



VCU

Virginia Commonwealth University
VCU Scholars Compass

Theses and Dissertations

Graduate School

2010

Power and Sample Size for Three-Level Cluster Designs

Tina Cunningham
Virginia Commonwealth University

Follow this and additional works at: <https://scholarscompass.vcu.edu/etd>



Part of the [Biostatistics Commons](#)

© The Author

Downloaded from

<https://scholarscompass.vcu.edu/etd/148>

This Dissertation is brought to you for free and open access by the Graduate School at VCU Scholars Compass. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of VCU Scholars Compass. For more information, please contact libcompass@vcu.edu.

© Tina Duong Cunningham

All Rights Reserved

Power and Sample Size for Three-Level Cluster Designs

**A dissertation submitted in partial fulfillment of the requirements for the degree
of Doctor of Philosophy at Virginia Commonwealth University**

By

Tina Duong Cunningham

Director

Robert E. Johnson

Associate Professor, Department of Biostatistics

Virginia Commonwealth University

Richmond, VA

October, 2010

Acknowledgement

The path to higher education is a painful but enjoyable one. As I was walking through it mile after mile during the last few years, I experienced hardship, disappointment, and frustration. Yet I also received tremendous love, trust, and encouragement. More than ever, I realized that I could not have reached this benchmark of my education without the help and support of many people.

I owe my deepest gratitude to my wonderful advisor, Dr. Robert E. Johnson. I am amazingly fortunate to have him as my supervisor and my research director. I would like to thank him for the hours he put into directing my work and the knowledge that he has taught me. His support helped me to overcome many difficulties and finish this dissertation on time. I thank him for inspiring in me the love of learning, the strength to never give up, and the desire to always strive for the better. He is my advisor, my colleague, and my true friend.

Special acknowledgement must go to my committee members. I would like to thank Dr. Viswanathan Ramakrishnan for his helpful suggestions and critical comments. I want to thank Dr. Roy Sabo, my next door “neighbor” who has always been there to give me some thoughts and ideas. I am grateful to Dr. D’arcy P. Mays for his insightful comments and constructive criticisms. I am indebted to Dr. Stephen F. Rothemich for his practical advice and encouragement.

My sincere thanks go to all the faculty, staff, and students in the VCU Biostatistics Department. I greatly appreciate their help and support during my time as a

graduate student in this department. More specifically, I would like to express my gratitude toward Dr. Best, Dr. Mukhopadhyay, Dr. Elswick, Dr. McClish, Dr. Thacker, and Mr. Boyle for their sound advice and input. Special thanks must be given to Ms. Yvonne Hargrove, our beloved office manager. She has always been there whenever I needed help.

Without the love and support from my family, I would not be able to finish this work. I am very grateful to my mother and my sister for their never-ending, unselfish love. They believe in me and trust that I will succeed. In our long-distance calls every weekend, they always tell me they are proud of me just for being who I am. I would like to thank my aunts who raised me up and taught me to hold myself to the high standard of working discipline and of being a good person overall. Although being half way around the world, my family has always been with me throughout this endeavor.

Last but not least, I am greatly indebted to my devoted husband, Glen Cunningham. He has been my source of constant love, support and strength during all these years. I thank him for encouraging me to go back to school and fulfill my learning desire. He helped me to pursue my dreams and to overcome many difficulties that I encountered. He gave me the confidence needed to succeed. His love and support without any complaint or regret have allowed me to finish this dissertation. Thank you, sweetheart, for all the love and patience that you give me during the past years and the years to come. This work is dedicated to you, indeed.

Vita

Tina Duong Cunningham was born on May 05, 1965. She worked for Zuellig Pharma, a pharmaceutical company in Vietnam. Upon imigrating to the US, she returned to school and received an Associate of Science degree from Richard Bland College in 2003. She graduated from The College of William and Mary in 2005. She entered the Ph.D. program in Biostatistics at Virginia Commonwealth University in the fall of 2006.

Table of Contents

List of Tables	xii
List of Figures	xiv
Abstract	xv
Chapter 1 Introduction	1
Chapter 2 Background and Significance	9
2.1 Review of Statistical Issues in Cluster Randomized Trials	10
2.1.1 Issues in Experimental Designs	10
2.1.2 Issues in the Selection of Appropriate Analysis Models	12
2.2 Review of Issues in Sample Sizes Determination.....	15
2.2.1 The Impact of ICC on Sample Size	17
2.2.2 Issues Involving the Number of Clusters and Cluster Size.....	18
2.2.3 Existing Methods for Computing Sample Sizes	20
2.3 Review of Statistical Software Computing Sample Size in Correlated Data.....	22
2.4 The Significance of Sample Size Determination in Three-Level Designs	24
2.4.1 The Gap in Study Designs with Three-Level Structure	25
2.4.2 The Gap in Methods Related to Computing Sample Size	26
2.4.3 The Gap in Analysis Software and Programs.....	27
2.4.4 Goals and Motivation	28

Chapter 3	Statistical Relevant Concepts	30
3.1	Statistical Power and Sample Size Assessment	31
3.1.1	Hypothesis Testing	31
3.1.2	Statistical Power	33
3.1.3	Sample Size Calculation in General	34
3.2	Intraclass Correlation	36
3.2.1	The Effect of Clustering in Two-Level Designs	37
3.2.2	The Effect of Clustering in Three-Level designs	40
3.3	Computing Sample Size in Two-Level Cluster Designs	41
3.3.1	Power in a Simple Clinical Trial	42
3.3.2	Power in a Two-Level Cluster Randomized Trial	43
3.3.3	Power and the Effect of Randomization	44
3.4	Generalized Linear Mixed Models	45
3.4.1	Linear Mixed Models	46
3.4.2	Generalized Linear Models (GLM)	48
3.4.3	Generalized Linear Mixed Models (GLMM)	50
Chapter 4	Power and Sample Size for Continuous Outcomes	52
4.1	Linear Mixed Model Approach	52
4.2	Randomize at Third Level	55
4.2.1	Estimate $\text{Var}(\hat{\beta})$	55
4.2.2	Derivation of $\mathbf{1}^T \mathbf{R}^{-1} \mathbf{1}$	57

4.3	Randomize at Second Level without Interaction Effect	59
4.3.1	Estimate $\mathbf{Var}(\hat{\boldsymbol{\beta}})$	59
4.3.2	Derivation of $\mathbf{X}_i^T \mathbf{R}^{-1} \mathbf{X}_i$	60
4.4	Randomize at Second Level with Interaction Effect	62
4.4.1	Estimate $\mathbf{Var}(\hat{\boldsymbol{\beta}})$	62
4.4.2	The structure of \mathbf{V}	63
4.5	Randomize at First Level without Interaction Effect	64
4.5.1	Estimate $\mathbf{Var}(\hat{\boldsymbol{\beta}})$	64
4.5.2	Derivation of $\mathbf{X}_i^T \mathbf{R}^{-1} \mathbf{X}_i$	65
4.6	Randomize at First Level with Interaction Effect	67
4.6.1	Estimate $\mathbf{Var}(\hat{\boldsymbol{\beta}})$	67
4.6.2	The Structure of \mathbf{V} with Treatment \times Level Three Interaction	68
4.6.3	The Structure of \mathbf{V} with Treatment \times Level Two Interaction.....	69
Chapter 5	Power and Sample Size for Binary Outcomes.....	70
5.1	Generalized Linear Mixed Models (GLMM) Approach	71
5.1.1	Basic Model.....	71
5.1.2	Pseudo-likelihood Method	72
5.2	Randomize at Third Level	76
5.3	Randomize at Second Level without Interaction Effect	78

5.4	Randomize at Second Level with Interaction Effect.....	79
5.5	Randomize at First Level without Interaction Effect.....	81
5.6	Randomize at First Level with Interaction Effect.....	83
5.6.1	The Structure of V_{zi} with Treatment \times Level Three Interaction.....	83
5.6.2	The Structure of V_{zi} with Treatment \times Level Two Interaction.....	84
Chapter 6	Simulation Study.....	86
6.1	Simulation Design.....	86
6.2	Simulation for Continuous Data.....	88
6.3	Simulation for Binary Data.....	96
6.4	Simulation Results.....	103
6.4.1	Simulation Results for Continuous Outcome	103
6.4.2	Simulation Results for Binary Outcome.....	104
Chapter 7	Application.....	106
7.1	The User-Interface Program.....	106
7.2	Application Example for Continuous Data.....	109
7.3	Application Example for Binary Data.....	115
7.4	Comments on the Application Examples.....	118
Chapter 8	Discussion and Future Work.....	121
8.1	Summary of Work.....	121
8.2	Discussion.....	123

8.3	Future work.....	128
8.4	Concluding Remarks.....	130
References	132
Appendix A	SAS Program User Manual.....	138
Appendix B	Simulation Results.....	144

List of Tables

Table 3.1	Probabilities Associated with a Statistical Test	33
Table 6.4.1	Summary of Simulation Results for Continuous Outcome	103
Table 6.4.2	Summary of Simulation Results for Binary Outcome	104
Table 7.1	Some Sample Size Combinations for the YouthMood Project Trial Randomizing at Level One without Interaction.....	113
Table 7.2	Some Sample Size Combinations for the YouthMood Project Trial Randomizing at Level Two with Interaction.....	115
Table 7.3	Some Sample Size Combinations for the Dutch Helping Hands Trial Randomization at Level Three.....	118
Table B.1	Simulation Results for Case 1 Randomize at Level Three—Continuous Outcome.....	143
Table B.2	Simulation Results for Case 2 Randomize at Level Two—Continuous Outcome.....	144
Table B.3	Simulation Results for Case 3 Randomize at Level Two with Interaction—Continuous Outcome.....	145
Table B.4	Simulation Results for Case 4 Randomize at Level One without Interaction—Continuous Outcome.....	146
Table B.5	Simulation Results for Case 5 Randomize at Level One with Treatment \times Level 3—Continuous Outcome.....	147
Table B.6	Simulation Results for Case 6 Randomize at Level One with Treatment \times Level 2—Continuous Outcome.....	148
Table B.7	Simulation Results for Case 7 Randomize at Level Three—Binary Outcome.....	149
Table B.8	Simulation Results for Case 8	

	Randomize at Level Two—Binary Outcome.....	150
Table B.9	Simulation Results for Case 9 Randomize at Level Two with Interaction—Binary Outcome.....	151
Table B.10	Simulation Results for Case 10 Randomize at Level One without Interaction—Binary Outcome.....	152
Table B.11	Simulation Results for Case 11 Randomize at Level One with Treatment \times Level 3—Binary Outcome.....	153
Table B.12	Simulation Results for Case 12 Randomize at Level One with Treatment \times Level 2—Binary Outcome.....	154

List of Figures

Figure 6.2.1	Simulation Algorithm Randomize at Level Three—Continuous Outcome.....	90
Figure 6.2.2	Simulation Algorithm Randomize at Level Two—Continuous Outcome.....	91
Figure 6.2.3	Simulation Algorithm Randomize at Level Two with Interaction—Continuous Outcome.....	92
Figure 6.2.4	Simulation Algorithm Randomize at Level One without Interaction—Continuous Outcome.....	93
Figure 6.2.5	Simulation Algorithm Randomize at Level One with Treatment \times Level Three —Continuous Outcome.....	94
Figure 6.2.6	Simulation Algorithm Randomize at Level One with Treatment \times Level Two —Continuous Outcome.....	95
Figure 6.3.1	Simulation Algorithm Randomize at Level Three—Binary Outcome.....	97
Figure 6.3.2	Simulation Algorithm Randomize at Level Two—Binary Outcome.....	98
Figure 6.3.3	Simulation Algorithm Randomize at Level Two with Interaction—Binary Outcome.....	99
Figure 6.3.4	Simulation Algorithm Randomize at Level One without Interaction—Binary Outcome.....	100
Figure 6.3.5	Simulation Algorithm Randomize at Level One with Treatment \times Level Three—Binary Outcome....	101
Figure 6.3.6	Simulation Algorithm Randomize at Level One with Treatment \times Level Two—Binary Outcome.....	102
Figure 7.1	Algorithm for the User-Interface Program	108

Abstract

Over the past few decades, Cluster Randomized Trials (CRT) have become a design of choice in many research areas. One of the most critical issues in planning a CRT is to ensure that the study design is sensitive enough to capture the intervention effect. The assessment of power and sample size in such studies is often faced with many challenges due to several methodological difficulties.

While studies on power and sample size for cluster designs with one and two levels are abundant, the evaluation of required sample size for three-level designs has been generally overlooked. First, the nesting effect introduces more than one intraclass correlation into the model. Second, the variance structure of the estimated treatment difference is more complicated. Third, sample size results required for several levels are needed.

In this work, we developed sample size and power formulas for the three-level data structures based on the generalized linear mixed model approach. We derived explicit and general power and sample size equations for detecting a hypothesized effect on continuous Gaussian outcomes and binary outcomes. To confirm the accuracy of the formulas, we conducted several simulation studies and compared the results. To establish a connection between the theoretical formulas and their applications, we developed a SAS user-interface macro that allowed the researchers to estimate sample size for a three-level design for different scenarios. These scenarios depend on which randomization level is assigned and whether or not there is an interaction effect.

Chapter I

Introduction

Over the last few decades, randomized controlled trials have become the design of choice in medical sciences and disease prevention research for the evaluation of new medical interventions. In such trials, an important issue in the design stage is the unit of randomization. This unit could be the individual patients, the physicians, a group of physicians within a practice, or several practices within a specific geographic area.

In some situations, randomization at the individual level has advantages because it minimizes the risk of covariate imbalances. For example, different practices tend to have different types of patients with their own characteristics. Furthermore, trials that randomize individual patients can use practices as blocks to reduce the dependency of the outcomes on practice level discrepancies, such as the degree of adherence to a study protocol or the difference in clinical skills found amongst the providers.

In other situations, however, the interventions are randomized not to individuals but to intact groups or clusters, such as medical practices, worksites, families, or communities. These studies, referred to as Cluster Randomized Trials (CRT), are being used increasingly because of many reasons. First, randomizing at cluster level might be the only feasible method of conducting a trial in certain fields. Examples include studies where interventions are targeted at a whole practice (the implementation of a new Electronic Health Record system) or a whole community (introduction of new water and sanitation schemes). Second, randomizing by cluster can be used to reduce the likelihood of contamination, which is likely to happen when

individuals receive an intervention within the same group share information with each other. For example, patients in an intervention group who are being educated on colon cancer screenings might talk to the control-allocated patients visiting the same practice, who in turn will ask for the screening themselves or might respond differently on the outcome measures due to this influence. Third, CRT is justified when the efficacy is established at individual level, but the primary goal is to measure the effectiveness when an intervention is applied at the cluster level, or when the desire is to capture the mass effect of an intervention on a large proportion of group members (Murray, 1998).

One of the most critical issues in designing a cluster randomized experiment is to ensure that the study design is sensitive enough to capture the effect of the intervention. This important task involves making a decision about power and sample sizes. Good study design requires the planning of sample sizes so that the test for the intervention effect has adequate statistical power to detect the smallest difference that is of scientific or practical interest. Insufficient sample sizes can lead to inadequate sensitivity, whereas excessive sample sizes can be a waste of time and valuable resources.

In the design stage of a cluster randomized study, researchers are required to answer three fundamental questions: (1) What are the minimum numbers of subjects in each cluster and the number of clusters that are required in order to attain a specified power for testing the treatment effect? (2) Is it better to investigate more clusters with fewer subjects in each cluster, or fewer clusters with more subjects in each cluster? (3) If there is a budget constraint, what is the optimum allocation of the sample sizes which will still allow for the desired power? To date, there are no definite guidelines that simultaneously answer the above questions.

The impact of randomization by cluster on the required sample sizes can be quite substantial. Campbell et al. (2004) provided a simple example to illustrate this impact. In a study to evaluate the effectiveness of educational training on asthma management for general practitioners, the intention of the design was to detect an increase in appropriate management of asthma patients from 40% to 60%, at a statistical power of 80% and significance level of 5%. From the researchers' calculation, an individual-based sample size formula yielded a total sample size of 194 patients. However, when appropriate adjustment for clustering by general practice was taken into account, even with a small clustering effect, the true sample size requirement was 400. This sample size is more than twice the sample size computed under the assumption of no clustering effect, a significant difference that cannot be ignored.

The assessment of power and sample size in studies involving clustering is faced with many challenges. First, the cluster effect must be taken into account. While standard sample size calculations operate under the assumption of independence between individuals, this assumption is violated in CRT. Second, in cluster studies, there is more than one component of variation: the variation among individuals within clusters, and the variation in the outcomes between clusters. It is essential that both sources of variation are taken into account in the design stage. Third, planning for a CRT involves more than just one sample size. The number of data units at each level required to reach a target level of statistical power must be determined. Furthermore, each sample size at a different level will affect the power differently. For example, in the study evaluating the appropriate management of asthma patients, the power of the test used to detect the intervention effect depends not only on the sample size of the patients within each practice, but also on the number of practices involved as well as the ratio of number of patients to number of practices.

Complicating matters further, the difficulty in computation of power in CRT arises from the complexity of the models. Much of the work on power and sample size requirements for cluster data is developed from the idea of multistage sampling, which creates a hierarchically nested structure with sample from one level nested within sample at another level. However, a sample taken from m clusters each of size n is not a simple random sample of mn individuals, and the sampling distribution of statistics on such clustered samples is not the same as that based on simple random samples of the same size.

Different nesting structures can be seen in practice. One common structure is the two-level design in which individual subjects are nested within clusters. For example, in trials where different interventions to improve quality of care are compared, the intervention is implemented at the level of healthcare professionals (e.g., clinicians, physicians, caregivers), while the effects are measured at the patient level. In such trials, patients are nested within healthcare professionals, creating the two-level nesting effects.

During the last two decades, the design and analysis techniques for CRT with two levels have been fairly well-developed in the literature (Donner and Klar, 2000; Murray, 1998). In addition, statistical software for the two-level design is currently available; some are free to the public (Hedges and Hedberg, 2007; Raudenbush & Liu, 2000; Snijders and Bosker, 1993). However, most work on power analysis and sample size determination for CRT has only emphasized two-level designs, and some only handled one particular type of outcome variable.

As a more recent development, randomized trials with three-level designs have become a common application in many different research areas, such as psychology, education, and medical sciences. Examples of three-level designs are abundant in the literature. Three of them will be listed here for illustration purpose.

In the first example, a cluster randomized controlled trial was designed to evaluate a school-based prevention program on tobacco and drug use. In this study, 170 schools from 9 centers from seven countries (Austria, Belgium, Germany, Greece, Italy, Spain, Sweden) were randomized to one of three arms of an intervention (basic curriculum, basic with peer involvement, and basic with parent involvement) or to a control group. Twenty seven schools dropped out after the allocation to the study arm, leaving 143 schools and 345 classes actually included in the study. The number of schools varied between centers, as did the number of classes. Of 7409 eligible students, 7079 (95.7%) participated in the baseline survey. The evaluation is based on a comprehensive social influence approach, and was delivered during the school year 2004–2005 to a population of 12 to 14-year-old students attending junior high school. An anonymous questionnaire administered before and after the intervention was used to track behavioral and attitudinal changes. This cluster randomized trial assumed a three level structure in that students, classes, and school served as the first, second, and third level in the hierarchy, respectively. Randomization took place at the third level (schools) (Faggiano et al., 2007).

In the second example, researchers in Spain conducted a cluster randomized study to assess the effectiveness of a new “Experimental Program for Physical Activity Promotion” (PEPAF) in increasing physical activity in inactive patients. PEPAF is a program designed to increase physical activity in patients who did not meet the recommended aerobic physical activity levels. Physicians who participated in PEPAF provided patients with advice on using health promotion websites and health educational materials. In addition, a 15 minute consultation on individualized physical activity plan was offered to patients who committed to increase their activity level. Control group physicians delivered standard care and delayed any new systematic

intervention related to physical activity until the end of the study, unless the reason for consultation or the patients' health problems were directly related to inactivity. Recruitment involved inviting all 15 research groups of the Health Promotion Primary Care Research Network to participate, with a collaboration of at least 4 physicians per center as the requirement for eligibility. Seventy family physicians from 13 primary care centers belonging to 8 research groups agreed to participate. After signing a collaboration consent form, all 70 physicians were randomized to either the PEPAF or usual care (control) arm of the trial in a 1:1 ratio. Twelve physicians dropped out before the start of the study because of technical complaints, and 2 physicians failed to participate. Finally, 56 physicians (29 allocated to the PEPAF arm and 27 to the control arm) performed the study at 13 primary care centers. Each family physician recruited 150 patients aged 20 to 80 years, who did not meet the recommended aerobic physical activity level. Physicians assessed the patients' physical activity with assistance of a computerized algorithm. The study was managed online using Web-based software designed to help physicians follow the research protocol and control the recruitment process of each eligible patient. This study is an example of a three-level design in which patients (level one) are nested within physicians (level two), and physicians are nested within centers (level three). Randomization took place at the second level (physicians) (Grandes et al., 2009).

In the third example, a randomized, controlled study was conducted by the Virginia Ambulatory Care Outcomes Research Network to compare the effect of an interactive web-based personalized healthcare record and the usual delivery of preventive services. Eight primary care practices in Northern Virginia were recruited. Practice size ranged from 2 to 35 clinicians. Of the 80,000 active study site patients, 4,500 active patients were randomly selected for study participation. The study sample was then randomly assigned to intervention (n=2,225) and

control (n=2,225) groups. Patients in the intervention group received up to three mailed requests from their clinician inviting them to use myPreventiveCare, a web-based personal health record that provides patients personalized prevention plans and individualized educational material about preventive services and chronic disease management. Control patients received “usual” preventive care and were not informed of myPreventiveCare. Outcome measurements included the percentage of intervention patients who visited myPreventiveCare, percentage of patients who were up-to-date with indicated preventive service, and an aggregated percentage of preventive services that were up to date. This study can be considered an example of a three-level design in which patients nested within physicians, and physicians nested within practices. Randomization took place at the first level (patients) (Krist et al., 2010).

The three examples above show that analyses for nested cluster data do not always consist of two levels. It is not uncommon that a three-level design is encountered in practice. Hence, it is important that the three-level structure is taken into account right at the design stage of the experiment.

In this work, we develop sample size formulas for the three-level data structures based on the generalized linear mixed model approach. We derive explicit and general power functions and sample size equations for detecting a hypothesized effect on continuous Gaussian outcomes and binary outcomes. In addition, we present a SAS macro that will allow the users to estimate sample size for three-level data structures in different scenarios depending on which level randomization is assigned and whether the interaction exists.

We begin by providing some background and significance of this work in Chapter 2. Chapter 3 reviews basic statistical properties of cluster data and the general concepts behind power and sample size computation. Chapter 4 derives the formulas for sample size and power

for continuous Gaussian outcome. Chapter 5 derives the formulas for sample size and power for binary outcome. Chapter 6 presents a simulation study to verify the accuracy of the proposed formulas. Chapter 7 explains the development of a SAS macro to compute power and sample size, together with some practical examples to illustrate the use of the macro. Finally, Chapter 8 includes a discussion of the current and future work.

Chapter 2

Background and Significance

Over the past three decades, the application of cluster randomized trials has led to a large body of methodological work and a growing literature that cuts across several areas of research. It has been well known that standard methods for controlled trials randomized at individual level cannot be applied directly to trials randomized at the cluster level. Thus, increased attention has been developed toward the design and analysis of CRT because of their special statistical characteristics.

Despite the large amount of literature and discussions dedicated to the topic of CRT, the selection of proper statistical approaches to determine sample size still remains a challenge for researchers in multiple disciplines. The literature on sample size computation for two-level data can be found consistently during the last two decades, and some publications on sample size for three-level design also appeared recently. It is important to take a general look at the previous work and evaluate the impact of recent developments in this topic.

This chapter provides a detailed review of several main issues in the development of the designs and analyses of CRT in general, and of the sample size estimation for multilevel data in particular. The chapter is structured as follows: Section 2.1 gives a general overview of the established literature in CRT. Section 2.2 presents issues concerning sample size determination and power assessment. Section 2.3 reviews statistical software that compute sample sizes in CRT. Finally, Section 2.4 discusses the significance and motivation behind this work.

2.1 Review of Statistical Issues in Cluster Randomized Trials

Although the concept of a cluster randomized study dates decades ago when Kish (1965) introduced the technique of cluster-sampling design in survey sampling theory, increased interest in its statistical features was brought to the general research community by Cornfield (1978). In his work, Cornfield pointed out that two primary concerns needed to be addressed in experiments where groups (or clusters) of subjects are randomized. First, the variance between clusters is the dominant factor in the variance of cluster means. Second, there will be fewer degrees of freedom to estimate the between cluster variance than to estimate the within cluster variance.

Over the years, issues involved in the design and analyses of clustered data have been discussed extensively by several researchers; examples include Klar and Donner (2001), Murray et al. (2008), and Dedrick et al. (2009). Based on the review of methodological and technical literature, three broad topics set the stage for discussion in the design and analysis of clustered data: (1) Issues in the experimental designs, (2) Issues in the selection of appropriate analytic models, and (3) Issues in the assessment of power and sample size determination.

We will discuss the first two items in this section. The last item, the review on power and sample size estimation, will be explored separately in the next two sections because it is the focus of this work.

2.1.1 Issues in Experimental Designs

A number of experimental designs have been specifically proposed for cluster randomized trials. The three most popular designs are matched-pair, stratified, and completely randomized trials. The last form of design is usually considered with no pre-stratification or matching on the baseline characteristics.

Freedman et al. (1990) explored the gain in efficiency obtained from the matched-pair design by estimating the relative efficiency of an unmatched versus a pair-matched design. Defining the relative efficiency as 1 minus the ratio between the variance of matched design over the variance of unmatched design, the authors showed that the matched design can be more efficient compared to unmatched design when high correlation between the pairs was created from the effect of matching.

One commonly adopted matching factor is cluster size where clusters are grouped into categories such as small, medium, or large. As Donner and Klar (2000) suggested, this method is appealing because it helps to avoid the imbalance in the number of subjects per treatment group. In addition, the cluster size itself might associate with factors associated with other baseline variables, such as socioeconomic status or access to healthcare resources.

The stratified design can be viewed as an extension of matched-pair design in which more than one cluster is randomly allocated within strata to each of the treatment arms. This design is preferred over the matched-pair design in situations when it is difficult to create close matches that correspond to important estimates. In addition, stratified design is more appealing because its allocation scheme is less rigid since it reduces many of the analytic limitations associated with matched-pair design ((Donner and Klar, 2000). In general, the choice between matching and stratification depends on the number of clusters, the accuracy of the matches, and the analytic plan (Murray, 1998).

A completely randomized design with no pre-stratification or matching is perhaps the most commonly adopted design in practice. In any case, for all forms of designs the issue arises in the choice of which level randomization is executed. Zucker (1990) discussed the difference in the context of two-level design, in which subjects are nested within clusters. The treatments can

be assigned in two different ways. In the first scheme, the assignment of the treatments is done on the individual basis within each cluster. In the literature, this is known as subject-level randomization. The second option is to randomly assign the clusters to different treatment arms. This is known as cluster-level randomization where the subjects are nested in clusters, which in turn are nested in the treatment groups.

2.1.2 Issues in the Selection of Appropriate Analysis Models

To date, there is no single agree-upon method to analyze data from CRT. According to Murray (1998), one major source of errors in the analysis of clustered data is model misspecification. The selection of proper statistical approaches in clustered design is difficult because the usual assumptions under familiar methods are not met. Methods involving general linear models or generalized linear models are inappropriate for the analysis of clustered data because they only allow for one source of random variation, usually by ignoring the cluster effect, or treating the cluster as a fixed effect in the model. Zucker's (1990) pointed out the pitfall in implementing such strategies by comparing what he defined as the fixed effect model (cluster is treated as fixed effect) and the mixed effect model (cluster is treated as random effect). By looking at the F-test, Zucker concluded that ignoring the cluster effect can inflate the Type I error and the level of inflation can be quite substantial when one mistakenly treats the clusters as a fixed effect.

One strategy that has been recommended for clustered data is the use of a multilevel model, also known as a hierarchical linear model, or random coefficient model. This type of model is common in educational and social research, originally applied for longitudinal data and growth data. In the last decade, this technique has been used in the context of CRT because of

the hierarchical structure of the data. In essence, multilevel modeling provides for components of variance for cluster and individual slopes as well as intercepts. These models can be specified by using equations for each level or combining equations at each level into one. An introduction to multilevel modeling was written by Raudenbush (1993). The theory and implementation of multilevel models were discussed in depth by Hox (2002).

The strength of the multilevel model lies on its ability to incorporate both fixed and random effects while allowing for unbalanced data and flexibility in handling missing data (Raudenbush, 1993). However, some issues still remain controversial in the application of multilevel models. For example, one issue involves the selection of predictors. As Dedrick et al. (2009) pointed out in their review, variable selection in multilevel models can be complicated and can easily become a source of error. Predictors can be selected for each level of the model, and interactions between predictors can be included at either one single level or across all levels. Another issue refers to the requirement of centering and the methods that accomplish this. One approach is to center around the grand mean of the predictor variable, another approach is to center around the cluster mean, or around a theoretically meaningful value. While the use of centering was not explicitly stated in many research studies, different methods have different implications for interpreting the parameter estimates (Raudenbush and Bryk, 2002).

Another technique to analyze CRT data involves the use of permutation tests. This method was first discussed by Gail et al. (1996) and was reviewed by Murray (1998). In this approach, the distribution of all possible allocations of the clusters was examined, and the intervention effect for each allocation is estimated under the null hypothesis. The observed intervention effect is regarded as just one possible intervention effect among many, and the probability of getting a more extreme result is the proportion of possible intervention effects that

are greater than the observed effect. The benefit of a permutation test is that it does not rely on statistical models for its validity. In their simulation work, Gail et al. reported that permutation tests had satisfactory type I and type II errors. However, the disadvantage of this method arises when covariates were included in the model. In such situations, this technique requires as many assumptions as other model-based methods (Murray, 1998).

Of all the statistical models being used to analyze cluster data, the most common one is the mixed model ANOVA/ANCOVA approach. This method is popular not only because of its computational simplicity, but also because of its ability to facilitate the estimations of models with fixed effects, random effects, or both. Of course, both mixed model ANOVA and ANCOVA are special cases of the generalized linear mixed models, which include random effects, random coefficients, and covariance patterns models. Murray (1998) illustrated the implementation of mixed model ANOVA/ANCOVA to analyze data collected from post-test only CRT, pre-test-post-test CRT, and CRT under both cross-sectional designs and cohort designs. In the cases of mixed model ANCOVA, regression adjustment for covariates was used to reduce bias and improve the precision. In their simulations, Murray and Wolfinger showed that the mixed model ANOVA/ANCOVA with one or two time intervals has a nominal type I error rate. They concluded that when properly executed, mixed model ANOVA/ANCOVA can provide a valid analysis for CRT (Murray and Wolfinger, 1994).

One should keep it in mind that no matter what statistical model is used to analyze the data, designs with more than one level will call for more complex variance structures. The questions are how to best specify the covariance structure for the model, is there an interaction effect, and which level varies randomly and which level is fixed. Dedrick et al. (2009) classified the covariance parameters into three groups. The first group includes those that are assumed to

be zero and can be ignored. The second group includes those that are not zero and are supposed to be estimated. The third group includes those for which the researchers are unsure. When many in-doubt variables are omitted, researchers are faced with the risk of biased results and incorrect conclusions. When too many questionable variance parameters are included, the model can become too complicated to estimate and estimation algorithms might not converge.

2.2 Review of Issues in Sample Sizes Determination

Although there is a large and growing literature on cluster randomized trials, the amount of publications on power and sample size is found to be much less than the literature that focused on model development and parameters interpretation. Let's explore the reasons behind the lack of discussion on power and sample size assessment for cluster data.

As Donner and Klar pointed out in their book (Donner and Klar, 2000), one obvious reason that computing of sample sizes is more complicated in CRT is because the total sample size for each level must be determined. In a simple randomized trial, power is a monotonic function of the sample size when other factors are constant. This is not the case for clustered data. For example, in situations with a large intraclass correlation coefficient, increasing the number of subjects in each cluster or the total number of subjects up to a certain point may not have a significant impact on power. In fact, several authors have discussed methods of sample sizes and power optimization. In these methods, usually a range of sample sizes at each level are recommended along with their corresponding estimated powers (Snijders and Boker, 1993; Raudenbush, 1997).

Another reason why much less research has been devoted toward sample size determination is because the cluster design must be taken into account. In cluster trials, the usual

standard sample size assumption is violated because data on subjects within the same cluster are no longer independent. This level of dependency is measured and explained via the concept of intraclass correlation coefficients (ICC), the ratio of the between cluster variability to the total variability. However, computing the ICC is problematic due to the difficulty in estimating the variance. Estimates of ICC values are rarely available, and not all published records of the ICC are reliable (Hedges and Hedberg, 2007). This is understandable, however, because the magnitude of the ICC depends largely on the study design and outcomes, the intervention, and the covariate adjustment (Guittet et al., 2005).

A third reason why there is much less discussion on sample size and power analysis is a misconception about the sample sizes itself, according to Donner and Klar (2000). Trials that enroll a large number of subjects usually give an impression of sufficient power, when in fact it may not be the case when the cluster effect is taken into consideration. For example, with a large ICC value the information drawn from any given individual in the cluster is redundant given the information available from other members in the same cluster. Thus, only adding more clusters will have an effect on power in this situation (Donner and Klar, 2000).

Despite the difficulties in estimating statistical power and the required sample sizes in cluster studies, in recent years there has been considerable effort devoted to this topic. The resulting literature is seen in many textbooks and journals. Some main topics can be identified, including (1) the importance of the ICC and how to estimate its value, (2) issues involving the number of clusters and cluster size, and (3) analytical methods that provide formulas to compute the sample sizes in different situations.

2.2.1 The Impact of ICC on Sample Size

It is established that the ICC has a direct relationship to the estimation of sample size and power analysis in a CRT. A number of authors have discussed this relationship (Hsieh, 1988; Murray, 1998; Donner and Klar, 2000). In a study on a complete cluster randomized design with normally distributed continuous outcome, Guittet et al. (2005) quantified the influence of the ICC on the power by examining power contour graphs under different values of the ICC. The authors concluded that underestimating the ICC can seriously under-power the trials. Together with lower number of clusters, the desired power might not be achieved. In another study, Maas and Hox (2005) explored the effect of the ICC on sufficient sample size and accurate estimation of parameters. The authors advocated that the estimated parameters and standard errors are biased downward when the cluster sizes are small (less than 30) and the ICC values are large (ranging from 0.10 to 0.30). These findings confirm the impact of ICC on sufficient sample size and parameters estimation, given the fact that most CRT used less than 30 clusters.

