University of California
Santa Barbara

Cluster-Robust Variance Estimators for Binary Observations in Heterogeneous Groups and Their Application to Psychometric Analyses of Repeated Measures

A dissertation submitted in partial satisfaction of the requirements for the degree

Doctor of Philosophy

in

Statistics and Applied Probability

by

Sarah Marie Marquis

Committee
Professor Andrew Carter, Chair
Professor Yuedong Wang
Professor Alex Petersen

December 2020

The Dissertation of Sarah Marie Marquis is approved.

Professor Andrew Carter

Professor Yuedong Wang

Professor Alex Peterson

December 2020

Acknowledgements

This thesis was made possible thanks to a multitude of teachers and friends.

In particular, I owe a great debt of gratitude to my committee advisor Professor Andrew Carter, who was a patient, kind, and insightful teacher at every step. As his teaching assistant for many years, I saw firsthand that he applies the same amount of care and attention to all of this classes and students as he did to our research, making complicated concepts comprehensible and prompting students to think logically and critically.

I would also like to thank Stefan Cano, PhD, CPsychol, AFBPsS, and Antoine Regnault, PhD for their priceless guidance throughout my research journey. I should probably thank them as well for the help and guidance they will continue to provide as I begin my professional career.

Thank you to my parents, who somehow always tell me exactly what I need to hear. Needless to say this thesis would not have been possible without them or the rest of my wonderful family. I would like to thank my friends Melissa Gordon and Emily Jensen who supported me with their light and energy as well as their psychometric and mathematical knowledge in the daily research grind.

And finally, thank you to Nick and Riley.

Vitae

Sarah Marquis

Education

- Ph.D.: Statistics And Applied Probability University Of California, Santa Barbara - Santa Barbara, CA
- MA: Statistics And Applied Probability University Of California, Santa Barbara - Santa Barbara, CA
- BA: Mathematics and Spanish, Tufts University - Medford, MA

Employment

- Teaching Associate, University Of California, Santa Barbara - Santa Barbara, CA
- Teaching Assistant, University Of California, Santa Barbara - Santa Barbara, CA
- Researcher, Modus Outcomes - Lyon France

Abstract

Cluster-Robust Variance Estimators for Binary Observations in Heterogeneous Groups and Their Application to Psychometric Analyses of Repeated Measures

by

Sarah Marquis

This dissertation is composed of a study of estimation methods in classical and test theories and the elaboration and application of a cluster-robust variance estimator. Variance estimators derived from generalized estimating equations are known to be robust to most covariance structures and are therefore well suited for psychometric analysis of longitudinal test data. However, the approximate normal distribution of the test statistic for clustered binary experiments breaks down when the variation between cluster variances is large. The degrees of freedom for the test statistic are smaller than the number of clusters in unbalanced experiments and closer to an effective number of clusters, G*, which we estimate as the degrees of freedom using Satterthwaite approximation. We calculate a bias bound as a function of G* to improve the coverage percentages of the test statistic. Simulations generated by a beta-binomial model and a Markov chain model show that the bias-adjusted cluster-robust variance estimator improves the test statistic and achieves a coverage percentage of at least 94% for highly heteroskedastic settings.

For conservative confidence intervals in even more unbalanced situations, t-scores with G* degrees of freedom can be used. When compared to a quasibinomial generalized linear model and a wild bootstrap estimator, the bias-adjusted CRVE is closer to the asymptotic distribution for low effective numbers of clusters and yields almost equivalent results to the other two estimators across simulations. Consistency conditions based on cluster heterogeneity are shown to be sufficient for convergence of a chi-square test for testing multiple probabilities across each cluster. We show that the chi-square statistic can be used to test for parallel scales, equivalent items, or time effects in classical test analysis of longitudinal data. Finally, we discuss the use of generalized estimating equations and the multivariate cluster-robust variance estimator in Rasch analysis of repeated measures.

Contents

List of Figures

Preface

The purpose of this thesis is to investigate current estimation methods in psychometric longitudinal analysis, develop conditions for asymptotic normality of a cluster-robust test statistic for binary responses, and explore its potential application in the world of test theory. We focus on the following statistical problem. Suppose the goal is to calibrate a test composed of binary questions which aim to measure the location of individuals on a latent trait of interest in an invariant way. A set of responses is collected from a sample of individuals. Considering a population of independent and identically distributed subjects, each question difficulty on the test can be estimated by taking the usual sample proportions. While this can be a reasonable assumption under well-balanced situations with a large sample size, the estimates depend on the given sample and are not population-invariant. Separating parameters in a logistic model with a set of ability parameters for each subject and item difficulties for each question is a common psychometric approach. This data has a fixed block structure, but unlike usual block-design

experiments, only one observation is collected for each cell. That is, each subject answers each item exactly once. This data is symmetric in form, but our interest is only in the item parameters; in other words, the level of each question, or the probability of "endorsing" each one for an average person. Many current solutions and papers use hierarchical models and focus on estimating subject parameters. How do we compare clustered populations? How do we model covariances? When is specifying covariances not necessary and potentially detrimental to the analysis? This dissertation aims to provide some answers to these questions. This thesis is separated into three parts.

