations, while the others contribute either pre- or post-intervention observations. In Chapter 3, we derive a closed-form sample size formula based on the generalized estimating equation (GEE) approach by treating the incomplete observations as missing data in a linear model. The proposed method properly accounts for the impact of mixed structure of observed data: a combination of paired and unpaired outcomes. The sample size formula is flexible to accommodate different missing patterns, magnitude of missingness, and correlation parameter values. In the presence of missing data, the proposed method would lead to a more accurate sample size estimate comparing with the crude adjustment. Simulation studies are conducted to evaluate the finite-sample performance of the GEE sample size formula. A real application example is presented for illustration.

In Chapter 4, we extend the method in Chapter 3 and propose closed-form sample size formulas for paired cluster design with both continuous and binary outcomes, based on the GEE approach in generalized linear models. The sample size formulas are flexible to accommodate different correlation structures and missing patterns. In the simulation studies, we use bias-corrected sandwich variance estimators to address the issue of inflated type I error when the number of clusters is small. A real application example about physical fitness in Ecuadorian adolescents is presented for illustration.
# TABLE OF CONTENTS

| LIST OF FIGURES | ix |
| LIST OF TABLES | x |

## CHAPTER

1. **INTRODUCTION** ................................................. 1  
   1.1. Background .................................................. 1  
   1.2. Sample size considerations for stratified CRTs ...................... 1  
   1.3. Sample size considerations for paired experimental design ........... 3  
   1.4. Sample size considerations for paired cluster experimental design ...... 6  

2. **SAMPLE SIZE CONSIDERATIONS FOR STRATIFIED CRTs** ............ 8  
   2.1. Methods ..................................................... 8  
       2.1.1. Sample size estimation for varying cluster size ................. 8  
       2.1.2. Estimation of clustering parameter .......................... 13  
   2.2. Simulation ................................................... 15  
   2.3. Example ....................................................... 17  
   2.4. Acknowledgments ............................................... 18  

3. **SAMPLE SIZE CONSIDERATIONS FOR PAIRED EXPERIMENTAL DESIGN** . 21  
   3.1. Statistical method and sample size estimation ......................... 21  
       3.1.1. Sample size based on the GEE approach .......................... 21  
       3.1.2. Sample size based on the paired t-test ......................... 25  
       3.1.3. Crude adjustment for incomplete observations .................... 26  
   3.2. Simulation ................................................... 27  
   3.3. Example ....................................................... 30  
   3.4. Acknowledgments ............................................... 31
### LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>The plot of relative change in sample size ((R)) versus ICC ((\rho)). Assuming the numbers of clusters and CVs to be equal across strata. The common CVs are set as (\gamma = 0.5, 0.75, ) and (1). The means of cluster sizes are set as (\mu: \mu_1 = (10, 30, 50)) and (\mu_2 = (20, 30, 40)).</td>
<td>19</td>
</tr>
<tr>
<td>3.1</td>
<td>The plot of sample size ratio against within-subject correlation coefficient under missing patterns (A) and (B), and different combinations of ((p_1, p_2)).</td>
<td>31</td>
</tr>
<tr>
<td>4.1</td>
<td>The plot of sample size ratio against within-subject correlation coefficient under the general missing ((\tau_1 = 0.3, \tau_2 = 0.1)) and the independent missing, with different missing combinations of ((s_1, s_2)).</td>
<td>48</td>
</tr>
<tr>
<td>4.2</td>
<td>Empirical type (I) error for three variance estimators. LZ: uncorrected sandwich variance; KC: KC-corrected sandwich variance; MD: MD-corrected sandwich variance.</td>
<td>48</td>
</tr>
<tr>
<td>4.3</td>
<td>Empirical power under misspecified missing pattern and true missing pattern.</td>
<td>49</td>
</tr>
</tbody>
</table>
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Sample size and empirical power from simulation for the proposed method,</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>the average cluster size method, the harmonic mean cluster size method, and</td>
<td></td>
</tr>
<tr>
<td></td>
<td>the minimum cluster size method. Note: Empirical power (number of clusters</td>
<td></td>
</tr>
<tr>
<td></td>
<td>per stratum per group) is presented in each cell under a combination of</td>
<td></td>
</tr>
<tr>
<td></td>
<td>design parameters.</td>
<td></td>
</tr>
<tr>
<td>3.1</td>
<td>Sample size (empirical power, empirical type I error) for simulation under</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>the missing pattern (A), type I error=0.05, power=0.8.</td>
<td></td>
</tr>
<tr>
<td>3.2</td>
<td>Sample size (empirical power, empirical type I error) for simulation under</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>the missing pattern (B), type I error=0.05, power=0.8.</td>
<td></td>
</tr>
<tr>
<td>4.1</td>
<td>Number of clusters (empirical type I error, empirical power) for simulation</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>under Independent missing, type I error = 5%, power = 80%.</td>
<td></td>
</tr>
<tr>
<td>4.2</td>
<td>Number of clusters (empirical type I error, empirical power) for simulation</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>under general missing when $m = 10$, type I error = 5%, power = 80%.</td>
<td></td>
</tr>
<tr>
<td>4.3</td>
<td>Number of clusters (empirical type I error, empirical power) for simulation</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>under general missing when $m = 20$, type I error = 5%, power = 80%.</td>
<td></td>
</tr>
</tbody>
</table>
I dedicate this dissertation to my grandmother, Daihua Liao.
CHAPTER 1
INTRODUCTION

1.1. Background

Sample size considerations are one of the important steps for a clinical trial, and are widely implemented in medical, epidemiological, and behavioural studies. Accurate sample size calculation is important to ensure the reliability. Inadequate sample size would lead to an underpowered study to detect clinically significant effects, while overestimated sample size would lead an overpowered study and a waste of research resource. Different study designs need different methods of sample size calculation and one formula cannot be used in all designs. In this dissertation, we investigate sample size calculations for three different study designs. Chapter 2 focuses on stratified randomization design with binary outcomes. Chapter 3 focuses on paired experimental design with continuous outcomes. In Chapter 4, we extend the method in Chapter 3 to paired cluster experimental design.

1.2. Sample size considerations for stratified CRTs

Cluster randomization trials (CRTs) have been increasingly employed in clinical and healthcare research [39, 43]. CRTs randomly assign clusters of individuals, rather than individuals themselves, to different intervention groups. Clusters can be geographical areas, health care districts, communities, clinics, providers, etc., and the cluster size can
vary substantially with the trial setting. Main reasons for conducting CRTs include the intervention by its nature has to be implemented at a cluster level and to avoid intervention contamination. Since individuals from the same cluster are not independent, both sample size estimation and subsequent statistical analysis must take into account the intracluster correlation caused by cluster randomization.

A fundamental design issue of CRTs is that a simple randomized CRT is often not as effective in reducing imbalance of baseline prognostic factors, threatening the internal validity of the trial [55]. This is because CRTs typically randomize a limited number of clusters that are heterogeneous in baseline covariates. To address this issue, stratified randomized designs are frequently implemented in CRTs. Different from simple CRTs, stratified CRTs use stratification to increase the balance in baseline covariates and may improve statistical power. In stratified CRTs, stratification factors adopted often include cluster size, cluster-level socio-economic status, geographic location, and categorized levels of prognostic factors. The stratified cluster randomization design can also be described as a randomized block design or a multi-site cluster randomization design with treatment at the cluster level, where the terminology of block or site is used instead of stratum. Due to the popularity, there has been a growing interest in methodological development on sample size estimation and power analysis for these types of designs [10, 24, 25, 37, 57], while most of the existing work assumes equal cluster size within each stratum and uses multilevel models. For continuous outcomes, Moerbeek and Teerenstra [37] provided a sample size method for multi-site cluster randomization trials with fixed cluster size. Konstantopoulos [23], Schochet [50], and Bloom and Spybrook [4] among others discussed power analysis for multi-site cluster randomization trials. Lewsey [25] compared statistical power between a simple cluster randomization design and a stratified cluster randomization design by a simulation study, assuming equal cluster size within each stratum. For binary outcomes, Moerbeek et al [38], Schochet [51], and Chan [7] investigated sample size estimation and statistical power for CRTs. Donner [10] proposed a sample size method for designs of stratified CRTs, where the cluster size itself may be a
stratifying factor. This approach generalizes the sample size formula developed by Woolson et al [62] to stratified categorical data under simple randomization, and it assumes equal cluster size within each stratum but different across strata. Such assumption of no variability in cluster size might be invalid because clusters are often naturally formed with random sizes, especially in real-world pragmatic trials. In presence of varying cluster size, one commonly used \textit{ad hoc} approach is to ignore the variability in cluster size and replace unequal cluster sizes by the average cluster size within each stratum in a sample size formula developed under equal cluster size [34] (average cluster size approximation). We may also replace unequal cluster sizes by the harmonic mean of cluster sizes within each stratum [44] (harmonic mean cluster size approximation) or by the smallest anticipated cluster size within each stratum (minimum cluster size approximation). As shown in later Chapter 2, when the intracluster correlation coefficient (ICC) is positive, the average (minimum) cluster size approximation would underestimate (overestimate) the actual required number of clusters for each group per stratum and lead to underpowered (overpowered) clinical trials, while the harmonic mean cluster size approximation would depend on the distribution of cluster sizes through its harmonic mean.

We propose a closed-form sample size formula for stratified CRTs with binary outcomes to deal with varying cluster size. There is some literature on sample size methods for CRTs with varying cluster size and Rutterford et al [48] provided a comprehensive review of the existing methods. For simple CRTs, Manatunga et al [34] developed a method to incorporate the variability in cluster size into sample size estimation for continuous outcomes and Kong et al [22] derived sample size estimation for binary outcomes. The coefficient of variation of cluster size is used to adjust for varying cluster size in the sample size formulas. Wang et al [57] extended this idea and proposed sample size estimation for cluster size-stratified CRTs with continuous outcomes, based on the generalized estimating equation approach.
1.3. Sample size considerations for paired experimental design

Paired experimental design, such as pre-post study design, matched case-control design and matched randomized trial design, has been widely used in clinical and health behavioural studies [13, 46, 52]. Typically, each study unit contributes a pair of observations, for example, one at pre-intervention and one at post-intervention in pre-post studies, or one from the subject under intervention and one from the subject under control in matched randomized trials. Paired design also arises from studies that involve natural pairs, such as twins, left and right ears or eyes. Statistical inference for these studies needs to account for the correlation between paired outcomes. The paired $t$-test is the most straightforward approach to determine the differential effect of intervention with paired continuous outcomes. Sample size estimation based on the paired $t$-test has served as a convenient method for designing paired experiments. When conducting these experiments, investigators often encounter incomplete observations in the data collected. Specifically, in pre-post studies, missing values occur since some study units contribute outcomes in either pre- or post-intervention periods. For example, a pre-post study on 300 breast cancer survivors was conducted to evaluate the impact of supportive care on improvement of quality-of-life. Among the 300 study participants, 160 completed both pre- and post-study questionnaires on quality-of-life, 97 completed only pre-study questionnaires and 43 completed only post-study questionnaires. In matched randomized trials, missing data may be due to patient drop-out in treatment or control group. As a result, some study units contribute complete pairs of outcomes from both treated and control subjects, some units contribute outcomes from subjects under treatment only, and some from subjects under control only. In practice, the data actually observed in paired experiments often present a mixed structure: a combination of paired and unpaired outcomes.

Missing data may attenuate the power of tests for detecting the intervention effect in paired experiments. In the presence of missing data, the default choice for most of the
available statistical software packages that analyze paired continuous outcomes is to discard incomplete pairs and conduct the paired \( t \)-test using the remaining complete pairs. However, simply discarding incomplete pairs or replacing missing data with values based on single imputation may lead to incorrect or inconsistent conclusions. Statistical methods for analyzing paired continuous outcomes with missing data have been proposed in literature. Lin and Stivers [30] developed a modified maximum likelihood estimator (MLE) of the mean difference between two comparison groups and proposed a \( t \)-test based on the asymptotic distribution of the modified MLE. Looney and Jones [32] introduced a corrected \( z \)-test for comparing two normal means that utilizes both complete and incomplete pairs. Mehrotra [36] considered a mixed effect model with restricted maximum likelihood approach to analyze paired continuous outcomes with incomplete data. Samawi and Vogel [49] suggested a weighted \( t \)-test and a pooled \( t \)-test to detect the mean difference for partially paired continuous data. Einsporn and Habtzghi [11] proposed a permutation test for such data. Despite the extensive development in statistical inference for paired experimental design with incomplete observations of continuous outcomes, sample size method for such study design is sparsely available. In practice, a common way to handle missing data in sample size determination for these studies is through a crude adjustment. First, assuming no missing data, investigators can estimate the sample size based on the paired \( t \)-test, denoted by \( n_{\text{pair}} \). Then, the sample size under incomplete observations is estimated by \( n_{\text{pair}}/q \), where \( q \) is the expected proportion of study units with complete pairs of outcomes. However, such adjustment method may fail to appropriately incorporate the impact of missing data on the power. In fact, as shown in Chapter 3, the impact of missing data on the sample size estimation depends on the missing pattern and correlation between paired outcomes. Correspondingly, the crude adjustment may lead to either the inefficient use of research resources by overestimating the sample size or an underpowered study by underestimating the sample size.

We propose to use the generalized estimating equation (GEE) method [63] to derive a sample size formula for such paired experiments by treating the incomplete observations
as missing data in a linear model. Note that we consider the paired experimental design where missing data can occur at either pre- or post-intervention. This is different from the typical repeated measurement design, where baseline (pre-intervention) measurements are observed in all subjects. GEE is commonly used to model correlated data from clustered and longitudinal studies due to its robustness against mis-specification of the true correlation structure and ability to accommodate missing data. Sample size estimation based on the GEE method has been studied by many researchers. Liu and Liang [31] developed a general sample size formula for studies with correlated outcomes based on a generalized score test. Jung and Ahu [18, 19] proposed sample size methods for comparing rates of change for repeated continuous and binary measurements between two treatment groups. Zhang and Ahu [64] and Lou et al [33] investigated sample size calculation for time-averaged differences for continuous and binary outcomes from repeated measurement studies in the presence of missing data. Zhu et al [67] developed sample size methods for split-mouth design with continuous and binary outcomes.

1.4. Sample size considerations for paired cluster experimental design

Paired cluster experimental designs, such as pre-post cluster study design and matched-pair cluster-randomized trial are becoming increasingly used in clinical and health behavioural studies [3, 6, 60, 61]. In such designs, clusters of individuals (formed by clinics, providers or schools, for example) are allocated to different treatment groups to avoid the intervention contamination. For a pre-post cluster study, each individual contributes a pair of observations, one at pre-intervention and one at post-intervention. For a matched-pair cluster-randomized trial, the matching would be both at an individual level and a cluster level. One recognized feature in a cluster design is that the outcomes within a cluster are more similar than those across clusters at one intervention. To quantify the similarity within a cluster, the intracluster correlation coefficient is commonly used to measure the correlation between any two individuals in the same cluster [48]. When pair of observa-
tions or matching occurred in different treatments, there are other two types of correlations we need to account for: the within-subject correlation and the inter-treatment correlation [28, 40, 42]. Generalized Estimating Equation (GEE) is one of the commonly used approaches to handle the complicated correlation structures in clustered and longitudinal studies due to its consistent estimation even with mis-specified correlation structure. Sample size calculations based on the GEE approach have been studied in many applications. Liu and Liang [31] derived a general sample size formula for studies involving correlated outcomes based on the score statistic. Zhang et al [65] and Zhu et al [66] developed sample size methods for paired experimental design for binary and continuous outcomes. Li et al [27, 28] investigated sample size determination for stepped wedge cluster randomized trials and cluster randomized crossover trials. Zhu et al [67] proposed sample size calculation for split-mouth design with continuous and binary outcomes. Another consideration in paired cluster experimental designs is missing data. Investigators often encounter incomplete observations of paired outcomes or similar withdrawals happened within a cluster in the data collected. For pairs of observations, some study units contribute complete pairs of outcomes from both treatment groups, while some study units only contribute one outcome from either intervention group or control group. In a cluster design, individuals within a cluster may withdraw and observed cluster sizes would be varied under different treatments [48, 53]. One common approach to handle such missing data in sample size calculation are through a crude adjustment that dividing the sample size by the expected proportion of complete observations. Such adjustment approach could overestimate or underestimate the required sample size. GEE approach can also be flexible to accommodate missing data [18, 26, 58].
In this chapter, we propose a closed-form sample size formula for stratified CRTs with binary outcomes and varying cluster size. We explore the connection between the proposed sample size formula and that in Donner [10] developed under equal cluster size, and investigate the impact of design parameters on the relative change in sample size due to varying cluster size. The remainder of the chapter is arranged as follows. In Section 2.1, we present the notations and the method for sample size estimation based on Cochran-Mantel-Haenszel statistic, adjusting for both clustering and varying cluster size. In Section 2.2, we conduct simulation studies to assess the performance of the proposed method and compare it with the average cluster size approximation, harmonic mean cluster size approximation, and minimum cluster size approximation, under various design configurations. In Section 2.3, we illustrate the proposed method with a real application example of a pragmatic stratified CRT of a triad of CKD, diabetes and hypertension.

2.1. Methods

2.1.1. Sample size estimation for varying cluster size

Suppose that in a CRT, clusters are grouped into $K$ strata and are randomly assigned to either the intervention or control group. For simplicity, we consider a balanced randomization, with equal number of clusters assigned to each group. Let $J_k$ denote the number of clusters for each group for stratum $k$, $k = 1, \ldots, K$. Let $n_{1jk}$ and $n_{2jk}$ denote the clus-
ter size for cluster \( j \) in stratum \( k \) for the intervention and control group, respectively, \( j = 1, \ldots, J_k, k = 1, \ldots, K \). We assume that \( n_{1jk}'s \) and \( n_{2jk}'s \) are independently and identically distributed (i.i.d.) samples with mean \( \mu_k \) and variance \( \tau_k^2 \), \( k = 1, \ldots, K \). The total number of subjects of a stratified CRT is \( N = \sum_{k=1}^{K} n_k = \sum_{k=1}^{K} (n_{1k} + n_{2k}) = \sum_{k=1}^{K} \sum_{j=1}^{J_k} (n_{1jk} + n_{2jk}) \), where \( n_k \) is the number of subjects required for stratum \( k \), \( n_{1k} = \sum_{j=1}^{J_k} n_{1jk} \) and \( n_{2k} = \sum_{j=1}^{J_k} n_{2jk} \) are the number of subjects for stratum \( k \) for the intervention group and control group, respectively.