Since knowledge of the ICC is essential for power and sample size determination in planning for a CRT, several researchers have attempted to address the issue of obtaining reasonable values for the ICC in realistic situations. One way to obtain this information is to use the ICC estimated by previous studies. For example, Murray et al. (2004) reported a summary of ICC values collected from 14 articles that provided information on the ICC for health-related outcomes. In educational research, a compilation of ICC values of academic achievement and related covariate effects was provided by Hedges, together with an illustration of how to use these values to compute sample sizes under hierarchical models (Hedges and Hedberg, 2007).

A reliable estimate of the ICC is certainly necessary to ensure robust sample size calculations. Several researchers dealt with the difficulty in getting accurate estimates of the

ICC, including methods to compute confidence intervals. One approach was presented by Feng and Grizzle, using the bootstrap method by simulating results from studies with the sample sizes that gave the observed estimates. The ICC computed from each simulation is substituted into the sample size formula and the distribution of powers is provided (Feng and Grizzle, 1992). In another paper, Turner et al., developed methods that allow for the uncertainty in previously obtained ICC under the use of prior distribution of the ICC in a Bayesian approach. The authors noticed that uncertainty in the ICC will lead to some inaccuracy in the power of the study. However, the risk of low power is low in a design based on a large number of clusters (Turner et al., 2004).

2.2.2 Issues Involving the Number of Clusters and Cluster Size

In multi-level design studies, sample sizes for more than one level need to be addressed. As Hox discussed in his book, the maximum likelihood methods used commonly in most multilevel data analysis rely on an asymptotic assumption, which imply a sufficiently large sample size (Hox, 2002). This begs the question what is the smallest sample size that the analyst would be able to accept without threatening the validity of the asymptotic assumption. In most cases, the problem comes from determining the number of clusters. First, as several authors have discussed (Murray, 1998; Donner and Klar, 2000), many CRT have a small number of clusters due to logistic and cost issues. Adding a new cluster (clinic, school, organization) to the study is more expensive than adding more individuals. Second, it is often the objective of the investigators to evaluate the intervention at the whole cluster level, for example, interventions that aim to manipulate the social or physical environment (Murray, 1998).

With respect to the impact of sample size of each level on power, a few simulation studies on two-level designs have confirmed that cluster level sample size is more important than the total sample size. Brown and Draper (2000) and Maas and Hox (2005) explored the influence of sample size on power. They showed that increasing the sample size at level 2 has a stronger impact on power than increasing sample size at level 1, under multiple ICC values.

As previously noted, the number of clusters available is usually limited, but how low can this limit go? There have been some efforts to develop a rule of thumb in estimating sample size. Donner and Klar (2004) suggested that once the number of subjects per cluster exceeds $1/(\text{value of ICC})$, the power will not increase significantly as the cluster size increases. Thus, when the ICC is about 0.05 then it is of little value to enroll more than 20 subjects per cluster. In terms of accuracy in standard errors of the parameter estimates, Hox (2002) reviewed a 30/30 rule, recommending that the researchers should recruit at least 30 clusters with at least 30 subjects per cluster in order to obtain statistical accuracy.

In addition to the sample sizes at each level, another issue is how to determine the sample size in cluster data where sample sizes are planned to be unbalanced. While many papers described the adopting of CRT, most of them assumed the same number of subjects per each cluster. The difference in cluster sizes is often ignored because there are very few appropriate and easy-to-use sample size formulas for this situation. Donner and Klar (2000) suggested replacing the number of individuals in each cluster m by an average number of individuals over all clusters \bar{m} into the sample size formula for the balanced case, or to replace m by m_{\max} , the largest anticipated cluster size in the sample. However, the authors admitted that there are some limitations in this method, in that it either underestimates or overestimates the sample size. Along the same line, Eldridge et al. (2006) studied sample size computation for both continuous

and binary data for unequal cluster sizes. In their paper, they defined the relationship between the design effect and the coefficient of variation of cluster size, which is the ratio of standard deviation of cluster sizes to the mean cluster size. Written in terms of the coefficient of variation, the appropriate design effect does not depend on the knowledge of individual cluster size. The researchers also concluded that when the coefficient of variation is less than 0.23, the effect of adjustment for unequal cluster size is negligible (Eldridge et al., 2006).

2.2.3 Existing Methods for Computing Sample Sizes

The increasing application of cluster randomized trials goes hand in hand with an increasing demand for power assessment and sample size determination. Although statistical methods to compute power and sample size have been developed years ago, most of these methods focus on the computation of power in studies with simple random samples (Cohen, 1977; Chow et al., 2008). However, there has been some literature on methods that are particularly pertinent to the computing of sample sizes for CRT and multilevel design studies. In this review, we restrict our attentions to these publications.

Several authors presented sample size formulas for intervention studies that use cluster as the unit of randomization. Hsieh (1988) provided sample size formulas and power contours for simple cluster randomization and stratified randomization in CRT with two levels. Hayes and Bennet (1999) proposed sample size calculation for CRT with three types of main outcomes: rates per person per year, proportions, and means. Donner and Klar (2000) considered sample size determination for randomized trials with continuous and binary data, under matched-pair and stratified designs. Eldridge et al. (2006) presented a series of sample size formulas for both continuous and binary data for unequal cluster sizes.

Power formulas for longitudinal data and repeated measured data have also been developed. Hedeker et al. (1999) derived formulas for sample size estimation and power assessment for longitudinal study. They compared the treatment means of two groups in terms of single degree of freedom contrasts across time. Roy et al. (2007) extended these results to cluster studies in which subjects are repeatedly measured over time, taking attrition rates into account. Their main interest is to test the treatment and time interaction for continuous Gaussian outcomes.

Computing sample size for cluster randomized studies with dichotomous data is another topic of discussion in the literature. In a pioneering paper in 1997, Shih presented a method to compute sample size for correlated binary data based on generalized estimating equations. Expanding Shih's work, Pan (2001) derived more explicit sample size formulas, allowing for different structures of covariance matrix. Dang et al. (2008), on the other hand, took a different direction and applied the generalized linear mixed model technique to compute the sample size for binary cluster data. This method has an advantage over previous approaches in that it allows the users to incorporate both the fixed effects and the random effects into the model.

Due to additional factors that are involved in computing sample size for cluster data, some researchers have developed a less direct strategy to address the power and sample size questions. Their general approach is to maximize power by choosing the optimal number of clusters and cluster size that produce a specified target standard error of a particular parameter estimator. This method is based on the work by Snijders and Bosker (1993). The author used asymptotic approximations via simulations to obtain the formulas for the covariance matrix of the regression estimators. These formulas are then used to derive approximately optimal sample sizes that produce the desired standard error. Methods to estimate the variance of the fixed

effects estimators were also presented by Raudenbush and Liu (2000). The key idea of this approach is based on an optimization process. The variance for the parameter of interest is estimated first, and the sample size is computed later based on the corresponding standard error. The variance formulas were derived via random coefficient models.

2.3 Review of Statistical Software Computing Sample Size in Correlated Data

Computer programs and software are available for calculating statistical power or sample size in cluster randomized trials. This section gives a brief introduction on those programs and software.

Most of the work reviewed in Section 2.2.3 was based on a basic principle to compute sample size. The logic of this principle is to compute the sample size for a subject level randomized trial and then inflate this sample size by the design effect to obtain the required sample size for cluster data. Campbell et al. (2004) use this key idea to develop a calculator that computes the appropriate design effect and thus the sample size, when the goals are comparing the means or the proportions in CRT. The underlying formulas were based on two-level completely randomized design with equal randomization and equal cluster sizes. The user is required to specify common factors addressing a sample size problem, including (1) the difference to be detected, (2) the standard deviation, and (3) the desired significance and power. The output is a table of number of clusters required for varying values of ICC and cluster size. The calculator can be downloaded at <http://.abdn.ac.uk/hsru/epp/samsize>.

With respect to repeated measures data, Hedeker et al. (1999) introduced a software application named RMASS2. This is a free web-based application, available at www.healthstats.org, that builds on the concept discussed in their paper (Hedeker et al., 1999).

The software calculates the sample size for a two-level repeated measures design. It also allows for attrition and a variety of variance-covariance structures for the repeated measures. To date, this is the only software that handles sample size determination for studies that involve both clustered and longitudinal data.

Along with their paper on determining the optimal design for a two-level design trial, Raudenbush and Liu (2000) presented the Optimal Design software to illustrate their ideas. In this free software (provided by the University of Michigan), restricted maximum likelihood method was used to estimate the variance of the treatment contrast. This variance is a function of sample size and will be minimized by different constraints. The software can compute sample size for a two-level design with both continuous and binary outcomes. Some extension was made to the three-level design with continuous data and treatment randomized at level 3. The software can be found at http://sitemaker.umich.edu/group-based/optimal_design_software.

Snijders and Bosker developed the PiNT (Power iN Two-level design) program which estimates sample size using simulations. In the related paper (Snijders and Bosker, 1993), the authors derived the asymptotic formulas for standard errors of the fixed effects, e.g. the treatment effect, in a two-level design. Since these formulas are complicated, PiNT was written to help the users with their calculation. As the name suggests, this software is restricted for two-level studies. Other assumptions include normal response model and equal cluster size. The program and its manual are available at <http://stat.gamma.rug.nl/snijders/multilevel.htm#progPiNT>.

Most recently, two software programs were made available to researchers in computing sample size for cluster data, both written at approximately the same time. The first is the ML-DEs (MultiLevel Design Efficiency using simulation) program, developed by Cools et al. (2008). ML-DEs is a sequence of R-scripts that set up simulation studies in order to compare the

efficiency of multilevel designs conditioned on a budget and with different sample sizes and different costs of sampling units. This is a free software package with password required from the authors. To run the software, however, the program R and another package called MLwiN are required. Information about ML-DEs can be found at

<https://perswww.kuleuven.be/~u0032822/ML-DEsimulation.html>.

Around the same time ML-DEs was introduced, MLPowSim was developed in 2008 and published by Browne and Golarizadeh in 2009. This is a free program designed to perform sample size/power calculations in multilevel models via a simulation method similar to PiNT. However, MLPowSim is more comprehensive than PiNT since it can handle continuous, binary, and count data for both two and three level designs with randomization at level three. The program and related information can be found at <http://seis.bris.ac.uk/~frwjb/esrc.html>.

2.4 The Significance of Sample Size Determination in Three-Level Designs

Despite the fast growth observed across different research areas, the evaluation of required sample sizes for three-level designs has been generally overlooked in the literature. Although some work has been devoted to this topic of interest, a review of past studies shows that there are at least three issues that need improvement and invite more discussion: (1) issues in study designs with three-level structure, (2) issues with the statistical methods, and (3) issues in analysis software and programs. A closer look at these gaps in the literature will explain the motivation behind our work.

2.4.1 The Gap in Study Designs with Three-Level Structure

It can be seen from previous sections of this chapter that although methods to compute sample size and power in cluster randomized studies have been widely discussed in the literature, most publications only focused on the two-level design. In reality, however, researchers might encounter situations in which the designs and data have more complicated structure.

Examples of three level-design experiments are seen often in practice. An experiment in education can involve measurements of students within classrooms and classrooms within schools. In medical and health intervention research, evaluations might be taken from patients within physicians and then physicians within centers. It is not uncommon that a researcher might choose to ignore the cluster effect of classrooms (level 2) and treat the educational study as a two-level design in which students are nested within schools. Similarly, one can “combine” the effects of physicians and centers into one and view the medical research example as two-level where clusters are the center-physician pair. The important point is that the three-level structure is inherent in the design and embedded in the data mentioned in the above situations — regardless of the fact that the analysts choose to take the three-level structure into consideration or not.

The issue here is not whether to ignore one level of the data or to combine two levels into one, but rather to ensure that the studies are planned and analyzed in the most appropriate way. The most prudent and appropriate strategy in any statistical work is to stay faithful to the data structure and to entertain as much as possible all the features driven by the study design. This same principle remains to be true when planning for sample size and undertaking power analysis in three-level studies. In other words, all three levels should be considered in the design and planning stage for such trials. It has been well known in the literature that ignoring the cluster

effect of one level can lead to inaccurate power and errors in parameter estimates (Murray, 1998; Donner and Klar, 2000).

Another point that lacks of discussion amongst researchers is that most sample size formulas seem to apply strictly to designs where randomization takes place at the highest cluster level. There are more choices of study designs in reality. These choices involve, for example, the allocation of treatments at the second or first level. Although the assignment to treatment at different levels has a strong impact on the sample sizes formulas, this issue in the study design is rarely addressed in the literature.

2.4.2 The Gap in Methods Related to Computing Sample Size

The technical issues and appropriate statistical models to compute sample size for three-level design studies are still in debate and need more development. Reviewing published literature reveals some room for further research. For example, in a thorough study, Konstantopoulos (2008) investigated different sample sizes for three-level randomized designs in three different situations when treatment assignment is given at the first, second, and third level. Based on mixed models ANOVA/ANCOVA, the author introduced sample size formulas for each different case, allowing for the use of covariates. Unfortunately, his method is restricted with continuous Gaussian data and has not been expanded to binary outcomes.

Research methodologists have also attempted to derive explicit formulas for sample size and power assessment in three-level studies. In their paper, Heo and Leon (2008) developed closed form power function and formulas for sample size to detect an intervention effect for a three-level data. Along the same line, Teerenstra et al. (2008) presented similar formulas but expressed the design effects in terms of Pearson correlation. Although both of the above studies

provide some guidelines in the choice of number of clusters and number of subjects per cluster for a three-level experiment, they only consider randomization at the third level and do not handle studies with binary outcomes.

Most recently, Teerenstra et al. (2010) published an interesting work in which they derived sample size formulas for three-level CRT using generalized estimating equations (GEE). Their approach is actually an extension of Shih's (1997) and Pan's (2001) methods, except that the design effect was derived in a more elegant way. The sample size formulas presented by Teerenstra et al.(2010) are convenient in that they can be used for both dichotomous data and continuous data. However, a drawback of these formulas is that they are based a population-average approach which is not useful when the objective is to make inference about the participants. In addition, the authors did not consider studies in which randomization takes place at the second or first level.

2.4.3 The Gap in Analysis Software and Programs

Although in some situations a simple spreadsheet program is sufficient enough to compute power and sample size, in many other cases the formulas presented for multilevel designs are fairly complicated. Few software programs specializing in sample size estimation are available for CRT. The problem is, as encountered in the methodology literature, most of these programs apply only to two-level designs. Examples include PiNT, RMASS2, and Campbell's sample size calculator.

To date, there are only a few statistical programs that allow for computing sample size of three-level models. However, there remain some limitations in these programs and some areas leave room for expansion. For example, although Raudenbush and Liu (2000) did incorporate a

short section about three-level data in their optimal design software manual, the tool they provided can only handle continuous normally distributed data with treatment given at the third level. In addition, the program does not compute sample size for three-level design with binary data.

The software MLPowSim presented recently by Browne and Golalizadeh (2009) seems to be the most comprehensive package that handles sample sizes in CRT for continuous, binary, and count data. However, the method relies heavily on simulations. It is complicated to use, and it requires several inputs from the users. Furthermore, this software does not compute sample size for binary data with three-level designs, and does not allow for randomization at different levels.

2.4.4 Goals and Motivation

The discussion above emphasizes the importance of sample size determination in three-level designs. Given the increasing use and the complexity of the three-level models, there is a need to extend the methodological issues considered and the type of applications examined.

Although in principle, the extension from the two-level to three-level designs seems to be straightforward, the task is not simple and trivial. Adding another level means adding more complexity to the design and another sample size to determine. Not only the resulting models will be difficult to follow from a conceptual point of view, but the sample size formulas might be difficult or even impossible to derive. Given the importance of the topic, it is worthwhile to explore different strategies and different applications to address the issue of computing sample size in three-level cluster data.

In the work reported here, we propose a method to derive sample size formulas using generalized linear mixed models specifically for three-level designs. We will consider situations in which the treatment is randomized at the first, second, and third levels for both continuous normally distributed outcome and binary outcome. We will calculate power and sample size for a two-group (treatment and control) comparison. These methods will be implemented in a series of menu-driven SAS macros. The intent is not to present a method that is superior to current research studies, but rather to present a more comprehensive approach to address the same research problem. Our primary goals are to (1) gain more understanding on the theory behind statistical power and sample size in CRT with three level designs, (2) develop an alternative form of the sample size formula using available information obtained from the researchers, and (3) provide a user-friendly program to compute sample size for cluster randomization with three levels.

Chapter 3

Relevant Statistical Concepts

Although it is clear that the assessment of sample size and power is study-specific, some basic requirements need to be established in order to answer most sample size problems. The computation of sample size depends on several factors, including the expectation of the researchers, the selection of the study design, the characteristics of the data, and the statistical methods chosen by the analyst. Without sound understanding of these factors, the estimation of sample size would be reduced to merely a guessing game.

Murray (1998) discussed specific requirements for sample size estimation, such as the form and magnitude of the intervention effect, the test statistic and its distribution under the null hypothesis, and the variance of the estimated intervention effect. Although these concepts generally apply to any sample size problem, sample size consideration in multi-level designs is further complicated by numerous factors, including the number of units for each level, the magnitude of the intraclass correlations (ICC), the presence or absence of the random effects, and the correlation structures specified in each model.

This chapter provides a brief review of some relevant statistical concepts that are essential to the estimation of sample size and power analysis in cluster studies. We begin with a refresher on the concept of statistical power and sample size assessment in Section 3.1, followed by some basic characteristics of the ICC in Section 3.2. Section 3.3 revisits the discussion in the literature on power and sample size in the simplest case of multilevel data, the two-level design.

Finally, Section 3.4 offers a review of generalized linear models and generalized mixed models, the statistical methods that we utilize in this work.

3.1 Statistical Power and Sample Size Assessment

The concept of statistical power and sample size determination has been studied extensively in the literature. Excellent discussions of this topic can be found in basic statistical works such as Cohen (1988), Murphy and Myos (2004), Chow et al. (2008). In his book, Cohen (1988) presented a detailed treatment on power analysis and the corresponding formulas to compute power, together with numerous tables of sample sizes for different situations. Murphy and Myors (2004) provided an overview and introduced a set of procedures for power analysis under the context of general linear model. Chow et al. (2008) discussed sample size calculations for various study designs in clinical research. Power and sample size considerations are also covered in textbooks written specifically for cluster randomized trials, such as Murray (1998) and Donner and Klar (2000). The discussion in this section relies heavily on these aforementioned publications.

3.1.1 Hypothesis Testing

One of the most important objectives of statistics is to make inferences about unknown population parameters based on information drawn from sampled data. Methods for drawing inferences about parameters can be classified into two types: (1) to estimate the values of the parameters and (2) to establish decision rules about parameter values and apply these decision rules based on the data. The later method is referred to as hypothesis testing, a topic that we will elaborate in this section.

A hypothesis is an assertion or conjecture concerning one or more populations of interest. In other words, it is a statement about population parameters (Casella and Berger, 2005). In hypothesis testing, research hypotheses are constructed first, leading next to the statistical hypotheses. A research hypothesis is a speculation or supposition created by the researchers based upon past observations or previous experimental outcomes. A statistical hypothesis is a statement that can be evaluated by appropriate statistical methods and techniques (Daniel, 2009).

Hypothesis tests are common in many situations in which a theory or a speculation is to be tested against observation. For example, a medical researcher might hypothesize that a new drug is more effective than a traditional drug in treating a certain disease, or an educator might claim that two methods of teaching are equally effective. The role of statistics in hypothesis testing is to make decisions when comparing the observed sample with the theory: does the sample agree with the researcher's hypothesis? Should we support the hypothesis? What is the probability that we will make a wrong decision? And particularly, how large is the sample size required to reduce the chance of error and to reach the decision correctly? (Cohen, 1988)

The structure of hypothesis testing is formulated with two statistical hypotheses: the null hypothesis and the alternative hypothesis. The null hypothesis is usually denoted by the symbol H_0 . Often, the value chosen in the null hypothesis is a historical value or a claim. For example, suppose that on average a traditional drug cures 50% patients from a certain disease, then we might use a null hypothesis $H_0 : p=0.5$ for a study comparing the effect of a new drug to a traditional drug. In general, the null hypothesis is set up for the purpose of being disproved. Thus, the complement of the conclusion that the researchers seek to prove is the null hypothesis. Any hypothesis that is opposite to the null hypothesis is called an alternative hypothesis, denoted by the symbol H_A . The alternative hypothesis is a statement of what the researchers believe to be

true if the sample data calls for a rejection of the null hypothesis (Daniel, 2009). For instant, in the new drug experiment if we believe that the new drug is more effective than the traditional one, the alternative hypothesis can be stated as $H_A: p > 0.5$.

3.1.2 Statistical Power

The purpose of any statistical test is to reach a conclusion. One could either reject or not reject the null hypothesis, but hypothesis testing does not lead to a proof of a hypothesis in general. It merely implies whether the alternative hypothesis is supported or not supported by the given data. If the null hypothesis is not rejected, it means that the data do not provide enough evidence to refute it. On the other hand, rejection of the null hypothesis simply indicates that the sample evidence is sufficient enough to support the alternative hypothesis.

What if we make a wrong decision? The states of nature can be partitioned into two options, either H_0 is true or H_0 is not true. Similarly, the decisions lie between the choices of whether to reject or not to reject H_0 . Suppose the researcher rejects the null hypothesis when it is in fact true, then a type I error has been committed. On the other hand, if the researcher mistakenly retains the null hypothesis when it is indeed false, then a type II error arises. The two types of errors are illustrated in Table 3.1.

Table 3.1: Probabilities Associated with a Statistical Test

Truth of H_0	Our Decision	
	Do not reject H_0	Reject H_0
H_0 is true	Correct decision ($1 - \alpha$)	Type I error (α)
H_0 is false	Type II error (β)	Correct decision ($1 - \beta = \text{Power}$)

The maximum type I error probability which a researcher is willing to risk is called the level of significance of a test, denoted by α . Usually α is specified in advance before the sample is drawn so that the results will not influence the level of significance of the test. The probability of making a type II error is denoted by β . In hypothesis testing, it is desirable to choose α and β as small as is practical.

Power is the probability of rejecting the null hypothesis when it is in fact false. To phrase differently, power is the probability that the test will be able to detect that the alternative hypothesis holds when that is the true hypothesis. It should be noted that both power and type I error probability are functions of the interested parameters. In Table 3.1, power is represented by the quantity $1 - \beta$, or one minus type II error probability.

One common question is how much power we can expect to achieve given a level of significance. A decision not to reject H_0 (first column) means the researchers either made a correct decision or committed a type II error. By the same token, a choice of rejecting H_0 (column 2) means the researchers either made a correct decision or to committed a type I error. Thus, it is incorrect to criticize a researcher who rejected H_0 to have committed a type II error. In contrast, a researcher who did not reject the null hypothesis is exposed to type II error and his study might be vulnerable to the lack of power. One cannot conclude a new drug is ineffective in a study with low power because such study would have little chance of detecting a statistical difference between the two drug treatments. (Donner and Klar, 2000).

3.1.3 Sample Size Calculation in General

Sample size calculation is a process of computation that yields the size of the sample required to test a specific statistical hypothesis stated by the researchers. In general, to determine

a sample size, researchers are required to specify four pieces of information (1) the significance level, (2) the desired power, (3) the effect size, and (4) the standard error of the effect size. We will discuss these concepts in the context of a clinical trial testing the effectiveness of a treatment between intervention and control groups.

As defined previously, the significance level represents the probability of mistakenly rejecting the null hypothesis, and the power reflects the probability of correctly rejecting the null hypothesis. In a typical experiment, a significance level of 5% is chosen. In the literature, a conventional desired power is often set at 80% or 90%.

The effect size is a measure of the difference between the two groups that is judged to be clinically important. In general, if the researchers are settled for a large effective size, a smaller sample size is needed. However, if the effective size is relatively small, a larger number of subjects will be required. Researchers should keep the structure of the effect size as simple as possible. For example, if several differences amongst the means are of interested, only one should be chosen as the primary outcome. Similarly, if there are different treatment groups, only two that relate to the contrast of primary interest should be considered (Murray, 1998).

Finally, knowledge of the standard deviation of the main outcomes is important for sample size estimation. As a general rule, a very precise method of measurement will detect any given difference with a much smaller sample size compared to the sample size required with a less precise method of measurement. In cluster randomized trials, the knowledge of standard deviation is usually connected to the expression of the variance of the estimate of the intervention effect, which we will discuss later.

Snijders and Bosker (1993) proposed the following formula to link the four quantities required in a sample size calculation together, under the normal distribution assumption for the variable associated with the effect size:

$$\frac{\delta}{SE(\delta)} = Z_{1-(\alpha/2)} + Z_{1-\beta}$$

In this formula, α is the significance level of the test, $1 - \beta$ is the power of the test, δ is the effect size, $SE(\delta)$ is the standard error of the estimated effect size, and $Z_{1-(\alpha/2)}$ and $Z_{1-\beta}$ are the quantiles of the standard normal distribution associated with the values of $1 - \alpha/2$ and $1 - \beta$.

It can be seen from the above formula that given three out of four factors, the fourth factor can be computed. In many types of designs in practice, this formula is a valid approximation that can be used to compute the sample size required for a certain level of power. In multilevel designs, however, more than one level of sample sizes is needed. Difficulty also arises in determining the standard error of the estimated treatment effect in cluster randomized trials. We will turn to the details of this discussion later in Section 3.3.

3.2 Intraclass Correlation

The sampling of subjects into experiments via clusters introduces special considerations that need to be addressed in sample size determination and power analysis. Outcomes measured from subjects within the same clusters tend to be similar and are not independent. In the literature, this level of dependency—the effect of clustering—is estimated by the measurement of the intraclass correlation. In this section, we will discuss the effect of clustering in two-level designs and then expand the idea to three-level designs.

3.2.1 The Effect of Clustering in Two-Level Designs

Consider an experiment in which sample of clusters of individuals are randomly assigned to two experimental groups: treatment and control. Let Y_{ijk} represent an observation from the i^{th} subject within the j^{th} cluster within the k^{th} treatment, and let the variable Y_{ijk} have a common total variance of σ^2 . Let all clusters have the same n number of subjects, and each of the two treatment arms have the same number of m clusters. The variance of a cluster mean is

$$\text{Var}(\bar{Y}_C) = \frac{\sigma^2}{n}$$

where σ^2 is the within cluster variance.

Furthermore, assuming constant variance across the clusters, the variance of the treatment mean can be written as:

$$\text{Var}(\bar{Y}_T) = \frac{\sigma^2}{mn}$$

Since the units of randomization are clusters, the dependency between individuals within the same cluster exists. In such situation, the total variance can be decomposed into two components, a between cluster variance σ_B^2 and a within cluster variance σ_W^2 , so that $\sigma^2 = \sigma_B^2 + \sigma_W^2$.

To take into account the effect of this dependency, we introduce the quantity intraclass correlation, denote by ρ , into the variance formula. Here, the parameter ρ can be viewed as the pair-wise correlation coefficient between any two measurements in the same cluster, meaning $\rho = \text{Corr}(Y_{ijk}, Y_{i'jk})$. With the assumption that ρ is positive, we can also interpret ρ as the fraction of the total variation in the data that is attributed to the unit of assignment. In other

words, ρ is the proportion of the total variance that is accounted for by the between cluster variation:

$$\rho = \frac{\sigma_B^2}{\sigma_W^2 + \sigma_B^2}$$

Methods to estimate the ICC were reviewed in Chapter 2. A classical approach is to apply the analysis of variance among and within cluster. In this method, the ICC is obtained by the following equation:

$$\hat{\rho} = \frac{MSC - MSW}{MSC + (n-1)MSW} = \frac{S_B^2}{S_B^2 + S_W^2}$$

Where:

- MSC and MSW are the mean square error between and within cluster respectively.
- $S_B^2 = \frac{(MSC - MSW)}{n}$ and $S_W^2 = MSW$ are the estimates of σ_B^2 and σ_W^2 respectively.

With clusters serving as the units of randomization, the variance of cluster mean now takes a different form that reflects the variation between groups and within members:

$$Var(\bar{Y}_c) = \frac{\sigma_W^2}{n} + \sigma_B^2$$

And the variance of treatment mean is:

$$Var(\bar{Y}_T) = \frac{\sigma_W^2}{mn} + \frac{\sigma_B^2}{m}$$

Written in terms of the total variance and ρ , the final formula yields:

$$Var(\bar{Y}_T) = \frac{\sigma^2}{mn} [1 + (n-1)\rho]$$

A few implications can be drawn from the above equation:

- When the estimated value of ρ is zero, the formula for the variance of treatment mean is reduced to the usual form of the variance of treatment mean under the assumption of independence amongst cluster members. On the other hand, when $\rho=1$, we have the total dependence situation. In this case, all measurements in a cluster are identical and the total information gained from a cluster is no more than information gained from a single member (Donner and Klar, 2000).
- The variance of the treatment mean in cluster randomized trials equals the variance of treatment mean under the assumption of independent errors multiplied by a quantity of $1+(n-1)\rho$ in two level-designs. This quantity is called the variance inflation factor, or the design effect. As we will see later, the design effect plays an important role in the estimation of sample size (Murray, 1998).
- Since the value of ρ is always positive, the design effect increases when ρ increases and when the cluster size n increases. When n is large the design effect can be large even with a small ICC. Thus, the variance of treatment mean is always larger in cluster designs than in studies without cluster effect and with equal total number of subjects (Donner and Klar, 2000).
- Analytical methods that treat clusters as the unit of analysis and take proper consideration of the cluster effect can provide the tests for the treatment effect. However, given all other factors constant, the tests might have lower power than would be obtained in independent cases (Hedges and Hedberg, 2007).

3.2.2 The Effect of Clustering in Three-Level Designs

The concept of the intraclass correlation as a measure of proportional variation can be extended to designs with more than two levels, the only difference is that we now have more variance components involved and there are different ways to define and interpret the intraclass correlations.

Consider an experiment involving a three level design in which patients are nested within physicians, and physicians are in turn nested within centers. The total variance in the outcome can be decomposed into three components: The between patients nested in physicians (level 1) variance σ_e^2 , the between physicians nested in centers (level 2) variance σ_p^2 , and the between centers (level 3) variance σ_c^2 . Thus, the total variance is $\sigma_T^2 = \sigma_c^2 + \sigma_p^2 + \sigma_e^2$. In the context of a three-level design, two definitions of the intraclass correlations are being circulated in the literature.

The first method specifies the intraclass correlation at the physician level as

$$r = \frac{\sigma_p^2}{\sigma_T^2}$$

and at the center level as

$$\rho = \frac{\sigma_c^2}{\sigma_T^2}$$

The second method defines the intraclass correlation at the physician level as

$$r = \frac{\sigma_p^2 + \sigma_c^2}{\sigma_T^2}$$

and at the center level as

$$\rho = \frac{\sigma_c^2}{\sigma_T^2}$$

The choice of which method to apply depends on the goal of the analysts. The first method should be used when our focus is on decomposing the variance components across all levels, or on estimating how much variation is explained by each level (Hox, 2002). On the other hand, the second method allows for an estimation of the correlation between two randomly chosen subjects in the same group. For example, in the second method r represents the correlation between two patients related to the same physician, and r also takes into account that two patients seeing the same physician are visiting in the same center. In the same context, ρ represents the correlation between two patients in different centers (and thus different physicians). Some authors view the intraclass correlation under the second method as an analogy to the Pearson correlation (Teerenstra et al., 2008).

Similar to what we have seen in the two-level designs, the variance of the treatment means in three-level situations can be written as a product of the variance that would be obtained if all outcomes are independent and a variance inflation factor. This statement will be discussed with further details in Chapter 4 and Chapter 5.

3.3 Computing Sample Size in Two-Level Cluster Designs

Before proposing the methods to compute sample size for three-level cluster randomized studies, a refresher on how the sample size formulas for two-level designs were derived would be helpful. First, we will review how sample sizes are computed in an individual randomized clinical trial. Next, we will discuss how the same formula is extended to a two-level cluster design.

3.3.1 Power in a Simple Clinical Trial

Consider a simple clinical trial in which $2N$ subjects are equally randomized into two treatment groups. Let Y_{ij} be the observation of the i^{th} subject in the j^{th} treatment. The model can be written:

$$Y_{ij} = \mu_j + \varepsilon_{ij}$$

where $j=1,2$ and $i=1,\dots,N$ and $\varepsilon_{ij} \sim N(0, \sigma^2)$.

Suppose our hypothesis is to test the difference between the treatment effects:

$$H_0 : \mu_1 = \mu_2 \quad \text{versus} \quad H_1 : \mu_1 \neq \mu_2$$

Denote the estimates of the difference in the two treatment means as

$$\hat{\Delta} = \hat{Y}_1 - \hat{Y}_2$$

Let $\hat{\sigma}_\Delta^2$ be the estimated variance of that difference under the null hypothesis.

Assuming equal variances in the two groups, then:

$$\hat{\sigma}_\Delta^2 = \frac{2\hat{\sigma}^2}{N}$$

A Z-statistic can be used to test the treatment mean difference.

$$z = \frac{\hat{\Delta}}{\sqrt{\hat{\sigma}_\Delta^2}}$$

Under the null hypothesis of no treatment effect, $Z \sim N(0,1)$. We reject the null hypothesis when

$$|Z| > Z_{\alpha/2}$$

Under the alternative hypothesis that $\mu_1 = \mu_2 + \delta$, then $Z \sim N(\mu^*, 1)$, where

$$\mu^* = \frac{\delta}{\sqrt{\hat{\sigma}_\Delta^2}}$$

The corresponding power is given by

$$\begin{aligned}
 P\{|Z| > Z_{\alpha/2}\} &\approx P\{Z > Z_{\alpha/2}\} \\
 &= 1 - \Phi\{Z_{\alpha/2} - \mu^*\} \\
 &= 1 - \beta
 \end{aligned}$$

In order to achieve the power of $(1 - \beta)100\%$, we need

$$Z_{\alpha/2} - \mu^* = -Z_{\beta} \quad \text{or} \quad \mu^* = Z_{\alpha/2} + Z_{\beta}$$

Substitute this into the equation for μ^* above, we get

$$\begin{aligned}
 \frac{\delta}{\sqrt{\hat{\sigma}_{\Delta}^2}} &= Z_{\alpha/2} + Z_{\beta} \\
 \frac{\delta}{\sqrt{\frac{2\hat{\sigma}^2}{N}}} &= Z_{\alpha/2} + Z_{\beta}
 \end{aligned}$$

From here, we obtain the usual sample size formula testing the difference between the two means

$$N = \frac{2\hat{\sigma}^2 (Z_{\alpha/2} + Z_{\beta})^2}{\delta^2}$$

3.3.2 Power in a Two-Level Cluster Randomized Trial

Now assume that a number of m clusters, rather than individuals, are randomized to each treatment group. Furthermore, assuming each cluster has n subjects and the clusters were chosen randomly from a larger population. Thus, the total sample size is $N=mn$. The model becomes:

$$Y_{ijk} = \mu + C_i + T_j + \varepsilon_{ijk}$$

- where $i=1, \dots, m, j=1, 2$, and $k=1, \dots, n$
- C_i represents the cluster effect
- T_j represents the treatment effect
- $\varepsilon_{ijk} \sim N(0, \sigma_e^2)$ and $C_i \sim Normal(0, \sigma_c^2)$

The intraclass correlation is defined as

$$\rho = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_e^2}$$

Where σ_c^2 is the between cluster variance and σ_e^2 is the within cluster variance.