Part I The first part of this dissertation is a literature review that covers all possible models for binary test data. Each model is examined, its assumptions and estimation techniques discussed. Two main branches of measurement theory are discussed, Classical Test theory (CTT) and Item Response Theory (IRT). As discussed at the end of Chapter 1, IRT is revealed to be statistically preferable to CTT and the selection of preferable models is narrowed down to the family of Rasch models. Its model structure, assumptions, statistical properties and estimation methods are discussed in Chapter 1. Chapter 2 extends the models discussed in the previous chapter to model longitudinal studies, which are quite common in education and pharmaceutical sciences. Many options exist and are currently

employed in clinical trials. We discuss data manipulation methods such as stacking, which organize the data in different ways to deal with potential temporal independence violations. The goal of these chapters is to give the reader an understanding of 1) the structure of test data and its particularities 2)ways to model survey responses and longitudinal extensions of those models, 3) Estimation methods for each model and 4) pros and cons of each model. Many interested in this topic come from a measurement-related background and will have used these models in their research. The programs they use perform complex estimation techniques, usually involving iterative numerical solutions, which are somewhat blindly trusted by many social scientists since they require an extensive statistical background to comprehend.

Test theories and related models aim to provide a stable framework for measuring and comparing populations in educational, psychological, and pharmaceutical sciences. It is therefore imperative that the measure itself as well as the handling of measurement errors be sensical and reliable. Following Occam's razor, the best model would be one that would explain the data reasonably well, without being overcomplicated. This delicate balance is a central theme throughout this thesis, and is a fundamental difference between the two main schools of thought in psychometrics, Classical Test Theory (CTT) and Item Response Theory (IRT). IRT encompasses a wide range of

models, including the Rasch model, which is the *only* model to have population-independent item estimates. However, the mathematical rigorousness of IRT models does not translate into blind trust of the resulting estimates. Psychometricians caution that making bold conclusions without proper understanding of background mechanics of IRT, can lead to erroneous conclusions.

Part II The second part of this thesis focuses on a cluster-robust variance estimator for Bernoulli random variables derived from generalized estimating equations. The CRVE has been previously established, but we develop a measure of cluster heterogeneity and show that asymptotic convergence and coverage percentages vary greatly as a function of the effective number of clusters. A bias based on consistency conditions and the effective number of clusters is derived and applied. We look into an approximation to the degrees of freedom for a potential t-distribution. This cluster-robust variance estimator is then extended to a multivariate setting to test multiple probabilities across clustered populations. Much of the theory translates directly and based on the same consistency conditions, we show that a chi-square test with the multivariate CRVE has asymptotic coverages of $1 - \alpha$ for reasonably few clusters, depending on the cluster size imbalance of the experiment. Simulation results from beta-binomial simulation help us verify our theory.

Part III The third and final part of this dissertation applies the cluster-robust variance estimator and corresponding chi-square tests to survey calibration problems. Estimates of item difficulties in longitudinal datasets involve possibly correlated binary or categorical random variables, rendering the estimating process even more complicated and delicate than it already is for IRT models. The adjusted cluster-robust variance estimator is extended for multivariate results in Chapter 6. Chapter 8 finally applies the MCRVE to longitudinal test data and its application to the Rasch model is discussed. The final chapter of this dissertation is a discussion on the results and takeaways from this study into psychometric analysis and new statistical approaches to longitudinal test data.

Part I

Psychometric and Statistical Solutions to Item Estimation in Longitudinal Test Data

"Science must be understood as a social phenomenon, a gusty human enterprise, not the work of robots programmed to collect pure information. " Mismeasure of Man, S.J.Gould

Chapter 1

Psychometric Models for Item Estimation

Psychometrics, the science of measurement of psychological qualities, aims to properly quantify a latent trait of interest on a subject or population with the help of an *instrument*, also called *measure*. When assessing a physical trait such as length, weight or temperature, the instrument is a ruler, scale, or thermometer, measuring a tangible quantity with a standardized unit. These tools provide a stable framework for comparing measurements across different experiments and subjects.

Psychological traits cannot be measured in the same explicit manner. Instead of observing an ability of interest directly, the instrument, or ruler, is a series of questions, or items, most often with binary or categorical responses. The concept that mental abil-

ities or feelings like discomfort can be measured with a metric of some sort through the use of questionnaires, in a way akin to that of physical traits, is at the foundation of psychometrics. Instruments that measure these latent traits must be modeled in a way that provides a frame of reference that should be as robust as for the measurement of physical traits. These instruments composed of questions, often referred to as *items*, are used to place subjects on an ordered scale, similar to the rulers used for physical measurements. The traits being quantified are inherently fluid and subjective. Intelligence or happiness, for example, cannot simply be measured by asking the subject where they are located on the ruler; "How intelligent are you?" or"How happy are you?" are interpreted differently from person to person. Designing a test that objectively and consistently measures the latent ability can therefore be quite and elaborate task.

Test theory was first developed by psychologists and educators (F. M. Lord, 1952) (Hambleton & Jones, 1993) and connected to probability measures mainly by (Zimmerman, 1975) and (F. Lord & Novick, 1968). Logistic models were introduced into the field of psychometrics about a decade later under the form of the Rasch model and item response theory (IRT), bringing with them the theory of linear regression and identifying separate parameters for the subject and the item. We now delve into the mathematical background of classical test theory and discuss measures of reliability.

1.1 Classical Test Theory

Classical Test Theory is governed by a simple set of rules and assumptions which have roots in probability theory. In CTT, the probability distribution of the response of a subject is test-wise. Correspondingly, subjects are compared by their overall test scores and item difficulties, which were derived mostly to be compared to IRT estimates, are determined by the total number of subjects who endorsed them. A test is then validated by checking the correlation between items and the overall test correlation with a reliability coefficient such as Cronbach's alpha. Validity is not discussed in this dissertation as the focus is the theoretical properties of test models and their resulting estimates and standard errors.

Test data is modeled with the simple formulation of three concepts: a true theoretical score, a test score (observed), and an error score. The true score of a person is their observed total score T on a test, with an additive *independent* mean zero error: $X = T + E$. The statistical assumptions were formulated by (F. Lord & Novick, 1968) and (Zimmerman, 1975):

1. test and error scores are uncorrelated,
2. the average error of this population is zero, and
3. error scores on parallel tests are uncorrelated.