Let \( Y_{1ijk} \) and \( Y_{2ijk} \) be the binary outcome variable (1 for response, 0 for no response) of subject \( i \) in cluster \( j \) of the stratum \( k \) for the intervention and control group, respectively. Let \( \pi_{1k} \) and \( \pi_{2k} \) be the response probability in stratum \( k \) for the intervention group and the control group, respectively, \( k = 1, \ldots, K \). Let \( d_k = p_{1k} - p_{2k} \) denote the difference in the estimated response probability between intervention and control groups in stratum \( k \), where \( p_{1k} = (\sum_{j=1}^{J_k} \sum_{i=1}^{n_{1jk}} Y_{1ijk})/(\sum_{j=1}^{J_k} n_{1jk}) \), and \( p_{2k} = (\sum_{j=1}^{J_k} \sum_{i=1}^{n_{2jk}} Y_{2ijk})/(\sum_{j=1}^{J_k} n_{2jk}) \). Let \( \rho \) be ICC quantifying the similarity of subjects within the same cluster [21, 68]. The odds ratio between intervention and response in stratum \( k \) is \( \theta_k = [\pi_{1k}(1 - \pi_{2k})]/[\pi_{2k}(1 - \pi_{1k})] \), \( k = 1, \ldots, K \). Let \( \theta \) denote the common odds ratio. Our interest is to estimate the sample size \( J_k, k = 1, \ldots, K \), for testing \( H_0: \theta = 1 \) versus \( H_a: \theta \neq 1 \) with a power of \( 1 - \gamma \) at a two-sided significant level of \( \alpha \).

Donner [10] generalized the sample size formula of Woolson et al [62] for stratified categorical data under simple randomization and proposed the sample size method for stratified cluster randomization with binary outcomes. Donner’s method assumes the cluster size to be equal within, but different across strata. It accounts for clustering in a stratified cluster randomization design, but not the variability in cluster size. We derive the sample size formula for stratified CRTs with varying cluster size. The Cochran-Mantel-Haenszel statistic [8] for testing the hypotheses: \( H_0: \theta = 1 \) versus \( H_a: \theta \neq 1 \) is

\[
C_0 = \frac{\sum_{k=1}^{K} w_k d_k}{\sqrt{\sum_{k=1}^{K} w_k^2 \text{var}(d_k)}},
\]
where \( w_k = n_{1k} n_{2k} / (n_{1k} + n_{2k}) \), and \( \text{var}(d_k) = \text{var}(p_{1k}) + \text{var}(p_{2k}) \). The large sample distribution of \( C_0 \) converges to a standard normal distribution under the null hypothesis \( H_0: \theta = 1 \). More generally, the statistic

\[
C_a = \frac{\sum_{k=1}^K w_k [(p_{1k} - p_{2k}) - (\pi_{1k} - \pi_{2k})]}{\sqrt{\sum_{k=1}^K w_k^2 \text{var}(d_k)}}
\]

converges to a standard normal distribution. The power for testing \( H_0: \theta = 1 \) versus \( H_a: \theta \neq 1 \) is \( Pr(|C_0| > Z_{1-\alpha/2} | \theta \neq 1) \), where \( Z_{1-\alpha/2} \) is the \( 100(1-\alpha/2) \)th percentile of the standard normal distribution.

The estimation of testing power and sample size requires the asymptotic variance of \( d_k = p_{1k} - p_{2k} \). Since \( Y_{ijk} \)'s are i.i.d. random samples with mean \( \pi_{1k} \) and variance \( \pi_{1k} (1 - \pi_{1k}) \), and \( n_{1jk} \)'s are i.i.d. random samples with mean \( \mu_k \) and variance \( \tau_k^2 \), we have

\[
\text{var}(p_{1k}) = \frac{\pi_{1k} (1 - \pi_{1k}) \sum_{j=1}^{J_k} n_{1jk} + 2 \sum_{j=1}^{J_k} (n_{1jk}^2) \pi_{1k} (1 - \pi_{1k}) \rho}{(\sum_{j=1}^{J_k} n_{1jk})^2} = \frac{\pi_{1k} (1 - \pi_{1k}) \sum_{j=1}^{J_k} (n_{1jk} + n_{1jk}^2 \rho - n_{1jk} \rho)}{(\sum_{j=1}^{J_k} n_{1jk})^2}
\]

\[
\lim_{J_k \to \infty} \frac{\pi_{1k} (1 - \pi_{1k}) E(n_{1jk}) + \rho E(n_{1jk}^2) - \rho E(n_{1jk})}{J_k E^2(n_{1jk})} = \frac{\pi_{1k} (1 - \pi_{1k})}{J_k \mu_k} [\rho \mu_k + \rho \gamma_k^2 \mu_k + (1 - \rho)],
\]

where \( \gamma_k = \tau_k / \mu_k \) is the coefficient of variation (CV). Let \( M_k = \rho \mu_k + \rho \gamma_k^2 \mu_k + (1 - \rho) \), then \( p_{1k} \) has the variance of \( M_k \pi_{1k} (1 - \pi_{1k}) / (J_k \mu_k) \), as \( J_k \to \infty \), where we consider \( M_k \) as a correction factor. Similarly, we have \( \text{var}(p_{2k}) = M_k \pi_{2k} (1 - \pi_{2k}) / (J_k \mu_k) \), as \( J_k \to \infty \). Therefore, as \( J_k \to \infty \), under \( H_0 \), the variance of \( d_k \) can be expressed as

\[
\text{var}_0(d_k) = \bar{p}_k (1 - \bar{p}_k) \left( \frac{1}{J_k \mu_k} + \frac{1}{J_k \mu_k} \right) M_k, \text{ where } \bar{p}_k = \frac{p_{1k} + p_{2k}}{2},
\]
and under $H_a$, the asymptotic variance of $d_k$ is

$$\text{var}_a(d_k) = \left[ \frac{p_{1k}(1-p_{1k})}{J_k\mu_k} + \frac{p_{2k}(1-p_{2k})}{J_k\mu_k} \right] M_k.$$  

Let $t_k$ denote the fraction of subjects in the trial belonging to stratum $k$, where $\sum_1^K t_k = 1$. Then, the sample size can be obtained by solving the following two equations:

\begin{equation}
Z_{1-\alpha/2} = \frac{\sum_{k=1}^K \frac{Nt_k}{4} (p_{1k} - p_{2k})}{\sqrt{\sum_{k=1}^K \frac{Nt_k}{4} \bar{p}_k (1 - \bar{p}_k) M_k}}, \tag{2.1}
\end{equation}

\begin{equation}
-Z_{1-\gamma} = \frac{\sum_{k=1}^K \frac{Nt_k}{4} (p_{1k} - p_{2k}) - \sum_{k=1}^K \frac{Nt_k}{4} (\pi_{1k} - \pi_{2k})}{\sqrt{\sum_{k=1}^K \frac{Nt_k}{8} [p_{1k}(1-p_{1k}) + p_{2k}(1-p_{2k})] M_k}}. \tag{2.2}
\end{equation}

Based on equations (2.1) and (2.2), the total number of subjects $N$ required for stratified cluster randomization design is given by

\begin{equation}
N = \frac{(Z_{1-\alpha/2}T + Z_{1-\gamma}U)^2}{V^2} \tag{2.3}
\end{equation}

where

\begin{align*}
T &= \frac{1}{2} \left\{ \sum_{k=1}^K t_k \left[ \rho\mu_k + \rho\gamma_k^2\mu_k + (1 - \rho) \right] \bar{\pi}_k (1 - \bar{\pi}_k) \right\}^{\frac{1}{2}}, \\
U &= \left\{ \frac{1}{8} \sum_{k=1}^K t_k \left[ \rho\mu_k + \rho\gamma_k^2\mu_k + (1 - \rho) \right] [\pi_{1k}(1-\pi_{1k}) + \pi_{2k}(1-\pi_{2k})] \right\}^{\frac{1}{2}}, \\
V &= \frac{1}{4} \sum_{k=1}^K t_k (\pi_{1k} - \pi_{2k}), \quad \bar{\pi}_k = (\pi_{1k} + \pi_{2k})/2.
\end{align*}

The estimated total number of subjects $N$ depends on the mean and variance of the distribution of cluster size, in addition to the expected group difference, ICC, and type $I$ and type $II$ error rates. We further consider a relatively simple case of equal number of clusters allocated to each stratum, where $t_k = \mu_k/\sum_{k=1}^K \mu_k$ and $J_1 = J_2 = \ldots = J_K = J$. Then, the number of clusters to be assigned to each group for each stratum is $J = \ldots = J_K = J$.
\[ N/(2 \sum_{k=1}^{K} \mu_k). \]

When there is no variability in cluster size, \( \gamma_k = 0 \) for \( k = 1, \ldots, K \), formula (2.3) reduces to

\[ N^* = \frac{(Z_{1-\alpha/2}T^* + Z_{1-\gamma}U^*)^2}{V^*}, \tag{2.4} \]

where

\[ T^* = \frac{1}{2} \left[ \sum_{k=1}^{K} t_k (\rho \mu_k + 1 - \rho) \pi_k (1 - \pi_k) \right]^{1/2}, \]

\[ U^* = \left\{ \frac{1}{8} \sum_{k=1}^{K} t_k (\rho \mu_k + 1 - \rho) [\pi_{1k}(1 - \pi_{1k}) + \pi_{2k}(1 - \pi_{2k})] \right\}^{1/2}, \]

\[ V^* = \frac{1}{4} \sum_{k=1}^{K} t_k (\pi_{1k} - \pi_{2k}), \]

which is consistent with the formula derived by Donner [10]. Then, the number of clusters to be assigned to each group in stratum \( k \) is \( J^*_k = N^* t_k/(2 \mu_k) \). Assuming equal number of clusters to be assigned to each stratum, we have \( J^* = N^*/(2 \sum_{k=1}^{K} \mu_k) \), and \( J^* \) is also the required number of clusters per group per stratum by the average cluster size approximation. Formula (2.4) can be modified to calculate the required number of clusters per group per stratum by the harmonic mean cluster size approximation and by minimum cluster size approximation, replacing \( \mu_k \) by the harmonic mean of cluster sizes and the smallest anticipated cluster size, respectively. The formulas for \( J^* \) and \( J \) have similar forms, except for the additional term involving \( \rho \gamma_k \mu_k^2 \). Let \( R \) be the relative change in the number of clusters per group per stratum due to varying cluster size, \( R = J/J^* - 1 \), assuming equal numbers of clusters per group across strata. When ICC is positive, \( R \) is positive, indicating that the variability in cluster size leads to an increased number of clusters per group per stratum, and the average cluster size approximation would underestimate the number of clusters per group per stratum, whereas when ICC is negative, the average cluster size approximation would overestimate the number of clusters per group.
per stratum. Although cases of negative ICC are uncommon, Hanley et al [15] discussed such cases which involved the birth weights of human twins and lung sizes of animal litter-mates. In these cases, with limited space or nutrition, nature allows considerable inequality among individual "competitors" therefore there exists a negative within-twin or within-litter correlation. To illustrate the impact of ICC and CV on the relative change in the number of clusters per group per stratum, we consider a simple scenario where both numbers of clusters and CVs are equal across strata, \( \gamma_1 = \ldots = \gamma_k = \gamma \), and plot the relative change in the number of clusters per group per stratum \( R \) versus ICC, \( \rho \), in Figure 2.1 under different combinations of \( (\gamma, \mu) \). We set the common CVs as \( \gamma = 0.5, 0.75 \) and 1, and choose two sets of \( \mu: \mu_1 = (10, 30, 50) \) and \( \mu_2 = (20, 30, 40) \). It shows that the relative change in the number of clusters per group per stratum \( R \) increases from 0 to \( \gamma^2 \), as \( \rho \) increase from 0 to 1, and \( R \) increases as CV increases. Also, \( R \) is greater for \( \mu_1 \), comparing with \( \mu_2 \). This is because that \( \mu_1 \) has a higher between stratum variability in cluster size than \( \mu_2 \), although their overall mean cluster sizes are the same.

2.1.2. Estimation of clustering parameter

The sample size for a stratified CRT depends on the value of ICC, which is not known in many cases. We can estimate ICC, \( \rho \), by the ANOVA method in the work of Donald and Donner [9]. Ridout et al [45] evaluated the performance of various estimators of \( \rho \) for clustered binary data through simulation, and they showed that the ANOVA estimator performed well. We estimate \( \rho \) using observed data from a stratified CRT as follows. Assuming that \( \rho \) is common to all clusters. We estimate \( \rho_{1k} \) and \( \rho_{2k} \) in the intervention and control groups for stratum \( k \) \( (k = 1, \ldots, K) \), respectively, and use the mean of these \( 2k \) estimates as the overall estimate of \( \rho \). For the intervention group in stratum \( k \), \( \rho_{1k} \) can be estimated by

\[
\hat{\rho}_{1k} = \frac{MSB - MSW}{MSB + (n_0 - 1)MSW},
\]

where
where the mean square for groups (MSB) and for error (MSW) are

$$MSB = \frac{\sum_{j=1}^{J_k} a^2_{jk}}{n_{1jk}} - \frac{\sum_{j=1}^{J_k} a^2_{jk}}{J_k - 1},$$

$$MSW = \frac{\left( a_k - \frac{\sum_{j=1}^{J_k} a^2_{jk}}{n_{1jk}} \right) - \left( \sum_{j=1}^{J_k} a^2_{jk} \right) - \frac{\sum_{j=1}^{J_k} a^2_{jk}}{n_{1jk}}}{\sum_{j=1}^{J_k}(n_{1jk} - 1)},$$

$$a_{jk} = \sum_{i=1}^{n_{1jk}} Y_{ij}^1, \quad a_k = \sum_{j=1}^{J_k} \sum_{i=1}^{n_{1jk}} Y_{ij}^1,$$

respectively, and \( n_{0k} \) is the adjusted mean cluster size

$$n_{0k} = \bar{n}_k - \frac{\sum_{j=1}^{J_k} (n_{1jk} - \bar{n}_k)^2}{(J_k - 1) \sum_{j=1}^{J_k} n_{1jk}}, \quad \bar{n}_k = \frac{\sum_{j=1}^{J_k} n_{1jk}}{J_k}.$$

Similarly, \( \rho_{2k} \) for the control group in stratum \( k \) can be estimated. We then estimate \( \rho \) by

$$\hat{\rho} = \frac{\sum_{k=1}^{K} \hat{\rho}_{1k} + \sum_{k=1}^{K} \hat{\rho}_{2k}}{2K}.$$

There are several methods for obtaining confidence intervals for the ICC estimates. Bonett [5] presented a formula for constructing confidence interval for the ANOVA estimator of ICC, based on \( F \)-distribution and can be implemented using statistical software PASS. Turner et al [56] presented Bayesian methods to construct confidence intervals. Ionan et al [17] evaluated three different methods for constructing confidence intervals: generalized confidence interval method, modified large sample method and Bayesian method, and discussed how to implement these methods using statistical software programs.
2.2. Simulation

We conduct simulation studies to assess the performance of the proposed sample size approach for stratified CRTs with varying cluster size, under various design configurations. In the simulation, we focus on the relatively simple case of equal number of clusters allocated to each stratum and calculate the sample size as the required number of clusters. We consider a stratified CRT where the clusters are stratified into \( K = 3 \) strata. Cluster sizes in the three strata are generated from discrete uniform distributions: \( DU(1, 10) \), \( DU(11, 30) \), \( DU(31, 100) \), respectively, corresponding to small, medium and large strata. We set the nominal levels of type I error at \( \alpha = 0.05 \) and power at \( 1 - \gamma = 90\% \). We consider three levels of ICC (\( \rho \)): 0.02, 0.05, 0.10, as well as two levels of common odds ratio \( \theta \): 1.25, 1.5. We set three distinct combinations of the response probability for the control group \( (\pi_{21}, \pi_{22}, \pi_{23}) \): \((0.2, 0.25, 0.3), (0.2, 0.3, 0.4) \) and \((0.3, 0.4, 0.5) \), and assume that odds ratios for each stratum are all equal to the common odds ratio.

The simulation procedure is as follows: (1) For each combination of \( (\rho, \theta, \pi_{21}, \pi_{22}, \pi_{23}) \), calculate the required number of clusters \( J \) by the proposed method, assuming the equal numbers of clusters across strata and a balanced randomization. (2) For computed \( J \) and each iteration \( s \) \((s = 1, ..., S)\), generate sets of \( 2J \) cluster sizes for stratum \( k \) from pre-specified discrete uniform distributions. Randomly select \( J \) clusters to the intervention group, denoted as \( (n_{11k}^{(s)} \ldots n_{1Jk}^{(s)}) \), and allocate the other \( J \) clusters to the control group, denoted as \( (n_{21k}^{(s)} \ldots n_{2Jk}^{(s)}) \). (3) For the intervention group in cluster \( j \) of stratum \( k \), generate correlated binary outcomes from beta-binomial distribution, \( \text{beta-binomial}(n_{1jk}^{(s)}, \pi_{1k}(1 - \rho)/\rho, (1 - \pi_{1k})(1 - \rho)/\rho) \), following the works of Donald and Donner [9]. Similarly, generate correlated binary outcomes for the control group in cluster \( j \) of stratum \( k \). (4) For each simulated data set, estimate \( \hat{\rho}^{(s)} \) and calculate the test statistics \( C_0^{(s)} \).

The empirical power is calculated as the proportion of times that \( H_0 \) is rejected under \( H_a \), \( \sum_{s=1}^{S} I\{|C_0^{(s)}| > Z_{1-\alpha/2}\}/S \), where \( I\{.\} \) is an indicator function. We set the total number
of iterations $S = 10,000$. We can use the same algorithm to compute the empirical type I error by setting the true odds ratio $\theta = 1$. For comparison, we carry out simulations where the number of clusters for each group per stratum is calculated by the proposed method, average cluster size approximation, harmonic mean cluster size approximation, and minimum cluster size approximation, respectively. The empirical power and type I error rate can be evaluated similarly.

Table 2.1 presents the estimated numbers of clusters for each group per stratum and empirical powers under different design configurations. With other factors fixed, the required number of clusters per group per stratum decreases as ICC decreases or the odds ratio increases. The empirical power values for the proposed method are all close to their nominal levels, suggesting that the proposed method allows stratified CRTs to be adequately powered in the presence of varying cluster sizes. If the variability in cluster size is ignored and ICC is positive as in the simulation, the required number of clusters per group per stratum calculated by the average cluster size approximation is underestimated, resulting in underpowered clinical trials. The harmonic mean cluster size approximation and proposed method yield empirical power close to the nominal level, although the empirical power by the proposed method is slightly closer to the nominal level than that by the harmonic mean cluster size approximation, while the required number of clusters per group per stratum calculated by the minimum cluster size approximation is overestimated, even up to 1.5 times of that by the proposed method, which would lead to an overpowered study and a waste of research resource. The excess of power increases as ICC decreases. When ICC is extremely small (i.e., $\rho = 0.02$), the estimated $\hat{\rho}$ might be negative which may result in lower empirical power. Note that the required numbers of clusters per group per stratum by the proposed method, average cluster size approximation, and harmonic mean cluster size approximation are rounded to the same values under certain combinations of design parameters, due to the integer constraint. In such cases, the empirical power values of the proposed method, average cluster size approximation, and harmonic mean cluster size approximation are about the same. We have explored
scenarios where the number of clusters per group per stratum by the proposed method can be as small as 12, and the corresponding empirical power remains relatively close to the nominal level of 0.9. This suggests that the proposed method derived based on the large sample theory is widely applicable to stratified CRTs with varying cluster size and maintains the desired power in scenarios where the sample size is relatively small.