The total variance of Y_{ijk} is $\sigma^2 = \sigma_e^2 + \sigma_c^2$

As derived in Section 3.2.1, the variance of the difference in estimate treatment means is

$$\hat{\sigma}_\Delta^2 = \frac{2\hat{\sigma}^2}{mn} [1 + (n-1)\rho]$$

Substitute this into the sample size equation

$$\frac{\delta}{\sqrt{\hat{\sigma}_\Delta^2}} = Z_{\alpha/2} + Z_\beta$$

The total sample size is now

$$N = \frac{2\hat{\sigma}^2 (Z_{\alpha/2} + Z_\beta)^2 (1 + (n-1)\rho)}{\delta^2}$$

Compared to the simple randomized clinical trial, the sample size in two-level cluster randomized trial is “inflated” by the variance inflation factor of $(1 + (n-1)\rho)$. Since this variance inflation factor depends on the variance of the treatment means, the problem of computing sample size in cluster data really boils down to the problem of deriving the variance of treatment effect.

3.3.3 Power and the Effect of Randomization

Randomization, an important component of experimental research, can be accomplished through two major types of design (1) a subject-randomized design and (2) a cluster-randomized

design. In Section 3.3.2, we discussed the derivation of sample size formula in two-level designs in which group of subjects (clusters) was assigned to different treatments. However, there are many situations in which for a given effect size, the randomizations can occur at different levels.

Unit of randomization plays an important role in computing sample size and power analysis. For example, the model presented in Section 3.3.2 is no longer true if the individuals, instead of clusters, are randomized. Raudenbush and Liu (2000) discussed this issue in a two-level design applied for multisite experiment, where persons within a site are randomly assigned to one or two more treatments. The sample size formula in this case is more complicated, since it takes into account the possibility that treatment effects can vary across different sites.

The situation is more complicated in three-level designs, when randomization could happen at any of the three levels. For instant, in the example discussed in Section 3.2.2, treatment assignment could be done at the level of patient, or provider, or practice. We will see in later chapters how the level chosen for random assignment will affect the variance-covariance structure of the statistic model and the corresponding sample size formulas.

3.4 Generalized Linear Mixed Models

In the course of this work, we will present methods to compute sample size and power for three-level cluster designs based on generalized linear models (GLM) and generalized linear mixed model (GLMM) approaches. To stimulate further interest, this section pulls together some basic concepts of the GLM and GLMM. The concepts introduced here will recur frequently in subsequent chapters. Understanding the definitions and their relevance is important, especially for looking at their applications in the context of three-level design. The discussions in this section are drawn mainly from Brown and Prescott (2006) and Littell et al. (2007).

3.4.1 Linear Mixed Models

Consider a normal linear model

$$y_i = \mu + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

In this model, the regression coefficients betas represent the fixed effects and the error term is the only random effect. In matrix notation, it can be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

Where $\mathbf{Y} = (y_1, y_2, \dots, y_n)'$ is the vector of the observed values

$\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)'$ is the vector of the fixed effects

$\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)'$ is the vector of the residuals

The normal mixed model extends the above fixed effect model by including additional random-effect, random coefficients, or covariance terms in the residual variance matrix. Normal mixed effects are often appropriate for representing clustered and dependent data. In normal mixed models, the random effects are assumed to follow a distribution and the fixed effects are considered constant. In matrix notation, a typical normal mixed model takes the form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

where

\mathbf{Y} is an $n \times 1$ vector of observed values

\mathbf{X} is an $n \times (p + 1)$ design matrix (of column rank p)

$\boldsymbol{\beta}$ is a $(p + 1) \times 1$ vector of fixed effects (including the overall mean μ)

\mathbf{Z} is an $n \times q$ second design matrix for the random effects $\boldsymbol{\gamma}$

$\boldsymbol{\gamma}$ is a $q \times 1$ vector of random effects, where $\boldsymbol{\gamma} \sim N_q(\mathbf{0}, \mathbf{G})$ and $\mathbf{G} = \{\sigma_c^2, \sigma_c^2, \dots, \sigma_c^2\}$

$\boldsymbol{\varepsilon}$ is an $n \times 1$ error vector where $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \mathbf{R})$ and $\mathbf{R} = \sigma^2 \mathbf{I}$

Under the above assumptions, the variance of the observation vector can be written as

$$V(\mathbf{y}) = \mathbf{V} = V(\mathbf{Z}\boldsymbol{\gamma}) + V(\mathbf{e}) = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$$

To give a general idea on the structure of \mathbf{V} , consider a simple example with a study of two centers, each with two patients, one in each treatment. Here we have $p=2$ and $q=2$ (thus the total sample size is 4).

$$\mathbf{y} = \begin{bmatrix} y_{111} \\ y_{221} \\ y_{312} \\ y_{422} \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_{111} \\ \varepsilon_{221} \\ \varepsilon_{312} \\ \varepsilon_{422} \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \beta_1 \\ \beta_2 \end{bmatrix}, \quad \mathbf{Z} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{\gamma} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix}$$

Then the variance covariance matrix \mathbf{V} in this case is

$$\mathbf{V} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \sigma_c^2 & 0 \\ 0 & \sigma_c^2 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} + \sigma^2 \mathbf{I}$$

$$= \begin{bmatrix} \sigma_c^2 & \sigma_c^2 & 0 & 0 \\ \sigma_c^2 & \sigma_c^2 & 0 & 0 \\ 0 & 0 & \sigma_c^2 & \sigma_c^2 \\ 0 & 0 & \sigma_c^2 & \sigma_c^2 \end{bmatrix} + \sigma^2 \mathbf{I} = \begin{bmatrix} \sigma_c^2 + \sigma^2 & \sigma_c^2 & 0 & 0 \\ \sigma_c^2 & \sigma_c^2 + \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma_c^2 + \sigma^2 & \sigma_c^2 \\ 0 & 0 & \sigma_c^2 & \sigma_c^2 + \sigma^2 \end{bmatrix}$$

Note that specifying the random effects is a convenient way to form the structure of the variance-covariance matrix. The above model is equivalent to a fixed effects model with the normally distributed residuals that share the same variance-covariance matrix as specified. In other words, we can write $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

With the same \mathbf{Y} and $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon} \sim N(0, \mathbf{V})$ where \mathbf{V} is defined as above.

3.4.2 Generalized Linear Models (GLM)

Normal linear mixed models work under two distinct assumptions: the errors and random effects are normally distributed, and the response variable is modeled directly as a linear combination of the fixed and random effects. These assumptions are not met in many practical situations where data are non-normal, such as studies with binary or count outcomes. In these situations, Generalized Linear Models are available for fitting non-normal fixed effect models with three basic components.

The first component is called the random component, referring to the response variable $\mathbf{Y} = (y_1, y_2, \dots, y_n)$. \mathbf{Y} belongs to the exponential family with a probability density function with the following general form:

$$f(y_i; \theta_i) = \exp\left\{\left[y_i \theta_i - b(\theta_i)\right] / a(\phi) + c(y_i, \phi)\right\}$$

where θ is a location parameter and ϕ is a dispersion parameter that only appears in distribution with two parameters (such as normal distribution). The forms of the function a , b , and c are different in different distributions.

For one-parameter distributions, the general form can be simplified to:

$$f(y_i; \theta_i) = \exp\left\{\left[y_i \theta_i - b(\theta_i)\right] / a + c(y_i)\right\}$$

It can be shown that the mean and variance of the random component can be written in terms of μ and θ as follows

$$E(y) = \mu = b'(\theta)$$

$$Var(y) = b''(\theta)$$

Hence, we can find θ by $\theta = b^{-1}(\mu)$.

The second component relates a vector $\boldsymbol{\eta} = (\eta_1, \eta_2, \dots, \eta_n)$ to the model fixed effects $\boldsymbol{\beta}$ and the predictors \mathbf{X} by a linear combination $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$. This linear combination of the fixed effects is called the linear predictor.

The third component is a link function that links the linear model to the mean of \mathbf{Y} through the formula

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} \quad \text{where } g(\cdot) \text{ is monotonic and differentiable}$$

$$\boldsymbol{\mu} = E(\mathbf{Y})$$

Note that the normal linear model is a special case of GLM in which the link function is the identity link $g(\mu) = \mu$.

In general, the GLM can be defined using the matrix notation similar to normal models as follows

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$$

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$$

Thus, the variance-covariance matrix of this model is

$$Var(\mathbf{y}) = Var(\boldsymbol{\varepsilon}) = \mathbf{V}$$

For a sample of n observations, the variance takes the form $\mathbf{V} = \mathbf{AB}$, where $\mathbf{A} = \text{diag}[a_i(\phi)]$

and $\mathbf{B} = \text{diag}[b_i(\theta)]$, both are $n \times n$ matrices. For example, in the binomial distribution with

$n=4$, we would have

$$\mathbf{A} = \begin{bmatrix} 1/n_1 & 0 & 0 & 0 \\ 0 & 1/n_2 & 0 & 0 \\ 0 & 0 & 1/n_3 & 0 \\ 0 & 0 & 0 & 1/n_4 \end{bmatrix} \text{ and } \mathbf{B} = \begin{bmatrix} \mu_1(1-\mu_1) & 0 & 0 & 0 \\ 0 & \mu_2(1-\mu_2) & 0 & 0 \\ 0 & 0 & \mu_3(1-\mu_3) & 0 \\ 0 & 0 & 0 & \mu_4(1-\mu_4) \end{bmatrix}$$

3.4.3 Generalized Linear Mixed Models (GLMM)

The Generalized Linear Mixed Models are the extended version of Generalized Linear Models to accommodate models with random effects. The extension is carried out through the link function. Here, the general model takes the same form of

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$$

However, the random effects are now added to the link function

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}$$

where

- \mathbf{Y} is an $n \times 1$ vector of observed values
- \mathbf{X} is an $n \times (p + 1)$ design matrix (of column rank p)
- $\boldsymbol{\beta}$ is a $(p + 1) \times 1$ vector of fixed effects (including the overall mean μ)
- \mathbf{Z} is an $n \times q$ second design matrix for the random effects $\boldsymbol{\gamma}$
- $\boldsymbol{\gamma}$ is a $q \times 1$ vector of random effects independent of $\boldsymbol{\varepsilon}$, and $\boldsymbol{\gamma} \sim N_q(\mathbf{0}, \mathbf{G})$
- $\boldsymbol{\varepsilon}$ is the vector of errors and $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\text{Var}(\boldsymbol{\varepsilon}) = \mathbf{AB}$ where \mathbf{A} and \mathbf{B} were defined previously.

The key difference between the normal mixed model and GLMM is that $\boldsymbol{\varepsilon}$ is not necessarily assumed to be normal.

The variance of \mathbf{y} is

$$V(\mathbf{y}) = \mathbf{V} = V(\boldsymbol{\mu}) + V(\boldsymbol{\varepsilon}) = V(\boldsymbol{\mu}) + \mathbf{A}\mathbf{B}.$$

To illustrate the structure of \mathbf{A} and \mathbf{B} , consider a simple data set with 6 *Bernoulli*(μ) observations, we then have:

$$\mathbf{B} = \begin{bmatrix} \mu_t(1-\mu_t) & 0 & 0 & 0 & 0 & 0 \\ 0 & \mu_t(1-\mu_t) & 0 & 0 & 0 & 0 \\ 0 & 0 & \mu_t(1-\mu_t) & 0 & 0 & 0 \\ 0 & 0 & 0 & \mu_c(1-\mu_c) & 0 & 0 \\ 0 & 0 & 0 & 0 & \mu_c(1-\mu_c) & 0 \\ 0 & 0 & 0 & 0 & 0 & \mu_c(1-\mu_c) \end{bmatrix}$$

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Discussion on how GLMM is applied in computing sample size for binary data will be presented with more details in Chapter 5.

Chapter 4

Power and Sample Size for Continuous Outcome

In Chapter 1, we introduced a few examples with three-level cluster design studies and different possible outcomes. In the simplest situation, these outcomes are continuous and normally distributed. In this chapter we present methods to compute power and sample size assuming a linear mixed model. We will derive the power formulas and the general forms of the variance structures for six different situations: (1) randomize at level three, (2) randomize at level two without interaction, (3) randomize at level two with interaction, (4) randomize at level one without interaction, (5) randomize at the level one with interaction between treatment and level three and (6) randomize at the level one with interaction between treatment and level two.

4.1 Linear Mixed Model Approach

Consider a three-level study design in which patients are nested within physicians, and physicians are in turn nested within centers. Let y_{ijk} be the response observed at k^{th} patient, j^{th} physician, and i^{th} center. Furthermore, let N be the total number of centers, p be the number of physicians in each center, and n be the number of patients visiting each physician (balanced design). The total sample size is $T=Npn$.

Assuming both level two (physician) and level three (center) are random effects and no covariates are included, the total variance in the outcome can be decomposed into three components: the within physicians and between patients (level 1) variance σ_e^2 , the between

physicians (level 2) and within centers (level 3) variance σ_p^2 , and the between centers (level 3) variance σ_c^2 .

The intraclass correlation between patients nested within the same physician and the same center can be specified by:

$$\text{corr}(y_{ijk}, y_{ijk'}) = \frac{\sigma_p^2 + \sigma_c^2}{\sigma_T^2} = r$$

The intraclass correlation between patients within the same center but under different physicians can be specified by:

$$\text{corr}(y_{ijk}, y_{ijk'}) = \frac{\sigma_c^2}{\sigma_T^2} = \rho.$$

Under mixed models notation, we can write

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

where

- \mathbf{Y} is the $(T \times 1)$ vector of the observed outcome data
- \mathbf{X} is a $(T \times m)$ fixed effects design matrix
- $\boldsymbol{\beta}$ is a $(m \times 1)$ vector of regression fixed effects coefficients
- \mathbf{Z} is a $(T \times q)$ random effects design matrix.
- $\boldsymbol{\gamma}$ is the $(q \times 1)$ vector of random effects, $\boldsymbol{\gamma} \sim N(\mathbf{0}, \mathbf{G})$
- $\boldsymbol{\varepsilon}$ is the error vector, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma_e^2 \mathbf{I})$

Follow the properties of Multivariate Normal distribution, we have $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{W})$ where

$$\mathbf{W} = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \sigma_e^2 \mathbf{I}$$

For any type of mixed model, the sample size and power estimates depend on the variances of the fixed effects, which can be found by

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^N \mathbf{X}_i^T \mathbf{V}^{-1} \mathbf{X}_i \right)^{-1} \quad (4.1)$$

where \mathbf{X}_i is the corresponding design matrix to center i^{th} and $\boldsymbol{\beta}$ is the vector of regression coefficients for the fixed effects.

Assuming the treatment effects are fixed and the model has no other covariates ($m=2$), then $\boldsymbol{\beta} = (\beta_0, \beta)^T$ where β_0 is the expected measurement for a patient in the control group and $\beta_0 + \beta$ is the expected measurement for patients in the intervention group. The hypothesis of interest can be written as

$$\begin{aligned} \mathbf{H}_0: & \beta = 0 \\ \mathbf{H}_A: & \beta = d \neq 0 \end{aligned}$$

To test this hypothesis, a Wald-type test based on asymptotically normal distributions can be used. The asymptotic variance of $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is determined by the right lower corner element of the estimated variance-covariance matrix $\hat{\Sigma} = \text{Var} \left[\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \right]$.

The power to detect a difference of size d with a two-sided type I error rate of α is:

$$\text{power} = 1 - \mathcal{T}_{t, \xi} \left(t_{\alpha/2, \xi} - d \sqrt{\frac{N}{\text{Var}(\hat{\boldsymbol{\beta}})}} \right) + \mathcal{T}_{t, \xi} \left(-t_{\alpha/2, \xi} - d \sqrt{\frac{N}{\text{Var}(\hat{\boldsymbol{\beta}})}} \right)$$

where $\mathcal{T}_{t, \xi}$ is the cumulative distribution function of the t-distribution with ξ degrees of freedom and $t_{\alpha/2, \xi}$ is the 100 α % percentile from the t-distribution with ξ degrees of freedom. The value of ξ depends on the analysis method and the level where randomization takes place.

In many sample size problems, computing the variance estimator $Var(\hat{\beta})$ is the main task. Once $Var(\hat{\beta})$ is obtained, we can easily plug its value into the above formula to estimate power or sample size. In three-level design, power depends on the sample size of all three levels, i.e. the values of N , p , and n . Although in the above formula n and p do not appear explicitly, their roles are embedded in the computation of $Var(\hat{\beta})$. The following sections derive the formulas for $Var(\hat{\beta})$ for different levels of randomization.

4.2 Randomize at Third Level

4.2.1 Estimate $Var(\hat{\beta})$

Assume that the centers are randomized such that πN centers are in the treatment arm and $(1-\pi)N$ centers are in the control arm. The number of patients allocated to the treatment arm is $T_1 = \pi Npn$, and the number of patients allocated to the control arm is $T_2 = (1-\pi)Npn$. The design matrix is $\mathbf{X}_i = \mathbf{X}_{treat} = (\mathbf{1}_{pn}, \mathbf{1}_{pn})$ if the i^{th} center is randomized into the treatment group, and $\mathbf{X}_i = \mathbf{X}_{control} = (\mathbf{1}_{pn}, \mathbf{0}_{pn})$ if the i^{th} center is randomized into the control group.

When randomization occurs at level three, the matrix \mathbf{V} can be written as:

$$\mathbf{V} = \mathbf{I}_p \otimes (\sigma_e^2 \mathbf{I}_n + \sigma_p^2 \mathbf{J}_n) + \sigma_c^2 \mathbf{J}_{pn}$$

For example, consider the i^{th} center with two physicians and each physician has two patients.

Then

$$\mathbf{V} = \begin{bmatrix} \sigma_T^2 & \sigma_c^2 + \sigma_p^2 & \sigma_c^2 & \sigma_c^2 \\ \sigma_c^2 + \sigma_p^2 & \sigma_T^2 & \sigma_c^2 & \sigma_c^2 \\ \sigma_c^2 & \sigma_c^2 & \sigma_T^2 & \sigma_c^2 + \sigma_p^2 \\ \sigma_c^2 & \sigma_c^2 & \sigma_c^2 + \sigma_p^2 & \sigma_T^2 \end{bmatrix}$$

Factor out the total variance σ_T^2 , we obtain $\mathbf{V} = \sigma_T^2 \mathbf{R}$, where

$$\mathbf{R} = \mathbf{I}_p \otimes \left[(1-r)\mathbf{I}_n + (r-\rho)\mathbf{J}_n \right] + \rho \mathbf{J}_{pn}$$

For the example above,

$$\mathbf{R} = \begin{bmatrix} 1 & r & \rho & \rho \\ r & 1 & \rho & \rho \\ \rho & \rho & 1 & r \\ \rho & \rho & r & 1 \end{bmatrix}$$

The robust variance estimator of $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is

$$\hat{\Sigma} = \lim_{N \rightarrow \infty} N \left(\sum_{i=1}^N \mathbf{X}_i^T \mathbf{V}^{-1} \mathbf{X}_i \right)^{-1}$$

$$\hat{\Sigma} = \sigma_T^2 \lim_{N \rightarrow \infty} N \left[N\pi (\mathbf{X}_{treat}^T \mathbf{R}^{-1} \mathbf{X}_{treat}) + N(1-\pi) (\mathbf{X}_{control}^T \mathbf{R}^{-1} \mathbf{X}_{control}) \right]^{-1}$$

$$\hat{\Sigma} = \sigma_T^2 \left[\pi (\mathbf{X}_{treat}^T \mathbf{R}^{-1} \mathbf{X}_{treat}) + (1-\pi) (\mathbf{X}_{control}^T \mathbf{R}^{-1} \mathbf{X}_{control}) \right]^{-1}$$

Following the derivation procedure presented by Shin (1997), $\hat{\Sigma}$ can be rewritten as

$$\hat{\Sigma} = \sigma_T^2 \left[\begin{pmatrix} 1 & \pi \\ \pi & \pi \end{pmatrix} (\mathbf{1}^T \mathbf{R}^{-1} \mathbf{1}) \right]^{-1}$$

$$\hat{\Sigma} = \frac{\sigma_T^2}{\mathbf{1}^T \mathbf{R}^{-1} \mathbf{1}} \begin{pmatrix} 1 & \pi \\ \pi & \pi \end{pmatrix} \begin{pmatrix} \pi & -\pi \\ -\pi & 1 \end{pmatrix}$$

$\text{Var}(\hat{\boldsymbol{\beta}})$ is the right lower corner element of the above variance-covariance matrix $\hat{\Sigma}$, which is

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \frac{\sigma_T^2}{\pi(1-\pi)(\mathbf{1}^T \mathbf{R}^{-1} \mathbf{1})}$$

The quantity $(\mathbf{1}^T \mathbf{R}^{-1} \mathbf{1})^{-1}$ is proportional to the design effect (variance inflation factor). This is a scalar with the value equal to the sum of all elements in the matrix \mathbf{R}^{-1} . The exact expression of $(\mathbf{1}^T \mathbf{R}^{-1} \mathbf{1})^{-1}$ will be derived next.

4.2.2 Derivation of $\mathbf{1}^T \mathbf{R}^{-1} \mathbf{1}$

This derivation is based on the work from Teerenstra (2010). The two following results from Henderson and Searle (1981) will be used recurrently in this section:

Result 1:

If A and B are arbitrary nonsingular square matrices of same dimension, then

$$(A+B)^{-1} = A^{-1} - A^{-1}B(I+A^{-1}B)^{-1}A^{-1} = \left[I - A^{-1}B(I+A^{-1}B)^{-1} \right] A^{-1}$$

Result 2:

$$(aI_q + bJ_q)^{-1} = \frac{1}{a} \left(I_q - \frac{b}{a+bq} J_q \right)$$

Recall that correlation matrix \mathbf{R} for a given center i^{th} takes the form $\mathbf{R} = A + B$, where

$$A = I_p \otimes [(1-r)I_n + (r-\rho)J_n] \quad \text{and} \quad B = \rho J_{pn}$$

Apply result 2, we obtain

$$A^{-1} = \left(\frac{1}{1-r} \right) I_p \otimes \left[I_n - \left(\frac{r-\rho}{1+(n-1)r-n\rho} \right) J_n \right]$$

Let $\gamma = 1+(n-1)r-n\rho$, then

$$A^{-1} = \left(\frac{1}{1-r} \right) I_p \otimes \left[I_n - \left(\frac{r-\rho}{\gamma} \right) J_n \right]$$

$$\begin{aligned}
A^{-1}B &= \left\{ \left(\frac{1}{1-r} \right) I_p \otimes \left[I_n - \left(\frac{r-\rho}{\gamma} \right) J_n \right] \right\} [\rho J_{pn}] \\
&= \frac{\rho}{\gamma(1-r)} (\gamma - nr + n\rho) J_{pn} \\
&= \frac{\rho}{\gamma} J_{pn}
\end{aligned}$$

Again, apply result 2 we get

$$(I + A^{-1}B)^{-1} = I_{pn} - \frac{\rho}{\gamma + pn\rho} J_{pn}$$

Denote $\varphi = \gamma + pn\rho = 1 + (n-1)r + n(p-1)\rho$, then

$$(I + A^{-1}B)^{-1} = I_{pn} - \frac{\rho}{\varphi} J_{pn}$$

Substitute the A^{-1} , $A^{-1}B$, and $(I + A^{-1}B)^{-1}$ in $R^{-1} = (A + B)^{-1}$ using result 1 we have

$$\mathbf{R}^{-1} = \frac{1}{1-r} \left[I_{pn} - \frac{\rho}{\varphi} J_{pn} \right] \left[I_p \otimes \left(I_n - \frac{r-\rho}{\gamma} J_n \right) \right]$$

Thus,

$$\mathbf{1}^T \mathbf{R}^{-1} \mathbf{1} = \frac{1}{1-r} \mathbf{1}^T \left[I_{pn} - \frac{\rho}{\varphi} J_{pn} \right] \left[I_p \otimes \left(I_n - \frac{r-\rho}{\gamma} J_n \right) \right] \mathbf{1}$$

$$\mathbf{1}^T \mathbf{R}^{-1} \mathbf{1} = \frac{1}{1-r} \left[1 - \frac{pn\rho}{\varphi} \right] \mathbf{1}^T \mathbf{1} \left[1 - \frac{n(r-\rho)}{\gamma} \right]$$

$$\mathbf{1}^T \mathbf{R}^{-1} \mathbf{1} = \frac{1}{1-r} \left[1 - \frac{pn\rho}{\varphi} \right] \left[1 - \frac{n(r-\rho)}{\gamma} \right] pn = \frac{pn}{\varphi}$$

Hence, when randomization takes place and the third level, the design effect is given by

$$\varphi = 1 + (n-1)r + n(p-1)\rho.$$

4.3 Randomize at Second Level without Interaction Effect

4.3.1 Estimate $\text{Var}(\hat{\boldsymbol{\beta}})$

Assume that for a center i^{th} , the physicians are randomized such that πp physicians are in the treatment arm and $(1-\pi)p$ physicians are in the control arm. Thus, the number of patients allocated to the treatment arm is $T_1 = \pi Npn$, and the number of patients allocated to the control arm is $T_2 = (1-\pi)Npn$. Furthermore, assume that there is no interaction effect between treatment and center, i.e., any difference due to treatment is the same in every center. Under these assumptions, the correlation matrix \mathbf{R} remains the same form as described in 4.1.2:

$$\mathbf{R} = \mathbf{I}_p \otimes [(1-r)\mathbf{I}_n + (r-\rho)\mathbf{J}_n] + \rho\mathbf{J}_{pn}$$

The covariate matrix \mathbf{X}_i for a center i^{th} is a $pn \times 2$ matrix. The first column of this matrix contains all ones, whereas the second column contains ones in the first πpn rows, and zero in the remaining $(1-\pi)pn$ rows. For example, consider a simple experiment in which each center i^{th} has four physicians, and each physician has two patients. Suppose the physicians in each center are randomized such that the first two physicians are in the treatment group and the last two are in the control group. For a given center i^{th} , the matrix \mathbf{X}_i takes the following form:

$$\mathbf{X}_i = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix}$$

The robust variance estimator of $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is

$$\hat{\Sigma} = \lim_{N \rightarrow \infty} N \left(\sum_{i=1}^N \mathbf{X}_i^T \mathbf{V}^{-1} \mathbf{X}_i \right)^{-1}$$

$$\hat{\Sigma} = \sigma_T^2 \lim_{N \rightarrow \infty} N \left[N \mathbf{X}_i^T \mathbf{R}^{-1} \mathbf{X}_i \right]^{-1}$$

$$\hat{\Sigma} = \sigma_T^2 \left[\mathbf{X}_i^T \mathbf{R}^{-1} \mathbf{X}_i \right]^{-1}$$

$Var(\hat{\beta})$ is the right lower corner element of the above variance-covariance matrix $\hat{\Sigma}$.

4.3.2 Derivation of $\mathbf{X}_i^T \mathbf{R}^{-1} \mathbf{X}_i$

Under the assumption of no interaction, the matrix \mathbf{R}^{-1} remain the same as derived in 4.1.2

$$\mathbf{R}^{-1} = \frac{1}{1-r} \left[I_{pn} - \frac{\rho}{\varphi} J_{pn} \right] \left[I_p \otimes \left(I_n - \frac{r-\rho}{\gamma} J_n \right) \right]$$

where $\gamma = 1 + (n-1)r - n\rho$ and

$$\varphi = \gamma + pn\rho = 1 + (n-1)r + n(p-1)\rho.$$

Expanding elements in the brackets, we have

$$\mathbf{R}^{-1} = \frac{1}{1-r} \left\{ I_p \otimes \left[I_n - \frac{r-\rho}{\gamma} J_n \right] - \frac{\rho}{\varphi} J_{pn} \left[I_m \otimes \left(I_n - \frac{r-\rho}{\gamma} J_n \right) \right] \right\}$$

For simplification, let $\frac{r-\rho}{\gamma} = a$ and $\frac{\rho}{\varphi} = b$

$$\mathbf{R}^{-1} = \frac{1}{1-r} \left\{ I_p \otimes [I_n - aJ_n] + b[na - 1] J_{pn} \right\}$$

The matrix $\mathbf{X}_i^T \mathbf{R}^{-1} \mathbf{X}_i$ is a 2 x 2 matrix with the following elements

$$\mathbf{X}_i^T \mathbf{R}^{-1} \mathbf{X}_i = \begin{pmatrix} \sum \text{all elements in } \mathbf{R}^{-1} & \sum \text{elements in first } \pi pn \text{ columns of } \mathbf{R}^{-1} \\ \sum \text{first } \pi pn \text{ rows of } \mathbf{R}^{-1} & \sum \text{elements in first } \pi pn \text{ columns and first } \pi pn \text{ rows of } \mathbf{R}^{-1} \end{pmatrix}$$

Denote s as the sum of all elements in \mathbf{R}^{-1} and write s in terms of $a, b, n,$ and r

$$\begin{aligned} s &= \frac{1}{1-r} \left[p(n - n^2a) + p^2n^2b(na - 1) \right] \\ &= \frac{1}{1-r} \left[pn - pn^2a + p^2n^3ab - p^2n^2b \right] \end{aligned}$$

Denote t as the sum of all elements that are in both first πpn columns and first πpn rows of \mathbf{R}^{-1} .

$$\begin{aligned} t &= \frac{1}{1-r} \left[p\pi(n - n^2a) + p^2n^2\pi^2b(na - 1) \right] \\ &= \frac{1}{1-r} \left[pn\pi - pn^2\pi a + p^2n^3\pi^2ab - p^2n^2\pi^2b \right] \end{aligned}$$

Since \mathbf{R}^{-1} is symmetric then

$$\mathbf{X}_i^T \mathbf{R}^{-1} \mathbf{X}_i = \begin{pmatrix} s & \pi s \\ \pi s & t \end{pmatrix}.$$

Add and subtract $pn^2\pi^2a - pn\pi^2$ to the inside of the brackets of t ,

$$\begin{aligned} t &= \frac{1}{1-r} \left[(pn\pi - pn^2\pi a + pn^2\pi^2a - pn\pi^2) + (pn\pi^2 - pn^2\pi^2a + p^2n^3\pi^2ab - p^2n^2\pi^2b) \right] \\ &= \frac{1}{1-r} \left[pn\pi(1-\pi)(1-na) \right] + \pi^2 \left[\frac{1}{1-r} (pn - pn^2a + p^2n^3ab - p^2n^2b) \right] \end{aligned}$$

Note that the terms in the second bracket equal s , the sum of all elements in \mathbf{R}^{-1} . Also, note that

$$1 - na = 1 - \frac{nr - n\rho}{\gamma} = \frac{\gamma - nr + n\rho}{\gamma} = \frac{1-r}{\gamma}$$

Substituting this into t we get

$$t = \frac{pn\pi(1-\pi)}{\gamma} + \pi^2s$$

Thus

$$\mathbf{X}_i^T \mathbf{R}^{-1} \mathbf{X}_i = \begin{pmatrix} s & \pi s \\ \pi s & \frac{pn\pi(1-\pi)}{\gamma} + \pi^2 s \end{pmatrix}$$

And

$$(\mathbf{X}_i^T \mathbf{R}^{-1} \mathbf{X}_i)^{-1} = \frac{\gamma}{spn\pi(1-\pi)} \begin{pmatrix} \frac{pn\pi(1-\pi)}{\gamma} + \pi^2 s & -\pi s \\ -\pi s & s \end{pmatrix}$$

The robust variance estimator of $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is

$$\hat{\Sigma} = \sigma_T^2 [\mathbf{X}_i^T \mathbf{R}^{-1} \mathbf{X}_i]^{-1}$$

The right lower corner element of the above variance-covariance matrix $\hat{\Sigma}$ is

$$\text{Var}(\hat{\beta}) = \frac{\sigma_T^2 \gamma}{pn\pi(1-\pi)}$$

Hence, when randomization takes place at the second level and assuming no interaction between center and treatment effects, the design effect is given by $\gamma = 1 + (n-1)r - n\rho$.

4.4 Randomize at Second Level with Interaction Effect

4.4.1 Estimate $\text{Var}(\hat{\boldsymbol{\beta}})$

Assume that in each particular center i^{th} , the physicians are randomized such that πp of them are in the treatment arm and $(1-\pi)p$ of them are in the control arm. Thus, the number of patients allocated to the treatment arm is $T_1 = \pi Npn$, and the number of patients allocated to the control arm is $T_2 = (1-\pi)Npn$. When the interaction between treatment and center exists, the robust variance estimator of $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ can be obtained by computing $\hat{\Sigma} = [\mathbf{X}_i^T \mathbf{V}^{-1} \mathbf{X}_i]^{-1}$. The

variance of the estimator $Var(\hat{\beta})$ is the right lower corner element of the above variance-covariance matrix $\hat{\Sigma}$.

The distinction between randomization at second level with interaction and without interaction is that a new variance component representing the interaction between treatment and center is added to the model. The introduction of this interaction term changes the structure of the variance matrix \mathbf{V} .

4.4.2 The structure of \mathbf{V}

Denote this variance of the interaction as σ_{ct}^2 , the total variance is now

$$\sigma_T^2 = \sigma_c^2 + \sigma_p^2 + \sigma_e^2 + \sigma_{ct}^2$$

σ_{ct}^2 allows for the impact of treatment on the outcome measurement to vary across centers.