Although greatly simplified, these properties stem from axioms

regarding conditional expectations as linear operators in Hilbert spaces. Formally defined by Zimmerman, a *test procedure* is a 5-tuple

$$T = (\Omega, \mathcal{U}, P, f, X),$$

in which Ω is the usual set of all possible observable outcomes, $f : \Omega \to \Phi$ is an assignment of individuals or experimental objects, and $X : \Omega \Rightarrow R$ is an assignment of scores (a realization of the test), such that $(\Omega, \mathcal{U}, P)$ is a probability space, f is a random point, and X is a random variable. In this way, every test procedure denotes a set of conditional random variables $\{X|f = \alpha\}$, $\alpha \in f(\Omega)$, which can be regarded as assignments of scores given particular subjects. The true score is the expected values of the conditional random variable $\{X|f = \alpha\}$, $t = \mathbb{E}(X|f)$. Although Zimmerman's notation suppresses the α here, it is actually inherent in this theory that t is a function of α (which is the subject). In other words, the expected value of the test depends on the subject. In fact, the errors are defined as the distance between the random variable and its conditional expectation

$$e_\alpha = [X|f = \alpha] - t(\alpha).$$

This enables the use of probability theory of projections on Hilbert spaces and therefore gives the desired assumption of uncorrelated,

mean zero errors:

$$\mathbb{E}(e_\alpha) = 0$$

and for two conditionally independent random variables $X_1|\alpha$ and $X_2|\alpha$,

$$\mathbb{E}(e_\alpha^{(1)} e_\alpha^{(2)}) = 0.$$

The random variables X_i described above are always assumed to be uncorrelated, but not necessarily identically distributed. Two realizations of a test which are conditionally independent and identically distributed are called *parallel tests*. Two test procedures T_1 and T_2 are *equivalent* if (f_1, X_1) and (f_2, X_2) are identically distributed, that is, the probability measure induced is the same:

$$P_{X_1|f_1} = P_{X_2|f_2}$$

For more details on the application of probability spaces and Hilbert spaces to classical test theory, see (Zimmerman, 1975).

The fact that classical test theory is "test-"based rather than item based is perhaps why nowhere in the literature review did I find an explicit statement that equated parallel dichotomous tests to a binomial distribution. Two tests scores may have the same probability distribution, but that does not explicitly mean that the

questions making up the test all have the same difficulty level. Independence is however inherently assumed across items so that the probability of a response vector for multiple items is the product of their individual probabilities, all given a constant ability α:

$$\mathbb{P}(X_{11} = x_{11}, \ldots, X_{1k} = x_{1k}) = \prod_{i=1}^{K} P_{1i,\alpha}^{x_{1i}} Q_{1i,\alpha}^{1-x_{1i}}$$

Items with the same difficulty level are often assumed in blocks, so that their probability of success is equal for all items keeping the ability constant. With this assumption, the number of endorsed items given ability α follows a binomial distribution.

Although this thesis is concerned with estimation of variances in particular, it is worth taking the time to understand a little bit about what reliability means in classical test theory, because it is used as a mathematical stamp of approval to show that a test will produce the same results under different circumstances; and the only quantity used to demonstrate this is, under some form or another, the reliability coefficient.

1.1.1 Reliability and Validity

Tests are evaluated based on two concepts, reliability and validity. Reliability addresses the re-test capabilities of a questionnaires, whereas validity focuses on whether the test is measuring what it intends to. Validity is a more philosophical question which we will

not address here. Reliability, however, is evaluated using statistics such as Cronbach's alpha:

$$\Big(\frac{K}{K-1}\Big)\Big(1-\frac{\sum S_i^2}{S_X^2}\Big)$$

the Spearman-Brown Formula, or the Kuder-Richardson 20 or 21. The KR-21 assumes all items have the same difficulty:

$$\Big(\frac{K}{K-1}\Big)\Big(1-\frac{M*(n-M)}{n\operatorname{Var}(X)}\Big)$$

where M is the mean score for the test, and n is the sample size. Statisticians will recognize the standard error of a sample proportion from a binomial experiment.

All are based on the *reliability coefficient* of a test procedure, defined as the ratio $\operatorname{Var} M_x/\operatorname{Var} X$, where $M_x = \mathbb{E}(X|f)$. If a test is taken multiple times, so that $X_1, X_2, \ldots$ are parallel, then the reliability coefficient is given by the correlation between the two, $\rho(X_1, X_2)$. As Zimmerman states, equivalent test procedures have the same reliability coefficient. A measure of reliability for a test which was taken two or more times is therefore to compare the reliability coefficient across test realizations.

It should be noted that reliability is a function of test length, so that adding questions to a test will increase its reliability.

1.1.2 Hypothesis Tests

Suppose a researcher assigned the same test to a placebo and a treatment group. A researcher interested in testing for treatment effect would simply assume two independent populations with normal distributions on the total test scores and do a t-test:

$$\frac{\bar{X} - \bar{Y}}{\sqrt{S_X^2 + S_Y^2}} > t_\alpha$$

The same would be assumed if the researcher wanted to test for a potential time change in a subject's score across two different time points. Again, any assumptions that would be verified prior to conducting these hypothesis tests would involve the reliability or validity of the questionnaire.

As noted above, there is often an assumption of parallel tests or block of equivalent items. Reliability is then checked using the KR-21. In the third part of this dissertation, we propose a chi-square test for these assumptions which deals with potential correlation between time points or subjects.