2.3. Example

Investigators plan to conduct a randomized pragmatic clinical trial of management of patients with a triad of CKD, diabetes and hypertension, in order to determine whether a new intervention (a clinical support model enhanced by technology support) can reduce one-year unplanned all-cause hospitalizations comparing with the standard medical care. A prospective stratified cluster randomization design will be employed. Patients are clustered by clinics, which are stratified by healthcare systems. The clinics are randomly allocated with equal probability to either the intervention or the standard medical care group within each stratum. We calculate the sample size based on the comparison of one-year unplanned all-cause hospitalization rates between the intervention group and the control group. From preliminary data, we observe that the rate of unplanned all-cause hospitalization during the 1-year follow-up period is 14% across all four large healthcare systems in the standard medical care group. We expect that the hospitalization rate in the intervention group will be 3% lower than that in the standard medical care group, corresponding to a common odds ratio $\theta = 0.76$ (intervention versus standard care). Electronic health records (EHR) show that the number of patients with coexistent CKD, hypertension and diabetes are 4419, 4738, 4175 and 1093 in the four large healthcare systems. The fractions of patients belonging to each healthcare system are $t_1 = 0.306$, $t_2 = 0.329$, $t_3 = 0.289$, and $t_4 = 0.076$ with the total number of patients equal to 14425. The numbers of clinics are 25, 40, 50, and 9 in the four large healthcare systems. The average number of patients for each clinic $\mu_k$ are 177, 119, 84 and 122, with standard deviation $\tau_k$ of 75, 53,
and 58, respectively. In the sample size calculation, we assume that the fractions of patients from each healthcare system in the trial remain the same as those in the EHR. Our preliminary data suggests that ICC is estimated as $\hat{\rho} = 0.008$ (95% confidence interval: 0.002 – 0.012, calculated by the ANOVA method). If we assume a conservative estimate, $\hat{\rho} = 0.015$, to detect a 3% difference in the rate of unplanned all-cause hospitalization, 11, 17, 22 and 4 clinics for each group from the four large healthcare systems (a total of 12387 patients) are needed to achieve 80% power at a two-sided 5% significance level by the proposed method. Here, 12387 patients are 85.9% of 14425 patients from the four healthcare systems participating in the study. By the average cluster size approximation, we would need to enroll 10, 16, 19, and 4 clinics for each group from the four large healthcare systems, corresponding to a total of 10991 patients. Note that the required numbers of clinics for each group from the fourth healthcare system by the proposed method and average cluster size approximation are the same and equal to 4, although the required numbers of patients for each group from the fourth healthcare system by the two methods differ (469 and 416, respectively), due to the issue of rounding up numbers of clinics in the calculation. Both the number of clinics for each group from the four large healthcare systems and the total number of patients required by the average cluster size approximation are smaller than that by the proposed method, which is likely to lead to an underpowered study without accounting for the variability in the size of clinic. Further, if stratification is ignored, the number of clinics for each group needed is 45, corresponding to a total of 11428 patients, by the sample size method proposed in Kong et al [22] that would only account for varying cluster size. This total number of patients is also smaller than that by the proposed method accounting for both stratification and varying cluster size.

### 2.4. Acknowledgments

This work was supported in part by the Cancer Center Support Grant from the National Cancer Institute (2P30CA142543), the National Center for Advancing Translational
Figure 2.1: The plot of relative change in sample size ($R$) versus ICC ($\rho$). Assuming the numbers of clusters and CVs to be equal across strata. The common CVs are set as $\gamma = 0.5$, 0.75, and 1. The means of cluster sizes are set as $\mu$: $\mu_1 = (10, 30, 50)$ and $\mu_2 = (20, 30, 40)$. 
Table 2.1: Sample size and empirical power from simulation for the proposed method, the average cluster size method, the harmonic mean cluster size method, and the minimum cluster size method. Note: Empirical power (number of clusters per stratum per group) is presented in each cell under a combination of design parameters.

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>$\theta$</th>
<th>Method</th>
<th>$(\pi_{21}, \pi_{22}, \pi_{23})$</th>
<th>$(0.2, 0.25, 0.3)$</th>
<th>$(0.2, 0.3, 0.4)$</th>
<th>$(0.3, 0.4, 0.5)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>1.25</td>
<td>Proposed</td>
<td>90.49% (147)</td>
<td>90.21% (133)</td>
<td>89.72% (124)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Average</td>
<td>88.29% (136)</td>
<td>87.44% (123)</td>
<td>88.01% (115)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Harmonic</td>
<td>88.28% (139)</td>
<td>88.47% (126)</td>
<td>88.38% (118)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Minimum</td>
<td>92.13% (160)</td>
<td>92.08% (144)</td>
<td>92.23% (136)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>Proposed</td>
<td>90.03% (43)</td>
<td>90.58% (40)</td>
<td>90.98% (38)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Average</td>
<td>87.77% (40)</td>
<td>88.03% (37)</td>
<td>88.12% (35)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Harmonic</td>
<td>88.45% (41)</td>
<td>87.85% (37)</td>
<td>88.50% (36)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Minimum</td>
<td>92.57% (47)</td>
<td>92.16% (43)</td>
<td>92.40% (41)</td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td>1.25</td>
<td>Proposed</td>
<td>89.87% (84)</td>
<td>89.58% (76)</td>
<td>89.50% (71)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Average</td>
<td>87.63% (79)</td>
<td>87.64% (71)</td>
<td>87.34% (67)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Harmonic</td>
<td>89.33% (82)</td>
<td>88.99% (74)</td>
<td>88.71% (70)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Minimum</td>
<td>94.97% (103)</td>
<td>94.51% (93)</td>
<td>95.28% (88)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>Proposed</td>
<td>89.98% (25)</td>
<td>89.96% (23)</td>
<td>90.04% (22)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Average</td>
<td>87.91% (23)</td>
<td>87.63% (21)</td>
<td>87.36% (20)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Harmonic</td>
<td>88.42% (24)</td>
<td>88.40% (22)</td>
<td>89.42% (21)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Minimum</td>
<td>94.40% (30)</td>
<td>94.51% (28)</td>
<td>95.03% (27)</td>
<td></td>
</tr>
<tr>
<td>0.02</td>
<td>1.25</td>
<td>Proposed</td>
<td>89.12% (47)</td>
<td>89.30% (42)</td>
<td>89.55% (40)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Average</td>
<td>88.26% (45)</td>
<td>87.89% (40)</td>
<td>87.57% (38)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Harmonic</td>
<td>89.57% (47)</td>
<td>89.39% (43)</td>
<td>89.39% (40)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Minimum</td>
<td>97.59% (69)</td>
<td>97.52% (62)</td>
<td>97.46% (59)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>Proposed</td>
<td>88.68% (14)</td>
<td>89.46% (13)</td>
<td>87.99% (12)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Average</td>
<td>86.62% (13)</td>
<td>85.76% (12)</td>
<td>87.50% (12)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Harmonic</td>
<td>88.11% (14)</td>
<td>88.55% (13)</td>
<td>87.37% (12)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Minimum</td>
<td>96.77% (20)</td>
<td>97.27% (19)</td>
<td>96.60% (18)</td>
<td></td>
</tr>
</tbody>
</table>
In this chapter, we present a closed-form sample size formula, which properly accounts for the mixed structure of observed data. Under complete data, we explore the connection between the sample size based on the GEE method and that based on the paired $t$-test. In the presence of missing data, we compare the sample size adjustment by the proposed GEE method with the crude adjustment method, and demonstrate how the correlation between paired outcomes may influence the corresponding sample size ratio.

The remainder of this chapter is arranged as follows. In Section 3.1, we present the statistical method and sample size estimation based on a marginal linear regression model and the GEE approach. In Section 3.2, we conduct simulation studies to investigate the performance of the GEE sample size formula. In Section 3.3, we illustrate the proposed sample size method with a real application example.

### 3.1. Statistical method and sample size estimation

#### 3.1.1. Sample size based on the GEE approach

In a paired experiment, let $y_{ij}$ be the continuous outcome variable for study unit $i$ under treatment $t_j$, for $i = 1, \ldots, n$ and $j = 1, 2$. For example, in a pre-post study, $t_1$ and $t_2$ represent the pre-intervention and post-intervention periods, respectively, and in a matched randomized trial, $t_1$ and $t_2$ represent the intervention and control, respectively. Let $\rho_{12} = \text{corr}(y_{i1}, y_{i2}) = \rho_{21} = \rho$ be the within-subject (study unit) correlation coefficient
and $\rho_{11} = \rho_{22} = 1$. We assume the outcomes to be independent across different study units, $\text{corr}(y_{ij}, y_{i'j'}) = 0$ for $i \neq i'$. To make an inference on the intervention effect on $y_{ij}$, we assume a linear regression model

$$ y_{ij} = \beta_1 + \beta_2 t_j + \epsilon_{ij}, \quad (3.1) $$

where $\beta_1$ is the intercept, $\beta_2$ is the intervention effect, and $\epsilon_{ij}$ is the zero-mean random error with $\text{var}(\epsilon_{ij}) = \sigma^2$. Our interest is to test the null hypothesis $H_0 : \beta_2 = 0$ versus the alternative hypothesis $H_a : \beta_2 \neq 0$.

We first consider the case with complete observations of paired outcomes and derive the sample size formula based on the GEE approach. Let $X_j = (1, t_j)'$ and $\beta = (\beta_1, \beta_2)'$. Under the independent working correlation structure, the GEE estimator of $\beta$, $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)'$, is obtained by solving the equation

$$ S_n(\beta) = n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{2} (y_{ij} - \beta' X_j) X_j = 0. $$

We have

$$ \hat{\beta} = \left( \sum_{i=1}^{n} \sum_{j=1}^{2} X_j X_j' \right)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{2} X_j y_{ij}. $$

By Zeger and Liang [63], $\sqrt{n}(\hat{\beta} - \beta)$ is approximately normal with mean zero and the variance is consistently estimated by $\Sigma_n = A_n^{-1} V_n A_n^{-1}$, where

$$ A_n = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{2} X_j X_j', $$

$$ V_n = n^{-1} \sum_{i=1}^{n} \left( \sum_{j=1}^{2} X_j \hat{\epsilon}_{ij} \right) \left( \sum_{j=1}^{2} X_j \hat{\epsilon}_{ij} \right)', $$

and $\hat{\epsilon}_{ij} = y_{ij} - \hat{\beta}' X_j$. We reject $H_0 : \beta_2 = 0$ if $|n^{1/2} \hat{\beta}_2 / \hat{\sigma}_2| > z_{1-\alpha/2}$, where $\hat{\sigma}_2^2$ is the $(2,2)$th element in $\Sigma_n$ and $z_{1-\alpha/2}$ is the $100(1 - \alpha/2)$th percentile of the standard normal
distribution. Let $A = \lim_{n \to \infty} A_n$ and $V = \lim_{n \to \infty} V_n$, then $\lim_{n \to \infty} \Sigma_n = A^{-1} V A^{-1} = \Sigma$. Let $\sigma_2^2$ be the (2,2)th element in $\Sigma$. Given the type I error $\alpha$, the power $1 - \gamma$, and the true value of the intervention effect $\beta_{20}$, the required sample size for paired experimental design with complete observations can be derived by solving the equation $-\sqrt{n} \beta_{20} \sigma_2 + z_{1-\alpha/2} = -z_{1-\gamma}$. Thus, the sample size is

$$n = \frac{\sigma_2^2(z_{1-\alpha/2} + z_{1-\gamma})^2}{\beta_{20}^2}.$$ (3.2)

Next, we consider the case that some study units may have incomplete observations of paired outcomes. The GEE sample size approach can be readily extended to accommodate missing data. In this section, we assume a missing completely at random (MCAR) mechanism [47]. That is, the occurrence of missing values is independent of the observed or unobserved outcomes. Let $\delta_{ij}$ be an indicator variable, where $\delta_{ij} = 1$ if the outcome of study unit $i$ is observed under treatment $t_j$ and 0 otherwise, for $j = 1, 2$. Under MCAR, $(\delta_{i1}, \delta_{i2})$ is independent of $(y_{i1}, y_{i2})$. In the presence of missing data, $A_n$ and $V_n$ become

$$A_n = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{2} \delta_{ij} \begin{pmatrix} 1 & t_j \\ t_j & t_j^2 \end{pmatrix}$$

and

$$V_n = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{2} \sum_{j'=1}^{2} \delta_{ij} \delta_{ij'} \hat{e}_{ij} \hat{e}_{ij'} \begin{pmatrix} 1 & t_{j'} \\ t_{j'} & t_{j'} t_{j'} \end{pmatrix},$$

respectively. Let $p_j = E(\delta_{ij})$ be the proportion of study units with outcomes under treatment $t_j$ for $j = 1, 2$, and $p_{jj'} = E(\delta_{ij} \delta_{ij'})$ be the proportion of study units with complete observations of paired outcomes $(p_{jj} = p_j)$. The proportion of contributing outcomes only under $t_j$ is $p_j - p_{12}$. For convenience of discussion, let $t_1 = 0$ and $t_2 = 1$. Then
\( A = \lim_{n \to \infty} A_n \) and \( V = \lim_{n \to \infty} V_n \) can be expressed as

\[
A = \begin{pmatrix}
\sum_{j=1}^{2} p_j & \sum_{j=1}^{2} p_j t_j \\
\sum_{j=1}^{2} p_j t_j & \sum_{j=1}^{2} p_j t_j^2
\end{pmatrix}
= \begin{pmatrix}
p_1 + p_2 & p_2 \\
p_2 & p_2
\end{pmatrix}
\tag{3.3}
\]

and

\[
V = \sigma^2 \begin{pmatrix}
\sum_{j=1}^{2} \sum_{j' = 1}^{2} p_{jj'} \rho_{jj'} & \sum_{j=1}^{2} \sum_{j' = 1}^{2} p_{jj'} \rho_{jj'} t_{jj'} \\
\sum_{j=1}^{2} \sum_{j' = 1}^{2} p_{jj'} \rho_{jj'} t_j & \sum_{j=1}^{2} \sum_{j' = 1}^{2} p_{jj'} \rho_{jj'} t_j t_{jj'}
\end{pmatrix}
= \sigma^2 \begin{pmatrix}
p_1 + p_2 + 2p_{12}\rho & p_{12}\rho + p_2 \\
p_{12}\rho + p_2 & p_2
\end{pmatrix},
\tag{3.4}
\]

respectively. From (3.3) and (3.4), the (2,2)th element of \( \Sigma = A^{-1} V A^{-1} \) is

\[
\sigma_2^2 = \frac{\sigma^2(p_1 + p_2 - 2p_{12}\rho)}{p_1 p_2}.
\]

Then from (3.2), the sample size formula based on the GEE approach that accounts for potential incomplete observations of paired outcomes is

\[
n_{GEE} = \frac{\sigma^2(p_1 + p_2 - 2p_{12}\rho)(z_{1-\alpha/2} + z_{1-\gamma})^2}{\beta_{20}^2 p_1 p_2}.
\tag{3.5}
\]

This closed-form sample size formula suggests that the required sample size depends on the missing data pattern and magnitude of missingness through parameters \((p_1, p_2, p_{12})\), and within-subject correlation coefficient \(\rho\), in addition to the variance of error terms \(\sigma^2\) and the true intervention effect \(\beta_{20}\). A stronger within-subject correlation would lead to a smaller sample size, with all the other factors fixed. That is, as \(\rho\) increases, the sample
size decreases. In this section, we specify the following two missing patterns: (A) Missing under treatment $t_1$ is independent of missing under treatment $t_2$. This implies that the proportion of study units with complete pairs of outcomes $p_{12} = p_1p_2$; (B) Each study unit contributes at least one outcome. That is, $P(\delta_{i1} = \delta_{i2} = 0) = 0$. Based on this, $p_{12} = p_1 + p_2 - 1$ and we assume that $p_1 + p_2 > 1$.

3.1.2. Sample size based on the paired $t$-test

When all study units contribute complete observations in paired experiments, the paired $t$-test is the most popular method to test the mean difference in paired continuous outcomes and the sample size can be determined accordingly. In the paired experiment, we assume that $y_{ij}$'s are independently and identically distributed (i.i.d.) random samples from a target population with mean $\mu_j$, for $j = 1, 2$. As in Section 3.1.1, we assume the equal variance of $\sigma^2$ for $y_{i1}$'s and $y_{i2}$'s, and $\rho_{12} = \text{corr}(y_{i1}, y_{i2}) = \rho$, for $i = 1, \ldots, n$. Then testing the intervention effect $\beta_2 = 0$ is equivalent to testing $\mu_1 = \mu_2$. Define $d_i = y_{i1} - y_{i2}$ as the difference in the paired outcome $(y_{i1}, y_{i2})$. Let $\bar{d} = \sum_{i=1}^{n} d_i/n$ and $S_d^2 = \sum_{i=1}^{n} (d_i - \bar{d})^2/(n-1)$ denote the sample mean and variance of $d_i$'s. The statistic for the paired $t$-test is

$$T = \frac{\bar{d}}{S_d/\sqrt{n}},$$

which follows a $t$-distribution with $n - 1$ degrees of freedom. Given the type I error rate $\alpha$, the power $1 - \gamma$, and the true values of the population means $\mu_{10}$ and $\mu_{20}$, the required sample size based on the paired $t$-test is

$$n_{\text{pair}} = \frac{\sigma_d^2(t_{n-1,1-\alpha/2} + t_{n-1,1-\gamma})^2}{(\mu_{10} - \mu_{20})^2} \approx \frac{\sigma_d^2(z_{1-\alpha/2} + z_{1-\gamma})^2}{(\mu_{10} - \mu_{20})^2},$$

25
where $\sigma^2_d$ is the variance of $d_i$’s and we have $\sigma^2_d = 2\sigma^2(1 - \rho)$. Thus under complete observations, the sample size based on the paired $t$-test is

$$n_{\text{pair}} = \frac{2\sigma^2(1 - \rho)(z_{1-\alpha/2} + z_{1-\gamma})^2}{(\mu_{10} - \mu_{20})^2}.$$ 

On the other hand, we can also obtain the GEE sample size estimator under complete observations, by setting $p_1 = p_2 = p_{12} = 1$ in the formula (5), and the sample size is

$$n_{\text{complete}} = \frac{2\sigma^2(1 - \rho)(z_{1-\alpha/2} + z_{1-\gamma})^2}{\beta_{20}^2},$$

where $\beta_{20} = \mu_{10} - \mu_{20}$ is the true intervention effect. It shows that under complete observations, the sample size based on the paired $t$-test is the same as the sample size based on the GEE approach ($n_{\text{pair}} = n_{\text{complete}}$).