Under mixed model theory, the structure of \mathbf{V} can be written as follows:

$\mathbf{V} = \text{block}(\mathbf{U}, \mathbf{T}) + J_{np}(\sigma_c^2)$, where:

$$\mathbf{U} = I_{\pi p} \otimes [I_n(\sigma_e^2) + J_n(\sigma_p^2)] + J_{\pi np}(\sigma_{ct}^2)$$

$$\mathbf{T} = I_{(1-\pi)p} \otimes [I_n(\sigma_e^2) + J_n(\sigma_p^2)] + J_{(1-\pi)np}(\sigma_{ct}^2)$$

To illustrate, suppose there are two centers, each center has three physicians and each physician has two patients. Furthermore, suppose randomization takes place at the physician level where the first two physicians are in treatment arm and the last physician is in control arm. For one center, the \mathbf{V} matrix is:

$$\mathbf{V} = \begin{bmatrix} \sigma_T^2 & \omega^2 & \nu^2 & \nu^2 & \sigma_c^2 & \sigma_c^2 \\ \omega^2 & \sigma_T^2 & \nu^2 & \nu^2 & \sigma_c^2 & \sigma_c^2 \\ \nu^2 & \nu^2 & \sigma_T^2 & \omega^2 & \sigma_c^2 & \sigma_c^2 \\ \nu^2 & \nu^2 & \omega^2 & \sigma_T^2 & \sigma_c^2 & \sigma_c^2 \\ \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_T^2 & \omega^2 \\ \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \omega^2 & \sigma_T^2 \end{bmatrix}$$

where $\nu^2 = \sigma_c^2 + \sigma_{ct}^2$ and $\omega^2 = \sigma_c^2 + \sigma_p^2 + \sigma_{ct}^2$.

There is no closed form for the right lower corner of the matrix $\hat{\Sigma} = [\mathbf{X}_i^T \mathbf{V}^{-1} \mathbf{X}_i]^{-1}$. Thus,

we will use the general formula specified in equation (4.1) to compute $\text{Var}(\hat{\beta})$.

4.5 Randomize at First Level without Interaction Effect

4.5.1 Estimate $\text{Var}(\hat{\beta})$

For a the j^{th} physician in the i^{th} center, suppose the patients are randomized such that πn patients are in the treatment arm and $(1-\pi)n$ patients are in the control arm. Thus, the total number of patients allocated to the treatment arm is $T_1 = \pi Npn$, and the number of patients allocated to the control arm is $T_2 = (1-\pi)Npn$. In addition, assume that there is no interaction effect, i.e., the treatment works the same in every center–physician pair. Under these assumptions, the correlation matrix \mathbf{R} remains the same form as described in 4.1.2:

$$\mathbf{R} = \mathbf{I}_p \otimes [(1-r)\mathbf{I}_n + (r-\rho)\mathbf{J}_n] + \rho\mathbf{J}_{pn}$$

The covariate matrix \mathbf{X}_i for the i^{th} center is a $pn \times 2$ matrix. The first column of this matrix contains all ones, whereas the second column contains ones in the πpn rows corresponding to those patients allocated to the treatment group, and zero in the $(1-\pi)pn$ rows corresponding to those patients allocated to the control group.

For example, consider a simple case with two centers, each center has two physicians, and each physician has four patients: two patients are in the treatment group and two are in the control group. The \mathbf{X}_i matrix for the i^{th} center is as follows:

$$\mathbf{X}_i = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \end{bmatrix}$$

The robust variance estimator of $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is

$$\hat{\Sigma} = \lim_{N \rightarrow \infty} N \left(\sum_{i=1}^N \mathbf{X}_i^T \mathbf{V}^{-1} \mathbf{X}_i \right)^{-1}$$

$$\hat{\Sigma} = \sigma_T^2 \lim_{N \rightarrow \infty} N \left[N \mathbf{X}_i^T \mathbf{R}^{-1} \mathbf{X}_i \right]^{-1}$$

$$\hat{\Sigma} = \sigma_T^2 \left[\mathbf{X}_i^T \mathbf{R}^{-1} \mathbf{X}_i \right]^{-1}$$

$\text{Var}(\hat{\boldsymbol{\beta}})$ is the right lower corner element of the above variance-covariance matrix $\hat{\Sigma}$.

4.5.2 Derivation of $\mathbf{X}_i^T \mathbf{R}^{-1} \mathbf{X}_i$

Under the assumption of no interaction, the matrix \mathbf{R}^{-1} remain the same as derived in 4.1.2

$$\mathbf{R}^{-1} = \frac{1}{1-r} \left[I_{pn} - \frac{\rho}{\phi} J_{pn} \right] \left[I_p \otimes \left(I_n - \frac{r-\rho}{\gamma} J_n \right) \right]$$

where $\gamma = 1 + (n-1)r - n\rho$ and

$$\varphi = \gamma + pn\rho = 1 + (n-1)r + n(p-1)\rho.$$

Expanding elements in the brackets, we have

$$\mathbf{R}^{-1} = \frac{1}{1-r} \left\{ I_p \otimes \left[I_n - \frac{r-\rho}{\gamma} J_n \right] - \frac{\rho}{\varphi} J_{pn} \left[I_m \otimes \left(I_n - \frac{r-\rho}{\gamma} J_n \right) \right] \right\}$$

For simplification, let $\frac{r-\rho}{\gamma} = a$ and $\frac{\rho}{\varphi} = b$

$$\mathbf{R}^{-1} = \frac{1}{1-r} \left\{ I_p \otimes [I_n - aJ_n] + b[na-1]J_{pn} \right\}$$

The matrix $\mathbf{X}_i^T \mathbf{R}^{-1} \mathbf{X}_i$ is a 2 x 2 matrix with the following elements

$$\mathbf{X}_i^T \mathbf{R}^{-1} \mathbf{X}_i = \begin{pmatrix} \sum \text{all elements in } \mathbf{R}^{-1} & \sum \text{elements in treatment columns} \\ \sum \text{elements in treatment rows} & \sum \text{elements in both treatment columns and treatment rows} \end{pmatrix}$$

Let s be the sum of all elements in \mathbf{R}^{-1} and write s in terms of a, b, n , and r

$$s = \frac{1}{1-r} \left[p(n - n^2a) + p^2n^2b(na-1) \right]$$

Let t be the sum of all elements appear in both treatment columns and treatment rows

$$t = \frac{1}{1-r} \left[p(n\pi - n^2\pi^2a) + p^2n^2\pi^2b(na-1) \right]$$

We then can write

$$\mathbf{X}_i^T \mathbf{R}^{-1} \mathbf{X}_i = \begin{pmatrix} s & \pi s \\ \pi s & t \end{pmatrix}$$

Thus

$$\left(\mathbf{X}_i^T \mathbf{R}^{-1} \mathbf{X}_i \right)^{-1} = \frac{1}{st - \pi^2 s^2} \begin{pmatrix} t & -\pi s \\ -\pi s & s \end{pmatrix}$$

The lower right element of the matrix $\hat{\Sigma} = \sigma_r^2 \left[\mathbf{X}_i^T \mathbf{R}^{-1} \mathbf{X}_i \right]^{-1}$ is $Var(\hat{\beta}) = \frac{\sigma_r^2}{t - \pi^2 s}$. Note that

$$t - \pi^2 s = \frac{1}{1-r} \left[p(n\pi - n^2\pi^2 a) + p^2 n^2 \pi^2 b(na-1) - \pi^2 p(n - n^2 a) - \pi^2 p^2 n^2 b(na-1) \right]$$

$$= \frac{pn\pi(1-\pi)}{1-r}$$

Thus, $Var(\hat{\beta}) = \frac{\sigma_r^2(1-r)}{np\pi(1-\pi)}$

Therefore, when randomization takes place at the first level without interaction effect, the design effect is $1-r$.

4.6 Randomize at First Level with Interaction Effect

4.6.1 Estimate $Var(\hat{\beta})$

For a the j^{th} physician in the i^{th} center, suppose the patients are randomized such that πn patients are in the treatment arm and $(1-\pi)n$ patients are in the control arm. Thus, the total number of patients allocated to the treatment arm is $T_1 = \pi Npn$, and the number of patients allocated to the control arm is $T_2 = (1-\pi)Npn$. The robust variance estimator of $\sqrt{N}(\hat{\beta} - \beta)$ can be obtained by computing $\hat{\Sigma} = [\mathbf{X}_i^T \mathbf{V}^{-1} \mathbf{X}_i]^{-1}$. The variance of the estimator $Var(\hat{\beta})$ is the right lower corner element of the above variance-covariance matrix $\hat{\Sigma}$.

We consider two different scenarios where the interaction effects might be found: (1) interaction between treatment and level three and (2) interaction between treatment and level two.

4.6.2 The Structure of \mathbf{V} with Treatment \times Level Three Interaction

The first scenario considers the situation where the treatment effect varies across different centers. Denote the variance of the interaction as σ_{ct}^2 , the total variance is now given by:

$$\sigma_T^2 = \sigma_c^2 + \sigma_p^2 + \sigma_e^2 + \sigma_{ct}^2$$

Under mixed model theory, the structure of \mathbf{V} can be written as follows:

$$\mathbf{V} = I_p \otimes \mathbf{U} + J_p \otimes \mathbf{T} - I_p \otimes \mathbf{T} \text{ where}$$

$$\mathbf{U} = \text{block} \left\{ \left[I_{n\pi} (\sigma_e^2) + J_{n\pi} (\sigma_{ct}^2) \right], \left[I_{n(1-\pi)} (\sigma_e^2) + J_{n(1-\pi)} (\sigma_{ct}^2) \right] \right\} + J_n (\sigma_c^2 + \sigma_p^2)$$

$$\mathbf{T} = \text{block} \left\{ J_{n\pi} (\sigma_{ct}^2), J_{n(1-\pi)} (\sigma_{ct}^2) \right\} + J_n (\sigma_c^2)$$

To illustrate, consider the i^{th} center with two physicians, each physician has four patients.

Suppose the patients are randomized equally to the two treatment groups, the \mathbf{V} matrix takes the following form:

$$\mathbf{V} = \begin{bmatrix} \sigma_T^2 & \omega^2 & \tau^2 & \tau^2 & \nu^2 & \nu^2 & \sigma_c^2 & \sigma_c^2 \\ \omega^2 & \sigma_T^2 & \tau^2 & \tau^2 & \nu^2 & \nu^2 & \sigma_c^2 & \sigma_c^2 \\ \tau^2 & \tau^2 & \sigma_T^2 & \omega^2 & \sigma_c^2 & \sigma_c^2 & \nu^2 & \nu^2 \\ \tau^2 & \tau^2 & \omega^2 & \sigma_T^2 & \sigma_c^2 & \sigma_c^2 & \nu^2 & \nu^2 \\ \nu^2 & \nu^2 & \sigma_c^2 & \sigma_c^2 & \sigma_T^2 & \omega^2 & \tau^2 & \tau^2 \\ \nu^2 & \nu^2 & \sigma_c^2 & \sigma_c^2 & \omega^2 & \sigma_T^2 & \tau^2 & \tau^2 \\ \sigma_c^2 & \sigma_c^2 & \nu^2 & \nu^2 & \tau^2 & \tau^2 & \sigma_T^2 & \omega^2 \\ \sigma_c^2 & \sigma_c^2 & \nu^2 & \nu^2 & \tau^2 & \tau^2 & \omega^2 & \sigma_T^2 \end{bmatrix}$$

Where $\tau^2 = \sigma_c^2 + \sigma_p^2$, $\nu^2 = \sigma_c^2 + \sigma_{ct}^2$, and $\omega^2 = \sigma_c^2 + \sigma_p^2 + \sigma_{ct}^2$.

There is no closed form for the right lower corner of the matrix $\hat{\Sigma} = [\mathbf{X}_i^T \mathbf{V}^{-1} \mathbf{X}_i]^{-1}$. Thus,

we will use the general formula specified in equation (4.1) to compute $\text{Var}(\hat{\beta})$.

4.6.3 The Structure of \mathbf{V} with Treatment \times Level Two Interaction

The second scenario considers the situation where the treatment effect varies across different physicians. Denote the variance of the interaction as σ_{pt}^2 , the total variance is now

$$\sigma_T^2 = \sigma_c^2 + \sigma_p^2 + \sigma_e^2 + \sigma_{pt}^2$$

Under mixed model theory, the structure of \mathbf{V} can be written as follows:

$$\mathbf{V} = I_p \otimes \mathbf{U} + J_p \otimes \mathbf{T} - I_p \otimes \mathbf{T} \text{ where}$$

$$\mathbf{U} = \text{block} \left\{ \left[I_{n\pi} (\sigma_e^2) + J_{n\pi} (\sigma_{pt}^2) \right], \left[I_{n(1-\pi)} (\sigma_e^2) + J_{n(1-\pi)} (\sigma_{pt}^2) \right] \right\} + J_n (\sigma_c^2 + \sigma_p^2)$$

$$\mathbf{T} = J_n (\sigma_c^2)$$

To illustrate the above structure, consider center i^{th} with two physicians, each physician has four patients. Suppose the patients are randomized equally to the two treatment groups, the \mathbf{V} matrix takes the following form:

$$\mathbf{V} = \begin{bmatrix} \sigma_T^2 & \mathcal{G}^2 & \tau^2 & \tau^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 \\ \mathcal{G}^2 & \sigma_T^2 & \tau^2 & \tau^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 \\ \tau^2 & \tau^2 & \sigma_T^2 & \mathcal{G}^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 \\ \tau^2 & \tau^2 & \mathcal{G}^2 & \sigma_T^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 \\ \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_T^2 & \mathcal{G}^2 & \tau^2 & \tau^2 \\ \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \mathcal{G}^2 & \sigma_T^2 & \tau^2 & \tau^2 \\ \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \tau^2 & \tau^2 & \sigma_T^2 & \mathcal{G}^2 \\ \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \tau^2 & \tau^2 & \mathcal{G}^2 & \sigma_T^2 \end{bmatrix}$$

where $\tau^2 = \sigma_c^2 + \sigma_p^2$ and $\mathcal{G}^2 = \sigma_c^2 + \sigma_p^2 + \sigma_{pt}^2$.

There is no closed form for the right lower corner of the matrix $\hat{\Sigma} = [\mathbf{X}_i^T \mathbf{V}^{-1} \mathbf{X}_i]^{-1}$. Thus,

we will use the general formula specified in equation (4.1) to compute $\text{Var}(\hat{\beta})$.

Chapter 5

Power and Sample Size for Binary Outcome

In Chapter 4, we presented power and sample size formulas for three-level designs with continuous and normally distributed outcomes. In practice, however, it is very common to encounter experiments with correlated binary data. For example, consider a three-level design medical trial in which patients are selected within physicians, and physicians are selected within centers. Suppose the outcome is the presence or absence of a disease. The data from such study are binary and are correlated in a twofold nested fashion: measurements on patients are correlated within the same physicians, which in turn correlated within centers.

Earlier, when working continuous Gaussian data, we derived the variance formulas for the estimated treatment effect based on linear mixed model theory. However, the vector of the expected means in binary outcome is typically not modeled as a linear function of the parameters. Thus, modification to the presented formulas for continuous data is required in order to answer the same power and sample size questions for binary data.

Details of the modification will be provided in this chapter. First we will explain how generalize linear mixed models can be applied to derive the power and sample size functions for binary outcome. Next, we will show how the variance of the treatment difference will be computed in light of this approach.

5.1 Generalized Linear Mixed Models (GLMM) Approach

5.1.1 Basic Model

Consider a three-level CRT design in which patients are nested within physicians and physicians in turn are nested within centers. Denote the binary outcome variable from the i^{th} center, j^{th} physician, and k^{th} patient as y_{ijk} . Let N be the total number of centers, p be the number of physicians in each center, and n be the number of patients visiting each physician (balanced design). The total sample size is $T=Npn$.

The Generalized Linear Mixed Model is based on extending the fixed effect Generalized Linear Model by including the random effects. The model takes the following form:

$$E(\mathbf{Y} | \boldsymbol{\gamma}) = g^{-1}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}) = g^{-1}(\boldsymbol{\eta}) = \boldsymbol{\mu}$$

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}$$

where:

- \mathbf{Y} is the $(T \times 1)$ vector of the observed data
- $\boldsymbol{\mu}$ is the vector of the expected means of \mathbf{Y}
- $g(\cdot)$ is a differentiable monotonic link function
- \mathbf{X} is a $(T \times m)$ fixed effects design matrix
- $\boldsymbol{\beta}$ is a $(m \times 1)$ vector of regression coefficients for the fixed effects
- \mathbf{Z} is a $(T \times q)$ random effects design matrix.
- $\boldsymbol{\gamma}$ is the $(q \times 1)$ vector of random effects, $\boldsymbol{\gamma} \sim N(\mathbf{0}, \mathbf{G})$
- $\boldsymbol{\eta}$ is the linear predictor, that is, $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}$.

Following GLMM theory, the variance of the observations conditional on the random effects can be written as:

$$\text{Var}(\mathbf{Y} | \boldsymbol{\gamma}) = \mathbf{A}\mathbf{B}$$

where $\mathbf{A} = \text{diag}\{a_i\}$ and $\mathbf{B} = \text{diag}\{b''(\theta)\} = \left(\frac{\partial g^{-1}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}\right)$. (See Chapter 3 for more details.)

The unconditional variance takes the form:

$$\mathbf{V} = \text{Var}(\mathbf{Y}) = \text{Var}(\boldsymbol{\mu}) + \mathbf{A}\mathbf{B}$$

\mathbf{V} is not easy to specify because in the binary case $\boldsymbol{\mu}$ is not a linear function of $\boldsymbol{\beta}$.

Solutions for GLMM depend on the form of the likelihood function. In many cases, it is difficult to maximize this function since it involves T integrals over the q dimensional random effects. Hence, techniques of numerical approximation are required to solve the problem.

5.1.2 Pseudo-likelihood Method

In order to derive the general form of the variance matrix \mathbf{V} , we consider the application of pseudo-likelihood, a linearization method that was tested and implemented in the SAS PROC GLIMMIX procedure. Pseudo-likelihood maximizes the quasi-likelihood by iteratively analyzing a linearized pseudo variable. The term “pseudo” is used because the likelihood function maximized is a function of a pseudo variable, not of the original data. Here, the pseudo variable is based on a first-order Taylor series expansion for $g(\mathbf{Y})$ about $\boldsymbol{\mu}$, which yields:

$$\begin{aligned} \mathbf{z} &= g(\boldsymbol{\mu}) + (\mathbf{Y} - \boldsymbol{\mu})\mathbf{B}^{-1} \\ &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + (\mathbf{Y} - \boldsymbol{\mu})\mathbf{B}^{-1} \end{aligned}$$

This new model can be viewed as a linear mixed model with the new pseudo-observation \mathbf{z} and the new random error term $(\mathbf{Y} - \boldsymbol{\mu})\mathbf{B}^{-1}$. The variance of \mathbf{z} can be formulated as (Brown and Prescott, 2006):

$$\mathbf{V}_z = \mathbf{ZGZ}^T + \mathbf{B}^{-1}(\mathbf{AB})\mathbf{B}^{-1}$$

$$\mathbf{V}_z = \mathbf{ZGZ}^T + \mathbf{B}^{-1}\mathbf{A}$$

The variance of this linear mixed pseudo model will be used to derive the sample size and power formulas for binary outcomes.

To specify \mathbf{V}_z we need two pieces of information. The first piece \mathbf{ZGZ}^T relates to the random effects in the model. As we will see in some explicit examples presented in the next sections, the components of \mathbf{ZGZ}^T consist of: (1) the between level two and within centers variance σ_p^2 , (2) the between level three variance σ_c^2 , and (3) the variance of the interaction terms when the interaction effect exists. The structure of \mathbf{ZGZ}^T depends on which level randomization takes place. The process of deriving \mathbf{ZGZ}^T for the binary response is the same with that of the continuous response.

The second piece $\mathbf{B}^{-1}\mathbf{A}$ refers to the within physicians and between patients variance of the first level measures. Unlike the case of continuous outcomes, this variance is now a function of the binary proportion and is not independent from the mean. Letting μ be the probability of “success” or the probability of observing a specific outcome and applying the logit link function, it can be seen from Section 5.1.1 that \mathbf{A} is simply an identity matrix and \mathbf{B}^{-1} is a diagonal matrix with elements of $\frac{1}{\mu(1-\mu)}$. We will see explicit examples of this in the next few sections.

Getting back to the three-level study design where the observed outcomes from the i^{th} center, j^{th} physician, and k^{th} patient are binary variables y_{ijk} , suppose the interest is to test the difference between the two treatment effects. Let the probability of an event in the treatment group be μ_t and in the control group be μ_c . The marginal distributions of y_{ijk} and the corresponding mean and variances are as follow

$y_{ijk} \sim \text{Bernoulli}(\mu_t)$ when y_{ijk} is in the treatment group

$$E(y_{ijk} | x_{ijk}) = \mu_t$$

$$\text{Var}(y_{ijk}) = \mu_t(1 - \mu_t)$$

$y_{ijk} \sim \text{Bernoulli}(\mu_c)$ when y_{ijk} is in the control group

$$E(y_{ijk} | x_{ijk}) = \mu_c$$

$$\text{Var}(y_{ijk}) = \mu_c(1 - \mu_c)$$

Assuming the treatment effects are fixed and the model has no other covariates, i.e. $m=2$ and

$\boldsymbol{\beta} = (\beta_0, \beta)^T$. Under logit model we can write

$$\mu_c = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}$$

$$\mu_t = \frac{\exp(\beta_0 + \beta)}{1 + \exp(\beta_0 + \beta)}$$

Thus, $\beta = \log\left(\frac{\mu_t}{1 - \mu_t}\right) - \log\left(\frac{\mu_c}{1 - \mu_c}\right)$, which is the log odds ratio of the corresponding response probabilities.

The hypothesis of

$$\mathbf{H}_0 : \mu_t = \mu_c$$

$$\mathbf{H}_0 : \mu_t \neq \mu_c$$

is now equivalent to

$$\mathbf{H}_0 : \beta = 0 \text{ and}$$

$$\mathbf{H}_A : \beta = b \neq 0$$

Once the structures of \mathbf{V}_{z_i} and of \mathbf{X}_i are defined, the sample size and power estimates depend on the variance estimates of the fixed effects, which can be found by

$$\text{Var}(\hat{\beta}) = \left(\sum_{i=1}^N \mathbf{X}_i^T \mathbf{V}_{z_i}^{-1} \mathbf{X}_i \right)^{-1}$$

The asymptotic variance of $\sqrt{N}(\hat{\beta} - \beta)$ is determined by the right lower corner element of the estimated variance-covariance matrix $\hat{\Sigma} = \widehat{\text{Var}} \left[\sqrt{N}(\hat{\beta} - \beta) \right]$.

The power to detect a difference of size d with a two-sided type I error rate of α can be obtained by

$$\text{power} = 1 - \mathcal{T}_{t,\xi} \left(t_{\alpha/2,\xi} - d \sqrt{\frac{N}{\text{Var}(\hat{\beta})}} \right) + \mathcal{T}_{t,\xi} \left(-t_{\alpha/2,\xi} - d \sqrt{\frac{N}{\text{Var}(\hat{\beta})}} \right)$$

where $T_{t,\xi}$ is the cumulative distribution function of the t-distribution with ξ degrees of freedom and $t_{\alpha/2,\xi}$ is the 100 $\alpha/2$ % percentile from the t-distribution with ξ degrees of freedom. The value of ξ depends on the analysis method and the level where randomization takes place.

It can be seen from the above equation that the most important step to determine power is the computation of the term $\text{Var}(\hat{\beta})$, which is the lower right hand corner of the estimated variance-covariance matrix $\hat{\Sigma}$. Since the structure of $\hat{\Sigma} = \widehat{\text{Var}} \left[\sqrt{N}(\hat{\beta} - \beta) \right]$ depends on the study design, the next following sections are devoted to explain how to compute $\hat{\Sigma}$ in particular cases

for binary outcomes depending on which level random assignment is performed and whether or not an interaction is taken into account.

5.2 Randomize at Third Level

Assume that the total N centers are randomized such that πN centers are in the treatment arm and $(1-\pi)N$ centers are in the control arm. Thus, the number of patients allocated to the treatment arm is $T_1=\pi Npn$, and the number of patients allocated to the control arm is $T_2=(1-\pi)Npn$. Furthermore, let the probability of an event in the treatment group be μ_t and in the control group be μ_c .

Since randomization occurs at the center level, the covariate matrix \mathbf{X}_i and the variance of the pseudo variable \mathbf{V}_{z_i} depend on to which treatment group a particular center is allocated. Denote the covariate matrix \mathbf{X}_i and the variance matrix of the pseudo variable \mathbf{V}_{z_i} for the i^{th} center. The robust variance estimator of $\sqrt{N}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$ is

$$\begin{aligned}\hat{\Sigma} &= \lim_{N \rightarrow \infty} N \left(\sum_{i=1}^N \mathbf{X}_i^T \mathbf{V}_{z_i}^{-1} \mathbf{X}_i \right)^{-1} \\ \hat{\Sigma} &= \lim_{N \rightarrow \infty} N \left[N\pi \left(\mathbf{X}_t^T \mathbf{V}_{z_t}^{-1} \mathbf{X}_t \right) + N(1-\pi) \left(\mathbf{X}_c^T \mathbf{V}_{z_c}^{-1} \mathbf{X}_c \right) \right]^{-1} \\ \hat{\Sigma} &= \left[\pi \left(\mathbf{X}_t^T \mathbf{V}_{z_t}^{-1} \mathbf{X}_t \right) + (1-\pi) \left(\mathbf{X}_c^T \mathbf{V}_{z_c}^{-1} \mathbf{X}_c \right) \right]^{-1}\end{aligned}$$

If the i^{th} center is randomized into the treatment group, we have

$$\begin{aligned}\mathbf{X}_i &= \mathbf{X}_t = (\mathbf{1}_{pn}, \mathbf{1}_{pn}) \\ \mathbf{V}_{z_i} &= \mathbf{V}_{z_t} = \mathbf{I}_p \otimes \left(\frac{1}{\mu_t(1-\mu_t)} \mathbf{I}_n + \sigma_p^2 \mathbf{J}_n \right) + \sigma_c^2 \mathbf{J}_{pn}\end{aligned}$$

If the i^{th} center is randomized into the control group, we have

$$\mathbf{X}_i = \mathbf{X}_c = (\mathbf{1}_{pn}, \mathbf{0}_{pn})$$

$$\mathbf{V}_{Zi} = \mathbf{V}_{Zc} = \mathbf{I}_p \otimes \left(\frac{1}{\mu_c(1-\mu_c)} \mathbf{I}_n + \sigma_p^2 \mathbf{J}_n \right) + \sigma_c^2 \mathbf{J}_{pn}$$

To illustrate, consider a simple experiment with two centers, each center has two physicians, and each physician has two patients. Suppose that the first center is randomized into the treatment group and the second center is randomized into the control group. Furthermore, denote the total variance for each treatment group as σ_{treat}^2 and $\sigma_{control}^2$, then

$$\sigma_{treat}^2 = \sigma_c^2 + \sigma_p^2 + \frac{1}{\mu_t(1-\mu_t)}$$

$$\sigma_{control}^2 = \sigma_c^2 + \sigma_p^2 + \frac{1}{\mu_c(1-\mu_c)}$$

The structure of \mathbf{X}_i and \mathbf{V}_{Zi} for the first center (treatment) are

$$\mathbf{X}_t = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{V}_{Zt} = \begin{bmatrix} \sigma_{treat}^2 & \sigma_c^2 + \sigma_p^2 & \sigma_c^2 & \sigma_c^2 \\ \sigma_c^2 + \sigma_p^2 & \sigma_{treat}^2 & \sigma_c^2 & \sigma_c^2 \\ \sigma_c^2 & \sigma_c^2 & \sigma_{treat}^2 & \sigma_c^2 + \sigma_p^2 \\ \sigma_c^2 & \sigma_c^2 & \sigma_c^2 + \sigma_p^2 & \sigma_{treat}^2 \end{bmatrix}$$

The structure of \mathbf{X}_i and \mathbf{V}_{Zi} for the second center (control) are

$$\mathbf{X}_c = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{V}_{Zc} = \begin{bmatrix} \sigma_{control}^2 & \sigma_c^2 + \sigma_p^2 & \sigma_c^2 & \sigma_c^2 \\ \sigma_c^2 + \sigma_p^2 & \sigma_{control}^2 & \sigma_c^2 & \sigma_c^2 \\ \sigma_c^2 & \sigma_c^2 & \sigma_{control}^2 & \sigma_c^2 + \sigma_p^2 \\ \sigma_c^2 & \sigma_c^2 & \sigma_c^2 + \sigma_p^2 & \sigma_{control}^2 \end{bmatrix}$$

The variance of the estimated treatment effect $\text{Var}(\hat{\beta})$ is the right lower corner element of the

variance-covariance matrix $\hat{\Sigma} = [\pi(\mathbf{X}_t^T \mathbf{V}_{Zt}^{-1} \mathbf{X}_t) + (1-\pi)(\mathbf{X}_c^T \mathbf{V}_{Zc}^{-1} \mathbf{X}_c)]^{-1}$

5.3 Randomize at Second Level without Interaction Effect

Assume that for the i^{th} center, the physicians are randomized such that πp of them are in the treatment arm and $(1-\pi)p$ of them are in the control arm. Thus, the number of patients allocated to the treatment arm is $T_1=\pi Npn$, and the number of patients allocated to the control arm is $T_2=(1-\pi)Npn$. Furthermore, let the binary proportion for the treatment group be μ_t and for the control group be μ_c and assume that there is no interaction effect between treatment and center, i.e., the treatment works the same in every center.

Since randomization takes place at the second level, all \mathbf{X}_i 's are the same and all \mathbf{V}_{z_i} 's are the same across the N centers. Thus

$$\hat{\Sigma} = N \lim_{N \rightarrow \infty} \left(\sum_{i=1}^N \mathbf{X}_i^T \mathbf{V}_{z_i}^{-1} \mathbf{X}_i \right)^{-1} = \left(\mathbf{X}_i^T \mathbf{V}_{z_i}^{-1} \mathbf{X}_i \right)^{-1}$$

In this study design, the covariate matrix \mathbf{X}_i for the i^{th} center is a $pn \times 2$ matrix. The first column of this matrix contains all ones, whereas the second column contains ones in the first πpn rows, and zero in the remaining $(1-\pi)pn$ rows. The structure of \mathbf{V}_{z_i} can be written as

$$\mathbf{V}_{z_i} = \text{block}(\mathbf{A}, \mathbf{B}) + J_{np}(\sigma_c^2), \text{ where}$$

$$\mathbf{A} = I_{\pi p} \otimes \left(\frac{1}{\mu_t(1-\mu_t)} I_n + J_n(\sigma_p^2) \right)$$

$$\mathbf{B} = I_{(1-\pi)p} \otimes \left(\frac{1}{\mu_c(1-\mu_c)} I_n + J_n(\sigma_p^2) \right)$$

For example, consider a simple experiment in which the i^{th} center has two physicians, and each physician has two patients. Suppose the first physician in the center is randomized into the

treatment group and the second physician is randomized into the control group. The structures of \mathbf{X}_i and \mathbf{V}_{z_i} for this particular center i^{th} take the following form

$$\mathbf{X}_i = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{V}_{z_i} = \begin{bmatrix} \sigma_{treat}^2 & \sigma_c^2 + \sigma_p^2 & \sigma_c^2 & \sigma_c^2 \\ \sigma_c^2 + \sigma_p^2 & \sigma_{treat}^2 & \sigma_c^2 & \sigma_c^2 \\ \sigma_c^2 & \sigma_c^2 & \sigma_{control}^2 & \sigma_c^2 + \sigma_p^2 \\ \sigma_c^2 & \sigma_c^2 & \sigma_c^2 + \sigma_p^2 & \sigma_{control}^2 \end{bmatrix}$$

where $\sigma_{treat}^2 = \sigma_c^2 + \sigma_p^2 + \frac{1}{\mu_i(1-\mu_i)}$ and

$$\sigma_{control}^2 = \sigma_c^2 + \sigma_p^2 + \frac{1}{\mu_c(1-\mu_c)}$$

5.4 Randomize at Second Level with Interaction Effect

The distinction between randomization at second level with interaction and without interaction is that a new variance term representing the interaction between treatment and center need to be added to the model. Although the covariate matrix \mathbf{X}_i defined in Section 5.2 remains to be the same, the existence of the interaction term cause some changes in the structure of the variance matrix \mathbf{V}_{z_i} . Denote the variance of the interaction as σ_{ct}^2 , \mathbf{V}_{z_i} now takes the following form

$$\mathbf{V}_{z_i} = \text{block}(\mathbf{A}, \mathbf{B}) + J_{np}(\sigma_c^2), \text{ where}$$

$$\mathbf{A} = I_{\pi p} \otimes \left(\frac{1}{\mu_i(1-\mu_i)} I_n + J_n(\sigma_p^2) \right) + J_{\pi pn}(\sigma_{ct}^2)$$

$$\mathbf{B} = I_{(1-\pi)p} \otimes \left(\frac{1}{\mu_c(1-\mu_c)} I_n + J_n(\sigma_p^2) \right) + J_{np(1-\pi)}(\sigma_{ct}^2)$$

For example, consider a simple experiment in which the i^{th} center has four physicians, and each physician has two patients. Suppose the first two physicians in the center are randomized into the treatment group and the next two physicians are randomized into the control group. Furthermore, suppose that the treatment effect varies across different centers, then for the

$$i^{\text{th}} \text{ center, we can write } \mathbf{X}_i = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix}$$

and \mathbf{V}_{Z_i} can be written as

$$\mathbf{V}_{Z_i} = \begin{bmatrix} \sigma_{treat}^2 & \omega^2 & \nu^2 & \nu^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 \\ \omega^2 & \sigma_{treat}^2 & \nu^2 & \nu^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 \\ \nu^2 & \nu^2 & \sigma_{treat}^2 & \omega^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 \\ \nu^2 & \nu^2 & \omega^2 & \sigma_{treat}^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 \\ \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_{control}^2 & \omega^2 & \nu^2 & \nu^2 \\ \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \omega^2 & \sigma_{control}^2 & \nu^2 & \nu^2 \\ \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \nu^2 & \nu^2 & \sigma_{control}^2 & \omega^2 \\ \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \nu^2 & \nu^2 & \omega^2 & \sigma_{control}^2 \end{bmatrix}$$

Where

$$\sigma_{treat}^2 = \sigma_c^2 + \sigma_p^2 + \sigma_{ct}^2 + \frac{1}{\mu_t(1-\mu_t)}$$

$$\sigma_{control}^2 = \sigma_c^2 + \sigma_p^2 + \sigma_{ct}^2 + \frac{1}{\mu_c(1-\mu_c)}$$

$$\nu^2 = \sigma_c^2 + \sigma_{ct}^2$$

$$\omega^2 = \sigma_c^2 + \sigma_p^2 + \sigma_{ct}^2$$

5.5 Randomize at First Level without Interaction Effect

Assume that there is no interaction effect, i.e., the treatment works the same in every center and every physician. For a the j^{th} physician in the i^{th} center, suppose the patients are randomized such that πn patients are in the treatment arm and $(1-\pi)n$ patients are in the control arm. Thus, the total number of patients allocated to the treatment arm is $T_1=\pi Npn$, and the number of patients allocated to the control arm is $T_2=(1-\pi)Npn$. Furthermore, let the probability of an event in the treatment group be μ_t and in the control group be μ_c . Since randomization takes place at the first level, \mathbf{X}_i and \mathbf{V}_{z_i} are the same across all centers. Thus

$$\hat{\Sigma} = N \lim_{N \rightarrow \infty} \left(\sum_{i=1}^N \mathbf{X}_i^T \mathbf{V}_{z_i}^{-1} \mathbf{X}_i \right)^{-1} = \left(\mathbf{X}_i^T \mathbf{V}_{z_i}^{-1} \mathbf{X}_i \right)^{-1}$$

The covariate matrix \mathbf{X}_i for the i^{th} center is a $pn \times 2$ matrix. The first column of this matrix contains all ones, whereas the second column contains ones in the πpn rows corresponding to those patients allocated to the treatment group, and zero in the $(1-\pi)pn$ rows corresponding to those patients allocated to the control group.