1.1.3 Criticisms of CTT

CTT is a long implemented straight-forward way to model and compare test scores. Its praised benefits include simplicity and intuitive analytical interpretation; however, it has been widely recog-

nized to have multiple shortcomings for multiple decades (Hambleton & Jones, 1993; Fischer & Molenaar, 1995; Borsboom, 2006; Magno, 2009; Sébille et al., 2010; Petrillo, Cano, McLeod, & Coon, 2015; Maul, 2017). As stated earlier, a stable frame of reference is key when creating a measure which will be used to compare populations, otherwise comparison is meaningless. CTT fails to provide such a robust measure due to the symmetric dependency of its estimates; more specifically, there are no person or item parameters, simply sum scores and correlations. As a result, any estimation of the difficulty level of the test will be dependent on its population and therefore a biased estimate of the item difficulty itself. This will become apparent with the parameterization in item response models in the next section.The dependency between test scores and population samples has been discussed repeatedly (Petrillo et al., 2015; Magno, 2009; Hambleton & Jones, 1993). This idea of relationship between the question and the subject was probably first stated by Frederic Lord , who realized that given the same population sample, the estimated sample distribution of the ability was different given two different tests, as he showed with this simple graph(F. M. Lord, 1952):

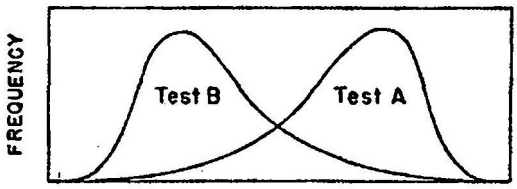


FIG. 1. Distributions of Test Scores

Figure 1.1: Scale dependence of ability estimates

Another shortcoming of CTT is its simplistic assumption that the distribution of the error scores is constant across all subjects, no matter their location on the latent trait. As noted in (Petrillo et al., 2015), there is more precision in the middle of the scale, and it therefore is "counterintuitive that patients' scores at the extremes of the scale (floor and ceiling) have the same level of precision as those in the middle of the scale".

Finally, it seems nearly impossible to find an example of a test which would be deemed *un*reliable. (Maul, 2017) demonstrate this when they create nonsense items which, by CTT standard, pass all the checks to show that the result is a reliable, valid measure. Maul attributes this nonsensical conclusion and observed high correlation to the potential tendency of any test taker to behave consistently in his or her way of answering the questions. This therefore begs the reader to wonder what an unreliable test would look like, and whether reliability is not an inherent part of a unidimensional test, which should be regarded as a default quality rather than evidence

of a reliable measure.

Despite these shortcomings, classical test theory is still a leading method of analyzing test data and creating test which are used as standardized measures of health in today's social sciences (ex: PROMIS Depression Scale). Many testing programs remain firmly rooted in classical measurement models and methods.(Hambleton & Jones, 1993). Somewhat surprisingly, there has been a statistically superior alternative to CTT since the 1960s.

In 1952, Lord brought forth the idea that observed scores and true scores from a test are not synonymous with ability parameters, which are more fundamental because they are test independent, whereas observed scores and true scores are test dependent. This led psychometricians to discuss models that might lead to descriptions of examinees that would be independent of the particular choice of items or assessment tasks that were used in a test as well as item estimates which were independent of the sample from which they were calculated. Item Response Theory came as a response to CTT. Logistic test models were introduced by Allen Birnbaum in technical papers published in 1957 and 1958. George Rasch published his work on the 1-parameter logistic test model in 1960 and presented a solution to sample-free item estimates by conditioning on the sufficient statistics for the subject parameters.

1.2 Item Response Theory

This section introduces the one, two, and three-parameter logistic test models. Item Response Theory was developed as an alternative to the classical approach, in hopes of item parameter estimates independent of subjects abilities. As stated by (Fischer & Molenaar, 1995), " IRT can do the same things better and can do more things, when it comes to modeling existing tests, constructing new ones, applying tests in non-standard settings, and above all interpreting the results of measurement". In contrast with CTT, IRT simultaneously estimates subject ability and item difficulty. For binary responses, this takes the form of a generalized linear models with the canonical link (*logit*) for the Bernoulli distribution. The systematic component is a linear combination of one or more item parameters, along with a latent subject ability. Many models belong to the IRT class, the Rasch model being the simplest (meaning it has the least parameters). It is also referred to as the 1 Parameter Logistic Model (1PL) because the difficulty of an item is represented by one parameter only. The 2-paremeter and 3-parameter models are also widely used in education (ex: standardized tests like the SATs) and health and are therefore discussed briefly below.

The general probability of success of an IRT model can be written

in the following way (De Boeck & Wilson, 2004):

$$\mathbb{P}(Y = 1|\alpha_i, \delta_i, c_i, \theta_j) = c_i + (1 - c_i)\frac{e^{\alpha_i(\theta_j - \delta_i)}}{1 + e^{\alpha_i(\theta_j - \delta_i)}} \tag{1.1}$$

where

- δ_i is the item difficulty level
- α_i is the item discrimination parameter - this regulates the slope of the systematic component, ie of the probability of correctly answering an item as a function of subject ability, given item difficulty.
- c_i is the guessing parameter for the ith item.
- θ_j is the subject ability level.

3 Parameter Logistic Model The 3PL is the only of the three IRT models we study which incorporates a guessing parameter. This model is an example of how a solution can quickly become over-parametrized and lose identifiability. For a set of $N \times L$ binary observations with only one response per subject per item, $3L + N$ parameters must be estimated. (Maris & Bechger, 2009) discuss identifiability issues in the 3 PL .

2 Parameter Logistic Model The 2PL, also called the Birnbaum model, is probably the most commonly used IRT model (for the

moment), because of its accommodating discrimination parameter. The guessing parameter is set to zero ($c_i = 0$ for each i) . Item discrimination is still modeled, but it is assumed that the proportion of guessed answers is negligible. Just like the 3PL, this model has identifiability issues (Fischer & Molenaar, 1995).