3.1.3. Crude adjustment for incomplete observations

In paired experiments, a common way to account for incomplete observations is to use a crude adjustment. This method estimates the sample size by $n_{\text{complete}}/q$, where $n_{\text{complete}}$ is the sample size under complete observations and $q$ is the proportion of study units with complete pairs of outcomes (here $q = p_{12}$). The sample size using the crude adjustment for incomplete observations is

$$n_{\text{crude}} = \frac{n_{\text{complete}}}{q} = \frac{2\sigma^2(1 - \rho)(z_{1-\alpha/2} + z_{1-\gamma})^2}{\beta_{20}^2 p_{12}}.$$

In the presence of missing data, the sample size ratio comparing the GEE approach with the crude adjustment is

$$R = \frac{n_{\text{GEE}}}{n_{\text{crude}}} = \frac{p_{12}(p_1 + p_2 - 2p_{12}\rho)}{2p_1p_2(1 - \rho)},$$
which is a function of parameters \((p_1, p_2, p_{12}, \rho)\). By some algebra, it can be shown that
the GEE method would lead to a smaller sample size comparing with the crude adjustment, \(n_{\text{GEE}} < n_{\text{crude}}\), when \(\rho < \rho_0 = (2p_1p_2 - p_1p_{12} - p_2p_{12})/2(p_1p_2 - p_{12}^2)\) and \(p_{12}^2 < p_1p_2\).
Such sample size saving is important in clinical practice on two fronts: (i) efficient use of research resources; (ii) smaller number of patients to be exposed to the potential risk of experimental intervention. We plot the sample size ratio \(R\) against the within-subject correlation coefficient \(\rho\) under missing patterns (A) and (B), and different combinations of \((p_1, p_2)\) in Figure 3.1. The cutoff point \(\rho_0\) for each combination is identified in Figure 3.1. It shows that the sample size ratio increases with the within-subject correlation coefficient, and the GEE method generally leads to a saving in sample size when \(\rho < \rho_0\). For example, under the missing pattern (B) and \(p_1 = p_2 = 0.85\), the crude adjustment would overestimate the sample size when \(\rho < 0.548\). With missing data, the proposed GEE method would lead to a more accurate sample size estimate comparing with the crude adjustment. When \(\rho < \rho_0\), the GEE-based sample size is smaller than the sample size by crude adjustment, therefore, the crude adjustment would lead to the inefficient use of research resources by overestimating the sample size. When \(\rho > \rho_0\), the GEE-based sample size is larger than the one by crude adjustment, therefore, the crude adjustment would result in an underpowered study by underestimating the sample size.

3.2. Simulation

Simulation studies are conducted to evaluate the performance of the proposed GEE sample size method for paired experimental design with incomplete observations of continuous outcomes, under various design configurations. The nominal levels of type I error and power are set at \(\alpha = 0.05\) and \(1 - \gamma = 0.8\), respectively. We consider missing patterns (A) with \(p_{12} = p_1p_2\) and (B) with \(p_{12} = p_1 + p_2 - 1\), and six combinations of \((p_1, p_2)\): \(m_1 = (p_1, p_2) = (1, 1)\) for no missing data, \(m_2 = (p_1, p_2) = (0.85, 0.85)\) for balanced distribution of missing values in outcomes under the intervention \(t_1\) and control.
$t_2$, $m_3 = (p_1, p_2) = (0.9, 0.8)$, $m_4 = (p_1, p_2) = (0.8, 0.9)$, $m_5 = (p_1, p_2) = (0.6, 0.7)$, and $m_6 = (p_1, p_2) = (0.5, 0.6)$ for unbalanced distribution of missing values and missing proportions varying from 10% to 50%. We choose the values of within-subject correlation coefficient $\rho = -0.3, 0, 0.3, 0.6, 0.9$, representing the negative correlation, independence, and positive correlation. We set the true values of regression coefficient $\beta_{20} = 0.1$ or 0.2 and variance $\sigma^2 = 0.5$ or 1, where $(\beta_{20}, \sigma^2) = (0.1, 1)$ indicates an effect size of 0.1 comparing the intervention with control. For each combination of $(p_1, p_2, \rho, \beta_{20}, \sigma^2)$, we calculate the required sample size $n_{\text{GEE}}$ based on the formula (5). Then we generate random samples of $n_{\text{GEE}}$ pairs of continuous outcomes from the model (1) with $\beta_1 = 1$. Simulation results are unchanged with different values of $\beta_1$. Correlated random errors $\epsilon_{ij}$’s are generated from the multivariate normal distribution with mean 0, variance $\sigma^2$ and within-subject correlation coefficient $\rho$. The incomplete data are imposed in outcomes under the intervention and control, according to pre-specified missing patterns and $(p_1, p_2)$. For the generated dataset, we test the null hypothesis $H_0 : \beta_2 = 0$ using the statistic $n^{1/2} \hat{\beta}_2 / \hat{\sigma}_2$. For each combination, we generate 5000 simulated datasets. The empirical type I error rate and empirical power are calculated as the proportion of times that $H_0$ is rejected under the null and alternative hypothesis, respectively.

Table 3.1 and Table 3.2 present the sample size estimate, empirical power and empirical type I error rate for simulation under missing patterns (A) and (B), respectively. Under the design configurations $(p_1, p_2, \rho, \beta_{20}, \sigma^2)$ that we have explored, the estimated sample size has a wide range, from 20 to 3349 for the missing pattern (A), and from 20 to 3035 for the missing pattern (B). The empirical power values and type I error rates are generally close to their nominal levels, which indicates a good performance of the proposed method. Since statistical inference under the GEE approach is based on a large sample approximation, it is important to assess the performance of the proposed method in some small sample size scenarios. We have explored scenarios where the sample size can be as small as 20 or 40, and the corresponding empirical power remains close to the nominal level of 0.8. This provides assurance to investigators that the proposed method is widely
applicable to paired experiments with incomplete observations, even when the sample size is relatively small. With all the other factors fixed, the required sample size increases as the intervention effect $\beta_{20}$ decreases or the variance $\sigma^2$ increases. The sample size increases as the within-subject correlation coefficient $\rho$ decreases. For example, under the missing pattern (A), $m_2 = (p_1, p_2) = (0.85, 0.85)$ and $(\beta_{20}, \sigma^2) = (0.1, 1)$, the required sample size is 1847 with no within-subject correlation ($\rho = 0$) and 905 with a positive correlation of 0.6, leading to a 51% sample size reduction. Therefore, at the design stage of paired experiments, it is imperative to obtain a reliable estimate of the within-subject correlation coefficient from literature review or preliminary data analysis. Comparing Table 3.1 and Table 3.2, the required sample size under the missing pattern (A) is smaller than that under the missing pattern (B) for a positive within-subject correlation, and is bigger for a negative within-subject correlation, although the sample sizes are generally very close under the two missing patterns. Lastly, the simulation results demonstrate that the proposed GEE sample size method leads to a more accurate estimate of sample size by appropriately taking into account the impact of missing data. Conversely, the crude adjustment may underestimate or overestimate the sample size. For example, under $\rho = 0.3$ and $(\beta_{20}, \sigma^2) = (0.1, 1)$, the sample size with complete observations is 1099. Under the missing pattern (B) with $p_{12} = p_1 + p_2 - 1$ and $p_1 = p_2 = 0.85$, we have the proportion of study units with complete observations $p_{12} = 0.7$ and thus the sample size by the crude adjustment for incomplete observations is $1099/0.7 = 1570$, which is bigger than the proposed GEE sample size of 1391. As $\rho$ increases to 0.6, the sample size with complete observations is 628. Then the sample size by the crude adjustment is $628/0.7 = 898$, which is smaller than the GEE sample size of 935. The result is consistent with that shown in Figure 3.1 (under the missing pattern B and $p_1 = p_2 = 0.85$), which suggests that the proposed method would lead to a saving in sample size if $\rho < 0.548$. 

29
3.3. Example

The Avon longitudinal study [12] enrolled women resident in Avon who were in the early stages of pregnancy with an expected date of delivery between April 1991 and December 1992. The study identified 13799 eligible women, among whom 12059(87%) completed at least one of the four questionnaires, and 9028(65%) completed all four. Symptom scores from the Edinburgh postnatal depression scale were measured at 18 and 32 weeks of pregnancy and 8 weeks and 8 months postpartum. The mean (standard deviation) Edinburgh postnatal depression scale scores were 6.62(4.66), 6.72(4.94), 5.84(4.65), and 5.25(4.61) for each period. It was found that mean scores were higher in pregnancy than postnatally.

Suppose we would like to conduct a similar study to determine whether there is a significant difference in Edinburgh postnatal depression scale scores for women between the stages at 32 weeks of pregnancy and 8 weeks of postpartum. Women undergo dramatic changes both physically and psychologically, during these two stages. Since 75% (= 9028/12059) of subjects contributed complete data in the previous study, we specify $p_{12} = 0.75$. We assume the incomplete data to be evenly distributed at both 32 weeks of pregnancy and 8 weeks of postpartum ($p_1 = p_2$). Missing patterns (A) and (B) are considered as in the previous sections. Then, we have $p_1 = p_2 = 0.866$ under the missing pattern (A) and $p_1 = p_2 = 0.875$ under the missing pattern (B). We assume an equal variance of $\sigma^2 = (4.94 + 4.65)/2 = 4.795$ between these two stages. To detect a difference of $\mu_{10} - \mu_{20} = \beta_{20} = 6.72 - 5.84 = 0.88$ in Edinburgh scores with 90% power and a two-sided 5% significance level, the sample sizes required are 907, 720, 533 and 346, under within-subject correlation coefficient $\rho = -0.3, 0, 0.3, \text{and} 0.6$, respectively, and the missing pattern (A). The corresponding sample sizes required are 896, 713, 530, and 347, under the missing pattern (B). The sample sizes calculated by the crude adjustment method are 1081, 832, 582, and 333, respectively. The result suggests that sample size estimates
under both missing patterns are close, and the proposed GEE sample size method could lead to a sample size saving when the correlation is small or moderate.

3.4. Acknowledgments

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported in part by the Cancer Center Support Grant from the National Cancer Institute (2P30CA142543) and the National Center for Advancing Translational Sciences of the National Institutes of Health (UL1TR001105).

Figure 3.1: The plot of sample size ratio against within-subject correlation coefficient under missing patterns (A) and (B), and different combinations of \((p_1, p_2)\).
Table 3.1: Sample size (empirical power, empirical type I error) for simulation under the missing pattern (A), type I error=0.05, power=0.8.

<table>
<thead>
<tr>
<th>$m$</th>
<th>$\rho = -0.3$</th>
<th>$\rho = 0$</th>
<th>$\rho = 0.3$</th>
<th>$\rho = 0.6$</th>
<th>$\rho = 0.9$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>($\beta_20, \sigma^2$) = (0.1, 0.5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m_1$</td>
<td>1021 (0.798, 0.055)</td>
<td>785 (0.800, 0.052)</td>
<td>550 (0.798, 0.049)</td>
<td>314 (0.793, 0.049)</td>
<td>79 (0.807, 0.052)</td>
</tr>
<tr>
<td>$m_2$</td>
<td>1159 (0.801, 0.050)</td>
<td>924 (0.802, 0.046)</td>
<td>688 (0.800, 0.046)</td>
<td>453 (0.805, 0.050)</td>
<td>217 (0.800, 0.050)</td>
</tr>
<tr>
<td>$m_3$</td>
<td>1163 (0.798, 0.046)</td>
<td>927 (0.798, 0.050)</td>
<td>692 (0.803, 0.053)</td>
<td>456 (0.797, 0.048)</td>
<td>221 (0.799, 0.050)</td>
</tr>
<tr>
<td>$m_4$</td>
<td>1163 (0.802, 0.053)</td>
<td>927 (0.799, 0.046)</td>
<td>692 (0.812, 0.054)</td>
<td>456 (0.812, 0.056)</td>
<td>221 (0.799, 0.047)</td>
</tr>
<tr>
<td>$m_5$</td>
<td>1451 (0.794, 0.049)</td>
<td>1215 (0.807, 0.048)</td>
<td>980 (0.802, 0.045)</td>
<td>744 (0.796, 0.051)</td>
<td>509 (0.800, 0.054)</td>
</tr>
<tr>
<td>$m_6$</td>
<td>1675 (0.800, 0.047)</td>
<td>1439 (0.795, 0.052)</td>
<td>1204 (0.802, 0.052)</td>
<td>969 (0.793, 0.048)</td>
<td>733 (0.797, 0.046)</td>
</tr>
<tr>
<td></td>
<td>($\beta_20, \sigma^2$) = (0.1, 1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m_1$</td>
<td>2041 (0.788, 0.048)</td>
<td>1570 (0.802, 0.051)</td>
<td>1099 (0.799, 0.050)</td>
<td>628 (0.807, 0.050)</td>
<td>157 (0.804, 0.049)</td>
</tr>
<tr>
<td>$m_2$</td>
<td>2318 (0.796, 0.052)</td>
<td>1847 (0.785, 0.047)</td>
<td>1376 (0.810, 0.049)</td>
<td>905 (0.796, 0.050)</td>
<td>434 (0.806, 0.052)</td>
</tr>
<tr>
<td>$m_3$</td>
<td>2325 (0.798, 0.051)</td>
<td>1854 (0.806, 0.053)</td>
<td>1383 (0.800, 0.054)</td>
<td>912 (0.806, 0.048)</td>
<td>441 (0.809, 0.052)</td>
</tr>
<tr>
<td>$m_4$</td>
<td>2325 (0.809, 0.055)</td>
<td>1854 (0.800, 0.048)</td>
<td>1383 (0.795, 0.054)</td>
<td>912 (0.800, 0.047)</td>
<td>441 (0.805, 0.052)</td>
</tr>
<tr>
<td>$m_5$</td>
<td>2901 (0.803, 0.050)</td>
<td>2430 (0.800, 0.048)</td>
<td>1959 (0.792, 0.048)</td>
<td>1488 (0.790, 0.051)</td>
<td>1017 (0.799, 0.055)</td>
</tr>
<tr>
<td>$m_6$</td>
<td>3349 (0.791, 0.050)</td>
<td>2878 (0.800, 0.045)</td>
<td>2407 (0.798, 0.049)</td>
<td>1937 (0.806, 0.052)</td>
<td>1466 (0.801, 0.047)</td>
</tr>
<tr>
<td></td>
<td>($\beta_20, \sigma^2$) = (0.2, 0.5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m_1$</td>
<td>256 (0.811, 0.056)</td>
<td>197 (0.803, 0.051)</td>
<td>138 (0.795, 0.053)</td>
<td>79 (0.796, 0.054)</td>
<td>20 (0.805, 0.054)</td>
</tr>
<tr>
<td>$m_2$</td>
<td>290 (0.796, 0.050)</td>
<td>231 (0.801, 0.053)</td>
<td>172 (0.794, 0.048)</td>
<td>114 (0.803, 0.047)</td>
<td>55 (0.809, 0.059)</td>
</tr>
<tr>
<td>$m_3$</td>
<td>291 (0.801, 0.048)</td>
<td>232 (0.801, 0.053)</td>
<td>173 (0.808, 0.051)</td>
<td>114 (0.808, 0.051)</td>
<td>56 (0.808, 0.051)</td>
</tr>
<tr>
<td>$m_4$</td>
<td>291 (0.798, 0.052)</td>
<td>232 (0.794, 0.051)</td>
<td>173 (0.809, 0.042)</td>
<td>114 (0.799, 0.054)</td>
<td>56 (0.808, 0.050)</td>
</tr>
<tr>
<td>$m_5$</td>
<td>363 (0.793, 0.050)</td>
<td>304 (0.803, 0.048)</td>
<td>245 (0.801, 0.050)</td>
<td>186 (0.807, 0.053)</td>
<td>128 (0.806, 0.047)</td>
</tr>
<tr>
<td>$m_6$</td>
<td>419 (0.793, 0.053)</td>
<td>360 (0.807, 0.054)</td>
<td>301 (0.804, 0.051)</td>
<td>243 (0.800, 0.052)</td>
<td>184 (0.806, 0.044)</td>
</tr>
<tr>
<td></td>
<td>($\beta_20, \sigma^2$) = (0.2, 1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m_1$</td>
<td>511 (0.802, 0.050)</td>
<td>393 (0.803, 0.051)</td>
<td>275 (0.803, 0.047)</td>
<td>157 (0.796, 0.045)</td>
<td>40 (0.801, 0.050)</td>
</tr>
<tr>
<td>$m_2$</td>
<td>580 (0.803, 0.049)</td>
<td>462 (0.803, 0.050)</td>
<td>344 (0.798, 0.053)</td>
<td>227 (0.813, 0.049)</td>
<td>109 (0.795, 0.054)</td>
</tr>
<tr>
<td>$m_3$</td>
<td>582 (0.798, 0.050)</td>
<td>446 (0.800, 0.051)</td>
<td>346 (0.806, 0.050)</td>
<td>228 (0.806, 0.050)</td>
<td>111 (0.802, 0.047)</td>
</tr>
<tr>
<td>$m_4$</td>
<td>582 (0.796, 0.047)</td>
<td>446 (0.808, 0.054)</td>
<td>346 (0.812, 0.051)</td>
<td>228 (0.796, 0.047)</td>
<td>111 (0.809, 0.050)</td>
</tr>
<tr>
<td>$m_5$</td>
<td>726 (0.801, 0.050)</td>
<td>608 (0.800, 0.047)</td>
<td>490 (0.802, 0.052)</td>
<td>372 (0.802, 0.048)</td>
<td>255 (0.800, 0.049)</td>
</tr>
<tr>
<td>$m_6$</td>
<td>838 (0.802, 0.053)</td>
<td>720 (0.802, 0.050)</td>
<td>602 (0.799, 0.044)</td>
<td>485 (0.801, 0.050)</td>
<td>367 (0.802, 0.052)</td>
</tr>
</tbody>
</table>
Table 3.2: Sample size (empirical power, empirical type I error) for simulation under the missing pattern (B), type I error=0.05, power=0.8.