The structure of \mathbf{V}_{z_i} can be written as

$$\mathbf{V}_{z_i} = \mathbf{I}_p \otimes \left[\text{block}(\mathbf{A}, \mathbf{B}) + \mathbf{J}_n (\sigma_c^2 + \sigma_p^2) \right] + \mathbf{J}_p \otimes \mathbf{C} - \mathbf{I}_p \otimes \mathbf{C}, \text{ where}$$

$$\mathbf{A} = \left(\frac{1}{\mu_t (1 - \mu_t)} \right) \mathbf{I}_{\pi n}$$

$$\mathbf{B} = \left(\frac{1}{\mu_c (1 - \mu_c)} \right) \mathbf{I}_{(1-\pi)n}$$

$$\mathbf{C} = \mathbf{J}_n (\sigma_c^2)$$

For example, suppose the i^{th} center has two physicians and each physician has four patients. Suppose the first two patients are randomized into the treatment group and the next two patients are randomized into the control group. Then, for the i^{th} center, we have

$$\mathbf{X}_i = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \end{bmatrix}$$

And the structure of \mathbf{V}_{Zi} is

$$\mathbf{V}_{Zi} = \begin{bmatrix} \sigma_{treat}^2 & \tau^2 & \tau^2 & \tau^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 \\ \tau^2 & \sigma_{treat}^2 & \tau^2 & \tau^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 \\ \tau^2 & \tau^2 & \sigma_{control}^2 & \tau^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 \\ \tau^2 & \tau^2 & \tau^2 & \sigma_{control}^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 \\ \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_{treat}^2 & \tau^2 & \tau^2 & \tau^2 \\ \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \tau^2 & \sigma_{treat}^2 & \tau^2 & \tau^2 \\ \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \tau^2 & \tau^2 & \sigma_{control}^2 & \tau^2 \\ \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \tau^2 & \tau^2 & \tau^2 & \sigma_{control}^2 \end{bmatrix}$$

Where

$$\sigma_{treat}^2 = \sigma_c^2 + \sigma_p^2 + \frac{1}{\mu_t(1-\mu_t)}$$

$$\sigma_{control}^2 = \sigma_c^2 + \sigma_p^2 + \frac{1}{\mu_c(1-\mu_c)}$$

$$\tau^2 = \sigma_c^2 + \sigma_p^2$$

5.6 Randomize at First Level with Interaction Effect

We consider two different scenarios where the interaction effects might be found: (1) interaction between treatment and level three and (2) interaction between treatment and level two.

5.6.1 The Structure of \mathbf{V}_{z_i} with Treatment \times Level Three Interaction

Under the same randomization scheme, what if the treatment effect varies across different centers? In this situation, an interaction effect between treatment and center should be taken into account.

Since the presence of the interaction effect does not lead to any change on covariate matrix \mathbf{X}_i , the difference in the computation of $\hat{\Sigma}$ relies solely on the modification in the structure of \mathbf{V}_{z_i} . This modification can be done by introducing a new variance term denoted by σ_{ct}^2 . \mathbf{V}_{z_i} now takes the following structure:

$$\mathbf{V}_{z_i} = \mathbf{I}_p \otimes \left[\text{block}(\mathbf{A}, \mathbf{B}) + \mathbf{J}_n (\sigma_c^2 + \sigma_p^2) \right] + \mathbf{J}_p \otimes \mathbf{C} - \mathbf{I}_p \otimes \mathbf{C}$$

$$\mathbf{A} = \left(\frac{1}{\mu_t (1 - \mu_t)} \right) \mathbf{I}_{\pi n} + \mathbf{J}_{\pi n} (\sigma_{ct}^2)$$

$$\mathbf{B} = \left(\frac{1}{\mu_c (1 - \mu_c)} \right) \mathbf{I}_{(1-\pi)n} + \mathbf{J}_{(1-\pi)n} (\sigma_{ct}^2)$$

$$\mathbf{C} = \text{block}(\mathbf{J}_{n\pi} (\sigma_c^2), \mathbf{J}_{n(1-\pi)} (\sigma_c^2)) + \mathbf{J}_n (\sigma_c^2)$$

For a simple example with two centers, each center has two physicians, and each physician has four patients: two are in the treatment group and two are in the control group. We then have:

$$\mathbf{X}_i = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} \text{ and } \mathbf{V}_{z_i} = \begin{bmatrix} \sigma_{treat}^2 & \omega^2 & \tau^2 & \tau^2 & \nu^2 & \nu^2 & \sigma_c^2 & \sigma_c^2 \\ \omega^2 & \sigma_{treat}^2 & \tau^2 & \tau^2 & \nu^2 & \nu^2 & \sigma_c^2 & \sigma_c^2 \\ \tau^2 & \tau^2 & \sigma_{control}^2 & \omega^2 & \sigma_c^2 & \sigma_c^2 & \nu^2 & \nu^2 \\ \tau^2 & \tau^2 & \omega^2 & \sigma_{control}^2 & \sigma_c^2 & \sigma_c^2 & \nu^2 & \nu^2 \\ \nu^2 & \nu^2 & \sigma_c^2 & \sigma_c^2 & \sigma_{treat}^2 & \omega^2 & \tau^2 & \tau^2 \\ \nu^2 & \nu^2 & \sigma_c^2 & \sigma_c^2 & \omega^2 & \sigma_{treat}^2 & \tau^2 & \tau^2 \\ \sigma_c^2 & \sigma_c^2 & \nu^2 & \nu^2 & \tau^2 & \tau^2 & \sigma_{control}^2 & \omega^2 \\ \sigma_c^2 & \sigma_c^2 & \nu^2 & \nu^2 & \tau^2 & \tau^2 & \omega^2 & \sigma_{control}^2 \end{bmatrix}$$

Where

$$\sigma_{treat}^2 = \sigma_c^2 + \sigma_p^2 + \sigma_{ct}^2 + \frac{1}{\mu_t(1-\mu_t)}$$

$$\sigma_{control}^2 = \sigma_c^2 + \sigma_p^2 + \sigma_{ct}^2 + \frac{1}{\mu_c(1-\mu_c)}$$

$$\tau^2 = \sigma_c^2 + \sigma_p^2$$

$$\nu^2 = \sigma_c^2 + \sigma_{ct}^2$$

$$\omega^2 = \sigma_c^2 + \sigma_p^2 + \sigma_{ct}^2$$

5.6.2 The Structure of \mathbf{V}_{z_i} with Treatment \times Level Two Interaction

Now, let us consider another scenario when the treatment effect varies across different physicians. In this scenario, an interaction effect between treatment and physician is required.

Again, the presence of the interaction effect does not lead to any change on covariate matrix \mathbf{X}_i and the difference in the computation of $\hat{\Sigma}$ only depends on the modification in the structure of \mathbf{V}_{z_i} . This modification is done by the introduction of the variance of the interaction term reflecting the interaction between physician and treatment σ_{pt}^2 . \mathbf{V}_{z_i} now takes the following

structure:

$$\mathbf{V}_{\mathbf{Z}_i} = \mathbf{I}_p \otimes \left[\text{block}(\mathbf{A}, \mathbf{B}) + \mathbf{J}_n (\sigma_c^2 + \sigma_p^2) \right] + \mathbf{J}_p \otimes \mathbf{C} - \mathbf{I}_p \otimes \mathbf{C}$$

$$\mathbf{A} = \left(\frac{1}{\mu_t (1 - \mu_t)} \right) \mathbf{I}_{\pi n} + \mathbf{J}_{\pi n} (\sigma_{pt}^2)$$

$$\mathbf{B} = \left(\frac{1}{\mu_c (1 - \mu_c)} \right) \mathbf{I}_{(1-\pi)n} + \mathbf{J}_{(1-\pi)n} (\sigma_{pt}^2)$$

$$\mathbf{C} = \mathbf{J}_n (\sigma_c^2)$$

For a simple example with two centers, each center has two physicians, and each physician has four patients: two are in the treatment group and two are in the control group. We then have

$$\mathbf{X}_i = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} \text{ and } \mathbf{V}_{\mathbf{Z}_i} = \begin{bmatrix} \sigma_{treat}^2 & \mathcal{G}^2 & \tau^2 & \tau^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 \\ \mathcal{G}^2 & \sigma_{treat}^2 & \tau^2 & \tau^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 \\ \tau^2 & \tau^2 & \sigma_{control}^2 & \mathcal{G}^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 \\ \tau^2 & \tau^2 & \mathcal{G}^2 & \sigma_{control}^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 \\ \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_{treat}^2 & \mathcal{G}^2 & \tau^2 & \tau^2 \\ \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \mathcal{G}^2 & \sigma_{treat}^2 & \tau^2 & \tau^2 \\ \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \tau^2 & \tau^2 & \sigma_{control}^2 & \mathcal{G}^2 \\ \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \tau^2 & \tau^2 & \mathcal{G}^2 & \sigma_{control}^2 \end{bmatrix}$$

where

$$\sigma_{treat}^2 = \sigma_c^2 + \sigma_p^2 + \sigma_{pt}^2 + \frac{1}{\mu_t (1 - \mu_t)}$$

$$\sigma_{control}^2 = \sigma_c^2 + \sigma_p^2 + \sigma_{pt}^2 + \frac{1}{\mu_c (1 - \mu_c)}$$

$$\tau^2 = \sigma_c^2 + \sigma_p^2$$

$$\mathcal{G}^2 = \sigma_c^2 + \sigma_p^2 + \sigma_{pt}^2$$

Chapter 6

Simulation Study

As presented in Chapters 4 and 5, our power and sample size calculations are based on asymptotic theory, particularly the derivation of the variance of the treatment effect. It is reasonable to suspect that a minor variation in the approach proposed for the intervention evaluation can cause substantial difference in power and sample size estimates. To assess the accuracy of our methods, we conducted a simulation study designed to verify sample size and power calculations. The results showed that the theoretical power estimates from our formulas are consistent to the empirical power computed from the simulated data.

This chapter explains the process that was used to generate the data for the different designs presented in Chapter 4 and Chapter 5. The simulation results will be then presented and discussed.

6.1 Simulation Design

We simulated data for cluster randomized trials with three-level design under the assumption of balanced sample sizes, i.e. the numbers of subjects per each cluster are equal across all clusters. For example, if the trial includes patient nested within physician and physician nested within center, then the numbers of patients for each physician are the same and the numbers of physicians for each center are the same across all centers. In all the settings, we assumed a completely randomized design that involved two arms: a treatment arm and a control arm. We considered six different scenarios for each type of outcomes: continuous and binary.

The six scenarios are dictated by which level the allocation of treatment takes place and whether there is an interaction effect or not. They include (1) randomize at level three, (2) randomize at level two without interaction, (3) randomize at level two with interaction, (4) randomize at level one without interaction, (5) randomize at level one with interaction between treatment and level three, and (6) randomize at level one with interaction between treatment and level two.

To verify the accuracy of the derived power and sample size functions, we compared the theoretical power given by our proposed formulas (denote by \mathcal{T}) to the empirical power given by the simulated data (denote by $\tilde{\mathcal{T}}$). The theoretical power was computed using SAS PROC IML, which specifies the forms of the variance matrices and yields the power based on our theory. The empirical power was obtained by fitting the simulated data sets with linear mixed models using SAS PROC MIXED for continuous outcome, and generalized linear mixed models using SAS PROC GLIMMIX for binary outcome. The p-values for testing the null hypothesis of no treatment effect were retained after each the model fitting process. Denote the p-value by p_r for the r^{th} repetition of the simulated data with a total of R repetitions, the empirical power $\tilde{\mathcal{T}}$ can be calculated by

$$\tilde{\mathcal{T}} = \frac{\sum_r (p_r < \alpha)}{R}$$

For comparison, we estimated the 95% confidence intervals for the empirical power and recorded the number of iterations where the confidence interval covered the theoretical power. In addition, the absolute difference between the two values of powers was also computed in order to assess the discrepancy between the theoretical power and the empirical power. The simulation methods are slightly different between the continuous outcome and the binary outcome. We will discuss the simulation processes for the two types of outcomes separately.

6.2 Simulation for Continuous Data

For continuous outcomes, the following parameters were pre-specified:

- The desired theoretical power \mathcal{T}
- The two-sided significance level α
- The proportion of allocation into treatment π
- The variances between level three σ_c^2 , the variance between level two σ_p^2 , and the variance between level one σ_e^2 .
- The variances of the interaction effect between treatment and level three σ_{ct}^2 or between treatment and level two σ_{pt}^2 (if the interaction at these levels exist).

The following parameters were allowed to vary

- The intervention effect μ_{int} and control effect μ_{ctrl}
- The sample sizes in each level c , p , and n

Our simulation programs were designed in such a way that the results can be collected for any chosen combination of the above parameters. In this report, we chose to fix the desired power $\mathcal{T} = 0.75$ and $\alpha = 0.05$. However, these two parameters can be set at different values.

To chose the values for the ICC, we looked at recent publications on sample size estimation for three-level designs in healthcare experiments. Our review suggested that the intraclass correlation on level three (ρ) is much smaller than that of level two (r). For example, in a recent study conducted by Heo and Leon (2009), the researchers examined a combination of $\rho = 0.01, 0.05, 0.10$ and $r = 0.4, 0.5, 0.6$. Taking the first combination of ρ and r , together

with a total variance of $\sigma_T^2 = 1$, our pre-specified variance parameters were $\sigma_c^2 = 0.01$, $\sigma_p^2 = 0.39$, $\sigma_e^2 = 0.60$, $\sigma_{ct}^2 = 0.05$, $\sigma_{pt}^2 = 0.02$.

The differences between intervention and control effects were derived from the corresponding sample sizes and the desired power of 0.75 from our theoretical formulas. Since the power is set at a fixed value and the variances are also set at the pre-specified values, the difference in treatment effect is a function of the variances and the selected power.

The sample sizes in each level were varied amongst large, medium, and small sizes. These sample sizes again were chosen based on common sample sizes reported from the literature of CRT. More specifically, we considered the following:

- Sample sizes for level three $c=10, 20, 30$
- Sample sizes for level two: $p=4, 8, 12$
- Sample sizes for level one $n=10, 20, 30$

This $3 \times 3 \times 3$ factorial design yielded a total of 27 combinations for each set of parameters. In order to allow for a workable compromise between the margin of error and the number of repetitions, we chose a margin of error of 0.035. With this, we arrived at a total number of repetitions of 784 data sets for each combination. Thus, altogether we had 21,168 sets of data across the six different scenarios with continuous outcomes.

For each different scenario, we performed different steps to simulate the corresponding data sets. Details of the steps as well as the simulation algorithms will be described in the next paragraphs.

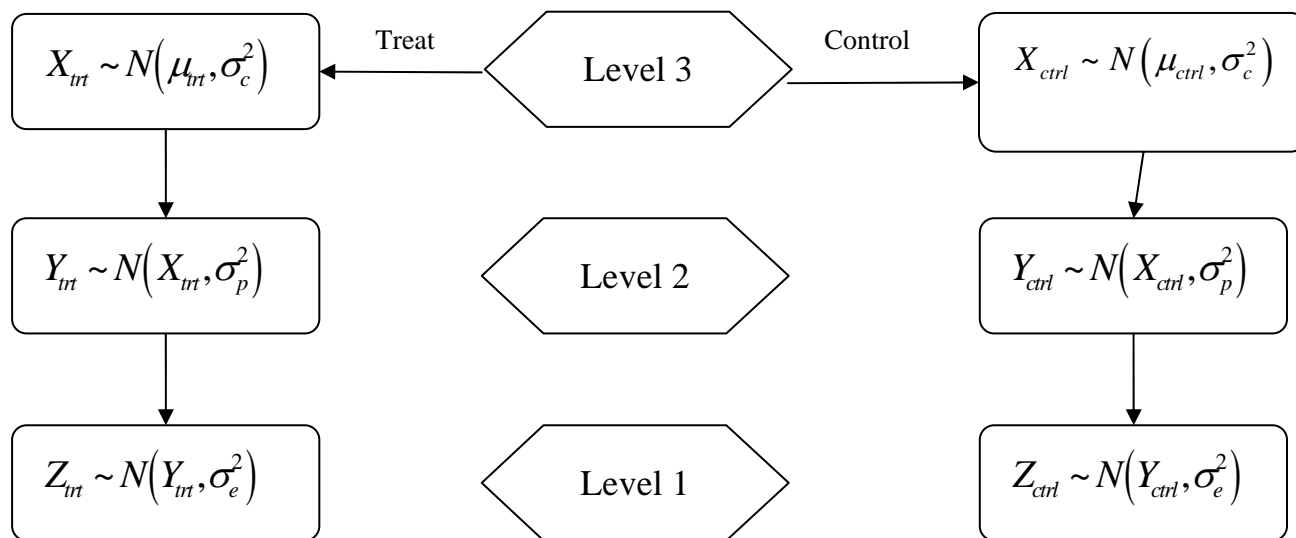
Case 1: Randomization at Level Three—Continuous Outcome

- Generate $X_i \sim N(\mu_{trt}, \sigma_c^2)$ independently, where $i=1,2,\dots, \pi c$ for the treatment
- Generate $X_i \sim N(\mu_{ctrl}, \sigma_c^2)$ independently, where $i=\pi c + 1, \pi c + 2, \dots, c$ for the control
- For each X_i , generate $Y_j \sim N(X_i, \sigma_p^2)$ independently, where $j=1,2,\dots, p$
- For each Y_j , generate $Z_k \sim N(Y_j, \sigma_e^2)$ independently, where $k=1,2,\dots, n$

The simulation algorithm for case 1 is illustrated in Figure 6.2.1

Figure 6.2.1: Simulation Algorithm

Randomize at Level Three—Continuous Outcome



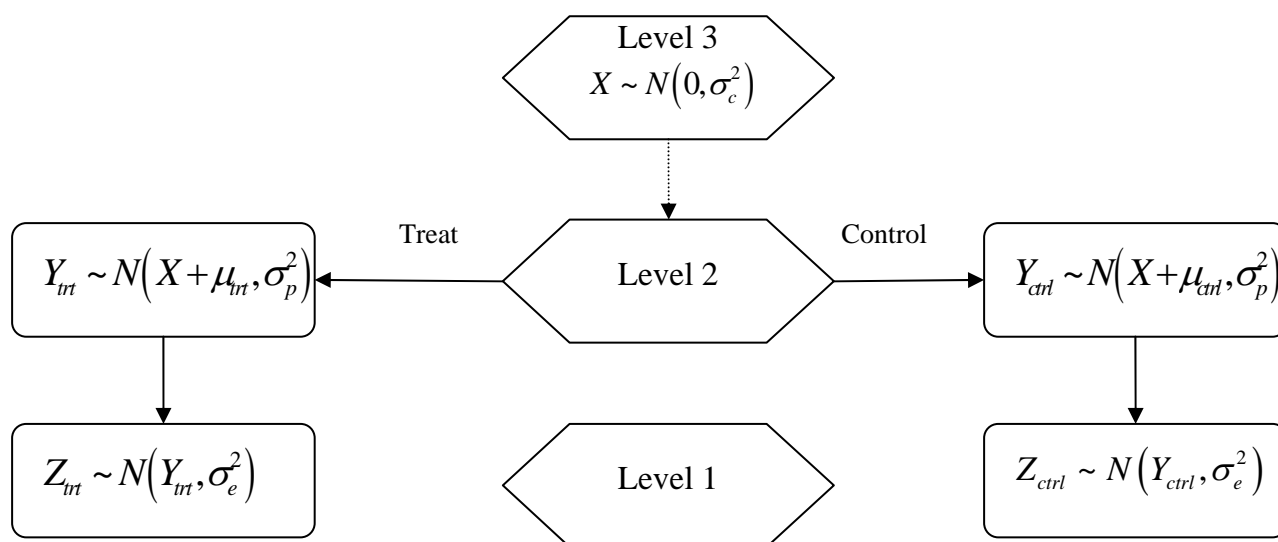
Case 2: Randomize at Level Two without Interaction—Continuous Outcome

- Generate $X_i \sim N(0, \sigma_c^2)$
- For each X_i , generate $Y_j \sim N(X_i + \mu_{tr}, \sigma_p^2)$ independently, where $j=1,2,\dots, \pi p$ for the treatment group and generate $Y_j \sim N(X_i + \mu_{ctrl}, \sigma_p^2)$ independently, where $j=\pi p+1, \pi p+2, \dots, p$ in the control group
- For each Y_j , generate $Z_k \sim N(Y_j, \sigma_e^2)$ independently, where $k=1,2,\dots,n$

The simulation algorithm for the case 2 is described in Figure 6.2.2

Figure 6.2.2: Simulation Algorithm

Randomize at Level Two without Interaction –Continuous Outcome



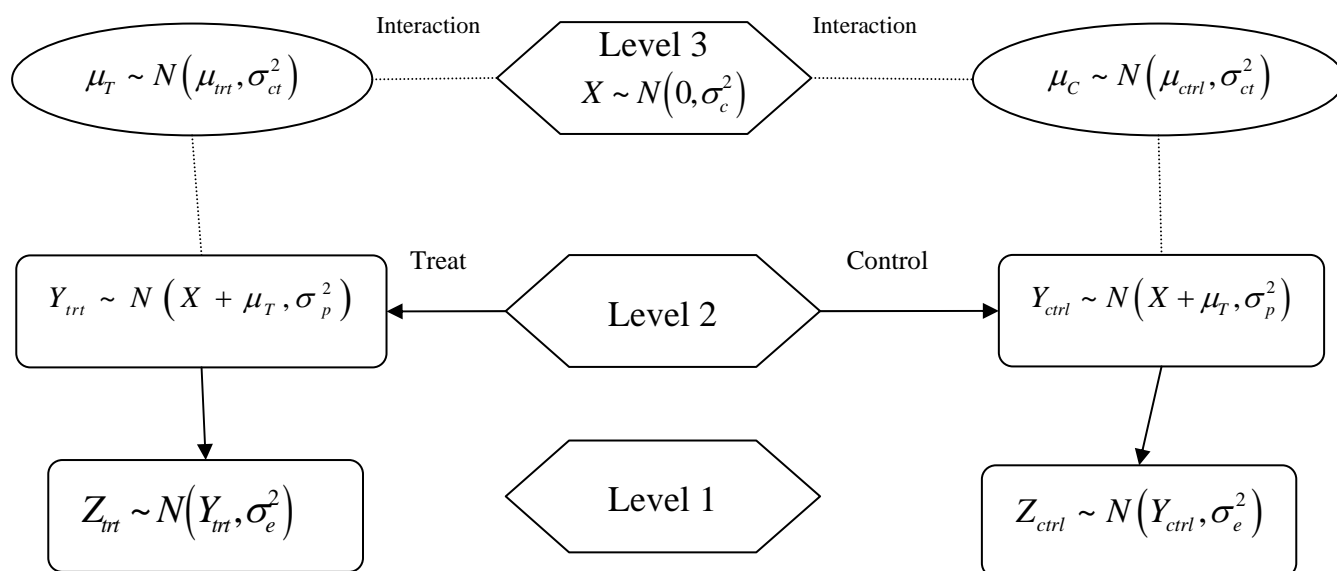
Case 3: Randomization at Level Two with Interaction—Continuous Outcome

- Generate $X_i \sim N(0, \sigma_c^2)$ independently
- Generate the treatment effects $\mu_T \sim N(\mu_{trt}, \sigma_{ct}^2)$ for the treatment group and $\mu_C \sim N(\mu_{ctrl}, \sigma_{ct}^2)$ for the control group.
- For each X_i , generate $Y_j \sim N(X_i + \mu_T, \sigma_p^2)$ independently, where $j=1,2,\dots, \pi p$ for the treatment group and generate $Y_j \sim N(X_i + \mu_C, \sigma_p^2)$ independently, where $j=\pi p+1, \pi p+2, \dots, p$ for the control group
- For each Y_j , generate $Z_k \sim N(Y_j, \sigma_e^2)$ independently, where $k=1,2,\dots,n$

The simulation algorithm for the case 3 is described in Figure 6.2.3

Figure 6.2.3: Simulation Algorithm

Randomize at Level Two with Interaction –Continuous Outcome



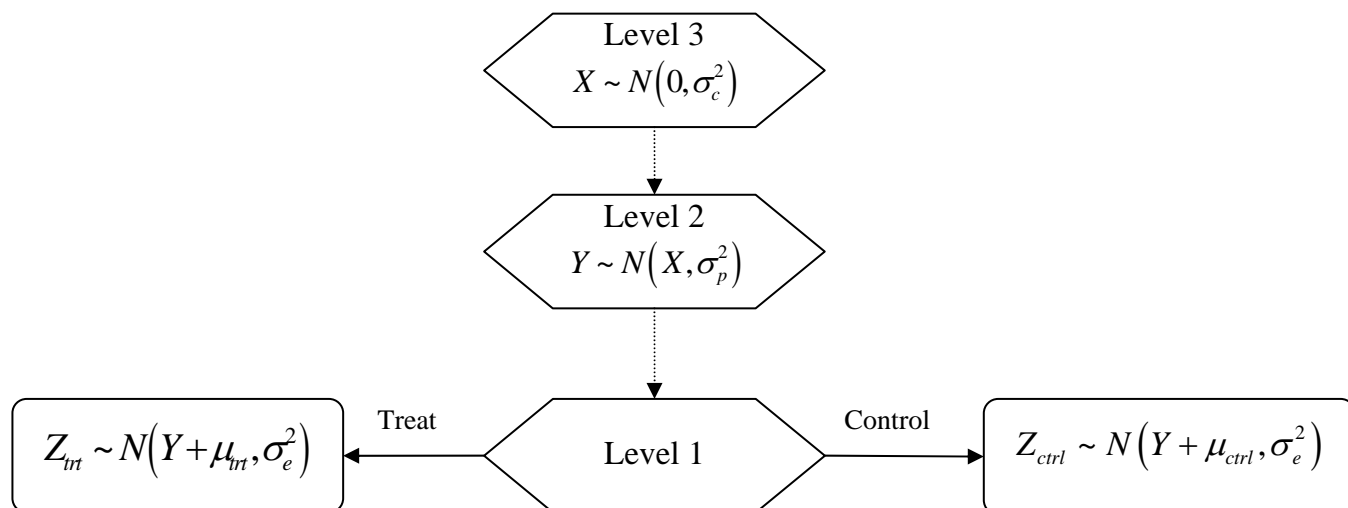
Case 4: Randomize at Level One without Interaction—Continuous Outcome

- Generate $X_i \sim N(0, \sigma_c^2)$
- For each X_i , generate $Y_j \sim N(X_i, \sigma_p^2)$ independently, where $j=1,2,\dots,p$
- For each Y_j , generate $Z_k \sim N(Y_j + \mu_{trt}, \sigma_e^2)$ independently, where $k=1,2,\dots, \pi n$ for the treatment group and generate $Z_k \sim N(Y_j + \mu_{ctrl}, \sigma_e^2)$ independently, where $k=\pi n+1, \pi n+2, \dots, n$ in the control group

The simulation algorithm for the case 4 is described in Figure 6.2.4

Figure 6.2.4: Simulation Algorithm

Randomize at Level One without Interaction –Continuous Outcome



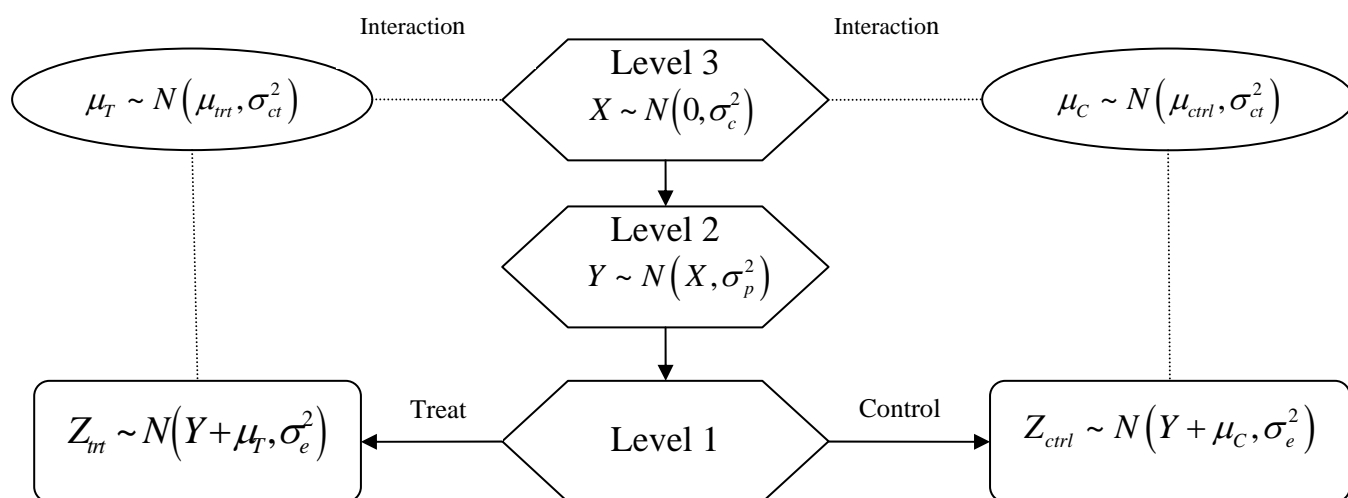
Case 5: Randomize at Level One with Treatment \times Level Three—Continuous Outcome

- Generate $X_i \sim N(0, \sigma_c^2)$ independently
- Generate the treatment effects $\mu_T \sim N(\mu_{tr}, \sigma_{ct}^2)$ for the treatment group and $\mu_C \sim N(\mu_{ctrl}, \sigma_{ct}^2)$ for the control group.
- For each X_i , generate $Y_j \sim N(X_i, \sigma_p^2)$ independently, where $j=1,2,\dots,p$
- For each Y_j , generate $Z_k \sim N(Y_j + \mu_T, \sigma_e^2)$ independently, where $k=1,2,\dots, \pi n$ for the treatment group and $Z_k \sim N(Y_j + \mu_C, \sigma_e^2)$ independently, where $k=\pi n+1, \pi n+2, \dots, n$ in the control group

The simulation algorithm for the case 5 is described in Figure 6.2.5

Figure 6.2.5: Simulation Algorithm

Randomize at Level One with Treatment \times Level Three –Continuous Outcome



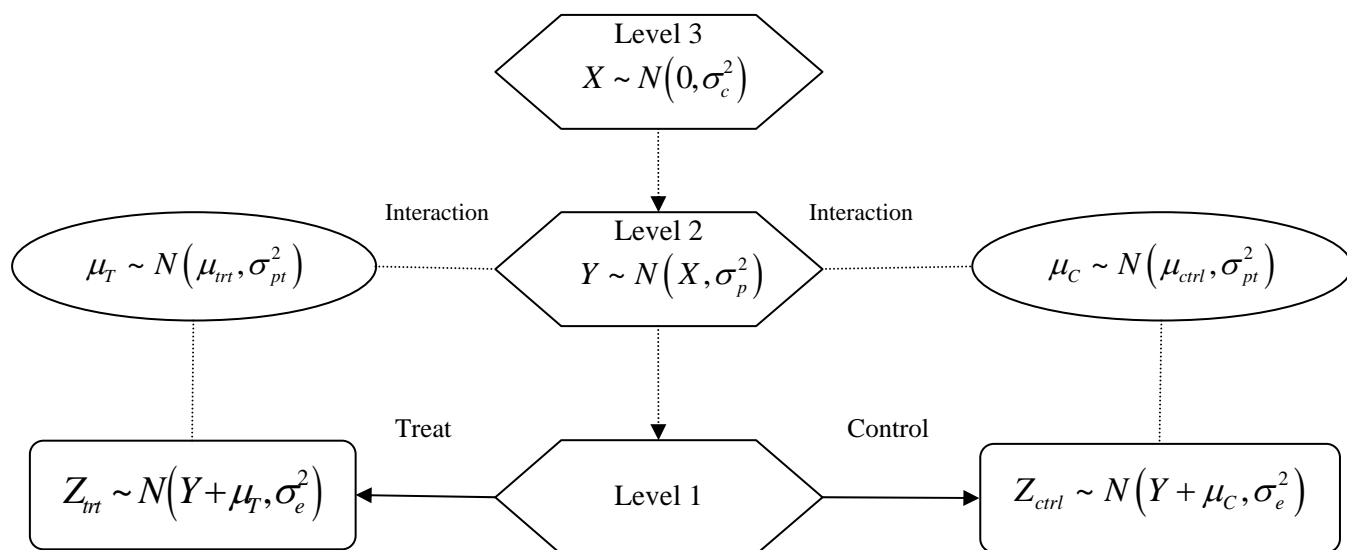
Case 6: Randomize at Level One with Treatment \times Level Two—Continuous Outcome

- Generate $X_i \sim N(0, \sigma_c^2)$ independently
- For each X_i , generate $Y_j \sim N(X_i, \sigma_p^2)$ independently, where $j=1,2,\dots,p$
- Generate the treatment effects $\mu_T \sim N(\mu_{trt}, \sigma_{pt}^2)$ for the treatment group and $\mu_C \sim N(\mu_{ctrl}, \sigma_{pt}^2)$ for the control group.
- For each Y_j , generate $Z_k \sim N(Y_j + \mu_T, \sigma_e^2)$ independently, where $k=1,2,\dots, \pi n$ for the treatment group and generate $Z_k \sim N(Y_j + \mu_C, \sigma_e^2)$ independently, where $k=\pi n+1, \pi n+2, \dots, n$ in the control group

The simulation algorithm for the case 6 is described in Figure 6.2.6

Figure 6.2.6: Simulation Algorithm

Randomize at Level One with Treatment \times Level Two –Continuous Outcome



6.3 Simulation for Binary Data

Similar to the continuous cases, the following parameters were pre-specified for binary data:

- The desired theoretical power \mathcal{T}
- The two-sided significance level α
- The proportion of allocation into treatment π
- The variances between level three σ_c^2 , the variance between level two σ_p^2
- The variances of the interaction effect between treatment and level three σ_{ct}^2 or between treatment and level two σ_{pt}^2 (if the interaction at these levels exist).