The log-likelihoods of the 2PL and 1PL are very similar, and since we derive it for the 1-Parameter model in the next section, we simply give the formula here to prove our point: item discrimination (α_i) and item difficulty parameters (δ_i) cannot be separated and hence neither are identifiable:

$$\begin{aligned} \ell(\boldsymbol{\beta}, \boldsymbol{\delta}|\mathbf{Y}) &= \sum_{j=1}^{N}\sum_{i=1}^{L}\left[\alpha_i(\beta_j - \delta_i)y_{ij} - \ln(1 + e^{\alpha_i(\beta_j - \delta_i)})\right] \\ &= \sum_{j=1}^{N}\sum_{i=1}^{L}\alpha_i\beta_j y_{ij} - \sum_{i=1}^{L}\alpha_i\delta_i y_{i.} - \sum_{j=1}^{N}\sum_{i=1}^{L}\ln(1 + e^{\alpha_i(\beta_j - \delta_i)}) \end{aligned} \tag{1.2}$$

1 Parameter Logistic Model The 1PL arises from setting both $c_i = 0$ and $\alpha_i = 1$ in Equation 1.1 for each item, eliminating both of those parameter vectors. As a result, we will see that this is the only model to have sufficient statistics for its parameters. Subjects' systematic components have parallel slopes on the same item and can therefore be compared. The 1PL is more commonly referred to as the Rasch model, which we heavily focus on in this literature review.

1.3 Rasch Model

The Rasch model is a specific type of generalized linear model for Bernoulli data. There are no informing covariates and no repeated observations, since test data consists of items answered once by each subject. This is similar to a complete block design set up with crossed effects. The standard Rasch model has one ability parameter for each subject, and one difficulty parameter for each question, both fixed. A common alternative is to consider a random distribution on the person ability, greatly reducing the number of parameters needed to estimate. This is equivalent to a generalized mixed effects model with a balanced design. Having a random distribution on a latent variable introduces a specific dependent structure within subject response vectors which is often difficult to evaluate given that this is not simply a linear model. There are three major estimation techniques for the Rasch model: joint, conditional, and marginal maximum likelihood. Although they have been shown to yield similar estimates (Linacre, 2004), all three have problematic pitfalls, which we discuss more below. However, only CML retains sufficiency and uses it to get consistent estimates by conditioning out the subject abilities. The Rasch model is presented in the same manner as David Andrich in *Rasch Models for Mea-*

surement (Andrich, 1988), rather than from a GLM perspective, to highlight the importance of conditional estimation in obtaining consistent estimates.

1.3.1 Structure of the model

Consider an individual answering L binary items on a test or survey. The Rasch model assumes an innate ability B, and a fixed difficulty level $D_i, \quad i = 1, \cdots, L$ for each item. Assuming that a higher ability leads to a higher probability of scoring 1 on more difficult items, the following ratio is set as the *odds of scoring 1 (vs 0) on item i*. In other words, we assume that the log-odds of the probability of success is a linear function of item difficulty and subject ability:

$$\frac{\mathbb{P}(X_i = 1)}{\mathbb{P}(X_i = 0)} = \frac{B}{D_i} = e^{\beta - \delta_i}$$

where $e^{\beta} = B, \quad e^{\delta_i} = D_i$.

This is just a general linear model for a Bernoulli distribution with the canonical logit link. (We briefly describe the structure of GLMs in section 3.) We can therefore rewrite the probability of response of a subject on item i as

$$\mathbb{P}(X_i = x_i) = \frac{e^{\theta_i x_i}}{1 + e^{\theta_i}} = \frac{e^{(\beta - \delta_i)x_i}}{1 + e^{\beta - \delta_i}} \tag{1.3}$$

Equation 1.3 above is the standard form for the Rasch model.

Independence is a key assumption of this model: a person's answers to different items are independent, and different subjects' responses to the same item are independent, given $\beta_1, \cdots, \beta_N$, and $\delta_1, \cdots, \delta_L$. Let $Y_1, Y_2, \cdots, Y_N$ represent the $(L \times 1)$ response vectors for each of N subjects. The observed data

$$\mathbf{Y} = \begin{bmatrix} Y_1 & Y_2 & \dots & Y_N \end{bmatrix}^T$$

has joint probability mass function

$$\mathbb{P}(\mathbf{Y} = \mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\delta}) = \prod_{j=1}^{N} \prod_{i=1}^{L} \frac{e^{(\beta_j - \delta_i)y_{ij}}}{1 + e^{\beta_j - \delta_i}} \tag{1.4}$$

In exponential family form:

$$\begin{aligned} \mathbb{P}(\mathbf{Y} = \mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\delta}) &= \prod_{j=1}^{N} \prod_{i=1}^{L} \exp\left[\theta_i y_{ij} + c(\theta_i)\right] \\ &= \exp\Big[\sum_{j=1}^{N} \sum_{i=1}^{L} \big(\theta_i y_{ij} + c(\theta_i)\big)\Big] \end{aligned} \tag{1.5}$$

From this, we obtain the complete **log-likelihood**:

$$\begin{aligned} \ell(\boldsymbol{\beta}, \boldsymbol{\delta}|\mathbf{Y}) &= \sum_{j=1}^{N} \sum_{i=1}^{L} \left[(\beta_j - \delta_i)y_{ij} - \ln(1 + e^{\beta_j - \delta_i})\right] \\ &= \sum_{j=1}^{N} \beta_j y_{.j} - \sum_{i=1}^{L} \delta_i y_{i.} - \sum_{j=1}^{N} \sum_{i=1}^{L} \ln(1 + e^{\beta_j - \delta_i}) \end{aligned} \tag{1.6}$$

Sufficiency: Expression 1.6 shows by direct consequence of the Fac-

torization Theorem that $(y_{.1}, \cdots, y_{.N}, y_{1.}, \cdots, y_{L.})$ are jointly minimally sufficient for the parameters $(\beta_1, \cdots, \beta_N, \delta_1, \cdots, \delta_L)$. We make the important distinction that $y_{i.}$ is a sufficient statistic for δ_i on its own when subjects are partitioned into groups according to their raw score $y_{.j}$. As described below, conditional maximum likelihood (CML) uses these sufficient statistics to obtain a score function independent of person ability parameters. We first discuss joint maximum likelihood.