<table>
<thead>
<tr>
<th>m</th>
<th>$\rho = -0.3$</th>
<th>$\rho = 0$</th>
<th>$\rho = 0.3$</th>
<th>$\rho = 0.6$</th>
<th>$\rho = 0.9$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$(\beta_{20}, \sigma^2) = (0.1, 0.5)$</td>
<td>$(\beta_{20}, \sigma^2) = (0.1, 1)$</td>
<td>$(\beta_{20}, \sigma^2) = (0.2, 0.5)$</td>
<td>$(\beta_{20}, \sigma^2) = (0.2, 1)$</td>
<td></td>
</tr>
<tr>
<td>$m_1$</td>
<td>1021 (0.804, 0.050)</td>
<td>785 (0.805, 0.045)</td>
<td>550 (0.799, 0.053)</td>
<td>314 (0.802, 0.055)</td>
<td>79 (0.808, 0.053)</td>
</tr>
<tr>
<td>$m_2$</td>
<td>1152 (0.795, 0.047)</td>
<td>924 (0.799, 0.053)</td>
<td>696 (0.799, 0.047)</td>
<td>468 (0.804, 0.047)</td>
<td>230 (0.808, 0.054)</td>
</tr>
<tr>
<td>$m_3$</td>
<td>1156 (0.800, 0.045)</td>
<td>927 (0.797, 0.048)</td>
<td>698 (0.795, 0.051)</td>
<td>469 (0.806, 0.054)</td>
<td>240 (0.788, 0.051)</td>
</tr>
<tr>
<td>$m_4$</td>
<td>1156 (0.808, 0.056)</td>
<td>927 (0.799, 0.049)</td>
<td>698 (0.793, 0.051)</td>
<td>469 (0.791, 0.051)</td>
<td>240 (0.804, 0.051)</td>
</tr>
<tr>
<td>$m_5$</td>
<td>1383 (0.799, 0.051)</td>
<td>1215 (0.796, 0.048)</td>
<td>1047 (0.797, 0.045)</td>
<td>879 (0.798, 0.049)</td>
<td>711 (0.800, 0.058)</td>
</tr>
<tr>
<td>$m_6$</td>
<td>1518 (0.803, 0.047)</td>
<td>1439 (0.803, 0.052)</td>
<td>1361 (0.799, 0.047)</td>
<td>1282 (0.799, 0.049)</td>
<td>1204 (0.805, 0.051)</td>
</tr>
<tr>
<td></td>
<td>$(\beta_{20}, \sigma^2) = (0.1, 1)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m_1$</td>
<td>2041 (0.801, 0.052)</td>
<td>1570 (0.797, 0.052)</td>
<td>1099 (0.801, 0.051)</td>
<td>628 (0.802, 0.041)</td>
<td>157 (0.801, 0.051)</td>
</tr>
<tr>
<td>$m_2$</td>
<td>2304 (0.794, 0.051)</td>
<td>1847 (0.798, 0.048)</td>
<td>1391 (0.800, 0.051)</td>
<td>935 (0.797, 0.051)</td>
<td>478 (0.798, 0.053)</td>
</tr>
<tr>
<td>$m_3$</td>
<td>2312 (0.805, 0.050)</td>
<td>1854 (0.804, 0.057)</td>
<td>1396 (0.797, 0.053)</td>
<td>938 (0.812, 0.044)</td>
<td>480 (0.795, 0.054)</td>
</tr>
<tr>
<td>$m_4$</td>
<td>2312 (0.795, 0.050)</td>
<td>1854 (0.800, 0.054)</td>
<td>1396 (0.805, 0.049)</td>
<td>938 (0.794, 0.051)</td>
<td>480 (0.797, 0.047)</td>
</tr>
<tr>
<td>$m_5$</td>
<td>2766 (0.789, 0.051)</td>
<td>2430 (0.802, 0.054)</td>
<td>2094 (0.799, 0.055)</td>
<td>1757 (0.798, 0.048)</td>
<td>1421 (0.805, 0.054)</td>
</tr>
<tr>
<td>$m_6$</td>
<td>3035 (0.804, 0.054)</td>
<td>2878 (0.815, 0.051)</td>
<td>2721 (0.800, 0.054)</td>
<td>2564 (0.798, 0.047)</td>
<td>2407 (0.809, 0.052)</td>
</tr>
<tr>
<td></td>
<td>$(\beta_{20}, \sigma^2) = (0.2, 0.5)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m_1$</td>
<td>256 (0.797, 0.046)</td>
<td>197 (0.812, 0.050)</td>
<td>138 (0.798, 0.046)</td>
<td>79 (0.804, 0.055)</td>
<td>20 (0.798, 0.053)</td>
</tr>
<tr>
<td>$m_2$</td>
<td>288 (0.803, 0.045)</td>
<td>231 (0.800, 0.054)</td>
<td>174 (0.798, 0.049)</td>
<td>117 (0.809, 0.057)</td>
<td>60 (0.794, 0.045)</td>
</tr>
<tr>
<td>$m_3$</td>
<td>289 (0.798, 0.050)</td>
<td>232 (0.788, 0.053)</td>
<td>175 (0.796, 0.047)</td>
<td>118 (0.797, 0.046)</td>
<td>60 (0.810, 0.053)</td>
</tr>
<tr>
<td>$m_4$</td>
<td>289 (0.803, 0.051)</td>
<td>232 (0.802, 0.049)</td>
<td>175 (0.806, 0.053)</td>
<td>118 (0.802, 0.049)</td>
<td>60 (0.798, 0.057)</td>
</tr>
<tr>
<td>$m_5$</td>
<td>346 (0.798, 0.049)</td>
<td>304 (0.806, 0.054)</td>
<td>262 (0.798, 0.049)</td>
<td>220 (0.801, 0.045)</td>
<td>178 (0.803, 0.049)</td>
</tr>
<tr>
<td>$m_6$</td>
<td>380 (0.798, 0.060)</td>
<td>360 (0.811, 0.050)</td>
<td>341 (0.801, 0.049)</td>
<td>321 (0.799, 0.044)</td>
<td>301 (0.796, 0.053)</td>
</tr>
<tr>
<td></td>
<td>$(\beta_{20}, \sigma^2) = (0.2, 1)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m_1$</td>
<td>511 (0.812, 0.048)</td>
<td>393 (0.806, 0.052)</td>
<td>275 (0.805, 0.049)</td>
<td>157 (0.788, 0.053)</td>
<td>40 (0.809, 0.049)</td>
</tr>
<tr>
<td>$m_2$</td>
<td>576 (0.803, 0.053)</td>
<td>462 (0.799, 0.046)</td>
<td>348 (0.803, 0.048)</td>
<td>234 (0.804, 0.050)</td>
<td>120 (0.805, 0.051)</td>
</tr>
<tr>
<td>$m_3$</td>
<td>578 (0.800, 0.053)</td>
<td>464 (0.799, 0.050)</td>
<td>349 (0.793, 0.052)</td>
<td>235 (0.807, 0.054)</td>
<td>120 (0.802, 0.048)</td>
</tr>
<tr>
<td>$m_4$</td>
<td>578 (0.801, 0.053)</td>
<td>464 (0.796, 0.051)</td>
<td>349 (0.796, 0.048)</td>
<td>235 (0.793, 0.051)</td>
<td>120 (0.806, 0.052)</td>
</tr>
<tr>
<td>$m_5$</td>
<td>692 (0.801, 0.048)</td>
<td>608 (0.803, 0.048)</td>
<td>524 (0.798, 0.048)</td>
<td>440 (0.794, 0.055)</td>
<td>356 (0.797, 0.048)</td>
</tr>
<tr>
<td>$m_6$</td>
<td>759 (0.801, 0.044)</td>
<td>720 (0.802, 0.046)</td>
<td>681 (0.798, 0.049)</td>
<td>641 (0.805, 0.046)</td>
<td>602 (0.801, 0.052)</td>
</tr>
</tbody>
</table>
CHAPTER 4
SAMPLE SIZE CONSIDERATIONS FOR PAIRED CLUSTER EXPERIMENTAL DESIGN

In this chapter, we propose to use GEE method to derive closed-form sample size formulas for paired cluster experiments with both continuous and binary outcomes in generalized linear models, which properly accounts for the complicated correlation structure and missing data. The remainder of this chapter is arranged as follows. In Section 4.1, we present the statistical method and sample size estimation with continuous outcomes based on a marginal linear regression model and the GEE approach. In Section 4.2, we conduct simulation studies of continuous outcomes to investigate the performance of the GEE sample size formula. In Section 4.3, we illustrate the proposed sample size method with a real application example about physical fitness in Ecuadorian adolescents. In Section 4.4, we extent this method to binary outcomes and propose the closed-form sample size formula.

4.1. Methods

4.1.1. Sample size estimation based on the GEE approach

In a paired cluster experiment, let $t_1$ and $t_2$ denote the two treatments, respectively. $t_1$ and $t_2$ can represent the pre- and post- intervention periods in a pre-post cluster study, respectively, or the intervention and control in a matched-pair cluster-randomized trial, respectively. Let $Y_{ijk}$ be the continuous outcome of study unit $k$ in cluster $j$ under treatment $i$, for $i = 1, 2$, $j = 1, ..., n$ and $k = 1, ..., m$, where $m$ is the cluster size. We further assume
a general correlation structure including three correlations: (1) $\rho_1 = \text{corr}(Y_{ijk}, Y_{ijk'})$ for $k \neq k'$, the within-treatment correlation that describing the similarity between responses from different study units within the same cluster under the same treatment, also named intracluster correlation; (2) $\rho_2 = \text{corr}(Y_{ijk}, Y_{i'jk})$ for $i \neq i'$, the within-subject (study unit) correlation that describing the similarity between responses from the same study unit under different treatments; (3) $\rho_3 = \text{corr}(Y_{ijk}, Y_{i'jk'})$ for $i \neq i'$ and $k \neq k'$, the inter-treatment correlation that describing the similarity between responses from different study units within the same cluster but under different treatments [28, 42]. If the matching only happen at the cluster level, the within-subject correlation $\rho_2$ would reduce to the inter-treatment correlation $\rho_3$.

To make an inference on the intervention effect, we assume a linear regression model,

$$
Y_{ijk} = \beta_1 + \beta_2 t_{ijk} + \epsilon_{ijk},
$$

(4.1)

where $\beta_1$ is the intercept, $\beta_2$ is the intervention effect, and $\epsilon_{ijk}$ is the random error with $E(\epsilon_{ijk}) = 0$ and $\text{Var}(\epsilon_{ijk}) = \sigma^2$. Our primary interest is to test the null hypothesis $H_0 : \beta_2 = 0$ versus the alternative hypothesis $H_a : \beta_2 \neq 0$ with a power of $1 - \gamma$ at a two-sided significant level of $\alpha$.

We first discuss the case with complete observations for both paired outcomes. Let $Y_j = (Y_{1j1}, \ldots, Y_{1jm}, Y_{2j1}, \ldots, Y_{2jm})^T$, an $2m \times 1$ vector of outcomes. Let $\epsilon_j = (\epsilon_{1j1}, \ldots, \epsilon_{1jm}, \epsilon_{2j1}, \ldots, \epsilon_{2jm})^T$, an $2m \times 1$ vector of random errors. For convenience of discussion, let $t_{1jk} = 0$ and $t_{2jk} = 1$, and $X_j$ can be an $2m \times 2$ design matrix

$$
X_j = \begin{pmatrix}
1_m & 0_m \\
1_m & 1_m
\end{pmatrix},
$$
where \( \mathbf{1}_m \) is an \( m \times 1 \) vector of 1's and \( \mathbf{0}_m \) is an \( m \times 1 \) vector of 0's. Let \( \beta = (\beta_1, \beta_2)^T \), and then the model (4.1) can be re-written as \( Y_j = X_j \beta + \epsilon_j \).

With the model assumption, the true correlation matrix is

\[
R_c = \begin{pmatrix}
(1 - \rho_1)I_m + \rho_1 J_m & (\rho_2 - \rho_3)I_m + \rho_3 J_m \\
(\rho_2 - \rho_3)I_m + \rho_3 J_m & (1 - \rho_1)I_m + \rho_1 J_m
\end{pmatrix},
\]

where \( I \) is an identity matrix and \( J \) is a square matrix of 1's. In Appendix A.1, we show that the correlation matrix \( R_c \) has four distinct eigenvalues,

\[
\lambda_1 = 1 - \rho_1 + \rho_2 - \rho_3, \quad \lambda_2 = 1 + (m - 1)\rho_1 + \rho_2 + (m - 1)\rho_3, \\
\lambda_3 = 1 - \rho_1 - \rho_2 + \rho_3, \quad \lambda_4 = 1 + (m - 1)\rho_1 - \rho_2 - (m - 1)\rho_3.
\]

The setting of \((\rho_1, \rho_2, \rho_3)\) should satisfy \( \min(\lambda_1, \lambda_2, \lambda_3, \lambda_4) > 0 \) so that the correlation matrix \( R_c \) is positive definite.

Under the independent working correlation structure, the GEE estimator \( \hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)^T \) is obtained by solving the equation

\[
S_n(\beta) = n^{-1/2} \sum_j (Y_j - X_j \beta)^T X_j = 0.
\]

We have

\[
\hat{\beta} = \sum_j Y_j^T X_j (\sum_j X_j^T X_j)^{-1},
\]

and can be simplified as \( \hat{\beta}_1 = (nm)^{-1} \sum_j \sum_k Y_{jk} \) and \( \hat{\beta}_2 = (nm)^{-1} \sum_j \sum_k (Y_{2jk} - Y_{1jk}) \).
By Zeger and Liang [63], \( n^{-1/2} (\hat{\beta} - \beta) \) is approximately normal with mean 0 and the variance is consistently estimated by \( \Sigma_n = A_n^{-1} V_n A_n^{-1} \), where

\[
A_n = n^{-1} \sum_j X_j^T X_j,
\]

\[
V_n = \sigma^2 \left( n^{-1} \sum_j X_j^T R c X_j \right).
\]

The \((2, 2)\)th element of \( \Sigma_n \) is the robust variance of \( n^{1/2} \hat{\beta}_2 \), denoted as \( \sigma^2_2 \). We reject \( H_0 : \beta_2 = 0 \) if \( |n^{1/2} \hat{\beta}_2 / \sigma_2| > z_{1-\alpha/2} \), where \( z_{1-\alpha/2} \) is the 100(1 - \( \alpha/2 \))th percentile of the standard normal distribution. Given the type I error rate \( \alpha \), power \( 1 - \gamma \), and the true value of the intervention effect \( \beta_2 \), we can solve \( n \) from equation

\[
P \left( \frac{|\hat{\beta}_2|}{\hat{\sigma}_2 / \sqrt{n}} > z_{1-\alpha/2} | H_a \right) = 1 - \gamma.
\]

Thus, the required total number of clusters with complete observations is

\[
n = \frac{\sigma^2_2 (z_{1-\alpha/2} + z_{1-\gamma})^2}{\beta^2_2}.
\]

(4.3)

We further extend the GEE sample size approach to accommodate the potential occurrence of missing data. This extension requires the assumption of missing completely at random (MCAR) [47]. That is, the missing events are independent both of the observed or unobserved outcomes. Let \( \Delta_{ijk} \) be an indicator variable, which \( \Delta_{ijk} = 1 \) if study unit \( jk \) has an outcome measurement under treatment \( t_i \) and 0 otherwise, for \( i = 1, 2 \). Under MCAR, \( (\Delta_{1jk}, \Delta_{2jk}) \) are independent of \( (Y_{1jk}, Y_{2jk}) \). In the presence of missing data, \( X_j \), \( Y_j \), \( \epsilon_j \) can be re-written as
Let $E(\Delta_{ijk} = 1) = s_{ijk}$ be the proportion of observing an outcome for study unit $k$ in cluster $j$ under treatment $i$. We assume that the study units under the same treatment share the same missing proportion, that $E(\Delta_{1jk} = 1) = s_1$ and $E(\Delta_{2jk} = 1) = s_2$. We further define the correlation coefficient for the missing data, $\tau_1 = \text{corr}(Y_{ijk}, Y_{ijk'})$ for $k \neq k'$, and $\tau_2 = \text{corr}(Y_{ijk}, Y_{i'jk})$ for $i \neq i'$. Therefore, we have $E(\Delta_{ijk}\Delta_{ijk'}) = s_1^2 + \tau_1 s_1 t_i$, $E(\Delta_{ijl}\Delta_{ijl}) = s_{12} = s_1 s_2 + \tau_2 \sqrt{s_1 t_1 s_2 t_2}$, and $E(\Delta_{ijl}\Delta_{ijl}) = E(\Delta_{ijl})$, where $t_i = 1 - s_i$.

And the missing correlation matrix can be written as

$$R_m = \begin{pmatrix}
(1 - \tau_1)I_m + \tau_1 J_m & \tau_2 I_m \\
\tau_2 I_m & (1 - \tau_1)I_m + \tau_1 J_m
\end{pmatrix}.$$  

The missing correlation matrix $R_m$ has four distinct eigenvalues,

$$\lambda_1' = 1 - \tau_1 + \tau_2, \quad \lambda_2' = 1 + (m - 1)\tau_1 + \tau_2,$$

$$\lambda_3' = 1 - \tau_1 - \tau_2, \quad \lambda_4' = 1 + (m - 1)\tau_1 - \tau_2.$$  

The setting of $(\tau_1, \tau_2)$ should satisfy $\min(\lambda_1', \lambda_2', \lambda_3', \lambda_4') > 0$ so that the correlation matrix $R_m$ is positive definite.
Therefore, as \( n \to \infty \), we have

\[
A_n^{-1} \xrightarrow{n \to \infty} \begin{pmatrix}
\frac{1}{ms_1} & -\frac{1}{ms_1} \\
-\frac{1}{ms_1} & \frac{1}{ms_1} + \frac{1}{ms_2}
\end{pmatrix}, \quad V_n \xrightarrow{n \to \infty} \sigma^2 \begin{pmatrix}
V_{11} & V_{12} \\
V_{21} & V_{22}
\end{pmatrix}.
\]

where

\[
V_{11} = [s_1 + s_2 + \rho_1(m - 1)(s_1^2 + s_2^2 + \tau_1 s_1 t_1 + \tau_1 s_2 t_2) + 2\rho_2(s_1 s_2 + \tau_2 \sqrt{s_1 t_1 s_2 t_2}) + 2\rho_3(m - 1)s_1 s_2]m,
\]

\[
V_{12} = V_{21} = [s_2 + \rho_1(m - 1)(s_2^2 + \tau_1 s_2 t_2) + \rho_2(s_1 s_2 + \tau_2 \sqrt{s_1 t_1 s_2 t_2}) + \rho_3(m - 1)s_1 s_2]m,
\]

\[
V_{22} = [s_2 + \rho_1(m - 1)(s_2^2 + \tau_1 s_2 t_2)]m.
\]

The \((2, 2)\)th element of \( \Sigma_n = A_n^{-1}V_nA_n^{-1} \) converges to

\[
\sigma^2 = \frac{\sigma^2}{m} \left[ \frac{1}{s_1} + \frac{1}{s_2} - 2\rho_2 + 2(m - 1)(\rho_1 - \rho_3) + \frac{\rho_1(m - 1)\tau_1(s_1 t_2 + s_2 t_1) - 2\rho_2\tau_2 \sqrt{s_1 t_1 s_2 t_2}}{s_1 s_2} \right].
\]

Thus, instead of the equation (4.3), the required total number of clusters that accounts for potential incomplete observations is given by

\[
n_{GEE} = \frac{\sigma^2(z_{1-\alpha/2} + z_{1-\gamma})^2}{m\beta_2^2} \left[ \frac{1}{s_1} + \frac{1}{s_2} - 2\rho_2 + 2(m - 1)(\rho_1 - \rho_3) + \frac{\rho_1(m - 1)\tau_1(s_1 t_2 + s_2 t_1) - 2\rho_2\tau_2 \sqrt{s_1 t_1 s_2 t_2}}{s_1 s_2} \right].
\]

(4.4)

The estimated total number of clusters \( n \) depends on the missing structure through parameters \((s_1, s_2, \tau_1, \tau_2)\), and the correlation combination \((\rho_1, \rho_2, \rho_3)\), in addition to the variance of error terms \( \sigma^2 \), the true intervention effect \( \beta_2 \), the cluster size \( m \), and type I and type II error. A stronger within-subject correlation coefficient \( \rho_2 \) or inter-treatment correlation coefficient \( \rho_3 \) would lead to a smaller sample size, while a stronger within-
treatment correlation coefficient $\rho_1$ would lead to a larger sample size, with all the other factors fixed. When $m \geq 2$, the changes for within-treatment correlation coefficient $\rho_1$ and inter-treatment correlation coefficient $\rho_3$ are more sensitive than that for within-subject correlation coefficient $\rho_2$.