The following parameters were allowed to vary

- The probability of an event μ_t in the treatment group and μ_c in the control group
- The sample sizes in each level c, p, and n

Again, we chose to fix the desired power $\mathcal{T} = 0.75$ and $\alpha = 0.05$. The pre-specified variance parameters were $\sigma_c^2 = 0.01$, $\sigma_p^2 = 0.39$, $\sigma_{ct}^2 = 0.05$, $\sigma_{pt}^2 = 0.02$. Setting the probability of the event in the control group as $\mu_c = 0.7$, the differences between intervention and control probabilities were computed from the corresponding sample sizes and the desired power of 0.75 given by the theoretical formulas. The sample sizes in each level were varied amongst large, medium, and small sizes. We considered $c=10, 20, 30$; $p=4, 8, 12$; and $n=10, 20, 30$, which yielded a total of 27 combinations for each set of parameters. For a margin of error of 0.035, we arrived at a total number of repetitions of 784 data sets for each combination. Thus, altogether we had 21,168 sets of data for across the six different scenarios with binary outcome.

Following the general linear mixed models approach, we assumed the random effects in level three and level two are normally (Gaussian) distributed. To generate binary data, we

followed a model that linearly relates level one to the higher level cluster effects by way of the logit of the probabilities. Details for the six cases are described below.

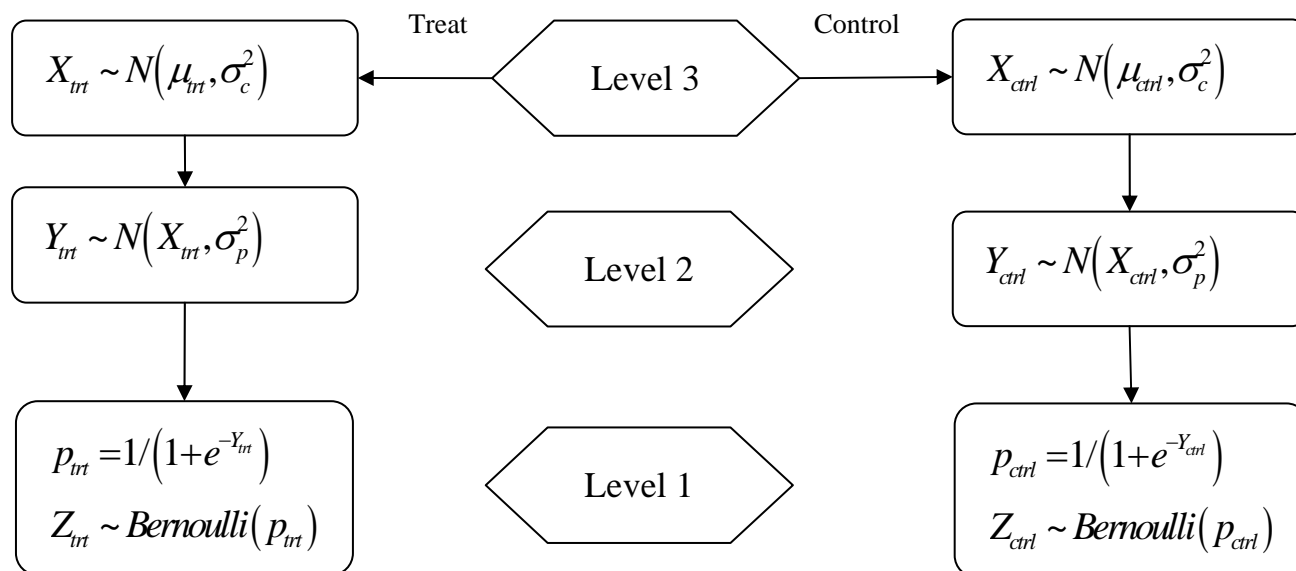
Case 7: Randomize at Level Three—Binary Outcome

- Generate $X_i \sim N(\mu_{trt}, \sigma_c^2)$ independently, where $i=1,2,\dots, \pi c$ for the treatment group and $X_i \sim N(\mu_{ctrl}, \sigma_c^2)$ independently, where $i=\pi c+1, \pi c+2, \dots, c$ for the control group
- For each X_i , generate $Y_j \sim N(X_i, \sigma_p^2)$ independently, where $j=1,2,\dots, p$
- For each Y_j , compute $p_j = 1/(1+e^{-Y_j})$ and generate $Z_k \sim \text{Bernoulli}(p_j)$ independently, where $k=1,2,\dots, n$

The simulation algorithm for case 7 is illustrated in Figure 6.3.1

Figure 6.3.1: Simulation Algorithm

Randomize at Level Three—Binary Outcome



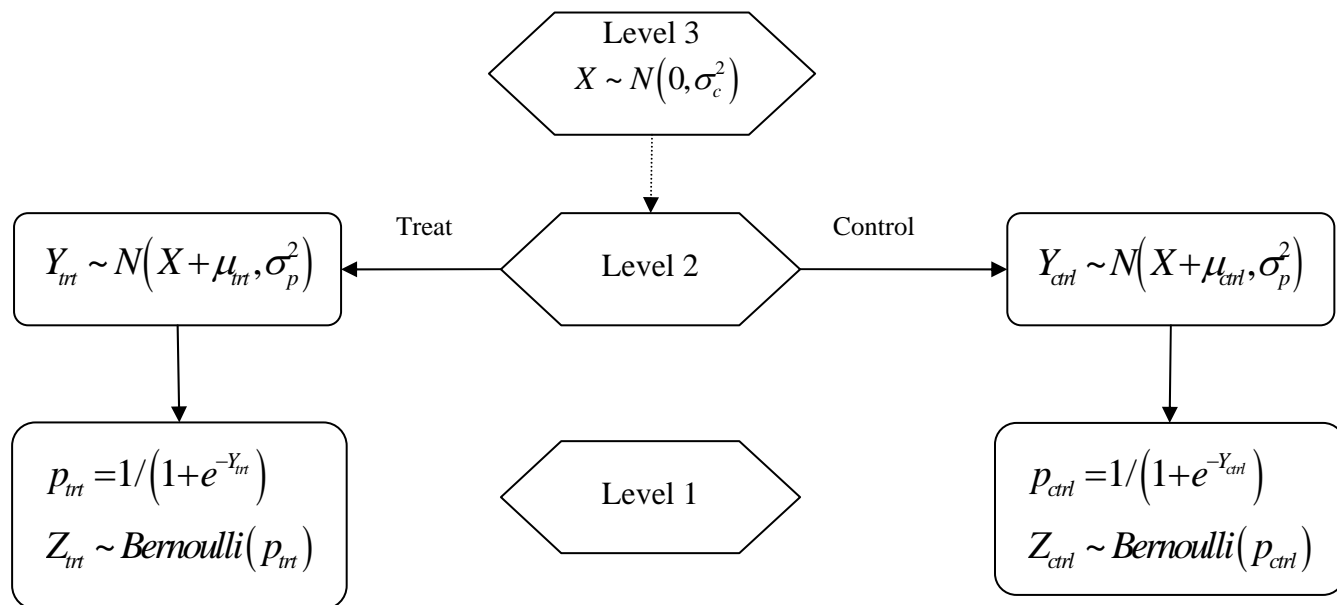
Case 8: Randomize at Level Two without Interaction—Binary Outcome

- Generate $X_i \sim N(0, \sigma_c^2)$ independently
- For each X_i , generate $Y_j \sim N(X_i + \mu_{tr}, \sigma_p^2)$ independently, where $j=1,2,\dots, \pi p$ for the treatment group and generate $Y_j \sim N(X_i + \mu_{ctrl}, \sigma_p^2)$ independently, where $j=\pi p+1, \pi p+2, \dots, p$ in the control group
- For each Y_j , compute $p_j = 1/(1+e^{-Y_j})$ and generate $Z_k \sim \text{Bernoulli}(p_j)$ independently, where $k=1,2,\dots,n$

The simulation algorithm for case 8 is described in Figure 6.3.2

Figure 6.3.2: Simulation Algorithm

Randomize at Level 2 without Interaction –Binary Outcome



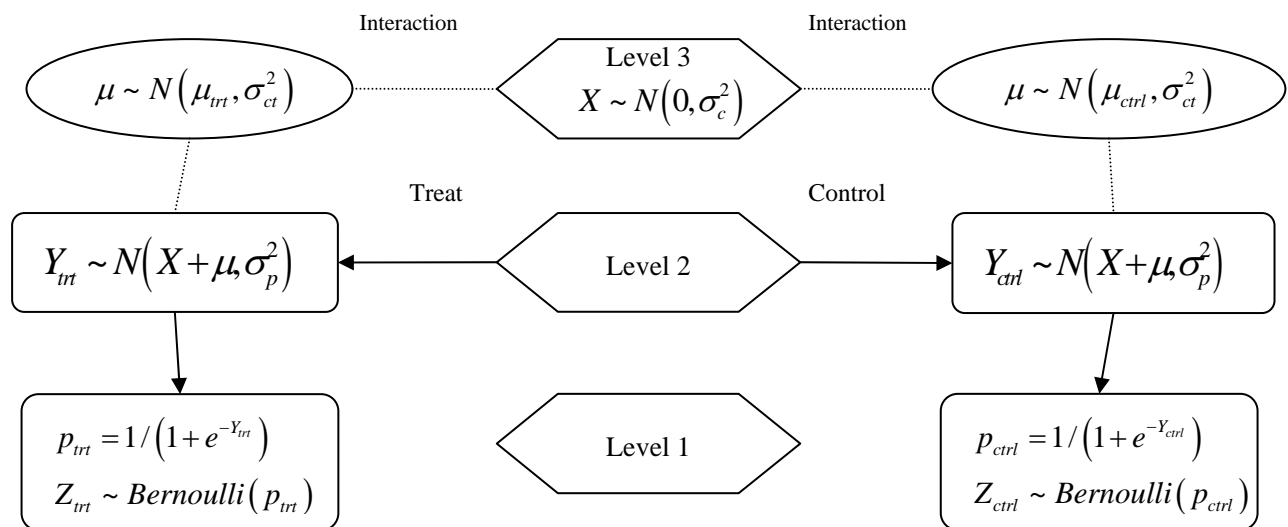
Case 9: Randomize at Level Two with Interaction—Binary Outcome

- Generate $X_i \sim N(0, \sigma_c^2)$ independently
- Generate the treatment effects $\mu_T \sim N(\mu_{tr}, \sigma_{ct}^2)$ for the treatment group and $\mu_C \sim N(\mu_{ctrl}, \sigma_{ct}^2)$ for the control group.
- For each X_i , generate $Y_j \sim N(X_i + \mu_T, \sigma_p^2)$ independently, where $j=1,2,\dots, \pi p$ for the treatment group and generate $Y_j \sim N(X_i + \mu_C, \sigma_p^2)$ independently, where $j=\pi p+1, \pi p+2, \dots, p$ for the control group
- For each Y_j , compute $p_j = 1/(1 + e^{-Y_j})$ and generate $Z_k \sim \text{Bernoulli}(p_j)$ independently, where $k=1,2,\dots,n$

The simulation algorithm for the case 9 is described in Figure 6.3.3

Figure 6.3.3: Simulation Algorithm

Randomize at Level 2 with Interaction –Binary Outcome



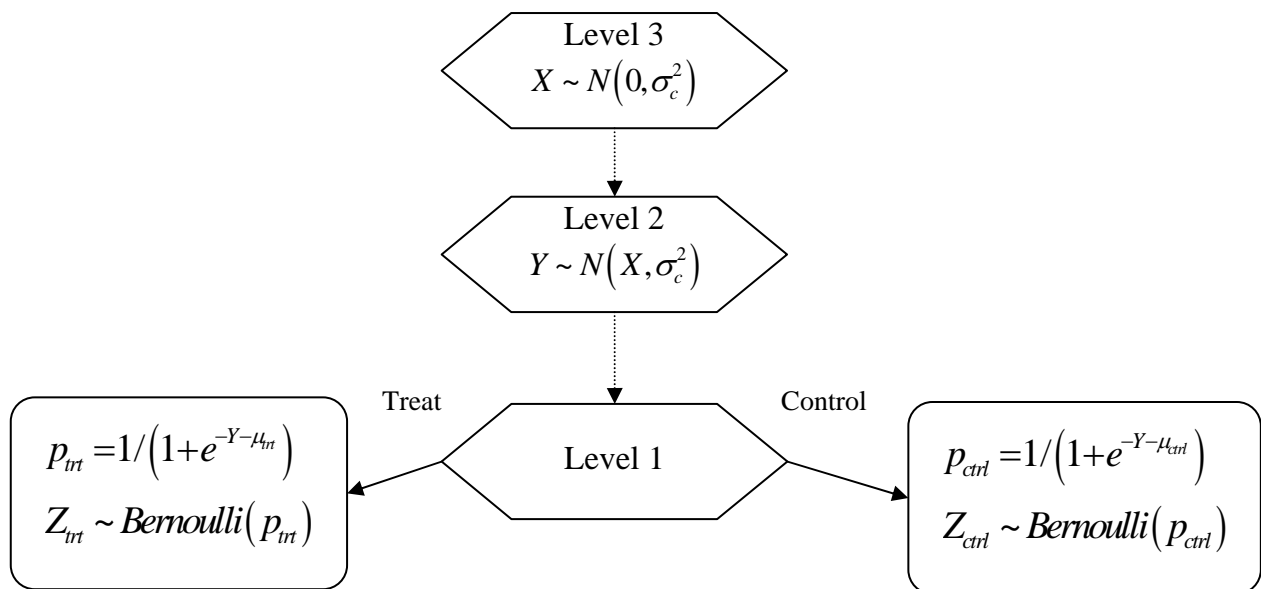
Case 10: Randomize at Level One without Interaction—Binary Outcome

- Generate $X_i \sim N(0, \sigma_c^2)$ independently
- For each X_i , generate $Y_j \sim N(X, \sigma_p^2)$ independently, where $j=1,2,\dots,p$
- For each Y_j , compute $p_{trt} = 1/(1 + e^{-Y_j - \mu_{trt}})$ and $p_{ctrl} = 1/(1 + e^{-Y_j - \mu_{ctrl}})$
- Generate $Z_{trt} \sim \text{Bernoulli}(p_{trt})$ independently, where $k=1,2,\dots, \pi n$ for the treatment group and generate $Z_{ctrl} \sim \text{Bernoulli}(p_{ctrl})$ independently, where $k=\pi n+1, \pi n+2, \dots, n$ in the control group.

The simulation algorithm for case 10 is described in Figure 6.3.4

Figure 6.3.4: Simulation Algorithm

Randomize at Level 1 without Interaction –Binary Outcome



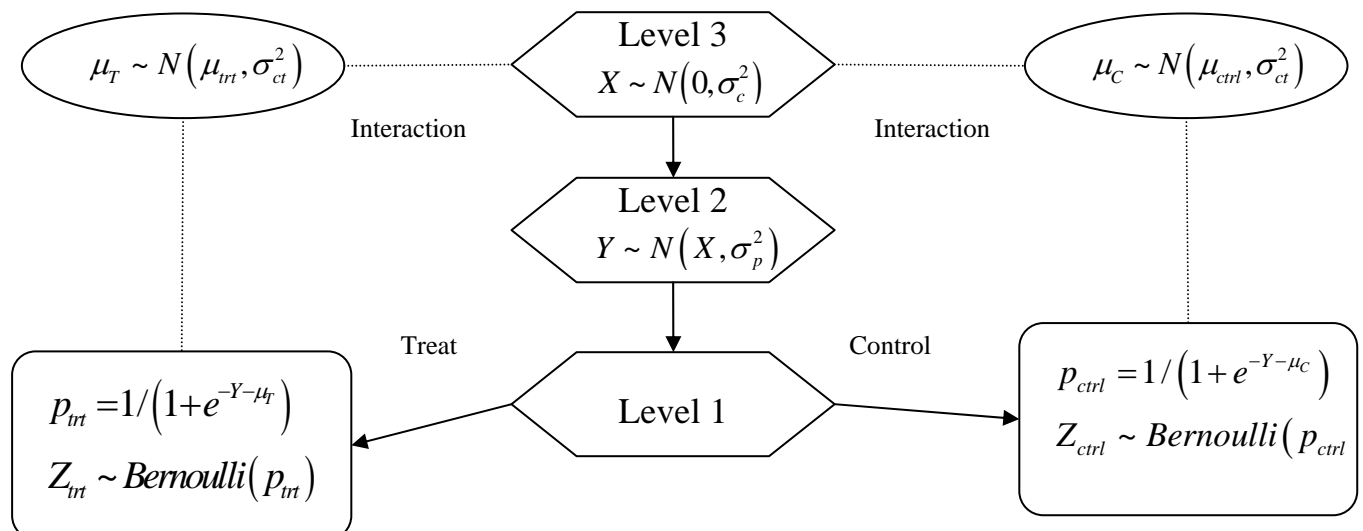
Case 11: Randomize at Level One with Treatment \times Level Three—Binary Outcome

- Generate $X_i \sim N(0, \sigma_c^2)$ independently
- Generate the treatment effects $\mu_T \sim N(\mu_{trt}, \sigma_{ct}^2)$ for the treatment group and $\mu_C \sim N(\mu_{ctrl}, \sigma_{ct}^2)$ for the control group.
- For each X_i , generate $Y_j \sim N(X, \sigma_p^2)$ independently, where $j=1,2,\dots,p$
- For each Y_j , compute $p_{trt} = 1/(1 + e^{-Y_j - \mu_T})$ and $p_{ctrl} = 1/(1 + e^{-Y_j - \mu_C})$
- Generate $Z_{trt} \sim \text{Bernoulli}(p_{trt})$ independently, where $k=1,2,\dots, \pi n$ for the treatment group and $Z_{ctrl} \sim \text{Bernoulli}(p_{ctrl})$ independently, where $k=\pi n+1, \pi n+2, \dots, n$ in the control group

The simulation algorithm for case 11 is described in Figure 6.3.5

Figure 6.3.5: Simulation Algorithm

Randomize at Level One with Treatment \times Level Three –Binary Outcome



Case 12: Randomize at Level One with Treatment \times Level Two—Binary Outcome

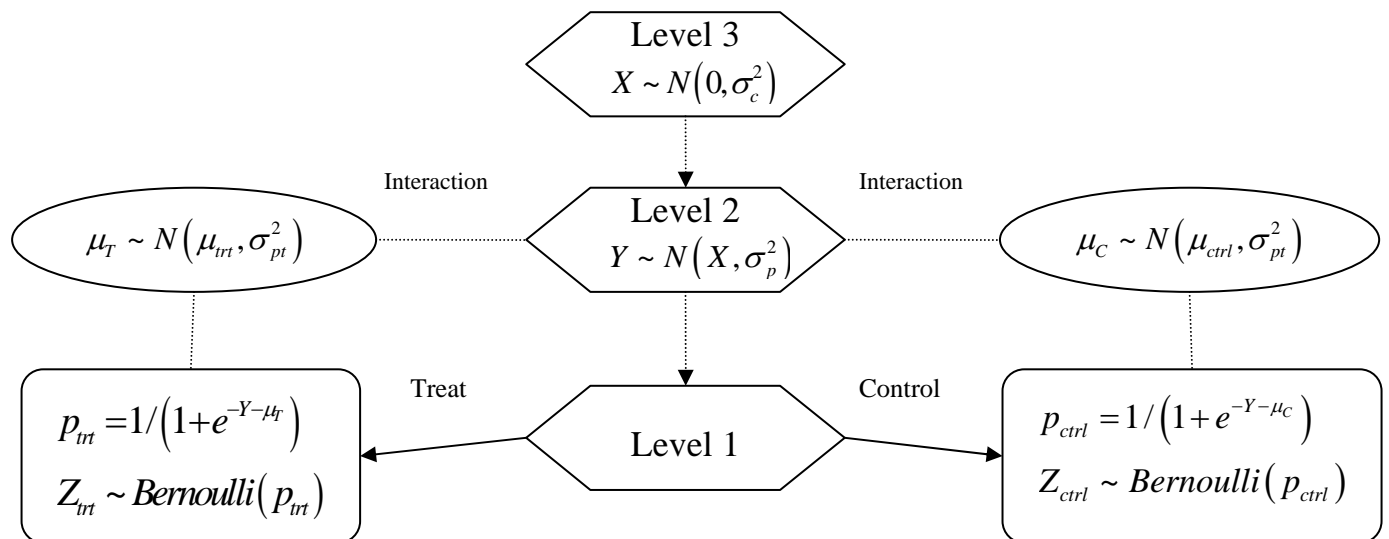
We performed the following:

- Generate $X_i \sim N(0, \sigma_c^2)$ independently
- For each X_i , generate $Y_j \sim N(X, \sigma_p^2)$ independently, where $j=1,2,\dots,p$
- Generate the treatment effects $\mu_T \sim N(\mu_{trt}, \sigma_{pt}^2)$ for the treatment group and $\mu_C \sim N(\mu_{ctrl}, \sigma_{pt}^2)$ for the control group.
- For each Y_j , compute $p_{trt} = 1/(1+e^{-Y_j-\mu_T})$ and $p_{ctrl} = 1/(1+e^{-Y_j-\mu_C})$
- Generate $Z_{trt} \sim Bernoulli(p_{trt})$ independently, where $k=1,2,\dots, \pi n$ for the treatment group and $Z_{ctrl} \sim Bernoulli(p_{ctrl})$ independently, where $k=\pi n+1, \pi n+2,\dots, n$ in the control group

The simulation algorithm for case 12 is described in Figure 6.3.6

Figure 6.3.6: Simulation Algorithm

Randomize at Level One with Treatment \times Level Two –Binary Outcome



6.4 Simulation Results

The main question addressed by the simulation study was that given a set of parameters, how well does the power from the derived sample size formulas agree with the empirical power generated by the simulation experiment. To answer this question, we examine the following results.

6.4.1 Simulation Results for Continuous Outcome

Table 6.4.1 shows the mean, minimum, maximum, and the standard errors of the empirical powers where results are combined for the six scenarios for continuous outcome.

Table 6.4.1: Summary of Simulation Results for Continuous Outcome

Design scenarios	Level 3	Level 2	Level 2 with Interaction	Level 1	Level 1 with Treat × Level 3	Level 1 with Treat × Level 2
Mean	0.75	0.75	0.72	0.75	0.73	0.73
Min	0.67	0.70	0.65	0.67	0.68	0.69
Max	0.85	0.79	0.77	0.82	0.76	0.77
SE	0.0065	0.0051	0.0061	0.0055	0.0042	0.0039

Tables B.1 to B.6 in Appendix B present in more detail the specified sample sizes for the three levels, the empirical power $\tilde{\mathcal{P}}$, and the absolute differences between the empirical powers and the theoretical powers for the continuous case. These tables show that in most of the cases the values of theoretical power are consistent with those of the empirical power regardless of the sample size combination.

It is noticed that in the cases with interaction, although the empirical powers were underestimated, most would consider the difference negligible. The rational for this is that Proc

MIXED tends to overestimate the interaction variance components, which in turn lowers the power because the estimated variance of the difference is larger.

Overall, the results for continuous data showed that the empirical powers varied around the theoretical powers at all sample sizes combination. In all combinations of the sample sizes considered, the theoretical the powers were covered by the 95% confidence intervals of the empirical powers in 93.8% (152/162) of the cases. This result is consistent for cases with and without interaction. The absolute differences between the empirical powers and the theoretical power were fairly small, with an average of 0.025 (SE=0.0284). Thus, the results for continuous data indicate that the power based on our derived formulas is nearly identical to the empirical power based on the simulated data.

6.4.2 Simulation Results for Binary Outcome

Table 6.4.2 displays the mean, the minimum, and the maximum, and the standard errors of the empirical powers for all the six different scenarios for binary outcome.

Table 6.4.2: Summary of Simulation Results for Binary Outcome

Design scenarios	Level 3	Level 2	Level 2 with Interaction	Level 1	Level 1 with Treat × Level 3	Level 1 with Treat× Level 2
Mean	0.75	0.74	0.73	0.76	0.73	0.73
Min	0.70	0.66	0.65	0.68	0.66	0.70
Max	0.84	0.85	0.78	0.80	0.77	0.77
SE	0.066	0.0074	0.0056	0.0053	0.0052	0.0043

Tables B.7 to B.12 in Appendix B present in more detail the specified sample sizes for the three levels, the empirical power $\tilde{\mathcal{T}}$, and the absolute differences between the empirical powers and the theoretical powers for the binary case.

Similar to the continuous case, the empirical powers were lower than the theoretical powers in the cases with interaction. This is due to the overestimate of the interaction variance components caused by PROC GLIMMIX.

Overall, the results for binary data showed that the empirical powers are consistent with the theoretical powers at all sample sizes combination. In combination of the sample sizes considered, the theoretical powers remain within the 95% confidence intervals of the empirical powers in 87.0% (141/162) of the cases. The absolute differences between the empirical powers and the theoretical powers were fairly small, with an average of 0.025 (SE=0.0202). Thus, the simulation results for binary data also indicate that the power based on our derived formulas is nearly identical to the empirical power based on the simulated data.

Chapter 7

Application

Up to this point, we have discussed the issues of sample size and power calculation for CRT with three-level designs. We derived the relevant formulas for important scenarios where random assignment takes place at different levels. However, these formulas are not easy to use since some are not straightforward and explicit. To make the connection between the theoretical methods and their practical applications, we composed a SAS program that allows the user to compute power and sample size using our methods. First, we will describe the basis of the program and its algorithm. Next, we will present a few hypothetical examples where the program can be used to answer practical power and sample size questions.

7.1 The User-Interface Program

To put theory into practice, we developed a macro program using the SAS software (version 9.2; SAS Institute Inc, Cary, North Carolina). The goals of the program are:

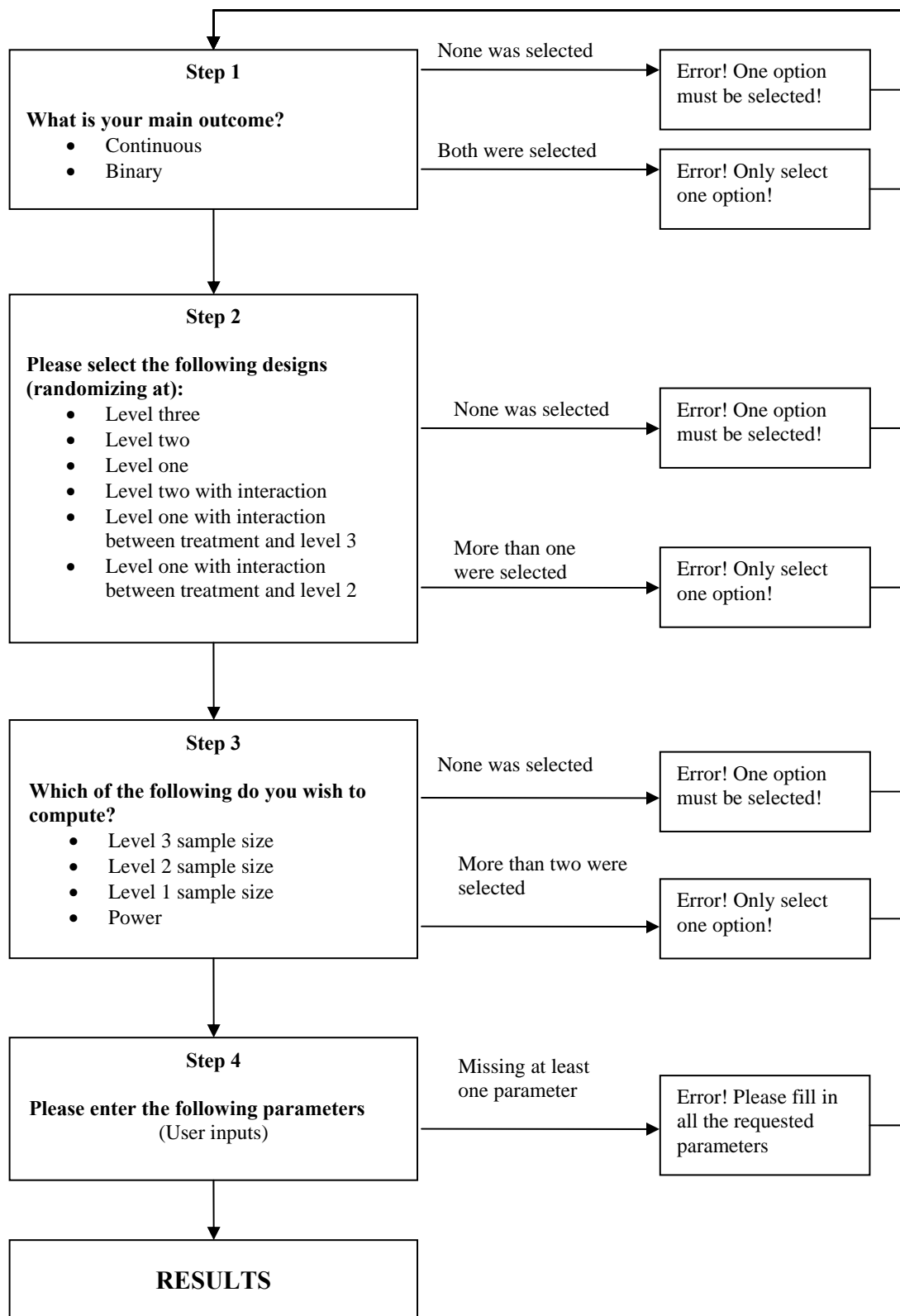
- (1) To compute sample size and power using the derived formulas.
- (2) To serve as a tool that enables users to input their own parameters and receive customized output.
- (3) To allow users to explore power and sample size results in different design scenarios.

The program consists of three sections. The first section uses SAS %WINDOW, which is part of the macro language included in the Base SAS system. The %WINDOW statement provides the functionality of prompting for user input, placing the input into macro variables, and resolving them to generate program code. Once defined, the macro variables remain in effect until the SAS session is terminated or until the user redefines them through additional input. In this program, the %WINDOW statement presents several groups of input fields. These groups include different design scenarios (depending on the user's choices), instructions for the next steps, and warnings about an input error.

The second section of the program consists of a series of macros, which are used to compute power or sample size for the six design settings for two types of outcomes, continuous and binary, as described in Chapter 4 and Chapter 5. Thus, altogether there are twelve different design scenarios. For each different design, four macros were written to compute (1) power, (2) sample size of level three, (3) sample size of level two, and (4) sample size of level one. The computations are carried out in SAS PROC IML based on the derived formulas. In addition, a few macros were composed to count the number of missing entries or to reset all macro variables once the users chose begin a new calculation.

The third part of the program put everything together. This section called for the information provided by the first section, ran the calculation from the macros in the second section, and provided the results using the %DISPLAY statement. To allow for a variety of reporting capabilities, the program followed an algorithm that depended on the user's input. A description of this algorithm is shown in Figure 7.1.

Figure 7.1: Algorithm for the User-Interface Program



7.2 Application Example for Continuous Data

To illustrate how power and sample size in a three-level study can be computed using our methods, we present a power analysis example with continuous data taken from a recent study. The example itself can be used as a template for other similar research designs.

Consider the cluster randomized trial conducted by Calear et al. (2009). The study was designed to investigate the effect of an online, self-directed cognitive behavior program in preventing and reducing the symptoms of anxiety and depression in an adolescent school-based population. The program was based on cognitive behavior therapy. It consisted of five interactive modules with information, animated demonstration, quizzes and exercise. The goal of the program was to change dysfunctional thoughts, improve interpersonal relationship, and teach the students important life skills.

The outcome variables of interest were anxiety and depressive symptoms. Both anxiety and depression quizzes were completed at the beginning and the end of the MoodGYM program. Anxiety was measured with the Revised Children's Manifested Anxiety Scale, which consists of 37-item self-report questionnaire. A total score was calculated by summing the "Yes" responses for the anxiety-related questions. Depressive symptoms were measured with the Center for Epidemiological Studies Depression Scale, which consists of 20 items. A total score was calculated by summing item scores and are assumed to have a continuous distribution.

The authors computed a power analysis from a two-level standpoint with students nested within classrooms (school effect was not taken into account). Under this assumption, the reported sample size was approximately 15 classes of 30 for each treatment arm. The calculation was based on detecting a post-intervention effect size of 0.30, with a power of 0.90 and $\alpha=0.05$. The ICC coefficient at the class level (between-classes variance divided by total

variance) was assumed to be 0.02 and the total variance was assumed to be 36 based on the report from the study.

We will perform a power analysis for the aforementioned study under the three-level framework. To do this, we first assume that the schools were randomly sampled from a larger population of schools. In addition, we also need to select plausible values for the clustering effect at the school level since this information was missing in the paper. We follow the report from the National Assessment of Educational Progress, which suggested that with a three-level design (students are nested within classrooms and classrooms are nested within schools), the clustering effect at classroom level is approximately $2/3^{\text{rds}}$ as large as the clustering effect at the school level (Konstantopoulos, 2008). Thus, given the classroom ICC of 0.02 in this study, we obtained an ICC of 0.03 for the school effect. This value represent the correlation between two students in the same school (and different classrooms).

It follows that the variances at school level and classroom level can be determined by

$$\sigma_c^2 = \rho \times \sigma_T^2 = 0.03 \times 36 = 1.08$$

$$\sigma_p^2 = r \times \sigma_T^2 = 0.02 \times 36 = 0.72$$

$$\sigma_e^2 = \sigma_T^2 - \sigma_c^2 - \sigma_p^2 = 36 - 1.08 - 0.72 = 34.2$$

Note that the authors defined $\rho = \frac{\sigma_c^2}{\sigma_T^2}$ and $r = \frac{\sigma_p^2}{\sigma_T^2}$. With an effect size of 0.3, the

difference measurement between the treatment and control groups is given by:

$$d = \delta \times \sqrt{\sigma_T^2} = 0.3 \times \sqrt{36} = 1.8$$

Case 1: Randomize at Level One without Interaction

Assume for each school the researchers plan to select 6 classrooms each with a class size of 30. Furthermore, the researchers will randomize all students equally into two study arms. The treatment effect is assumed to be the same across all schools (no interaction effect).

We propose to compute sample sizes under the case of randomizing at level one without interaction using our methods. Given the above information, our first task is to compute the sample size of level three (number of schools). Following the algorithm in Figure 7.1, the steps to be taken are as follows:

- Step 1: Continuous outcome
- Step 2: Randomizing at level one
- Step 3: Computing sample size at level three
- Step 4: The parameters entered are
 - Proportion randomized to intervention arm $\pi = 0.5$,
 - Expected power $1 - \beta = 0.90$
 - Significance level $\alpha = 0.05$
 - Sample size of level two $p = 6$
 - Sample size of level one $n=30$
 - Variance between schools $\sigma_c^2 = 1.08$
 - Variance between classes $\sigma_p^2 = 0.72$
 - Residual variance $\sigma_e^2 = 34.2$
 - The treatment difference needed to detect $d = 1.8$

Entering the above parameters in the program, we obtain the sample size of $c=3$ for level three. Thus, given the class size of 30 and 6 classrooms per school, we will need 3 schools to enroll in the study in order to achieve a power of 90% with $\alpha = 0.05$.