1.3.2 Estimation Methods

Joint Maximum Likelihood (JML)

Maximizing Equation 1.6 gives the solution to the score function:

$$y_{.j} = \sum_{i=1}^{L} \frac{e^{\beta_j - \delta_i}}{1 + e^{\beta_j - \delta_i}}, \qquad y_{i.} = \sum_{j=1}^{N} \frac{e^{\beta_j - \delta_i}}{1 + e^{\beta_j - \delta_i}} \tag{1.7}$$

The estimates which solve these equations are referred to as joint maximum likelihood estimates to differentiate them from the conditional and marginal estimates, described below. The JML estimates are therefore the "usual" MLE's which solve Equation 1.7. When both subject and item parameters are unknown, Fischer scoring is used to find a solution.

Ideally, and theoretically, an instrument created should be centered and scaled with respect to person ability. This means no item should be too easy or too difficult for everyone, and no subject should pass or fail every item. Going back to the example of an entire class getting an A on a test: This indicates that the *location* of the items is lower than the average person ability, and therefore, the measure is not centered on its target population. Understandably, problems with estimation occur if these situations arise: if an item was failed by everyone, then it seems reasonable, though impractical, that its estimate should be infinite. JML gives $\pm\infty$ as a solution for any item or subject with a null or perfect score. Since this is not a practical value, it is practice to say that, for example, if an item was failed by all subjects, it is more difficult than all of the subject abilities, but its exact difficulty level is not estimable.

Andersen (Andersen, 1973) showed that when L remains fixed, JML estimates are not consistent as N goes to infinity, which also implies that the estimator is inconsistent. JML estimates are almost twice as over-dispersed as CML estimates (Andersen, 1973), (Linacre, 2004). Consistency for JML can only be established when $N \to \infty, \quad L \to \infty$, and $N/L \to \infty$. This is a major shortcoming since in any applicable situation, the number of test questions is fixed.; often a researcher would like to apply this model to smaller tests and groups. Therefore, adding additional subjects will not help

in obtaining more precise estimates (Fischer & Molenaar, 1995).

Marginal Maximum Likelihood (MML)

For a large sample of subjects, it seems reasonable to suppose that each person's ability is drawn from a random probability distribution, normal distribution being the most common choice. This method is considered when item parameters are of interest, which is the case when attempting to create a measure: item calibration. This leads to a mixed generalized linear model- or GLMM. We discuss estimation methods in the GLMM section of this literature review.

Conditional Maximum Likelihood (CML)

In this scenario, the researcher is interested in item calibration. Subject abilities are therefore first regarded as nuisance parameters. Because a subject's total score $Y_{.j}$ is minimally sufficient for β_j, conditioning on that subject's total score results in a distribution independent of β_j (by definition of a sufficient statistic). Andersen (Andersen, 1971) proved that for a two parameter model, under a set of reasonable assumptions, consistent estimators could be obtained by conditioning on sufficient statistics for one parameter. That is, he showed the consistency of the conditional maximum likelihood estimator for the Rasch model.

Consider an individual answering two items. His total score $Y_{\cdot j} \in \{0, 1, 2\}$ is only helpful in differentiating δ_1 from δ_2 if it is 1. If he answered both incorrectly, or both correctly, then his score contains no information as to which item might be more difficult. The following shows that the probability of scoring on the first item given that the total score is 1 does not depend on subject ability:

$$\mathbb{P}(Y_1 = 1, Y_2 = 0 | Y_{\cdot} = 1) = \frac{\frac{B}{B+D_1}\frac{D_2}{B+D_2}}{\frac{B}{B+D_1}\frac{D_2}{B+D_2} + \frac{D_1}{B+D_1}\frac{B}{B+D_2}}$$

$$= \frac{BD_2}{BD_2 + D_1 B} = \frac{D_2}{D_1 + D_2} = \frac{e^{\delta_2}}{e^{\delta_1} + e^{\delta_2}} = \frac{1}{1 + e^{\delta_1 - \delta_2}}$$

We can obtain $\mathbb{P}(Y_1 = 0, Y_2 = 1 | Y_{\cdot} = 1)$ similarly so that

$$\mathbb{P}(Y_1 = y_1, Y_2 = y_2 | Y_{\cdot} = 1) = \frac{D_1 y_2 + D_2 y_1}{D_1 + D_2}.$$

Recall that $D_i = e^{\delta_i}$, $B_j = e^{\beta_j}$, and therefore:

$$\mathbb{P}(Y_1 = y_1, Y_2 = y_2 | Y_{\cdot} = 1) = \frac{y_1 e^{\delta_2} + y_2 e^{\delta_1}}{e^{\delta_1} + e^{\delta_2}} = \frac{e^{-(y_1\delta_1 + y_2\delta_2)}}{e^{-\delta_1} + e^{-\delta_2}},$$

where the last part is rewritten for convenience. We can therefore obtain a likelihood independent of the β's by conditioning on the

independent subjects who scored 1:

$$L(\delta_1, \delta_2 | Y_{.1} = 1, Y_{.2} = 1, \cdots, Y_{.N} = 1) = \prod_{j=1}^{N} \mathbb{P}(Y_{1j} = y_{1j}, Y_{2j} = y_{2j} | Y_{.j} = 1)$$
$$= \prod_{j=1}^{N} \frac{e^{-(y_{1j}\delta_1 + y_{2j}\delta_2)}}{e^{-\delta_1} + e^{-\delta_2}}$$