As a special case, we consider the independent missing, where the missing for different study units in the same cluster are independent ($\tau_1 = 0$), and the missing under treatment $t_1$ is independent of missing under treatment $t_2$ ($\tau_2 = 0$). Then equation (4.4) reduces to

$$n_{\text{GEE}}^* = \frac{\sigma^2(z_1 - \alpha/2 + z_1 - \gamma)^2}{m\beta_2^2} \left[ \frac{1}{s_1} + \frac{1}{s_2} - 2\rho_2 + 2(m - 1)(\rho_1 - \rho_3) \right]. \quad (4.5)$$

On the other hand, if we assume no missing data ($s_1 = s_2 = 1$), we can estimate the total number of clusters by

$$n_{\text{complete}} = \frac{\sigma^2(z_1 - \alpha/2 + z_1 - \gamma)^2}{m\beta_2^2} [2 - 2\rho_2 + 2(m - 1)(\rho_1 - \rho_3)]. \quad (4.6)$$

4.1.2. Crude adjustment for incomplete observations

In paired cluster experiments, one common ad hoc approach for incomplete observations is to use the crude adjustment. This approach first estimates the total number of study units under complete pairs of outcomes, denoted by $N_{\text{complete}} = mn_{\text{complete}}$, where $m$ is cluster size. And then, calculate the total number of study units under incomplete observations by $N_{\text{crude}} = N_{\text{complete}}/q$, where $q$ is the proportion of study units with complete pairs of outcomes. Under the general missing and the independent missing, $q = s_1 s_2 + \tau_2 \sqrt{s_1 t_1 s_2 t_2}$ and $q^* = s_1 s_2$, respectively. Let $R$ be the sample size ratio comparing the proposed GEE method with the crude adjustment under the general missing,
where
\[
R = \frac{N_{\text{GEE}}}{N_{\text{crude}}} = \frac{\frac{1}{s_1} + \frac{1}{s_2} - 2\rho_2 + 2(m - 1)(\rho_1 - \rho_3)}{2 - 2\rho_2 + 2(m - 1)(\rho_1 - \rho_3)} \cdot \frac{\rho_1(m-1)\tau_1(s_1t_2+s_2t_1)-2\rho_2\tau_2\sqrt{s_1t_1s_2t_2}}{s_1s_2}.
\]

Let \( R^* \) be the sample size ratio comparing the proposed GEE method with the crude adjustment under the independent missing, where

\[
R^* = \frac{N'_{\text{GEE}}}{N'_{\text{crude}}} = \frac{\frac{1}{s_1} + \frac{1}{s_2} - 2\rho_2 + 2(m - 1)(\rho_1 - \rho_3)}{2 - 2\rho_2 + 2(m - 1)(\rho_1 - \rho_3)} \cdot s_1s_2.
\]

Under the independent missing, when \( \rho_2 < \rho^* = \frac{s_1+s_2-2}{2(s_1s_2-1)+(m-1)(\rho_1-\rho_3)} \), the crude adjustment would lead to a larger sample size, whereas when \( \rho_2 > \rho^* \), the crude adjustment would lead to a smaller sample size comparing with the proposed GEE method. We can also find the corresponding cutoff point \( \rho \) under the general missing assumption. To illustrate the impact of within-subject correlation on the sample size ratio, we plot the sample size ratio \( R(R^*) \) versus within-subject correlation coefficient \( \rho_2 \) in Figure 4.1 under different missing patterns and combinations of missing proportions \( (s_1, s_2) \), fixed \( (\rho_1, \rho_3) = (0.01, 0.005) \) and \( m = 10 \). The cutoff point \( \rho(\rho^*) \) for each combination is marked in Figure 4.1. It shows that the sample size ratio \( R(R^*) \) increases as \( \rho_2 \) increases, and the range of the sample size ratio \( R(R^*) \) becomes bigger with increased missing proportion.

For incomplete observations, the sample size estimating by the proposed GEE method would be more accurate comparing with the sample size by the crude adjustment. When within-subject correlation coefficient \( \rho_2 \) is smaller (greater) than the cutoff point, the crude adjustment would overestimate (underestimate) the sample size. If we choose small to moderate within-subject correlation, the proposed GEE method would mostly lead to a saving in sample size.
4.2. Simulation

We conduct simulation studies to assess the performance of the proposed GEE method for paired cluster experimental design with incomplete observations of continuous outcomes, under various design configurations. The nominal levels of type I error and power are set at $\alpha = 5\%$ and $1 - \gamma = 80\%$, respectively. We consider both independent missing pattern with $(\tau_1, \tau_2) = (0, 0)$ and general missing pattern with $(\tau_1, \tau_2) = (0.3, 0.1)$ or $(0.5, 0.2)$, and six combinations of missing proportions $(s_1, s_2)$: $S1 = (s_1, s_2) = (1, 1)$ representing no missing data; $S2 = (s_1, s_2) = (0.85, 0.85)$ representing balanced distribution of missing values in the outcomes under the treatment $t_1$ and treatment $t_2$; $S3 = (s_1, s_2) = (0.9, 0.8), S4 = (s_1, s_2) = (0.8, 0.9), S5 = (s_1, s_2) = (0.6, 0.7), S6 = (s_1, s_2) = (0.5, 0.6)$ representing unbalanced distribution of missing values and missing proportions varying from $10\%$ to $50\%$. In the cluster design, cluster-level missing is a more common problem [48], so we assume $\tau_1 > \tau_2$. We set the cluster size $m = 10$ or $20$, the true values of regression coefficients $\beta = (\beta_1, \beta_2)^T = (0.3, 0.15)^T$ and variance $\sigma^2 = 0.75$ or $1$, where $(\beta_2, \sigma^2) = (0.15, 1)$ indicates an effect size of $0.15$ comparing the intervention with control. We choose six distinct combinations of the correlation values $(\rho_1, \rho_2, \rho_3)$: $(0.01, 0.15, 0.005), (0.01, 0.3, 0.005), (0.05, 0.15, 0.005), (0.05, 0.15, 0.025), (0.05, 0.3, 0.005)$ and $(0.05, 0.3, 0.025)$.

The simulation procedure is as follows: (i) For each combination of $(m, \rho_1, \rho_2, \rho_3, s_1, s_2, \tau_1, \tau_2, \sigma^2)$, calculate the required number of clusters $n_{\text{GEE}}$ based on equation (4.4). (ii) For each iteration $c (c = 1, ..., C)$, generate $n_{\text{GEE}}$ correlated random errors $\epsilon_j^{(c)}$'s from the multivariate normal distribution with mean $0_{2m}$ and variance-covariance matrix $\sigma^2 R_m$. (iii) Obtain the continuous outcomes $Y_j^{(c)}$ from model (4.1), where $\beta_2 = 0$ under the null hypothesis $H_0$ and $\beta_2 = 0.15$ under the alternative hypothesis $H_a$, respectively. (iv) According to pre-specified missing patterns, generate the missing indicator variable $\Delta_{ijk}^{(c)}$, and combine the indicator variable $\Delta_{ijk}^{(c)}$ with $Y_{ijk}^{(c)}, \epsilon_{ijk}^{(c)}$ and $t_{ijk}$ one on one. (v) Estimate
\( \hat{\beta}^{(c)} \) based on equation (4.2) and calculate \( \epsilon_j^{(c)} \) by \( Y_j^{(c)} - X_j^{(c)} \hat{\beta}^{(c)} \) to estimate \( \hat{\Sigma}_n^{(c)} \) by

\[
\hat{\Sigma}_n^{(c)} = \left( n^{-1} \sum_j X_j^{T(\cdot)} X_j^{(\cdot)} \right)^{-1} \left( n^{-1} \sum_j X_j^{T(\cdot)} \epsilon_j^{(c)} \epsilon_j^{T(\cdot)} X_j^{(\cdot)} \right) \left( n^{-1} \sum_j X_j^{T(\cdot)} X_j^{(\cdot)} \right)^{-1},
\]

and then calculate the test statistic by \( |\sqrt{n_{\text{GEE}}} \hat{\beta}^{(\cdot)} / \hat{\sigma}^{(\cdot)}| \). The empirical type I error rate and the empirical power are calculated as proportion of times that \( H_0 \) is rejected under \( H_0 \) and \( H_a \), respectively. We set the total number of iterations \( C = 5,000 \).

Table 4.1 presents the estimate of number of clusters, empirical type I error and empirical power for simulation under independent missing, and Tables 4.2 and 4.3 present the simulation results under general missing with the cluster size \( m = 10 \) and \( m = 20 \), respectively. Under the design configurations \( (m, \rho_1, \rho_2, \rho_3, s_1, s_2, \tau_1, \tau_2, \sigma^2) \), the estimated number of clusters changes from 21 to 157. The empirical power is preserved to the nominal level for all the scenarios, suggesting that the proposed method allows paired cluster experiment to be adequately powered in the presence of missing data. With all the other factors fixed, the required number of clusters increases as the variance \( \sigma^2 \) increases. The required number of clusters increases as the within-subject correlation coefficient \( \rho_2 \) and inter-treatment correlation coefficient \( \rho_3 \) decrease, or as the within-treatment correlation coefficient \( \rho_1 \) increases. The missing correlation coefficients have the similar effects on the required number of clusters that within-treatment \( \tau_1 \) (within-subject \( \tau_2 \)) missing correlation coefficient leads to a larger (smaller) number of clusters.

The ability to control the empirical type I error rates performs generally well with moderate to large sample sizes. When the sample size is small, the empirical type I error rates are slightly higher than the nominal level of 5\%. This is because that the sandwich variance estimator tends to be biased downwards and may inflate the type I error rate when the sample size is small, typically \( n < 40 \) \[29, 59\]. To address this issue, we further use the bias-corrected variance estimators to adjust the bias for the scenario \( (\rho_1, \rho_2, \rho_3) = (0.01, 0.3, 0.005) \) and \( m = 20 \) under both missing patterns, and the corre-
sponding empirical type I error rates are showed in Figure 4.2, where KC and MD represent the bias-corrected variance estimators proposed by Kauermann and Corroll [20] and Mancl and DeRouen [35], respectively, and LZ is the uncorrected sandwich estimator of Zeger and Liang [63]. With the same generated dataset, the variances of these methods are LZ < KC < MD [41], therefore, KC tends to give a moderate adjustment and MD provides the most conservative results. The empirical type I error rate based on KC is lower than that based on LZ but still slightly higher than the nominal level (less than 7%), while MD has a better performance of controlling the type I error rate resulting in a loss of power within 2.6%. Also, with the increased number of clusters, the difference of both error rates among these three methods become smaller. Moreover, we conduct a sensitivity analysis to investigate how the misspecified missing pattern affects the empirical power and the required number of clusters. We estimate the required number of clusters using formula (4.5) under independent missing, while the data are generated under general missing with \((\tau_1, \tau_2) = (0.3, 0.1)\) or \((0.5, 0.2)\) and \((s_1, s_2) = (0.85, 0.85)\) or \((0.5, 0.6)\).

We then calculate and plot the empirical powers in Figure 4.3 under misspecified missing pattern and true missing pattern for different scenarios. Figure 4.3 suggests that, with the increased missing proportion, the effect of misspecified missing pattern becomes bigger, especially for the cluster size \(m = 20\), and larger missing correlation coefficients would increase the effect.

4.3. Example

Andrade et al [1] reported a matched-pair cluster-randomized trial of improving physical fitness in Ecuadorian adolescents. Ten pairs of schools were enrolled with the average school size 72 and the schools were randomly assigned to either the intervention group or control group. The students discontinued/missing rates were 21.4% and 28.0% for intervention group and control group, respectively. One primary physical fitness outcome was speed shuttle run time, and the average (standard deviation) time were 1.89(2.09) and
2.69(3.44) seconds for intervention group and control group, respectively. In this trial, the intervention effect and the variance of random error were estimated as $\hat{\beta}_2 = -0.72$ and $\hat{\sigma}^2 = 5$, respectively, with the within-treatment correlation estimated as $\hat{\rho}_1 = 0.15$.

An investigator would like to conduct a similar study to detect whether there is a significant difference in speed shuttle run time under a new intervention comparing with a control. Based on the preliminary data, we specify the design factors are $m = 72$, $\beta_2 = -0.72$, $\sigma^2 = 5$, $\rho_1 = 0.15$, $s_2 = 1 - 21.4\%$ and $s_1 = 1 - 28.0\%$. We assume a small value of inter-treatment correlation coefficient $\rho_2 = \rho_3 = 0.005$. Under the general missing pattern with $(\tau_1, \tau_2) = (0.3, 0)$, the required number of clusters by the proposed GEE method are 27 per group to achieve 80% power at a two-sided 5% significance level. Under the independent missing pattern with $(\tau_1, \tau_2) = (0, 0)$, the required number of clusters decreases to 24 per group, which would lead to an underpowered study. Moreover, the number of clusters calculated by the crude adjustment are 42, suggesting that the proposed GEE method would lead to a saving in sample size when the within-subject correlation is small.

4.4. Future direction

The GEE approach for paired cluster experimental design with incomplete observations can be also extended to binary outcomes, where $Y_{ijk}$ is the binary outcome of study unit $k$ in cluster $j$ under treatment $i$, and other settings are same as the continuous outcomes in Section 4.1. To make an inference on the intervention effect on $Y_{ijk}$, we assume the following logistic regression model: $Y_{ijk} \sim \text{Bernoulli}(p_{ijk})$ and

$$\text{logit}\{Pr(Y_{ijk} = 1)\} = \log\left(\frac{p_{ijk}}{1-p_{ijk}}\right) = \beta_1 + \beta_2 t_{ijk},$$

where $\beta_1$ is the log-transformed odds for the pre time point or control group, and $\beta_2$ is the log-transformed odds ratio between treatment $t_1$ and treatment $t_2$, representing the
treatment difference on the outcome. The primary interest is to test the null hypothesis $H_0 : \beta_2 = 0$ accounting for the missing patterns and the correlation structure, with a power of $1 - \gamma$ at a two-sided significant level of $\alpha$. Let $\beta = (\beta_1, \beta_2)^T$ and $Z_{ijk} = (1, t_{ijk})^T$. The model (4.7) can be written as

$$p_{ijk} = \frac{e^{\beta Z_{ijk}}}{1 + e^{\beta Z_{ijk}}}.$$ 

Under the independent working correlation structure, the GEE estimator $\hat{\beta}$ is obtained by solving the equation

$$S_n^*(\beta) = n^{-1/2} \sum_{j=1}^{n} \sum_{i=1}^{2} \sum_{k=1}^{m} \{Y_{ijk} - p_{ijk}(\beta)\} Z_{ijk} = 0.$$ 

The equation is solved by the Newton-Raphson algorithm. At the $t$th iteration,

$$\hat{\beta}^{(t)} = \hat{\beta}^{(t-1)} + n^{-1/2} A_n^{-1}\left(\hat{\beta}^{(t-1)}\right) S_n^*\left(\hat{\beta}^{(t-1)}\right),$$

where

$$A_n(\beta) = -n^{-1/2} \frac{\partial S_n^*(\beta)}{\partial \beta} = n^{-1} \sum_{j=1}^{n} \sum_{i=1}^{2} \sum_{k=1}^{m} p_{ijk}(1 - p_{ijk}) \begin{pmatrix} 1 & t_{ijk} \\ t_{ijk} & t_{ijk}^2 \end{pmatrix}.$$ 

By Zeger and Liang [63], $n^{-1/2}(\hat{\beta} - \beta)$ is approximately normal with mean 0 and the variance is consistently estimated by $\Sigma_n^* = A_n^{-1}\left(\hat{\beta}\right) V_n\left(\hat{\beta}\right) A_n^{-1}\left(\hat{\beta}\right)$, where

$$V_n\left(\hat{\beta}\right) = n^{-1} \sum_{j=1}^{n} \left( \sum_{i=1}^{2} \sum_{k=1}^{m} \hat{\epsilon}_{ijk} Z_{ijk} \right)^{\otimes 2}, \hat{\epsilon}_{ijk} = Y_{ijk} - p_{ijk}(\hat{\beta}).$$

Here, $c^{\otimes 2} = cc^T$ for a vector $c$. The $(2, 2)$th element of $\Sigma_n^*$ is the robust variance of $n^{1/2}\hat{\beta}_2$, denoted as $\sigma_{2*}^2$. We reject $H_0 : \beta_2 = 0$ if $|n^{1/2}\hat{\beta}_2/\sigma_{2*}| > z_{1-\alpha/2}$. Denote the true success rates at treatment $t_1$ and treatment $t_2$ as $P_1 = e^{\beta_1}/(1 + e^{\beta_1})$ and $P_2 = e^{\beta_1+\beta_2}/(1 + e^{\beta_1+\beta_2})$, respectively. Define $Q_1 = 1 - P_1$ and $Q_2 = 1 - P_2$. In Appendix A.2, accounting for
the missing patterns and the correlation structure, as \( n \to \infty \), we show that the \((2, 2)\)th element of \( \Sigma_n^* \) converges to

\[
\sigma_{2*}^2 = \frac{s_1 + (m - 1)(s_1^2 + \tau_1 s_1 t_1)\rho_1}{ms_1^2 P_1 Q_1} + \frac{s_2 + (m - 1)(s_2^2 + \tau_1 s_2 t_2)\rho_1}{ms_2^2 P_2 Q_2} - \frac{2\sqrt{P_1 P_2 Q_1 Q_2} [(s_1 s_2 + \tau_2 \sqrt{s_1 s_2 t_1 t_2})\rho_2 + (m - 1)s_1 s_2 \rho_3]}{ms_1 s_2 P_1 P_2 Q_1 Q_2}.
\]

(4.8)

Thus, the required total number of clusters that accounts for potential incomplete observations is given by

\[
n_{GEE^*} = \frac{\sigma_{2*}^2 (z_{1-\alpha/2} + z_{1-\gamma})^2}{\beta_2^2}.
\]

(4.9)

In future work, we will conduct simulation studies to assess the performance of the proposed GEE sample size approach for paired cluster experimental design with incomplete observations of binary outcomes under various design configurations.

### 4.5. Acknowledgments

This work was supported in part by the Cancer Center Support Grant from the National Cancer Institute (2P30CA142543) and the National Center for Advancing Translational Sciences of the National Institutes of Health (UL1TR001105).
Figure 4.1: The plot of sample size ratio against within-subject correlation coefficient under the general missing ($\tau_1 = 0.3, \tau_2 = 0.1$) and the independent missing, with different missing combinations of $(s_1, s_2)$.

Figure 4.2: Empirical type I error for three variance estimators. LZ: uncorrected sandwich variance; KC: KC-corrected sandwich variance; MD: MD-corrected sandwich variance.
Figure 4.3: Empirical power under misspecified missing pattern and true missing pattern.
Table 4.1: Number of clusters (empirical type I error, empirical power) for simulation under Independent missing, type I error = 5%, power = 80%.