What if the number of classrooms per school is larger or smaller? Let us go back to step 4 and increase the sample size of level 2 from 6 to 10. The result from the program showed that we now only need 2 schools instead of 3 schools. On the other hand, if we reduce the sample size of level two from 6 to 4, the program then indicated that we now will need 4 schools instead of 3 schools.

Now, fixing the total number of classrooms per school at 6, how many schools will we need if the class size is either larger or smaller? To answer this question, we go back to step 4 of the program and increase the sample size at level one (number of students per class) from 30 to 40. The result suggests that we will need only 2 schools instead of 3 schools in this case. On the contrary, if we reduce the class size from 30 to 20, the result showed that we then need 4 schools instead of 3 schools in order to achieve the same power.

Table 7.1 displays the results of a few different sample size combinations to achieve a power of 90% with $\alpha = 0.05$ for the parameters given in the current study.

Case 2: Randomize at Level Two with Interaction

For further illustration, now let us assume that instead of randomizing at the student level, the researchers decide to randomly assign the classrooms in each school to the two study arms. In addition, we also assume that the treatment effect is not the same across schools. We now have a different study design where randomization takes place at the second level and an interaction effect between treatment and schools exists.

Table 7.1:
Some Sample Size Combinations for the YouthMood Project Trial
Randomizing at Level One without Interaction

Class size	Number of classrooms	Number of Schools
10	4	12
	6	8
	8	6
	10	5
20	4	6
	6	4
	8	3
	10	3
30	4	4
	6	3
	8	2
	10	2
40	4	3
	6	2
	8	2
	10	2

Assuming all other parameters remain the same, denote the variance of the interaction effect by σ_{ct}^2 . Suppose the magnitude of this new variance term is about 20% of the variance across schools, we obtain the following:

$$\sigma_{ct}^2 = 0.20 \times \sigma_c^2 = 0.216$$

$$\sigma_e^2 = \sigma_T^2 - \sigma_c^2 - \sigma_p^2 - \sigma_{ct}^2 = 36 - 1.08 - 1.72 - 0.216 = 33.98$$

Again, assume for each school the researchers plan to select 6 classrooms each with a class size of 30. Given the algorithm in Figure 7.1, the steps in the program now can be modified as follows

- Step 1: Continuous outcome
- Step 2: Randomizing at level one with interaction between treatment and level three

- Step 3: Computing sample size at level three
- Step 4: The parameters entered are
 - Proportion randomized to intervention arm $\pi = 0.5$,
 - Expected power $1 - \beta = 0.90$
 - Significance level $\alpha = 0.05$
 - Sample size of level two $p=6$
 - Sample size of level one $n=30$
 - Variance between schools $\sigma_c^2 = 1.08$
 - Variance between classes $\sigma_p^2 = 0.72$
 - Variance of the interaction between treatment and school $\sigma_{ct}^2 = 0.216$
 - Residual variance $\sigma_e^2 = 33.98$
 - The treatment difference needed to detect $d = 1.8$

Entering the above into the program, we obtain a total sample size of 8 schools in order to achieve a power of 90% with $\alpha = 0.05$.

How would this sample size change if the number of classrooms per school changes? Going back to step 4 and increase the sample size of level two from 6 to 10, we see that only 7 schools are needed instead of 8 schools. On the hand, if we reduce the sample size of level two from 6 to 4, the program then indicates that we need 10 schools instead of 8 schools as initially stated.

To examine the effect of class size, we fix the total number of classrooms per school at 6 and vary the number of students per classroom. Modifying step 4 of the program by increasing the sample size at level one (number of students per class) from 30 to 40. The result suggests that

we will only need 7 schools in this case. However, reducing the sample size at level one from 30 to 20 indicates that we will need 9 schools instead of 6. Table 7.2 displays the results of a few sample size combinations to achieve a power of 90% with $\alpha = 0.05$ for the parameters given in this hypothesized scenario.

Table 7.2:

Some Sample Size Combinations for the YouthMood Project Trial

Randomizing at Level Two with Interaction

Class size	Number of classrooms	Number of Schools
10	4	17
	6	13
	8	11
	10	9
20	4	12
	6	9
	8	8
	10	7
30	4	10
	6	8
	8	7
	10	7
40	4	9
	6	7
	8	7
	10	6

7.3 Application Example for Binary Data

Consider the Dutch Helping Hands Trial, where researchers studied methods to reduce hospital-acquired infections through a preventive measure –hand hygiene (Teerenstra et al.,

2010). In this study, two methods to improve adherence to hand hygiene guidelines in hospitals were compared. Both methods targeted on changing the nurse behavior.

The first method focused on the nurses (education, feedback) and the wards (facilities). The second method added other elements such as social influence in groups (norm and target setting within the nurse team). Random assignment took place at the ward level, i.e. each method is randomized to a subset of wards (level three), nurses (level two), and evaluations (level one). Evaluations were binary outcomes reflecting whether the guidelines are followed for each hand hygiene opportunity.

In their power analysis, the researchers forecasted that the probabilities of adherence in the standard group and the group with extended strategy to be 0.60 and 0.70 respectively. It is assumed that the behavior of an individual nurse is fairly consistent and nurses within the same ward share some common working environment. Based on past studies, the intraclass correlation between the wards was set as $\rho = 0.3$ and the intraclass correlation between nurses (within the same ward) was set as $r = 0.6$. The suggested sample sizes were 3 evaluations on approximately 15 nurses on each ward. Assuming the total variance was approximated at $\sigma_T^2 = 0.1$ based on the

study report, and since the authors define $\rho = \frac{\sigma_c^2}{\sigma_T^2}$ and $r = \frac{\sigma_c^2 + \sigma_p^2}{\sigma_T^2}$, the variances of each level

can be computed by:

$$\sigma_c^2 = \rho \times \sigma_T^2 = 0.3 \times 0.1 = 0.03 \quad \text{and} \quad \sigma_c^2 + \sigma_p^2 = r \times \sigma_T^2 = 0.6 \times 0.1 = 0.06$$

$$\sigma_p^2 = 0.06 - 0.03 = 0.03$$

Using the parameters given above, we propose to compute the sample size under the context of three-level design using the formulas derived from our methods. The following steps in the program will be taken:

- Step 1: Binary outcome
- Step 2: Randomizing at level three
- Step 3: Computing sample size at level three
- Step 4: The parameters entered are
 - Proportion randomized to intervention arm $\pi = 0.5$,
 - Expected power $1 - \beta = 0.80$
 - Significance level $\alpha = 0.05$
 - Sample size of level two $p=15$
 - Sample size of level one $n=3$
 - Variance between wards $\sigma_c^2 = 0.03$
 - Variance between nurses $\sigma_p^2 = 0.03$
 - The probability of event in treatment group $p\text{-treat}=0.7$
 - The probability of event in control group $p\text{-control}=0.6$

Inputting the above parameters into the sample size program, we arrive at a total sample size of 24 wards for a power of 80% and $\alpha = 0.05$.

How would variation on the number of nurses per each ward affect this sample size results? Suppose the number of nurses (sample size at level 2) is increased from 15 to 20 nurses per each ward. Making this modification in the program shows that we now need only 20 wards. On the other hand, reducing the number of nurses per ward from 15 down to 10 brought the total wards required to achieve the same level of power up to 32 wards instead of 24.

Suppose the number of nurses in each ward is fixed at 15. Adjusting sample size at the first level (number of evaluations) also affects the sample size results for level three. Suppose

the number of evaluations is increased from 3 to 5 per nurse. This modification shows that we now need only 18 wards. On the other hand, reducing the number of evaluation per nurse from 3 down to 2 brought the total wards required to achieve the same level of power up to 32 wards instead of 24 wards. Table 7.3 shows the sample size combinations for the given parameters described in step 4 above.

Table 7.3:

Some Sample Size Combinations for the Dutch Helping Hands Trial
Randomization at Level Three

Number of Evaluations	Number of Nurses	Number of Wards
2	5	80
	10	44
	15	32
	20	26
4	5	44
	10	26
	15	20
	20	18
6	5	32
	10	20
	15	16
	20	14
8	5	26
	10	18
	15	14
	20	12

7.4 Comments on the Application Examples

Above are a few examples to illustrate the use of the user-interface program and how our derived formulas can be applied to compute sample size in different study design settings. The

values chosen in Table 7.1, Table 7.2, and Table 7.3 were based on a range of typical sample sizes that were seen in similar studies. The program, however, was designed to be flexible and allow the users to specify any combination sets of parameters. For example, although only sample sizes at level one and level two were allowed to change, it should be noted that similar analyses can easily be done by looking at different sets of combinations of the desired power and different values of the intraclass correlation as well as varying the sample size at level three.

The goal of the application examples was to demonstrate the use of our formulas and our program to provide the researchers with several trade-off options to achieve the appropriate sample sizes. A user manual is provided in Appendix A for more details on how to implement the SAS program.

Chapter 8

Discussion and Future Work

This chapter presents the principal findings, discusses the results, and gives a list of future work that can build on the work presented here. We will start by providing a brief summary of the work that was detailed in previous chapters. We then will discuss the significance of this work by interpreting the results in light of our current knowledge about sample size and power in cluster randomized trials. Finally, we will suggest extensions for future work.

8.1 Summary of Work

Motivated by the need of a proper study design, this work provides sample size and power calculations for three-level cluster randomized trials. We considered a two-group comparison involving one treatment arm and one control arm. We assumed no covariates were included in the model. We proposed the relevant power functions in six important settings for two of the most popular outcomes: continuous and binary. The six scenarios depend on which level the allocation of treatment takes place and whether there is an interaction effect. They are:

- (1) Randomizing at level three
- (2) Randomizing at level two without interaction
- (3) Randomizing at level two with interaction between treatment and level three
- (4) Randomizing at level one without interaction
- (5) Randomizing at level one with interaction between level three and treatment

(6) Randomizing at level one with interaction between level two and treatment.

We derived our power functions under the framework of generalized linear mixed model (GLMM) theory. For continuous outcomes, we used a robust variance estimator to derive the variance of the treatment effects. For binary outcomes, we applied the pseudo-likelihood approach to approximate the variance of the estimated treatment effects using a first-order Taylor series expansion. In the cases where the inverse of the correlation matrix could be written explicitly, we derived close forms for the design effects and for the power function respectively. For situations where a closed form solution does not exist, we proposed general formulas that yield satisfactory results for every setting.

To assess the accuracy of our formulas, we conducted a simulation study in which we compared the empirical power and the estimated power. The simulation programs were written such that any set of parameters can be tested. We reported the simulation results for 27 combinations of parameters per each design settings. We simulated 784 sets of data for each combination based on a margin of error of 0.035. Thus, altogether we had 21,168 sets of data for each of the six different scenarios per each type of outcome. The simulation results confirmed that our derived formulas for sample size and power are accurate under the conditions that we examined.

Finally, we demonstrated the application of our power and sample size through the development of a user-interface SAS program. This program allows the researchers to calculate approximate sample sizes and power for a variety of design settings described above. To make the calculation, certain information is needed about the specific parameters. These parameters include the difference in the treatment effects, the variances of the random effects, the level of

significance alpha, and the proportion of the sample that is allocated into each treatment arm. We illustrated the use of the SAS program through some practical examples taken from past studies.

8.2 Discussion

In this work, we presented the relevant formulas to compute sample size and power for several important three-level study designs. Our formulas were derived based on the GLMM method, which has been proven to be a powerful technique to handle correlated outcomes. GLMM combines the properties of the two popular statistical frameworks, linear mixed models (which allows for random effects) and generalized linear models (which handle non-normal data via the use of the link functions and exponential family).

As discussed in Chapter 2, statistical models that have been used for data with nested correlation structure are classified into two categories: the population-averaged (marginal) models or the cluster-specific (conditional) models. Each approach has its own advantages. Since GLMM is a subject-specific approach, our formulas estimated the parameters of the random effects for each subject, and estimated the fixed effect as a common factor. Thus, it makes specific use of the within-subject information as oppose to population-average approach.

It should be noticed that our power functions were applicable only for two-arm trials. More specifically, we focused on the test for treatment contrast drawn from two treatment groups, not from multiple treatment effects. The reason behind this is that in most published literature, multilevel designs were chosen to achieve the power of a particular treatment contrast. Even when several treatments are being compared, typically there is always a contrast that plays the most important role to the researchers. Calculations for sample size and power would then be based on this main treatment contrast.

In some previous similar work, the sample size formulas were constructed from the Z-test under asymptotic normal distribution (Campbell et al., 2004; Teerenstra et al., 2008; Dang et al., 2008). In multilevel design, however, power depends on the magnitude of the variance components that are usually unknown. Thus, a better choice of for the sampling distribution of the test statistic should be a t-distribution, which has a thicker tail than the normal distribution. To gain better accuracy for our results, we derived the power functions based on the CDF of a t-distribution, adjusting for appropriate degrees of freedom. This helped to avoid the overestimation of power from the same critical value based on the standard normal distribution.

In this work we discussed different design situations where randomization can take place at any level. In the cases where an interaction might occur, we considered only the two-way interaction effects. That is, we only looked at the interactions between treatment and level three or treatment and level two. While conceptionally a three-way interaction effect can happen, in reality this model is not practical. Firstly, from our observation the convergence rate of GLMM model with three-way interaction is very slow, especially for data with three-level structure. Secondly, in the design stage it is very difficult to pre-specify the variance of this three-way interaction term. Lastly, it is hard to avoid the confounding effect in models with three-way interaction. For example, a treatment effect that varies between physicians will also vary between practices, since physicians are nested within practices.

Our power functions were derived from the unadjusted models where no covariate was included. However, sample sizes that include covariates other than treatment effect can be incorporated by following the same procedure. Our methods can be extended so that the correlation matrix will account for the presence of the covariates. Nevertheless, adding covariates into the models should not bring any significant changes in the magnitude of the

treatment effect, since in principle the covariates should not be related to the treatment. However, including covariates will affect the variance of each level and eventually will change the total variance. Generally, adding other covariates besides treatment will decrease the residual variance and increase the power. It will also reduce the degrees of freedom for the test statistics (Murray, 1998). Thus, the sample size formulas we presented here yield a more conservative result as oppose to methods where covariates are included. In addition, it is common for power and sample size software to account only for the treatment effect and not include the covariates.

As always, one needs to specify the underlying correlation structure to compute power in a multilevel design study. In this work, we assumed the most convenient and the most common choice—the compound symmetry correlation structure. The nested exchangeable structure we used for the three-level design is a direct generalization of the exchangeable structure that is commonly used in two-level models. This structure is suitable when the units in level one are exchangeable within the level two units, and the level two units are exchangeable within the level three units. Although more complex correlation structures such as AR1 or Toeplitz can be applied, we argue that at the early design stage, it is doubtful that we know the true correlation structure unless some sensitivity analyses were conducted beforehand. In addition, a more complex correlation structure introduces theoretical difficulties that are not necessary at the design stage.

Knowledge of plausible values of the variances or the two intraclass correlations is required for our computation. However, the ability to *a priori* postulate the intraclass correlation when designing a cluster randomized trial is limited. Such estimates are sparse in the literature, especially for three-level designs. To overcome this problem, a range of plausible values of the two intraclass correlations should be considered in order to assess the sensitivity of the results to

the misspecification of the parameters. Teerenstra et al. (2008) suggested that the researchers should examine the two-level intraclass correlations that are currently available in the literature and apply these values to the three-level models. Another way to facilitate the estimation of the intraclass correlation is to base on intermediate assessment of the within and between variances. Vierron and Giraudeau (2007) suggested that the recruitment can be flexible enough to be shortened or extended depending on the results of the internal pilot study. Thus, adjustment of the sample sizes to the intraclass correlation estimates can be done accordingly.

Our methods made the assumption that all designs were balanced. Although in practice it is a natural consequence of recruitment process that the sample sizes can be unequal, it has been suggested that imbalance designs should be avoided as much as possible. The connection between imbalanced cluster sizes and power computation is intuitively obvious. Firstly, the estimates of small cluster sizes will be less accurate than those of larger cluster sizes in imbalanced designs. Secondly, the addition of more subjects into larger clusters does not compensate for the lost of precision in smaller clusters (Eldridge et al., 2006). Thus, our study was based on a balanced design, since imbalance might lead to biased estimates and decrease the reliability of the results.

In this work we only examined power in the setting of a completely randomized design. Although other designs such as matched pairs and stratified randomized designs are also available, the completely randomized design is chosen for pragmatic reasons. First, it is the simplest and the most popular design in CRT. It allows for a wide number of statistical methods to be applied in the analysis steps and serves as a basis for deriving another sample size algorithm for more complex designs. Second, sensitivity analyses based on other designs can always be explored when needed.

In our simulation study, we examined a number of combinations defined by different values of the variances and sample sizes at each level. Although the number of parameters was limited, these parameters were selected from analysis of data of previous CRTs encountered in the literature. Therefore, the number of units on each level and the intraclass correlation were similar to those observed in this area of research. In addition, the simulation program was design in such a way that the same finding can be examined across a much larger number of scenarios. We speculate that similar conclusions would be obtained in these different settings.

Most existing sample size calculators and software rely heavily on simulation algorithms. Such method is limited for the GLMM approach, since the convergence rate is very low for data with small or medium sample sizes. Besides, software based on simulations usually requires the users to specify the variance-covariance structures, which sometimes can be confusing and complicated. Our sample size program focuses on the user-friendly aspect. It computes sample sizes and power directly from SAS IML and provided the results much faster than does simulation.

In most studies with a two-level design, either the clusters or the subjects within clusters can be randomized into different treatment conditions, suggesting that random assignment can be done at any level. The same logic holds for three-level designs, except that the nested structure is more complicated. While most previous work in power and sample size for three-level studies focused mainly on designs where random assignment takes place at the highest level, our study considered other situations that might be encountered in practice. Although the formulas we provided were model-specific, we present the results for a wide variety of different models corresponding to different designs. We allowed for the fact that randomization can occur at any level. We also considered the cases when interaction effects occurred, including the interaction

between treatment with each of the higher levels, level three and level two. By addressing the vast different designs, we provide a versatile tool to aid researchers in making their sample size decisions.

8.3 Future work

Taken as a whole, our study provided the general statistical formulas to compute power and sample size in three-level cluster randomized data under GLMM approach. It should be pointed out that the work presented here is just an initial step with many further steps to follow. Additional methodological and computational work is necessary and important in this research area.

First, our methods only focused on continuous and binary data, whereas other types of data can be encountered in practice. By following the same path, we believe that relevant formulas for different types of data can be derived using the similar methods. For example, under the framework of GLMM different link functions can be selected and different variance structures can be constructed. Thus, future work on this topic can include extending the same set of formulas to ordinal, nominal, or count data.

The extensions of this work to longitudinal correlated data at level 1 would be useful in practice. Although in our earlier discussion, we advocated for the use of compound symmetry variance-covariance structure in the design stage, there are certainly situations when other structures are more appropriate. For example, another area that we should investigate in our future work is to apply similar findings to repeated measures or longitudinal studies. Following the same procedure we can develop formulas that yield sample sizes requirement under other variance-covariance structures such as first-order autoregressive, toeplitz, or other more complex

structures with different variances at different time points. One of the advantages of our finding is the fact that as long as the variance structure is correctly specified, proper modifications can be implemented under GLMM models to address the sample size problem. Hence, future work for repeated measures and longitudinal studies is feasible.

Our formulas were based on the assumption that all levels were treated as random. This assumption implies that the sample sizes of each level are selected randomly from a larger population. In practice, this does not always happen. In designing CRT, it is often the case that available number of clusters is somewhat limited. For example, consider a three-level CRT of center—physician—patient, where only three centers are available in the community. In this situation, if all three centers are selected in the study then it is more appropriate to treat center as a fixed effect rather than a random effect. Thus, another extension of our work is to examine situations where any level(s) can be treated as a fixed effect when the corresponding units of that level are not randomly selected.

Even though our work provided methods for computing sample sizes, it did not account for the cost involved in the design stage. In any nested study, the units at each level have a cost associate with them and the researchers need to decide on a sample size which will minimize cost or be within a certain budget. Obviously, increasing sample sizes of any level can be a trade-off between the gain in statistical power and the recruitment costs. Optimizing the choice of number of units per each level under certain budget constraints is another practical issue that we should examine. A cost function can be derived and used when costs for sampling an additional unit of any level can be quantified. In addition, minimum budget required for a given power or precision should also be examined separately for each design.

Although our user-interface program is a versatile tool for computing sample size and power for the designs under examination, it is by no means a final product. Several features can be added to the program, for example, the addition of a help menu, a printout option for final report, and more graphical descriptions of the power functions. The program can also serve as a foundation for the development of independent software – not necessarily on the SAS platform – to compute power and sample size for CRT.

Finally, a number of complications in computing sample size and power in individual randomized studies can equally be seen in CRT. To name a few, these issues include: designs with more than one treatment group, designs with unequal sample sizes between treatment groups, matched-pair designs and stratified designs, trials with losses to follow-up and missing data, longitudinal designs with different attrition rates. These issues can be topics for future research.

8.4 Concluding Remarks

In the early stage of planning any clinical trial or experiment, it has been widely accepted that the evaluation of sample size is crucial. Sample size determination is an important task, as insufficient sample size can lead to inadequate power and inaccurate findings, whereas excessive sample size is a waste of resources.

We contend that this work serves as a useful and practical application for sample size and power planning in three-level study designs. Our formulas and our user interface SAS program provides users with a quick tool to estimate sample size and power in a variety of different design settings.

Finally, calculation of power and sample size involves more than simply making the best guesses about the parameters of a proposed analysis. It requires assumptions that typically cannot

be tested until the actual data have been collected. To obtain accurate results from our methods, researchers should be able to make reasonable estimates and assumptions on preliminary information that are available. If the inputs are inappropriate for the model under examination or if the wrong method is used, the wrong answer will emerge and the sample size results will no longer be appropriate for the study of interest.

References

- Brown, W. J., and Draper, D. (2000), "Implementation and Performance Issues in the Bayesian and Likelihood Fitting of Multilevel Models," *Computational Statistics*, 15, 391-420.
- Brown, H., and Prescott, R. (2006). *Applied Mixed Models in Medicine*. Great Britain: John Willey and Sons.
- Calear, A.L., Christensen H., Mackinnon A., Griffiths, K.M., and O'Kearney, R. (2009), "The YouthMood Project: a Cluster Randomized Controlled Trial of an Online Cognitive Behavioral Program with Adolescents," *Journal of Consulting and Clinical Psychology*, 77,1021-1032.
- Campbell, MK., Thomson, S., Ramsay, R.C., MacLennan, GS., and Grimshaw, JM. (2004), "Sample Size Calculator for Cluster Randomized Trials," *Computers in Biology and Medicine*, 34, 113-125.
- Casella, G., and Berger, L.R. (2005), *Statistical Inference*, China: Thomson Learning.
- Chow, S., Shao, J., and Wang, H. (2008), *Sample Size Calculations in Clinical Research*, New York: Chapman & Hall.
- Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences*, New York: Academic Press.
- Cools, W., Noortgate, W.V., and Onghena, P. (2008), "ML-DEs: A program for designing efficient multilevel studies," *Behavior Research Methods*, 40, 236-249.
- Cornfield, J. (1978), "Randomization by Group: a Formal Analysis," *American Journal of Epidemiology*, 108, 101-102.

- Daniel, W.W. (2009), *Biostatistics, a Foundation for Analysis in the Health Science (8th edition)*, New Jersey: John Willey and Sons.
- Dang, Q., Mazumdar, S., and Houck, P.R. (2008), "Sample Size and Power Calculations Based on Generalized Linear Mixed Models with Correlated Binary Outcomes," *Computer Methods and Programs in Biomedicine*, 91, 122-127.
- Dedrick, F.R., Ferron, M.J., Hess, R.M., Hogarty, Y.K., Kromrey, D.J., Lang, R.T., Nile, J.D., and Lee, S.R. (2009), "Multilevel Modeling: A Review of Methodological Issues and Applications," *Review of Educational Research*, 1, 69-102.
- Donner, A., and Klar, N. (2000), *Design and Analysis of Cluster Randomization Trials in Health Research*. London: Arnold.
- Donner A., and Klar N. (2004), "Pitfalls of and Controversies in Cluster Randomized Trials," *American Journal of Public Health*, 94, 416-422.
- Eldridge, MS., Ashby, D., and Kerry, S. (2006), "Sample Size for Cluster Randomized Trials: Effect of Coefficient of Variation of Cluster Size and Analysis Method," *International Journal of Epidemiology*, 35, 1292-1300.
- Faggiano, F., Richardson, C., Bohrn, K., and Galanti, M. R. (2007), "A cluster randomized controlled trial of school-based prevention of tobacco, alcohol and drug use: The EU-Dap design and study population," *Preventive Medicine*, 44, 170-173.
- Feng, Z., and Grizzle, J.E. (1992), "Correlated Binomial Variates: Properties of Estimator of Intraclass Correlation and Its Effect on Sample Size Calculation," *Statistics in Medicine*, 11, 1607-1614.
- Freedman, L.S., Green, S.B., and Byar, D.P. (1990), "Assessing The Gain in Efficiency due to Matching in A Community Intervention Study," *Statistics in Medicine*, 9, 943-953.

- Gail, M.H., Mark, S.D., Carroll, R.J., Green, S.B., and Pee, D. (1996), "On Design Considerations and Randomization-Based Inference for Community Intervention Trials," *Statistics in Medicine*, 15,1069 -1092.
- Browne, W.J., Golalizadeh Lahi, M. and Parker, R.M.A. (2009), "A Guide to Sample Size Calculations for Random Effect Models via Simulation and the MLPowSim Software Package,". University of Bristol.
- Grandes, G., Sanchez, A., Sanchez-Pinilla R. O., Torcal J., Montoya, I., Lizarraga, K., and Serra, J. (2009), "Effectiveness of Physical Activity Advice and Prescription by Physicians in Routine Primary Care," *Archives of Internal Medicine*, 7: 694-701.
- Guittet, L., Giraudeau, B., and Ravaud, P. (2005) "A Priori Postulated and Real Power in Cluster Randomized Trials: Mind the Gap," *BMC Medical Research Methodology*, 5, 25.
- Hayes, R.J., and Bennet, S. (1999), "Simple Sample Size Calculation for Cluster-Randomized Trials," *Journal of International Epidemiological Association*, 28, 319-326.
- Hedeker, D., Gibbons, R.D., and Waternaux, C. (1999), "Sample Size Estimation for Longitudinal Designs With Attrition: Comparing Time-Related Contrasts between Two Groups," *Journal of Educational and Behavioral Statistics*, 24, 70-93.
- Hedges, L., and Hedberg, E.C. (2007), "Intraclass Correlation Values for Planning Group-Randomized Trials in Education," *Journal of Educational Evaluation and Policy Analysis*, 29, 60-87.
- Henderson, H.V. and Searle, S.R. (1981), "On Deriving The Inverse of a Sum of Matrices," *Siam Review*, 23, 53-60.

- Heo, M., and Leon A.C. (2009), "Sample Size Requirements to Detect an Intervention by Time Interaction in Longitudinal Cluster Randomized Clinical Trials," *Statistics in Medicine*, 28, 1017-1027.
- Hox, J. (2002), *Multilevel Analysis Techniques and Applications*. USA: Lawrence Erlbaum Associates Publishers.
- Hsieh, F.Y. (1988), "Sample Size Formulas for Intervention Studies with the Cluster as Unit of Randomization," *Statistics in Medicine*, 8, 1195-1201.
- Kish, L. (1965), *Survey Sampling*. New York: John Wiley.
- Klar, N., and Donner, A. (2001), "Current and Future Challenges in the Design and Analysis of Cluster Randomization Trials," *Statistics in Medicine*, 20, 3729–3740.
- Konstantopoulos, S. (2008), "The Power of the Test for Treatment Effects in Three-Level Cluster Randomized Designs," *Journal for Research on Educational Effectiveness*, 1, 66-88.
- Krist A.H., Woolf, S.H., Johnson R.E., Rothemich S.F., Cunningham T.D., Longo, R.D., Peele, E., Matzke, G. "The Effect of a Personalized Healthcare Record on the Delivery of Preventive Care" (in review).
- Littell, R.C, Milliken, G.A., Stroup, W.W, Wolfinger, R.D, and Schabeneger,O. (2007), *SAS for Mixed Models, Second Edition*. Cary, NC: SAS Institute Inc.
- Maas, C. J. M., and Hox, J. J. (2005), "Sufficient Sample Sizes for Multilevel Modeling," *European Journal of Research Methods for the Behavioral & Social Sciences*, 1, 85-91.
- Murphy, K. R., Myers, B. (2004), *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests (2nd ed.)*, Mahwah, NJ: Erlbaum.

- Murray, D.M. (1998), *Design and Analysis of Group-Randomized Trials*. New York: Oxford University Press
- Murray, D.M., Pals, S.L., Blitstein, J.L., Alfano, C.M., and Lehman, J. (2008), "Design and Analysis of Group-Randomized Trials in Cancer: A Review of Current Practices." *Journal of National Cancer Institute*, 100: 483-492.
- Murray, D. M., Varnell, S. P., and Blitstein, J. L. (2004), "Design and Analysis of Group-Randomized Trials: A Review of Recent Methodological Developments," *American Journal of Public Health*, 94, 423-432.
- Murray, D.M. , and Wolfinger, R.D. (1994), "Analysis Issues in the Evaluation Community Trials: Progress toward Solutions in SAS/STAT MIXED," *Journal of Community Psychology CSAP*. Special Issue, 140-154
- Pan,W. (2001), "Sample Size and Power Calculations with Correlated Binary Data," *Control Clinical Trials*, 22, 211-227.
- Raudenbush, S.W. (1993), *Hierarchical linear models and experimental design*. In Lynne K., and Edwards, *Applied analysis of variance in behavioral science*, New York: Marcel Dekker.
- Raudenbush, S.W. (1997), "Statistical Analysis and Optimal Design for Cluster Randomized Trials," *Psychological methods*, 2.2, 173-185.
- Raudenbush, S.W., and Liu, X. (2000), "Statistical Power and Optimal Design for Multisite Randomized Trials," *Psychological methods*, 5.2, 99-213.
- Raudenbush, S.W., and Bryk, A.S. (2002), *Hierrarchial Linear Models: Applications and Data Analysis Methods*, Newbury Park, CA: Sage.

- Roy, A., Bhaumik, D.K., Aryal, S., and Gibbons, D.R. (2007), "Sample Size Determination for Hierarchical Longitudinal Designs with Differential Attrition Rates," *Biometrics*, 63, 699-707.
- Shih, W.J. (1997), "Sample size and power calculations for periodontal and other studies with clustered sample using the method of generalized estimating equations," *Biometrical Journal*, 39, 899-908.
- Snijders, T., and Bosker, R. (1993), "Standard Errors and Sample Sizes for Two-Level Research," *Journal of educational statistics*, 18, 237-259.
- Teerenstra, S., Moerbeek, M., Achterberg, T., Pelzer, B.J., Borm, G.F. (2008) "Sample size Calculations for Three-Level Cluster Randomized Trials," *Clinical Trials*, 5, 486-95.
- Teerenstra, S., Lu, B., Preisser, J.S., Achterberg, T.V., and Borm, G.F. (2010), "Sample Size Considerations for GEE Analyses of Three-Level Cluster Randomized Trials," *Biometrics*. 2010 (in press)
- Turner, R.M., Prevost, A.T., and Thompson, S.G. (2004), "Allowing for Imprecision of the Intracluster Correlation Coefficient in the Design Of Cluster Randomized Trials," *Statistics in Medicine*, 23, 1195–1214.
- Vierron, E., Giraudeau, B. (2007), "Sample Size Calculation for Multicenter Randomized Trial: Taking the Center Effect into Account," *Contemporary Clinical Trials*, 28, 451-458.
- Zucker, M.D. (1990), "An Analysis of Variance Pitfall: The Fixed Effects Analysis in a Nested Design," *Educational and Psychological Measurement*, 50, 731—738.

Appendix A

User's Guide for the SAS Program

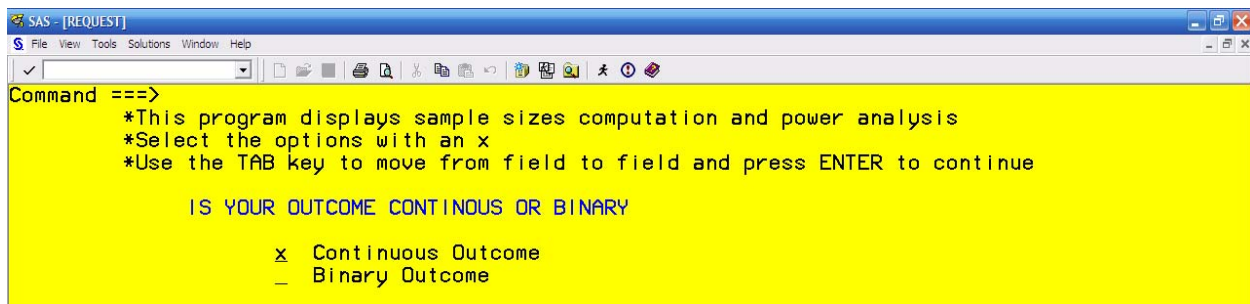
This manual provides the researchers with a guide to effectively use the interface SAS program that was written to demonstrate the application of our work. The program was developed to address sample size and power questions for three-level study designs with continuous and binary outcomes. The formulas adopted in this program were derived in Chapter 4 and Chapter 5. Application examples were presented in Chapter 7.