The log-likelihood:

$$\begin{aligned} \ell(\delta_1, \delta_2 | \mathbf{Y}) &= \sum_{j=1}^{N} -y_{1j}\delta_1 - y_2 j \delta_2 - \ln\left(e^{-\delta_1} + e^{-\delta_2}\right) \\ &= -y_{1.}\delta_1 - y_{2.}\delta_2 - N \ln\left(e^{-\delta_1} + e^{-\delta_2}\right) \end{aligned} \tag{1.8}$$

The resulting maximum likelihood equations for δ_1 and δ_2 are:

$$y_{1.}/N = \hat{\pi_1} \qquad y_{2.}/N = \hat{\pi_2} \tag{1.9}$$

where $\pi_i = \mathbb{P}(Y_i = 1 | Y_1 + Y_2 = 1) = \frac{e^{-\delta_i}}{e^{-\delta_1} + e^{-\delta_2}}, \quad i = 1, 2$. Note that this is just a binomial distribution on a subset of the sample, where each independent subject either scored on item 1, or did not (and therefore scored on item 2). Therefore the sample proportions are the best estimators for π_1 and π_2. $\delta_i - \delta_j$ can be retrieved by taking the logit of π_i and numerical solutions are found thanks to the imposed constraint $\sum \delta_i = 0$. This is similar to a sign-test, where only observations that differ in the two categories are used.

The possible combinations that make up an individual's total score increase faster as the number of items grows. Extending this theory to L items, we partition the subjects according to their individual total scores. Subjects with scores of 0 or L will not help in estimating item difficulties. For each total score $R = 1, 2, \cdots, L-1$, let n_R be the number of subjects with that score. For instance, for three items, excluding 0 and perfect scores, the remaining ones are partitioned into two groups, $R = 1; \{(1,0,0),(0,1,0),(0,0,1)\}$ and $R = 2; \quad \{(1,1,0),(1,0,1),(0,1,1)\}$. Focusing on the group with success on 2 items out of 3, the conditional PMF is:

$$\mathbb{P}(Y_{1j} = y_1, Y_{2j} = y_2, Y_{3j} = y_3 | Y_{.j} = R = 2) = \frac{e^{-(y_1\delta_1 + y_2\delta_2 + y_3\delta_3)}}{e^{-\delta_1-\delta_2} + e^{-\delta_1-\delta_3} + e^{-\delta_2-\delta_3}} \tag{1.10}$$

For L items, we obtain a maximum likelihood equation for *each* of the observed acceptable ($R \neq L, 0$) total scores , with a maximum of $\binom{L}{1} + \cdots + \binom{L}{L-1} = 2^{L-1}$ possible combinations. This partitions the subject population according to total score. Since there are no repetitions, responses are still independent. For one given score, the

likelihood is:

$$L(\delta_1, \cdots, \delta_L | \mathbf{Y}, R = r) = \prod_{j=1}^{n_r} \mathbb{P}(Y_{1j} = y_{1j}, \cdots, Y_{Lj} = y_{LJ} | y_{\cdot j} = r), \quad y_{ij} \in \{0, 1\}$$

$$= \prod_{j=1}^{n_r} \frac{e^{-\sum_{i=1}^{L} y_{ij}\delta_i}}{\sum_{(y)|r,k=1}^{K} \gamma_{rk}} = \frac{e^{-\sum_{i=1}^{L} y_{i\cdot}\delta_i}}{\left(\sum_{(y)|r,k=1}^{K} \gamma_{rk} \right)^{n_r}}. \tag{1.11}$$

where the $\gamma_{rk} = e^{-\delta_1^* - \cdots - \delta_r^*}$ correspond to the $k = 1, \cdots, K = \binom{L}{r}$ possible combinations, and the δ^*'s represent the items answered correctly for that kth possible combination. These are the elementary symmetric functions of order r. To obtain the final conditional likelihood, we use the fact that we have partitioned the subject pool into independent groups by raw score and combine them:

$$L(\boldsymbol{\delta} | \mathbf{Y}, R) = \prod_{r=1}^{L-1} L(\delta_1, \cdots, \delta_L | \mathbf{Y}, R = r) = \frac{e^{-\sum_{i=1}^{L} y_{i\cdot}\delta_i}}{\prod_{r=1}^{L-1} \left(\sum_{(y)|r,k=1}^{K} \gamma_{rk} \right)^{n_r}}. \tag{1.12}$$

Let $\pi_{ri} = \mathbb{P}(Y_i = 1 | Y_\cdot = r)$. Then Equation 1.13 is a general solution equation to differentiating Equation 1.12 and setting it equal

to zero.

$$y_{i.} = \sum_{r=1}^{L-1} n_r \hat{\pi_{ri}}, \qquad i = 1, \cdots, L. \tag{1.13}$$

These equations must be solved with numerical iterations under the constraint that $\sum_{i=1}^{L} \delta_i = 0$. The solution algorithm, dubbed "CON procedure", has been implemented in R and is used over JML methods despite its complexity, because of the consistency of the estimators (Andrich, 1988).

In his review of estimation methods for Rasch models, (Linacre, 2004) cautions against CML for small samples, pointing out that JML estimates can in some cases have less bias than their CML counterparts. This is due to the fact that the likelihood in CML is computed using only non-extreme scores, and for a low number of items, this eliminates many possible response vectors. For more concrete examples, see (Linacre, 2004).