<table>
<thead>
<tr>
<th></th>
<th>(0.01, 0.15, 0.005)</th>
<th>(0.01, 0.3, 0.005)</th>
<th>(0.05, 0.15, 0.005)</th>
<th>(0.05, 0.15, 0.025)</th>
<th>(0.05, 0.3, 0.005)</th>
<th>(0.05, 0.3, 0.025)</th>
</tr>
</thead>
<tbody>
<tr>
<td>m = 10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>σ² = 1</td>
<td>S1  62 (5.64%, 80.52%)</td>
<td>52 (5.32%, 81.06%)</td>
<td>88 (5.32%, 80.00%)</td>
<td>75 (5.62%, 80.46%)</td>
<td>77 (5.36%, 80.30%)</td>
<td>65 (4.66%, 81.32%)</td>
</tr>
<tr>
<td></td>
<td>S2 75 (5.26%, 81.10%)</td>
<td>64 (5.44%, 79.66%)</td>
<td>100 (5.84%, 79.48%)</td>
<td>87 (5.70%, 78.94%)</td>
<td>89 (5.10%, 80.50%)</td>
<td>77 (5.54%, 80.32%)</td>
</tr>
<tr>
<td></td>
<td>S3 75 (5.66%, 79.26%)</td>
<td>65 (6.04%, 80.06%)</td>
<td>100 (4.96%, 80.42%)</td>
<td>88 (5.40%, 80.50%)</td>
<td>90 (4.72%, 80.16%)</td>
<td>77 (6.20%, 80.42%)</td>
</tr>
<tr>
<td></td>
<td>S4 75 (5.42%, 80.04%)</td>
<td>65 (5.88%, 80.34%)</td>
<td>100 (5.24%, 79.80%)</td>
<td>88 (5.46%, 79.84%)</td>
<td>90 (5.46%, 79.70%)</td>
<td>77 (5.10%, 80.00%)</td>
</tr>
<tr>
<td></td>
<td>S5 101 (5.30%, 80.48%)</td>
<td>90 (5.22%, 79.86%)</td>
<td>126 (5.40%, 79.64%)</td>
<td>113 (5.22%, 79.84%)</td>
<td>115 (5.08%, 80.26%)</td>
<td>103 (5.84%, 79.78%)</td>
</tr>
<tr>
<td></td>
<td>S6 121 (5.58%, 79.74%)</td>
<td>110 (5.46%, 80.28%)</td>
<td>146 (5.38%, 80.74%)</td>
<td>133 (5.46%, 80.34%)</td>
<td>135 (4.98%, 80.56%)</td>
<td>123 (5.86%, 79.58%)</td>
</tr>
<tr>
<td>σ² = 0.75</td>
<td>S1 47 (6.06%, 80.46%)</td>
<td>39 (6.10%, 80.72%)</td>
<td>66 (5.50%, 80.84%)</td>
<td>56 (5.76%, 81.10%)</td>
<td>58 (5.76%, 80.54%)</td>
<td>48 (5.82%, 79.78%)</td>
</tr>
<tr>
<td></td>
<td>S2 56 (5.08%, 80.96%)</td>
<td>48 (5.48%, 79.90%)</td>
<td>75 (5.68%, 80.68%)</td>
<td>65 (5.76%, 80.44%)</td>
<td>67 (5.00%, 80.50%)</td>
<td>58 (5.52%, 80.56%)</td>
</tr>
<tr>
<td></td>
<td>S3 56 (5.90%, 79.40%)</td>
<td>48 (5.62%, 80.16%)</td>
<td>75 (5.34%, 79.74%)</td>
<td>66 (5.56%, 80.62%)</td>
<td>67 (5.58%, 79.38%)</td>
<td>58 (5.44%, 80.22%)</td>
</tr>
<tr>
<td></td>
<td>S4 56 (5.28%, 80.20%)</td>
<td>48 (5.26%, 80.44%)</td>
<td>75 (5.36%, 80.16%)</td>
<td>66 (5.48%, 81.22%)</td>
<td>67 (5.22%, 80.76%)</td>
<td>58 (5.76%, 81.04%)</td>
</tr>
<tr>
<td></td>
<td>S5 75 (5.42%, 79.04%)</td>
<td>68 (6.18%, 80.18%)</td>
<td>94 (5.10%, 80.48%)</td>
<td>85 (5.34%, 79.94%)</td>
<td>86 (5.54%, 80.18%)</td>
<td>77 (5.14%, 80.28%)</td>
</tr>
<tr>
<td></td>
<td>S6 90 (5.50%, 78.52%)</td>
<td>83 (6.10%, 80.70%)</td>
<td>109 (5.58%, 79.66%)</td>
<td>100 (5.40%, 80.46%)</td>
<td>101 (5.68%, 80.70%)</td>
<td>92 (5.98%, 80.80%)</td>
</tr>
<tr>
<td>m = 20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>σ² = 1</td>
<td>S1 33 (6.20%, 80.20%)</td>
<td>28 (6.40%, 81.08%)</td>
<td>59 (5.84%, 79.84%)</td>
<td>46 (5.90%, 80.94%)</td>
<td>54 (5.40%, 81.26%)</td>
<td>41 (5.72%, 79.92%)</td>
</tr>
<tr>
<td></td>
<td>S2 39 (6.40%, 79.52%)</td>
<td>34 (6.22%, 79.80%)</td>
<td>66 (5.78%, 80.44%)</td>
<td>52 (5.84%, 80.18%)</td>
<td>60 (5.24%, 79.76%)</td>
<td>47 (5.36%, 79.32%)</td>
</tr>
<tr>
<td></td>
<td>S3 39 (6.12%, 80.10%)</td>
<td>34 (5.60%, 80.50%)</td>
<td>66 (5.84%, 80.46%)</td>
<td>53 (5.94%, 80.90%)</td>
<td>61 (6.40%, 80.62%)</td>
<td>47 (5.68%, 80.38%)</td>
</tr>
<tr>
<td></td>
<td>S4 39 (6.68%, 80.10%)</td>
<td>34 (5.88%, 79.96%)</td>
<td>66 (5.32%, 80.96%)</td>
<td>53 (5.68%, 79.84%)</td>
<td>61 (5.22%, 80.30%)</td>
<td>47 (5.56%, 81.08%)</td>
</tr>
<tr>
<td></td>
<td>S5 52 (6.04%, 80.26%)</td>
<td>47 (5.92%, 80.16%)</td>
<td>79 (5.48%, 79.40%)</td>
<td>65 (5.42%, 80.54%)</td>
<td>73 (5.18%, 79.58%)</td>
<td>60 (5.80%, 79.94%)</td>
</tr>
<tr>
<td></td>
<td>S6 62 (4.88%, 80.22%)</td>
<td>57 (5.42%, 81.56%)</td>
<td>89 (5.50%, 80.70%)</td>
<td>75 (5.90%, 78.64%)</td>
<td>83 (6.00%, 79.82%)</td>
<td>70 (5.36%, 80.22%)</td>
</tr>
<tr>
<td>σ² = 0.75</td>
<td>S1 25 (7.50%, 81.44%)</td>
<td>21 (7.60%, 81.50%)</td>
<td>45 (6.46%, 81.08%)</td>
<td>35 (6.48%, 81.24%)</td>
<td>41 (5.60%, 81.42%)</td>
<td>31 (5.96%, 80.38%)</td>
</tr>
<tr>
<td></td>
<td>S2 29 (6.68%, 79.66%)</td>
<td>25 (6.52%, 79.92%)</td>
<td>49 (5.48%, 81.12%)</td>
<td>39 (5.88%, 80.16%)</td>
<td>45 (5.92%, 79.68%)</td>
<td>35 (6.14%, 80.92%)</td>
</tr>
<tr>
<td></td>
<td>S3 29 (6.60%, 80.74%)</td>
<td>26 (6.36%, 81.42%)</td>
<td>49 (6.02%, 79.80%)</td>
<td>39 (5.94%, 79.80%)</td>
<td>45 (5.92%, 79.42%)</td>
<td>35 (6.02%, 80.36%)</td>
</tr>
<tr>
<td></td>
<td>S4 29 (6.38%, 80.16%)</td>
<td>26 (6.66%, 82.20%)</td>
<td>49 (6.60%, 78.94%)</td>
<td>39 (5.32%, 80.52%)</td>
<td>45 (6.36%, 80.20%)</td>
<td>35 (6.02%, 79.70%)</td>
</tr>
<tr>
<td></td>
<td>S5 39 (6.28%, 79.80%)</td>
<td>35 (5.84%, 80.60%)</td>
<td>59 (5.22%, 80.56%)</td>
<td>49 (6.08%, 80.96%)</td>
<td>55 (5.52%, 79.52%)</td>
<td>45 (5.88%, 81.16%)</td>
</tr>
<tr>
<td></td>
<td>S6 47 (6.38%, 80.26%)</td>
<td>43 (6.02%, 81.16%)</td>
<td>66 (5.62%, 80.32%)</td>
<td>56 (5.90%, 79.88%)</td>
<td>62 (5.58%, 79.48%)</td>
<td>53 (5.38%, 80.18%)</td>
</tr>
</tbody>
</table>
Table 4.2: Number of clusters (empirical type I error, empirical power) for simulation under general missing when \( m = 10 \), type I error = 5%, power = 80%.

<table>
<thead>
<tr>
<th>((\rho_1, \rho_2, \rho_3))</th>
<th>(\sigma^2 = 1)</th>
<th>(\sigma^2 = 0.75)</th>
<th>(\sigma^2 = 1)</th>
<th>(\sigma^2 = 0.75)</th>
<th>(\sigma^2 = 1)</th>
<th>(\sigma^2 = 0.75)</th>
</tr>
</thead>
<tbody>
<tr>
<td>((0.01, 0.3, 0.005))</td>
<td>((0.01, 0.3, 0.005))</td>
<td>((0.01, 0.15, 0.005))</td>
<td>((0.01, 0.3, 0.005))</td>
<td>((0.01, 0.15, 0.025))</td>
<td>((0.05, 0.3, 0.005))</td>
<td>((0.05, 0.3, 0.025))</td>
</tr>
<tr>
<td>((\tau_1, \tau_2) = (0.3, 0.1))</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S1</td>
<td>62 (5.46%, 80.76%)</td>
<td>52 (5.62%, 80.46%)</td>
<td>88 (5.52%, 80.18%)</td>
<td>75 (6.10%, 79.36%)</td>
<td>77 (5.36%, 81.00%)</td>
<td>65 (5.88%, 80.68%)</td>
</tr>
<tr>
<td>S2</td>
<td>75 (5.56%, 80.60%)</td>
<td>64 (4.84%, 80.94%)</td>
<td>101 (5.40%, 79.70%)</td>
<td>89 (5.00%, 80.32%)</td>
<td>91 (5.06%, 79.86%)</td>
<td>78 (5.62%, 79.70%)</td>
</tr>
<tr>
<td>S3</td>
<td>75 (5.76%, 80.16%)</td>
<td>65 (6.06%, 80.18%)</td>
<td>102 (5.98%, 80.28%)</td>
<td>89 (5.42%, 80.52%)</td>
<td>91 (5.12%, 79.56%)</td>
<td>78 (6.20%, 80.18%)</td>
</tr>
<tr>
<td>S4</td>
<td>75 (6.28%, 80.58%)</td>
<td>65 (6.60%, 80.62%)</td>
<td>102 (5.34%, 80.24%)</td>
<td>89 (5.68%, 80.92%)</td>
<td>91 (5.48%, 81.28%)</td>
<td>78 (5.86%, 79.58%)</td>
</tr>
<tr>
<td>S5</td>
<td>101 (5.10%, 80.06%)</td>
<td>90 (4.72%, 80.60%)</td>
<td>130 (6.00%, 79.32%)</td>
<td>118 (6.02%, 80.70%)</td>
<td>119 (5.48%, 81.28%)</td>
<td>107 (4.98%, 78.94%)</td>
</tr>
<tr>
<td>S6</td>
<td>121 (5.22%, 79.62%)</td>
<td>110 (5.44%, 80.74%)</td>
<td>153 (5.60%, 79.78%)</td>
<td>140 (5.36%, 80.04%)</td>
<td>141 (5.20%, 79.80%)</td>
<td>129 (5.16%, 79.66%)</td>
</tr>
<tr>
<td>((\tau_1, \tau_2) = (0.5, 0.2))</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S1</td>
<td>47 (5.76%, 80.50%)</td>
<td>39 (5.66%, 80.34%)</td>
<td>66 (6.12%, 80.16%)</td>
<td>56 (5.50%, 79.44%)</td>
<td>58 (5.74%, 80.26%)</td>
<td>48 (6.20%, 79.84%)</td>
</tr>
<tr>
<td>S2</td>
<td>56 (5.68%, 80.54%)</td>
<td>56 (5.86%, 79.34%)</td>
<td>76 (6.50%, 79.76%)</td>
<td>68 (5.24%, 80.80%)</td>
<td>69 (5.60%, 79.82%)</td>
<td>60 (5.60%, 80.70%)</td>
</tr>
<tr>
<td>S3</td>
<td>56 (5.82%, 79.18%)</td>
<td>56 (7.62%, 80.30%)</td>
<td>76 (5.86%, 79.86%)</td>
<td>68 (5.92%, 80.24%)</td>
<td>69 (5.82%, 80.00%)</td>
<td>59 (5.70%, 79.80%)</td>
</tr>
<tr>
<td>S4</td>
<td>76 (5.82%, 79.62%)</td>
<td>68 (5.14%, 79.96%)</td>
<td>98 (5.98%, 80.10%)</td>
<td>90 (5.18%, 80.68%)</td>
<td>92 (5.24%, 79.76%)</td>
<td>59 (5.70%, 80.00%)</td>
</tr>
<tr>
<td>S5</td>
<td>75 (5.76%, 80.16%)</td>
<td>56 (6.04%, 79.92%)</td>
<td>100 (5.76%, 80.16%)</td>
<td>90 (5.52%, 80.18%)</td>
<td>91 (5.18%, 79.54%)</td>
<td>59 (5.52%, 79.76%)</td>
</tr>
<tr>
<td>S6</td>
<td>91 (5.48%, 80.16%)</td>
<td>82 (6.00%, 79.42%)</td>
<td>115 (5.18%, 80.94%)</td>
<td>105 (5.76%, 80.24%)</td>
<td>106 (5.82%, 79.80%)</td>
<td>97 (5.36%, 80.12%)</td>
</tr>
</tbody>
</table>
Table 4.3: Number of clusters (empirical type I error, empirical power) for simulation under general missing when \( m = 20 \), type I error = 5%, power = 80%.

<table>
<thead>
<tr>
<th>((\tau_1, \tau_2)) = (0.3, 0.1)</th>
<th>((0.01, 0.15, 0.005))</th>
<th>((0.01, 0.3, 0.005))</th>
<th>((0.05, 0.15, 0.005))</th>
<th>((0.05, 0.15, 0.025))</th>
<th>((0.05, 0.3, 0.005))</th>
<th>((0.05, 0.3, 0.025))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\sigma^2 = 1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S1</td>
<td>33 (6.38%, 81.64%)</td>
<td>28 (6.64%, 81.06%)</td>
<td>59 (5.16%, 79.56%)</td>
<td>46 (5.80%, 79.92%)</td>
<td>54 (5.92%, 80.62%)</td>
<td>41 (5.64%, 80.40%)</td>
</tr>
<tr>
<td>S2</td>
<td>39 (5.54%, 80.24%)</td>
<td>34 (6.22%, 80.92%)</td>
<td>67 (5.44%, 80.18%)</td>
<td>54 (6.26%, 80.64%)</td>
<td>62 (5.62%, 80.28%)</td>
<td>49 (5.62%, 80.70%)</td>
</tr>
<tr>
<td>S3</td>
<td>40 (6.00%, 80.92%)</td>
<td>34 (6.66%, 80.98%)</td>
<td>67 (5.62%, 80.24%)</td>
<td>54 (6.16%, 80.00%)</td>
<td>62 (5.62%, 81.26%)</td>
<td>49 (6.00%, 80.38%)</td>
</tr>
<tr>
<td>S4</td>
<td>40 (5.88%, 81.70%)</td>
<td>34 (5.74%, 80.58%)</td>
<td>67 (5.50%, 80.36%)</td>
<td>54 (5.44%, 80.04%)</td>
<td>62 (5.42%, 80.12%)</td>
<td>49 (5.84%, 79.86%)</td>
</tr>
<tr>
<td>S5</td>
<td>53 (5.96%, 79.88%)</td>
<td>47 (5.96%, 78.96%)</td>
<td>84 (5.62%, 80.02%)</td>
<td>70 (6.22%, 79.22%)</td>
<td>78 (5.66%, 79.38%)</td>
<td>65 (4.92%, 80.16%)</td>
</tr>
<tr>
<td>S6</td>
<td>63 (5.64%, 79.98%)</td>
<td>58 (5.24%, 79.16%)</td>
<td>83 (5.54%, 79.42%)</td>
<td>73 (5.90%, 78.64%)</td>
<td>81 (5.34%, 80.54%)</td>
<td>77 (5.54%, 79.92%)</td>
</tr>
<tr>
<td>(\sigma^2 = 0.75)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S1</td>
<td>25 (6.48%, 81.74%)</td>
<td>21 (6.80%, 81.90%)</td>
<td>45 (5.74%, 80.32%)</td>
<td>35 (5.36%, 80.10%)</td>
<td>41 (5.94%, 81.78%)</td>
<td>31 (6.26%, 82.00%)</td>
</tr>
<tr>
<td>S2</td>
<td>30 (6.24%, 82.64%)</td>
<td>26 (6.94%, 81.44%)</td>
<td>50 (5.76%, 79.50%)</td>
<td>41 (6.30%, 81.72%)</td>
<td>46 (6.56%, 78.96%)</td>
<td>37 (6.10%, 81.16%)</td>
</tr>
<tr>
<td>S3</td>
<td>30 (6.38%, 80.98%)</td>
<td>26 (6.52%, 81.40%)</td>
<td>51 (6.06%, 80.98%)</td>
<td>41 (5.92%, 81.52%)</td>
<td>47 (6.14%, 80.12%)</td>
<td>37 (6.42%, 80.66%)</td>
</tr>
<tr>
<td>S4</td>
<td>30 (5.84%, 81.26%)</td>
<td>26 (5.98%, 80.58%)</td>
<td>51 (5.58%, 79.16%)</td>
<td>41 (5.96%, 80.94%)</td>
<td>47 (6.32%, 79.82%)</td>
<td>37 (6.70%, 81.56%)</td>
</tr>
<tr>
<td>S5</td>
<td>40 (6.22%, 81.20%)</td>
<td>36 (6.30%, 81.30%)</td>
<td>63 (6.06%, 80.38%)</td>
<td>53 (6.14%, 80.40%)</td>
<td>59 (5.84%, 80.62%)</td>
<td>49 (6.24%, 79.40%)</td>
</tr>
<tr>
<td>S6</td>
<td>47 (6.04%, 79.72%)</td>
<td>43 (6.44%, 79.56%)</td>
<td>72 (5.42%, 80.70%)</td>
<td>62 (5.92%, 79.74%)</td>
<td>68 (5.36%, 79.52%)</td>
<td>58 (6.24%, 81.16%)</td>
</tr>
<tr>
<td>(\tau_1, \tau_2) = (0.5, 0.1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\sigma^2 = 1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S1</td>
<td>33 (6.38%, 81.64%)</td>
<td>28 (7.34%, 81.36%)</td>
<td>59 (6.04%, 80.22%)</td>
<td>46 (5.58%, 80.88%)</td>
<td>54 (5.86%, 79.94%)</td>
<td>41 (5.56%, 80.78%)</td>
</tr>
<tr>
<td>S2</td>
<td>40 (6.38%, 80.40%)</td>
<td>34 (6.30%, 79.48%)</td>
<td>68 (6.08%, 79.70%)</td>
<td>55 (5.94%, 79.62%)</td>
<td>63 (6.12%, 80.22%)</td>
<td>50 (5.36%, 80.96%)</td>
</tr>
<tr>
<td>S3</td>
<td>40 (6.04%, 80.04%)</td>
<td>34 (6.42%, 79.68%)</td>
<td>69 (5.50%, 79.82%)</td>
<td>55 (6.36%, 79.40%)</td>
<td>63 (6.20%, 80.44%)</td>
<td>50 (5.64%, 81.12%)</td>
</tr>
<tr>
<td>S4</td>
<td>40 (5.76%, 81.14%)</td>
<td>34 (5.54%, 80.32%)</td>
<td>69 (5.56%, 80.52%)</td>
<td>55 (5.64%, 81.18%)</td>
<td>63 (4.80%, 81.06%)</td>
<td>50 (5.78%, 80.40%)</td>
</tr>
<tr>
<td>S5</td>
<td>53 (6.24%, 80.42%)</td>
<td>48 (5.86%, 80.38%)</td>
<td>87 (6.04%, 79.58%)</td>
<td>74 (5.84%, 79.34%)</td>
<td>81 (6.12%, 80.16%)</td>
<td>68 (5.62%, 81.34%)</td>
</tr>
<tr>
<td>S6</td>
<td>64 (5.62%, 80.44%)</td>
<td>58 (5.70%, 80.42%)</td>
<td>102 (5.50%, 80.06%)</td>
<td>88 (5.66%, 80.12%)</td>
<td>95 (5.48%, 79.94%)</td>
<td>82 (5.82%, 79.78%)</td>
</tr>
<tr>
<td>(\sigma^2 = 0.75)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S1</td>
<td>25 (6.58%, 81.78%)</td>
<td>21 (7.68%, 82.24%)</td>
<td>45 (5.88%, 81.02%)</td>
<td>35 (5.72%, 80.12%)</td>
<td>41 (5.76%, 81.10%)</td>
<td>31 (6.24%, 81.70%)</td>
</tr>
<tr>
<td>S2</td>
<td>30 (6.12%, 81.50%)</td>
<td>26 (6.68%, 81.56%)</td>
<td>51 (5.68%, 79.62%)</td>
<td>41 (5.66%, 80.76%)</td>
<td>47 (5.72%, 80.14%)</td>
<td>37 (5.72%, 80.50%)</td>
</tr>
<tr>
<td>S3</td>
<td>30 (5.68%, 82.24%)</td>
<td>26 (6.72%, 81.64%)</td>
<td>51 (5.80%, 80.06%)</td>
<td>42 (6.34%, 80.20%)</td>
<td>47 (6.88%, 80.00%)</td>
<td>37 (6.06%, 80.44%)</td>
</tr>
<tr>
<td>S4</td>
<td>30 (6.24%, 80.98%)</td>
<td>26 (6.32%, 81.86%)</td>
<td>51 (6.14%, 79.54%)</td>
<td>42 (6.00%, 80.88%)</td>
<td>47 (5.96%, 80.00%)</td>
<td>37 (5.88%, 80.20%)</td>
</tr>
<tr>
<td>S5</td>
<td>40 (6.08%, 80.80%)</td>
<td>36 (5.56%, 80.90%)</td>
<td>65 (5.10%, 80.00%)</td>
<td>55 (6.10%, 80.30%)</td>
<td>61 (5.80%, 79.88%)</td>
<td>51 (5.90%, 80.40%)</td>
</tr>
<tr>
<td>S6</td>
<td>48 (6.08%, 80.40%)</td>
<td>43 (5.88%, 79.42%)</td>
<td>76 (5.54%, 80.22%)</td>
<td>66 (5.34%, 80.00%)</td>
<td>72 (5.72%, 80.64%)</td>
<td>62 (6.36%, 79.82%)</td>
</tr>
</tbody>
</table>
CHAPTER 5
CONCLUSIONS