When the users first run the program, a brief description will be shown on the screen to give basic instructions as to how to choose the options and how to move around in the program. The program consists of four steps. The users will:

- Choose an option by typing x into the blank
- Use TAB key to move from one field to another field
- Press ENTER to move to the next step

A.1 First step

In the first step, the users will be asked to choose the type of outcomes for their experiment. The choices are either continuous or binary:



The screenshot shows a SAS window titled 'SAS - [REQUEST]'. The command prompt displays the following text:

```

Command ==>
  *This program displays sample sizes computation and power analysis
  *Select the options with an x
  *Use the TAB key to move from field to field and press ENTER to continue

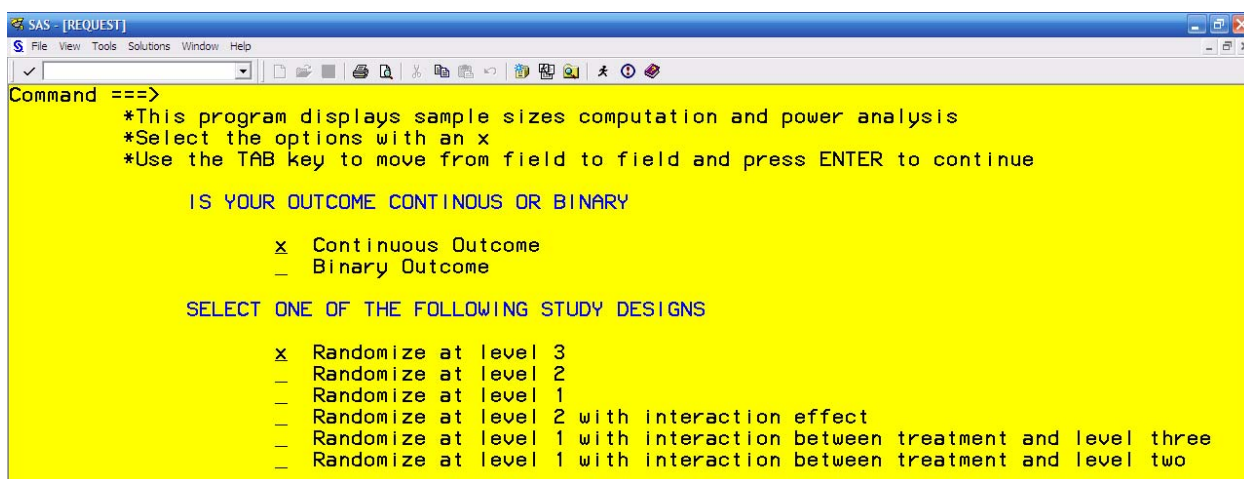
  IS YOUR OUTCOME CONTINUOUS OR BINARY

      x Continuous Outcome
      _ Binary Outcome
  
```

The selection is done by typing an x into the blank provided in front of the options. Here the users are required to select only one of the two options. If neither of the options were selected or both options were selected, an error message will appear to ask the users to re-enter their input. If the selection is done correctly, the program will move to the next step after the users hit ENTER. The above screen shows an example when continuous outcome was chosen.

A.2 Second step

In the second step the users will be asked to choose the study design. There are six study designs available. The following screen will appear:



```

SAS [REQUEST]
File View Tools Solutions Window Help
Command ==>
*This program displays sample sizes computation and power analysis
*Select the options with an x
*Use the TAB key to move from field to field and press ENTER to continue

IS YOUR OUTCOME CONTINUOUS OR BINARY

    x Continuous Outcome
    _ Binary Outcome

SELECT ONE OF THE FOLLOWING STUDY DESIGNS

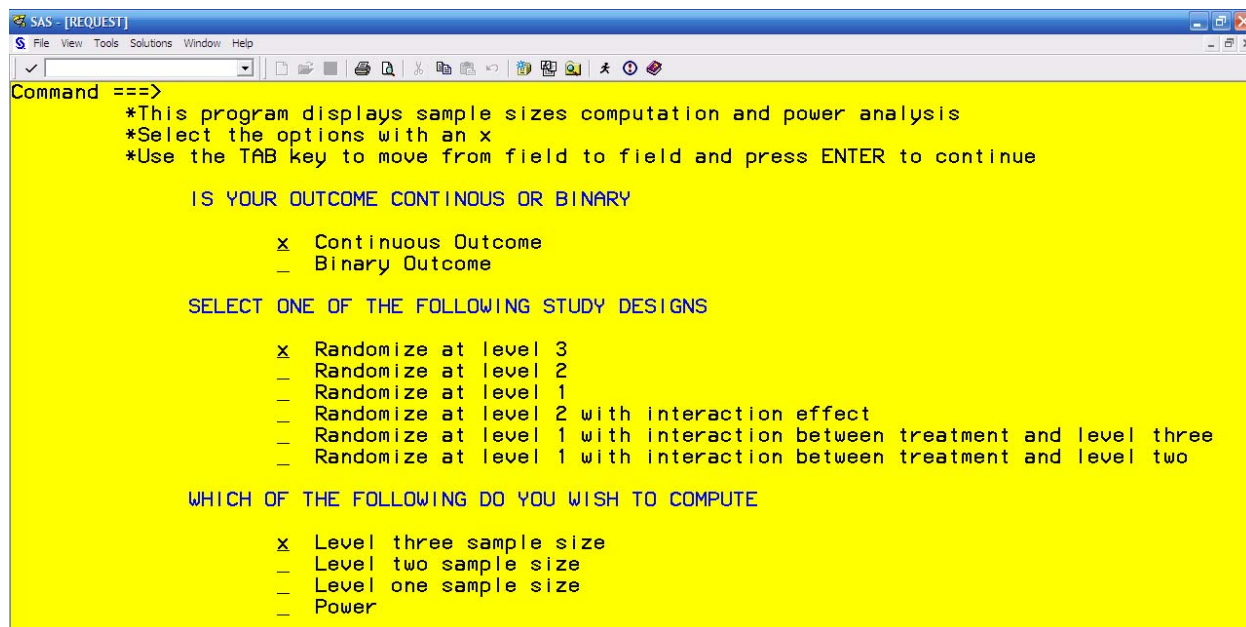
    x Randomize at level 3
    _ Randomize at level 2
    _ Randomize at level 1
    _ Randomize at level 2 with interaction effect
    _ Randomize at level 1 with interaction between treatment and level three
    _ Randomize at level 1 with interaction between treatment and level two
  
```

Again, the users are required to select one of the six options by typing an x in the blanks. An error message will appear if none of the options was selected or more than one were selected. The above screen shows an example when randomization at level three was chosen. The users will hit ENTER to move to the next step.

A.3 Third step

In the third step, the users will be asked to select what they desire to compute. The choices include sample sizes for the three levels or power. Again, one and only one option must be selected in order to move to the next step. An error message will appear if the instructions were not followed properly.

For demonstration, the following screen will appear if the users decide to compute level three sample size.



```

SAS [REQUEST]
File View Tools Solutions Window Help
Command ==>
  *This program displays sample sizes computation and power analysis
  *Select the options with an x
  *Use the TAB key to move from field to field and press ENTER to continue

  IS YOUR OUTCOME CONTINUOUS OR BINARY

    x Continuous Outcome
    _ Binary Outcome

  SELECT ONE OF THE FOLLOWING STUDY DESIGNS

    x Randomize at level 3
    _ Randomize at level 2
    _ Randomize at level 1
    _ Randomize at level 2 with interaction effect
    _ Randomize at level 1 with interaction between treatment and level three
    _ Randomize at level 1 with interaction between treatment and level two

  WHICH OF THE FOLLOWING DO YOU WISH TO COMPUTE

    x Level three sample size
    _ Level two sample size
    _ Level one sample size
    _ Power
  
```

A.4 Fourth step

Hitting ENTER after the third step, the users will then be asked to enter certain parameters. The number of parameters depends on the options chosen in previous steps. The following lists all possible parameters and instruction on how to input them:

Proportion of sample size allocated to treatment:

By default, this proportion is set at 0.5 (equal allocation). However, any proportion can be chosen as long as the product of the chosen sample size at the corresponding level and this proportion is an integer. For example, suppose randomization takes place at level three and the researchers decide on a proportion of 0.6 for treatment arm. A sample size of 8 for level three will be an illegal input in this case since $0.6 \times 8 = 4.8$ is not an integer. However, a sample size of 10 for level three will be a valid input since $0.6 \times 10 = 6$ is an integer.

The power expected to achieve

Here the users are required to type in the value for the power expected to achieve in their study. These values typically range from 0.75 to 0.90. The input is required to be in decimal format.

Level one (two, three) sample size

For these parameters, the users are expected to type in their anticipated sample sizes for the corresponding levels. The numbers are supposed to be integers.

Variance between level one (two, three)

These variances are usually computed from a pilot study or selected from previous studies. The variances are expected to be in integer or decimal formats.

Variance of the interaction

If a design with interaction effect was selected in previous steps, the program will be prompted to ask for the variance of the interaction effects. Again, these variances are usually computed from a pilot study or selected from previous studies. They are expected to be in integer or decimal formats.

Treatment difference

If the task is to calculate sample size or power for continuous outcome, users will be asked to input the difference of the two treatment means to be detected. Values of treatment difference are expected to be in integer or decimal formats.

The probabilities of event in treatment group and control group

If the task is to calculate sample size or power for binary outcome, users will be asked to input the expected control group proportion and the expected proportion in the intervention group. These proportions should be entered in decimal format.

Alpha

The level of significance alpha is set at a default value of 0.05. However, the default can be changed if the users wish to do so.

The following screen shows an example of the parameters input for step 4 when the requirement was to compute sample size at level three for continuous outcome where randomization takes place at level three. The sample size result is 8, after hitting ENTER after all parameters were provided.

```

SAS - [REQUEST]
File View Tools Solutions Window Help
Command ==>
*This program displays sample sizes computation and power analysis
*Select the options with an x
*Use the TAB key to move from field to field and press ENTER to continue

IS YOUR OUTCOME CONTINDOUS OR BINARY
  x Continuous Outcome
  - Binary Outcome

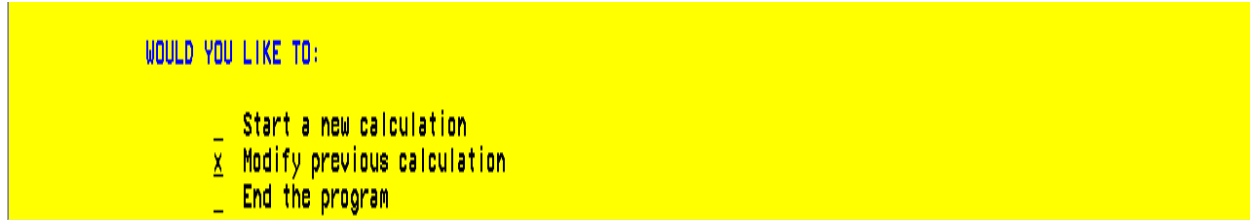
SELECT ONE OF THE FOLLOWING STUDY DESIGNS
  x Randomize at level 3
  - Randomize at level 2
  - Randomize at level 1
  - Randomize at level 2 with interaction effect
  - Randomize at level 1 with interaction between treatment and level three
  - Randomize at level 1 with interaction between treatment and level two

WHICH OF THE FOLLOWING DO YOU WISH TO COMPUTE
  x Level three sample size
  - Level two sample size
  - Level one sample size
  - Power

TO COMPUTE SAMPLE SIZE AT LEVEL TRHEE, PLEASE PROVIDE THE FOLLOWING INFORMATION
  Proportion of sample size allocated to treatment  0.5
  The power expected to achieve  0.80
  Level two sample size  10
  Level one sample size  10
  Variance between level three  0.01
  Variance between level two  0.39
  Variance between level one  0.60
  Treatment difference  0.70
  Alpha  0.05

THE SAMPLE SIZE FOR LEVEL THREE IS      8
  
```

After completing step 4, the users will be prompted to three options as follows



To start a new calculation

In choosing this option, the users will be brought to step one of a new calculation. All displays of the old screen will disappear.

Modify previous calculation

In choosing this option, the users will be brought back to the beginning of step 4 of the same calculation. All displays from the old screen will remain the same. The users are allowed to modify the parameters in step 4 and obtain the new result.

End the program

When this option is selected, the SAS program will stop running and the interface screen will be closed.

Appendix B

Simulation Results

The following tables in Appendix B showed results of the simulation study for different randomization schemes. For all tables, the notations are defined as follows:

c: Sample size of level three

p: Sample size of level two

n: Sample size of level one T

d: The difference in treatment effect

$\widehat{\mathcal{T}}$: Estimated power based on simulation

\mathcal{T} : Theoretical power based on the derived formulas

π : Proportion of sample size allocated to the treatment group

μ_c : Probability of the event of interest in control group (for binary cases)

σ_c^2 : Variance between level three

σ_p^2 : Variance between level two

σ_e^2 : Variance between level one

σ_{ct}^2 : Variance of the interaction between level three and treatment

σ_{pt}^2 : Variance of the interaction between level three and treatment

**Table B.1: Simulation Results for Case 1
Randomize at Level Three—Continuous Outcome**

c	p	n	d	$\hat{\mathcal{T}}$	95% CI of $\hat{\mathcal{T}}$
10	4	10	0.67	0.74	(0.65, 0.75)
		20	0.64	0.71	(0.61, 0.79)
		30	0.64	0.76	(0.67, 0.85)
	8	10	0.49	0.74	(0.65, 0.83)
		20	0.48	0.77	(0.69, 0.85)
		30	0.47	0.67	(0.58, 0.76)
	12	10	0.41	0.74	(0.65, 0.83)
		20	0.40	0.75	(0.69, 0.77)
		30	0.40	0.76	(0.54, 0.79)
20	4	10	0.44	0.72	(0.63, 0.81)
		20	0.42	0.73	(0.64, 0.82)
		30	0.42	0.72	(0.63, 0.81)
	8	10	0.32	0.79	(0.72, 0.88)
		20	0.31	0.71	(0.62, 0.80)
		30	0.31	0.76	(0.74, 0.76)
	12	10	0.27	0.71	(0.62, 0.80)
		20	0.26	0.85	(0.78, 0.92)
		30	0.26	0.73	(0.64, 0.82)
30	4	10	0.35	0.76	(0.63, 0.81)
		20	0.34	0.76	(0.75, 0.78)
		30	0.33	0.72	(0.63, 0.81)
	8	10	0.26	0.73	(0.64, 0.82)
		20	0.25	0.77	(0.61, 0.79)
		30	0.25	0.73	(0.64, 0.82)
	12	10	0.22	0.79	(0.71, 0.87)
		20	0.21	0.72	(0.63, 0.81)
		30	0.21	0.77	(0.74, 0.90)

$\sigma_c^2=0.01, \sigma_p^2=0.39, \sigma_e^2=0.60, \mathcal{T}=0.75, \pi=0.5$

Table B.2: Simulation Results for Case 2
Randomize at Level Two without Interaction—Continuous Outcome

c	p	n	d	$\hat{\mathcal{T}}$	95% CI of $\hat{\mathcal{T}}$	
10	4	10	0.58	0.72	(0.68, 0.76)	
		20	0.56	0.70	(0.66, 0.77)	
		30	0.55	0.71	(0.67, 0.75)	
	8	10	10	0.40	0.75	(0.71, 0.78)
			20	0.39	0.73	(0.72, 0.79)
			30	0.38	0.73	(0.69, 0.76)
		12	10	0.33	0.74	(0.71, 0.81)
			20	0.31	0.78	(0.75, 0.80)
			30	0.31	0.70	(0.65, 0.75)
20	4	10	0.40	0.79	(0.76, 0.81)	
		20	0.39	0.78	(0.75, 0.80)	
		30	0.38	0.74	(0.70, 0.77)	
	8	10	10	0.28	0.77	(0.74, 0.79)
			20	0.27	0.76	(0.72, 0.79)
			30	0.27	0.76	(0.72, 0.79)
		12	10	0.23	0.76	(0.72, 0.79)
			20	0.22	0.75	(0.67, 0.82)
			30	0.22	0.74	(0.69, 0.82)
30	4	10	0.33	0.73	(0.69, 0.76)	
		20	0.32	0.76	(0.72, 0.79)	
		30	0.31	0.78	(0.75, 0.80)	
	8	10	10	0.23	0.78	(0.75, 0.80)
			20	0.22	0.73	(0.69, 0.76)
			30	0.22	0.75	(0.72, 0.82)
		12	10	0.19	0.70	(0.65, 0.72)
			20	0.18	0.73	(0.69, 0.76)
			30	0.18	0.73	(0.69, 0.76)

$\sigma_c^2=0.01$, $\sigma_p^2=0.39$, $\sigma_e^2=0.60$, $\mathcal{T}=0.75$, $\pi=0.5$

**Table B.3: Simulation Results for Case 3
Randomize at Level Two with Interaction—Continuous Outcome**

c	p	n	d	$\hat{\mathcal{T}}$	95% CI of $\hat{\mathcal{T}}$	
10	4	10	0.69	0.70	(0.65, 0.77)	
		20	0.67	0.74	(0.74, 0.79)	
		30	0.67	0.75	(0.71, 0.76)	
	8	10	10	0.53	0.66	(0.61, 0.78)
			20	0.52	0.71	(0.66, 0.76)
			30	0.52	0.76	(0.71, 0.76)
		12	10	0.47	0.76	(0.71, 0.76)
			20	0.46	0.69	(0.64, 0.76)
			30	0.45	0.68	(0.63, 0.80)
20	4	10	0.46	0.70	(0.65, 0.77)	
		20	0.45	0.70	(0.65, 0.77)	
		30	0.44	0.72	(0.67, 0.75)	
	8	10	10	0.35	0.75	(0.70, 0.75)
			20	0.35	0.74	(0.69, 0.76)
			30	0.34	0.65	(0.60, 0.77)
		12	10	0.31	0.71	(0.66, 0.76)
			20	0.30	0.74	(0.69, 0.75)
			30	0.30	0.77	(0.72, 0.79)
30	4	10	0.37	0.74	(0.69, 0.74)	
		20	0.36	0.73	(0.71, 0.76)	
		30	0.35	0.74	(0.71, 0.76)	
	8	10	10	0.28	0.73	(0.72, 0.78)
			20	0.28	0.74	(0.69, 0.75)
			30	0.27	0.68	(0.63, 0.77)
		12	10	0.25	0.73	(0.72, 0.79)
			20	0.24	0.74	(0.73, 0.78)
			30	0.24	0.77	(0.74, 0.79)

$$\sigma_c^2=0.01, \sigma_p^2=0.39, \sigma_e^2=0.55, \sigma_{ct}^2=0.05, \mathcal{T}=0.75, \pi=0.5$$

Table B.4: Simulation Results for Case 4
Randomize at Level One without Interaction—Continuous Outcome

c	p	n	d	$\hat{\mathcal{I}}$	95% CI of $\hat{\mathcal{I}}$	
10	4	10	0.20	0.77	(0.76, 0.78)	
		20	0.14	0.82	(0.7, 0.82)	
		30	0.12	0.71	(0.69, 0.75)	
	8	10	10	0.14	0.76	(0.74, 0.78)
			20	0.10	0.78	(0.74, 0.78)
			30	0.08	0.67	(0.64, 0.84)
		12	10	0.12	0.77	(0.75, 0.78)
			20	0.08	0.74	(0.72, 0.76)
			30	0.07	0.78	(0.75, 0.79)
20	4	10	0.14	0.76	(0.74, 0.78)	
		20	0.10	0.76	(0.74, 0.78)	
		30	0.08	0.74	(0.72, 0.76)	
	8	10	10	0.10	0.74	(0.72, 0.76)
			20	0.07	0.73	(0.71, 0.76)
			30	0.06	0.73	(0.71, 0.76)
		12	10	0.08	0.72	(0.68, 0.75)
			20	0.06	0.74	(0.72, 0.76)
			30	0.05	0.77	(0.75, 0.78)
30	4	10	0.12	0.76	(0.73, 0.77)	
		20	0.08	0.77	(0.76, 0.78)	
		30	0.07	0.73	(0.71, 0.76)	
	8	10	10	0.08	0.75	(0.72, 0.77)
			20	0.06	0.76	(0.74, 0.78)
			30	0.05	0.73	(0.70, 0.76)
		12	10	0.07	0.78	(0.75, 0.79)
			20	0.05	0.76	(0.72, 0.81)
			30	0.04	0.74	(0.72, 0.81)

$\sigma_c^2=0.01, \sigma_p^2=0.39, \sigma_e^2=0.60, \mathcal{I}=0.75, \pi=0.5$

Table B.5: Simulation Results for Case 5
Randomize at Level One with Treatment x Level Three—Continuous Outcome

c	p	n	d	$\hat{\mathcal{T}}$	95% CI of $\hat{\mathcal{T}}$	
10	4	10	0.37	0.75	(0.73, 0.77)	
		20	0.34	0.69	(0.67, 0.79)	
		30	0.32	0.73	(0.72, 0.78)	
	8	10	10	0.34	0.72	(0.70, 0.74)
			20	0.32	0.71	(0.70, 0.78)
			30	0.31	0.75	(0.73, 0.77)
		12	10	0.32	0.70	(0.69, 0.76)
			20	0.31	0.76	(0.71, 0.86)
			30	0.31	0.74	(0.72, 0.79)
20	4	10	0.25	0.72	(0.71, 0.79)	
		20	0.22	0.71	(0.71, 0.8)	
		30	0.22	0.70	(0.64, 0.78)	
	8	10	10	0.22	0.74	(0.74, 0.78)
			20	0.21	0.75	(0.73, 0.77)
			30	0.21	0.74	(0.72, 0.76)
		12	10	0.22	0.75	(0.73, 0.81)
			20	0.21	0.71	(0.70, 0.78)
			30	0.20	0.73	(0.71, 0.76)
30	4	10	0.20	0.72	(0.68, 0.75)	
		20	0.18	0.75	(0.74, 0.78)	
		30	0.17	0.71	(0.69, 0.82)	
	8	10	10	0.18	0.68	(0.66, 0.75)
			20	0.17	0.72	(0.68, 0.75)
			30	0.17	0.73	(0.71, 0.77)
		12	10	0.17	0.71	(0.70, 0.80)
			20	0.17	0.70	(0.70, 0.79)
			30	0.16	0.74	(0.72, 0.76)

$\sigma_c^2=0.01, \sigma_p^2=0.39, \sigma_e^2=0.55, \sigma_{ct}^2=0.05, \mathcal{T}=0.75, \pi=0.5$

Table B.6: Simulation Results for Case 6
Randomize at Level One with Treatment x Level Two—Continuous Outcome

c	p	n	d	$\hat{\mathcal{T}}$	95% CI of $\hat{\mathcal{T}}$	
10	4	10	0.23	0.75	(0.61, 0.77)	
		20	0.17	0.69	(0.65, 0.79)	
		30	0.15	0.73	(0.69, 0.77)	
	8	10	10	0.16	0.74	(0.71, 0.76)
			20	0.12	0.74	(0.71, 0.77)
			30	0.10	0.72	(0.67, 0.75)
		12	10	0.13	0.70	(0.69, 0.77)
			20	0.10	0.74	(0.72, 0.76)
			30	0.08	0.72	(0.71, 0.78)
20	4	10	0.16	0.77	(0.75, 0.77)	
		20	0.12	0.70	(0.64, 0.79)	
		30	0.10	0.74	(0.73, 0.83)	
	8	10	10	0.11	0.71	(0.68, 0.75)
			20	0.08	0.74	(0.71, 0.76)
			30	0.07	0.73	(0.72, 0.77)
		12	10	0.09	0.74	(0.71, 0.75)
			20	0.07	0.69	(0.68, 0.77)
			30	0.06	0.74	(0.71, 0.80)
30	4	10	0.13	0.71	(0.70, 0.76)	
		20	0.10	0.72	(0.71, 0.77)	
		30	0.08	0.73	(0.73, 0.79)	
	8	10	10	0.09	0.75	(0.71, 0.77)
			20	0.07	0.72	(0.72, 0.77)
			30	0.06	0.71	(0.71, 0.78)
		12	10	0.07	0.74	(0.71, 0.77)
			20	0.06	0.70	(0.70, 0.77)
			30	0.05	0.75	(0.75, 0.79)

$$\sigma_c^2=0.01, \sigma_p^2=0.39, \sigma_e^2=0.55, \sigma_{p_i}^2=0.02, \mathcal{T}=0.75, \pi=0.5$$

**Table B.7: Simulation Results for Case 7
Randomize at Level Three—Binary Outcome**

c	p	n	d	$\hat{\mathcal{T}}$	95% CI of $\hat{\mathcal{T}}$
10	4	10	0.16	0.74	(0.73, 0.76)
		20	0.14	0.77	(0.76, 0.86)
		30	0.13	0.72	(0.71, 0.74)
	8	10	0.12	0.72	(0.70, 0.73)
		20	0.11	0.71	(0.70, 0.80)
		30	0.10	0.74	(0.73, 0.79)
	12	10	0.10	0.75	(0.74, 0.76)
		20	0.09	0.71	(0.70, 0.83)
		30	0.09	0.75	(0.74, 0.97)
20	4	10	0.11	0.74	(0.71, 0.77)
		20	0.10	0.70	(0.69, 0.80)
		30	0.09	0.78	(0.77, 0.85)
	8	10	0.08	0.75	(0.71, 0.76)
		20	0.07	0.80	(0.79, 0.91)
		30	0.07	0.76	(0.75, 0.84)
	12	10	0.07	0.74	(0.72, 0.92)
		20	0.06	0.75	(0.73, 0.77)
		30	0.06	0.84	(0.74, 0.93)
30	4	10	0.09	0.75	(0.73, 0.82)
		20	0.08	0.74	(0.71, 0.84)
		30	0.07	0.76	(0.72, 0.80)
	8	10	0.07	0.71	(0.68, 0.85)
		20	0.06	0.73	(0.71, 0.75)
		30	0.06	0.84	(0.74, 0.94)
	12	10	0.06	0.77	(0.73, 0.81)
		20	0.05	0.76	(0.75, 0.85)
		30	0.05	0.79	(0.76, 0.81)

$$\sigma_c^2=0.01, \sigma_p^2=0.39, \mu_c=0.70, \mathcal{T}=0.75, \pi=0.5$$

Table B.8: Simulation Results for Case 8
Randomize at Level Two without Interaction—Binary Outcome

c	p	n	d	$\hat{\mathcal{T}}$	95% CI of $\hat{\mathcal{T}}$
10	4	10	0.14	0.74	(0.71, 0.76)
		20	0.12	0.73	(0.7, 0.75)
		30	0.12	0.73	(0.69, 0.75)
	8	10	0.10	0.70	(0.66, 0.72)
		20	0.09	0.76	(0.73, 0.79)
		30	0.09	0.73	(0.67, 0.78)
	12	10	0.09	0.76	(0.73, 0.78)
		20	0.08	0.78	(0.74, 0.82)
		30	0.07	0.75	(0.73, 0.87)
20	4	10	0.11	0.74	(0.70, 0.83)
		20	0.09	0.76	(0.70, 0.83)
		30	0.09	0.69	(0.65, 0.72)
	8	10	0.08	0.78	(0.75, 0.80)
		20	0.07	0.66	(0.46, 0.68)
		30	0.06	0.68	(0.66, 0.76)
	12	10	0.06	0.72	(0.65, 0.79)
		20	0.05	0.73	(0.67, 0.78)
		30	0.05	0.79	(0.75, 0.82)
30	4	10	0.09	0.73	(0.70, 0.75)
		20	0.08	0.73	(0.68, 0.76)
		30	0.07	0.75	(0.60, 0.77)
	8	10	0.06	0.71	(0.55, 0.79)
		20	0.05	0.73	(0.67, 0.77)
		30	0.05	0.85	(0.81, 0.88)
	12	10	0.05	0.67	(0.60, 0.73)
		20	0.04	0.75	(0.70, 0.79)
		30	0.04	0.74	(0.57, 0.76)

$$\sigma_c^2=0.01, \sigma_p^2=0.39, \mu_c=0.70, \mathcal{T}=0.75, \pi=0.5$$

**Table B.9: Simulation Results for Case 9
Randomize at Level Two with Interaction—Binary Outcome**

c	p	n	d	$\hat{\mathcal{T}}$	95% CI of $\hat{\mathcal{T}}$
10	4	10	0.16	0.72	(0.71, 0.73)
		20	0.14	0.69	(0.67, 0.70)
		30	0.13	0.75	(0.74, 0.76)
	8	10	0.12	0.74	(0.72, 0.76)
		20	0.11	0.72	(0.70, 0.75)
		30	0.11	0.76	(0.72, 0.78)
	12	10	0.11	0.73	(0.72, 0.77)
		20	0.10	0.74	(0.68, 0.78)
		30	0.09	0.75	(0.66, 0.78)
20	4	10	0.11	0.77	(0.74, 0.78)
		20	0.10	0.76	(0.74, 0.78)
		30	0.09	0.70	(0.70, 0.76)
	8	10	0.09	0.73	(0.71, 0.82)
		20	0.08	0.78	(0.72, 0.83)
		30	0.07	0.75	(0.74, 0.84)
	12	10	0.08	0.74	(0.68, 0.78)
		20	0.07	0.74	(0.68, 0.79)
		30	0.07	0.75	(0.65, 0.82)
30	4	10	0.09	0.75	(0.69, 0.76)
		20	0.08	0.70	(0.70, 0.80)
		30	0.08	0.74	(0.71, 0.87)
	8	10	0.07	0.73	(0.65, 0.75)
		20	0.06	0.71	(0.75, 0.86)
		30	0.06	0.73	(0.65, 0.81)
	12	10	0.06	0.70	(0.61, 0.78)
		20	0.06	0.65	(0.57, 0.73)
		30	0.05	0.69	(0.51, 0.86)

$$\sigma_c^2=0.01, \sigma_p^2=0.39, \mu_c=0.70, \sigma_{ct}^2=0.05, \mathcal{T}=0.75, \pi=0.5$$

Table B.10: Simulation Results for Case 10
Randomize at Level One without Interaction—Binary Outcome

c	p	n	d	$\hat{\mathcal{T}}$	95% CI of $\hat{\mathcal{T}}$
10	4	10	0.11	0.76	(0.68, 0.77)
		20	0.08	0.75	(0.72, 0.82)
		30	0.07	0.77	(0.76, 0.92)
	8	10	0.08	0.74	(0.73, 0.89)
		20	0.06	0.74	(0.73, 0.90)
		30	0.05	0.77	(0.75, 0.95)
	12	10	0.07	0.73	(0.71, 0.73)
		20	0.05	0.77	(0.61, 0.80)
		30	0.04	0.72	(0.69, 0.75)
20	4	10	0.08	0.79	(0.67, 0.86)
		20	0.06	0.80	(0.74, 0.79)
		30	0.05	0.73	(0.69, 0.82)
	8	10	0.06	0.68	(0.63, 0.78)
		20	0.04	0.74	(0.64, 0.88)
		30	0.04	0.76	(0.71, 0.77)
	12	10	0.05	0.76	(0.74, 0.77)
		20	0.04	0.78	(0.74, 0.79)
		30	0.03	0.73	(0.67, 0.84)
30	4	10	0.07	0.78	(0.65, 0.81)
		20	0.05	0.75	(0.73, 0.86)
		30	0.04	0.79	(0.68, 0.85)
	8	10	0.05	0.72	(0.67, 0.82)
		20	0.04	0.75	(0.70, 0.83)
		30	0.03	0.76	(0.72, 0.77)
	12	10	0.04	0.79	(0.65, 0.89)
		20	0.03	0.79	(0.65, 0.89)
		30	0.02	0.74	(0.69, 0.84)

$$\sigma_c^2=0.01, \sigma_p^2=0.39, \mu_c=0.70, \mathcal{T}=0.75, \pi=0.5$$

Table B.11: Simulation Results for Case 11
Randomize at Level One with Treatment x Level Three—Binary Outcome

c	p	n	d	$\hat{\mathcal{T}}$	95% CI of $\hat{\mathcal{T}}$
10	4	10	0.13	0.73	(0.72, 0.77)
		20	0.10	0.73	(0.73, 0.74)
		30	0.09	0.75	(0.75, 0.77)
	8	10	0.10	0.73	(0.73, 0.84)
		20	0.09	0.71	(0.70, 0.83)
		30	0.08	0.74	(0.71, 0.82)
	12	10	0.09	0.73	(0.71, 0.77)
		20	0.08	0.74	(0.73, 0.79)
		30	0.07	0.70	(0.69, 0.78)
20	4	10	0.09	0.72	(0.70, 0.87)
		20	0.07	0.71	(0.69, 0.75)
		30	0.06	0.68	(0.65, 0.77)
	8	10	0.07	0.77	(0.75, 0.80)
		20	0.06	0.77	(0.72, 0.87)
		30	0.05	0.76	(0.68, 0.85)
	12	10	0.06	0.70	(0.69, 0.76)
		20	0.05	0.76	(0.69, 0.83)
		30	0.05	0.73	(0.68, 0.87)
30	4	10	0.08	0.74	(0.73, 0.76)
		20	0.06	0.71	(0.69, 0.86)
		30	0.05	0.74	(0.74, 0.75)
	8	10	0.06	0.70	(0.65, 0.77)
		20	0.05	0.73	(0.71, 0.87)
		30	0.04	0.75	(0.72, 0.86)
	12	10	0.05	0.74	(0.70, 0.80)
		20	0.04	0.66	(0.61, 0.87)
		30	0.04	0.75	(0.74, 0.77)

$$\sigma_c^2=0.01, \sigma_p^2=0.39, \sigma_{ct}^2=0.05, \mu_c=0.70, \mathcal{T}=0.75, \pi=0.5$$

Table B.12: Simulation Results for Case 12
Randomize at Level One with Treatment x Level Two—Binary Outcome

c	p	n	d	$\hat{\mathcal{T}}$	95% CI of $\hat{\mathcal{T}}$
10	4	10	0.12	0.76	(0.72, 0.76)
		20	0.09	0.73	(0.73, 0.76)
		30	0.08	0.75	(0.75, 0.77)
	8	10	0.09	0.74	(0.73, 0.77)
		20	0.07	0.71	(0.70, 0.73)
		30	0.06	0.76	(0.71, 0.79)
	12	10	0.08	0.74	(0.71, 0.77)
		20	0.06	0.75	(0.73, 0.79)
		30	0.05	0.70	(0.69, 0.78)
20	4	10	0.09	0.72	(0.70, 0.87)
		20	0.06	0.75	(0.69, 0.76)
		30	0.05	0.73	(0.71, 0.78)
	8	10	0.06	0.70	(0.70, 0.79)
		20	0.05	0.73	(0.71, 0.77)
		30	0.04	0.74	(0.71, 0.76)
	12	10	0.05	0.70	(0.69, 0.76)
		20	0.04	0.76	(0.72, 0.78)
		30	0.03	0.70	(0.69, 0.77)
30	4	10	0.07	0.77	(0.73, 0.78)
		20	0.05	0.72	(0.70, 0.81)
		30	0.04	0.74	(0.74, 0.75)
	8	10	0.05	0.76	(0.75, 0.77)
		20	0.04	0.73	(0.71, 0.77)
		30	0.03	0.75	(0.72, 0.76)
	12	10	0.04	0.74	(0.70, 0.80)
		20	0.03	0.70	(0.70, 0.78)
		30	0.03	0.74	(0.73, 0.77)

$$\sigma_c^2=0.01, \sigma_p^2=0.39, \sigma_{pi}^2=0.02, \mu_c=0.70, \mathcal{T}=0.75, \pi=0.5$$