In this dissertation, we discuss sample size calculations for three different study designs: (1) Stratified randomization design with binary outcomes; (2) Paired experimental design with continuous outcomes; (3) Paired cluster experimental design with both continuous and binary outcomes. Based on each proposed sample size method, we investigate how parameters in the sample size formula affect the required sample size and conduct simulation studies to evaluate the finite-sample performance of the proposed sample size method under various design configurations. And we also present how to use the proposed sample size methods in the real applications for all these three study designs.

5.1. Sample size considerations for stratified CRTs

In Chapter 2, we present a closed-form sample size formula based on Cochran-Mantel-Haenszel statistic, for stratified cluster randomization design with binary outcomes and varying cluster size. The proposed methodology can be used generally for stratified CRTs and multi-site CRTs, where the stratifying factor can be cluster size, geographic location, categorized levels of prognostic factors, or study site. The variability in cluster size is incorporated into sample size estimation by using the CV of cluster size, without specifying the distribution of cluster size. The closed-form formula provides insight into the impact of various design parameters (odds ratio, ICC, mean and CV of cluster size, etc.) on the sample size. We show that the proposed sample size is the same as that in the work of Donner [10] under equal cluster size, and theoretically derive the rela-
tive change in the number of clusters per group per stratum due to varying cluster size. The relative change in the number of clusters per group per stratum increases with ICC and CV increase. When ICC is positive (negative), the average cluster size approximation would underestimate (overestimate) the required number of clusters per group per stratum, whereas the minimum cluster size approximation would overestimate (underestimate) the required number of clusters per group per stratum; therefore, a good estimate of ICC is essential. The harmonic mean cluster size approximation has some uncertainty and would depend on the specific design configuration. We discuss the estimation of ICC for stratified CRTs with binary outcomes. The proposed sample size method is developed based on the asymptotic approximation, which might not be satisfactory when the number of clusters per group per stratum is small. Our simulation demonstrates that the nominal power is preserved when the number of clusters per group per stratum is as small as 12.

Missing data are often encountered in clinical trials due to reasons such as loss to follow-up, missing visits or patient dropout, and may attenuate statistical power if not appropriately adjusted. A crude adjustment of missing data in sample size estimation is to inflate the sample size determined under the assumption of no missing data by the expected missing proportion. Unfortunately, such adjustment may fail to incorporate the impact of missing data on sample size and power, which largely depends on the missing pattern and correlation within cluster. One interesting extension of the current work is to incorporate the missing data into sample size consideration. Clinical trials with multiple intervention arms are becoming increasingly popular, as an attractive way of optimizing resources and simultaneously testing various intervention strategies. Stratified CRTs with multiple arms can be more complex in their design and analysis. Complications of such trials are directly related to the number of intervention arms and the number of possible comparisons [2], in addition to stratification and correlation within cluster. It is our intention in future research to develop sample size methods for designing multiple-arm stratified CRTs. Finally, we have considered stratified cluster randomization design with binary outcomes. In future studies, we will extend the proposed method to stratified cluster
randomization design with categorical, count, and survival outcomes as well as longitudi-
dinal measurements. The aforementioned extensions require significant methodological
development and will be pursued in separate studies.

5.2. Sample size considerations for paired experimental design

In Chapter 3, we propose a closed-form sample size formula based on the GEE ap-
proach for paired experimental design with incomplete observations of continuous out-
comes. The formula is flexible to accommodate different missing patterns, magnitude of
missingness, and correlation parameter values. We show that the proposed GEE sample
size estimate is the same as that based on the paired $t$-test under complete observations.
In the presence of missing data, the proposed method would lead to a more accurate
sample size estimate comparing with the crude adjustment. The closed-form sample size
formula allows us to theoretically derive the condition under which the proposed method
is superior to the crude adjustment for missing data. Moreover, in the paired experimental
design, missing data can occur at either pre- or post-intervention. This is different from the
typical repeated measurement design, where baseline (pre-intervention) measurements
are observed in all subjects. The proposed sample size formula is very general and can
be applied to pre-post studies with dropout only at post-intervention by setting $p_1 = 1$ and
$p_{12} = p_2$.

Simulation studies demonstrate that the proposed sample size method preserves the
nominal levels of power and type I error under various design configurations and over a
wide range of sample sizes. The simulation also suggests that values of correlation and
missing probability parameters have a substantial influence on the sample size estima-
tion. In practice, when the information concerning correlation and missing probabilities
is absent, a sample size re-estimation in the middle of the study may be recommended
based on the observed values of correlation and missing probability parameters. We
consider scenarios of the negative within-subject correlation, independence, and positive within-subject correlation in simulation. Although examples of negative within-subject correlation coefficient $\rho$ are uncommon, Hanley et al. [15] discussed cases of negative $\rho$ which involved the birth weights of human twins and lung sizes of animal litter-mates. In these cases, with limited space or nutrition, nature allows considerable inequality among individual “competitors” therefore a negative within-twin or within-litter correlation.

The performance of the proposed GEE sample size method depends on the normal approximation, which might be unsatisfactory when the sample size is small. In this case, sample size estimation based on non-parametric exact methods could be considered and is a topic of our future research. We assume that the missing data arise from a MCAR mechanism in the development of the sample size method. When the missing data mechanism is informative, more sophisticated model-based methods are needed to account for the non-MCAR mechanism. Simulation studies designed for a specific study and missing mechanism could be used to determine the sample size. Moreover, the “regression-to-the-mean (RTM)” phenomenon affects pre-post, single-arm design where participants may be selected on the basis of an extreme, usually low but sometimes high, pre-test score [16]. Therefore, caution must be exercised when using pre-post, single-arm studies to inform sample size calculations for randomized clinical trials, because such studies may overestimate the treatment effects and lead to an underestimation of the sample size [54]. The issue of RTM can be addressed by introducing a randomized control group. Then, the RTM effect will affect both groups similarly and will be cancelled out when comparing groups. Another way of reducing the RTM effect is to undertake repeated multiple baseline measurements till a stable score is achieved so as to reduce the variability in the selection process. Nevertheless, this section proposes a general sample size method for paired experimental design with missing data, including matched case-control design, matched randomized trial design, and pre-post design. Finally, in this section we have considered paired experimental design with incomplete observations of continuous outcomes. In future studies, we will extend the proposed method to paired design with
incomplete observations of categorical, count and survival outcomes.

5.3. Sample size considerations for paired cluster experimental design

In Chapter 4, we propose closed-form sample size formulas for paired cluster design with both continuous outcomes and binary outcomes, based on the generalized estimating equation approach by treating incomplete observations as missing data in generalized linear models. The sample size formulas are flexible to accommodate different correlation structures, missing patterns and magnitude of missingness. In the presence of missing data, the proposed methods would lead to a more accurate sample size (i.e., number of clusters) estimation comparing with the crude adjustment. Simulation studies demonstrate that the proposed sample size method preserves the nominal levels of power under various design configurations. We use bias-corrected sandwich variance estimators to address the issue of inflated type I error rate when the number of clusters is small. The simulation also suggests the missing pattern have a substantial influence on the sample size estimation, especially when the missing proportion is high.

We assume the missing mechanism to be missing completely at random to derive the sample size formula. One interesting future work is to investigate whether the proposed GEE sample size formula could be applied under a weaker missing assumption, such as missing at random or missing not at random. Moreover, we have considered paired cluster experimental design with incomplete observations of continuous outcomes. In future studies, we will extend the proposed method to paired cluster experimental design with binary, categorical and count outcomes.
A.1. Eigenvalues of the correlation matrix $R_c$

Find the eigenvalues $\lambda$ for $R_c$ by $|R_c - \lambda I_{2m}| = 0$, where

$$R_c - \lambda I_{2m} = 
\begin{pmatrix}
(1 - \rho_1 - \lambda)I_m + \rho_1 J_m & (\rho_2 - \rho_3)I_m + \rho_3 J_m \\
(\rho_2 - \rho_3)I_m + \rho_3 J_m & (1 - \rho_1 - \lambda)I_m + \rho_1 J_m
\end{pmatrix}.$$ 

We know if $A$ and $B$ are square matrix, we have

$$|A B| = |A + B| |A - B|.$$ 

Therefore,

$$|R_c - \lambda I_{2m}| = |(1 - \rho_1 - \lambda + \rho_2 - \rho_3)I_m + (\rho_1 + \rho_3)J_m|$$

$$|I - \rho_1 - \lambda - \rho_2 + \rho_3)I_m + (\rho_1 - \rho_3)J_m|.$$ 

In Theorem 8.4.4 by Graybill [14], for the $k \times k$ exchangeable matrix $C = (a - b)I + bJ$, the determinant is given by $|C| = (a - b)^{k-1}[a + (k - 1)b]$. So we have

$$|R_c - \lambda I_{2m}| = (1 - \rho_1 - \lambda + \rho_2 - \rho_3)^{m-1} [1 - \lambda + \rho_2 + (m - 1)(\rho_1 + \rho_3)]$$

$$(1 - \rho_1 - \lambda - \rho_2 + \rho_3)^{m-1} [1 - \lambda - \rho_2 + (m - 1)(\rho_1 - \rho_3)].$$
Hence, The correlation matrix $R$, has four distinct eigenvalues,

$$\lambda_1 = 1 - \rho_1 + \rho_2 - \rho_3, \quad \lambda_2 = 1 + (m - 1)\rho_1 + \rho_2 + (m - 1)\rho_3,$$

$$\lambda_3 = 1 - \rho_1 - \rho_2 + \rho_3, \quad \lambda_4 = 1 + (m - 1)\rho_1 - \rho_2 - (m - 1)\rho_3.$$
A.2. The robust variance $\sigma^2_{2*}$ with binary outcomes (4.8)

Under MCAR, we separate $A_n(\hat{\beta})$ into two parts for the treatment $t_1$ and the treatment $t_2$ as

\[
A_n(\hat{\beta}) = \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{2} \sum_{k=1}^{m} \Delta_{ijk} \cdot p_{ijk}(\hat{\beta}) \left( 1 - p_{ijk}(\hat{\beta}) \right) \begin{pmatrix} 1 & t_{ijk} \\ t_{ijk} & t_{ijk}^2 \end{pmatrix}
\]

\[
= \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{2} \sum_{k=1}^{m} \Delta_{ijk} \cdot p_{ijk}(\hat{\beta}) \left( 1 - p_{ijk}(\hat{\beta}) \right) \begin{pmatrix} 1 & t_{ijk} \\ t_{ijk} & t_{ijk}^2 \end{pmatrix} + o_p(1)
\]

\[
= \frac{1}{n} \sum_{j=1}^{n} \sum_{k=1}^{m} \Delta_{1jk} \cdot p_{1jk}(1 - p_{1jk}) \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \frac{1}{n} \sum_{j=1}^{n} \sum_{k=1}^{m} \Delta_{2jk} \cdot p_{2jk}(1 - p_{2jk}) \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} + o_p(1).
\]

As $n \to \infty$, $A_n(\hat{\beta})$ converges to

\[
A = \begin{pmatrix}
    m \left[ s_1 P_1 Q_1 + s_2 P_2 Q_2 \right] & m s_2 P_2 Q_2 \\
    m s_2 P_2 Q_2 & m s_2 P_2 Q_2
\end{pmatrix},
\]

and

\[
A^{-1} = \frac{1}{m s_1 s_2 P_1 P_2 Q_1 Q_2} \begin{pmatrix}
    s_2 P_2 Q_2 & -s_2 P_2 Q_2 \\
    -s_2 P_2 Q_2 & s_1 P_1 Q_1 + s_2 P_2 Q_2
\end{pmatrix}.
\]

60
And then, $V_n(\hat{\beta})$ can be written as

$$V_n(\hat{\beta}) = \frac{1}{n} \sum_{j=1}^{n} \left( \sum_{i=1}^{2} \sum_{k=1}^{\frac{m}{2}} \Delta_{ijk} \hat{\epsilon}_{ijk} Z_{ijk} \right)^{\otimes 2}$$

$$= \frac{1}{n} \sum_{j=1}^{n} \left\{ \sum_{k=1}^{\frac{m}{2}} \left( \Delta_{1jk} \epsilon_{1jk} + \Delta_{2jk} \epsilon_{2jk} \right) \right\} + o_p(1)$$

$$= \frac{1}{n} \sum_{j=1}^{n} \sum_{k=1}^{\frac{m}{2}} \sum_{k'=1}^{\frac{m}{2}} \left( \begin{array}{c} (\Delta_{2jk} \epsilon_{2jk}) (\Delta_{2jk'} \epsilon_{2jk'}) \\ (\Delta_{1jk} \epsilon_{1jk}) (\Delta_{1jk'} \epsilon_{1jk'}) \\ (\Delta_{2jk} \epsilon_{2jk}) (\Delta_{1jk'} \epsilon_{1jk'}) \\ (\Delta_{1jk} \epsilon_{1jk}) (\Delta_{2jk'} \epsilon_{2jk'}) \\ (\Delta_{2jk} \epsilon_{2jk}) (\Delta_{2jk'} \epsilon_{2jk'}) \\ (\Delta_{1jk} \epsilon_{1jk}) (\Delta_{2jk'} \epsilon_{2jk'}) \end{array} \right) + o_p(1),$$

As $n \to \infty$, $V_n(\hat{\beta})$ converges to

$$V = \begin{pmatrix} C_1 + 2C_2 + C_3 & C_1 + C_2 \\ C_1 + C_2 & C_1 \end{pmatrix},$$

where

$$C_1 = ms_2 P_2 Q_2 + m(m - 1)(s_2^2 + \tau_1 s_2 t_2) \rho_1 P_2 Q_2,$$

$$C_2 = m(s_1 s_2 + \tau_2 \sqrt{s_1 s_2^2 t_1^2 t_2}) \rho_2 \sqrt{P_1 P_2 Q_1 Q_2} + m(m - 1)s_1 s_2 \rho_3 \sqrt{P_1 P_2 Q_1 Q_2},$$

$$C_3 = ms_1 P_1 Q_1 + m(m - 1)(s_1^2 + \tau_1 s_1 t_1) \rho_1 P_1 Q_1.$$
Therefore, the \((2, 2)\)th element of \(\Sigma^* = A^{-1} V A^{-1}\) is

\[
\sigma_{22}^2 = \frac{s_1 + (m - 1)(s_1^2 + \tau_1 s_1 t_1)\rho_1}{m s_1^2 P_1 Q_1} + \frac{s_2 + (m - 1)(s_2^2 + \tau_1 s_2 t_2)\rho_1}{m s_2^2 P_2 Q_2} - \frac{2\sqrt{P_1 P_2 Q_1 Q_2}}{m s_2^2 P_1 P_2 Q_1 Q_2} \left[ (s_1 s_2 + \tau_2 \sqrt{s_2^2 t_1 t_2})\rho_2 + (m - 1)s_1 s_2 \rho_3 \right].
\]