Pairwise estimation All pairs of items are collected into a data matrix, with the $(i, j)^{th}$ entry representing the number of subjects who answered item i correctly and item j incorrectly. Conditioning on the total score being equal to 1 for each pair, a pseudo-likelihood (Eq. 1.14) is then evaluated and shown to be consistent at some cost of efficiency. It is not the likelihood of the data because clearly, the pairs are not independent. This approach is

similar to the generalized estimating equations, discussed later in this review in that it uses the "wrong" likelihood to obtain consistent estimates. The standard errors, however, are estimated using JML, which leaves room for improvement since estimation methods for the mean and variances estimates have entirely different solution algorithms. Taking the Rasch model as in Equation 1.3, let $f_{ij} = \frac{e^{\delta_i}}{e^{\delta_i}+e^{\delta_j}} = \mathbb{P}(Y_i = 1, Y_j = 0 | Y_i + Y_j = 1)$. Then the pseudo-likelihood is simply:

$$\ell(\boldsymbol{\delta}) = \sum_{j=1}^{N} \sum_{i<j}^{L} y_{ij}(1 - y_{ij}) \log(f_{ij}) + (1 - y_{ij}) \log(1 - f_{ij}) \quad (1.14)$$

(Andrich, 1988; Zwinderman, 1995)

Comparing Estimation Methods

Joint, marginal and conditional maximum likelihood methods all lead to complicated algorithms. JML, which uses Newton-Raphson to maximize the joint likelihood, has been shown to have consistency problems; in other words, its estimates are not guaranteed to converge to their true values for small samples or small subject-item ratios. CML uses sufficient statistics to obtain consistent estimators of the item difficulties and is therefore the only method which yields population-independent estimates. Conditioning and partitioning by total score however yields multiple estimates for the same parameter, and the number of estimates only grows with the number

of items on a test. Pairwise estimation greatly simplifies calculations by using a pseudo-likelihood between all item pairs, treating them as independent. However, the methods with which standard errors for CML estimates are being estimated in current software do not match the estimation method for the parameters, which seems ad hoc at best.

Both joint and conditional maximum likelihood methods first exclude all null and perfect scores from the set of observations, reducing the sample size, since they result in no information for CML and infinite parameter estimates for JML. This burden is not shared by marginal maximum likelihood estimation. Computing methods are also different in MML since the likelihood is a Gaussian integral. The downfall of marginal maximum likelihood estimation is that of a a generalized linear mixed model. The estimates of the fixed effects have been shown to not be robust to misspecification of the random effect component (Greene, 2002; Hubbard et al., 2010). In their discussion on estimation methods (Fischer & Molenaar, 1995) recommend using CML estimates: "MML requires to estimate or postulate a distribution for the latent trait, and if this is wrongly estimated or postulated, the MML estimates may be inferior. Moreover, CML stays closer to the concept of person-free item assessment." Although current software like R have implemented CML, MML seems more common, and pairwise CML is only available in the Rasch software RUMM.

1.3.3 Subject parameter estimation

Once items have been estimated, the test can be administered to new subjects to evaluate their ability in this particular trait, considering the items parameters to be known. Various estimators exist for person ability in this case: the usual MLE, Bayes model estimator (BME), weighted likelihood estimator (WLE), and Bayes expected a posteriori (EAP) (Fischer & Molenaar, 1995). We list these methods but do not plan on focusing on them. Rather, we focus on item parameter estimation and look at IRT in a GLMM context, with cluster-robust methods in mind.

The models discussed in this chapter model the application of one measure to a sample of individuals exactly once, but longitudinal studies and repeated measures are a frequent occurrence in social psychometric sciences. How should psychometricians deal with longitudinal data? On one hand, repeated measures imply a larger sample size. On the other hand, there is very good reason to expect some sort of dependence between time points within subject responses, and treating observations as independent will lead to underestimated standard errors in the presence of positive correlation. The next chapter explores these questions.

Chapter 2

Psychometric Approaches to Longitudinal Data and Repeated Measures

This chapter extends the literature review in Chapter 1 to a search for longitudinal adaptations or approaches to model repeated measures. We first discuss solutions practiced in classical test theory.

2.1 Classical Test Theory and Mixed Models

In order to explain the evolution of scores, an approach in CTT is to use a mixed effects model directly on the score:

$$\begin{aligned} \boldsymbol{S}_i &= X\beta + \mathbf{e}_i, \\ \text{Var}(\boldsymbol{e}_i) &= \Sigma, \\ \boldsymbol{S}_i &\sim N_{n_i}(X\beta, \Sigma), \end{aligned} \tag{2.1}$$

where $\boldsymbol{S}_i$ is the score on the i^{th} individual and $S_i^{(t)} = \sum_j y_{ij}^{(t)}$. Possible covariance structures include 1) unstructured, 2) AR(1), assuming correlation decreases over time but variance is constant, 3) ARH(1), for modeling correlations that decrease over time but constant variances, and 4) CSH, which assumes heterogeneous variances are not equal but constant correlation over time. AIC can be used to chose best covariance structure, as for a common linear regression problem.

Hypothesis Test for Time Effect

The following F-test can be used to detect any potential time effect. $\mu^{(t)} = (\mu_1, \mu_2, \mu_3)' = \mathbf{X}\beta$

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu \Leftrightarrow L\beta = 0, \quad L = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix}$$

$$H_1 : \text{ not } H_0$$

$$F_L = \left(L\hat{\beta}\right)\prime\left(L\widehat{\boldsymbol{V}}_\beta L\prime\right)^{-1}\left(L\hat{\beta}\right)/rank(L) \overset{H_0}{\underset{appr}{\sim}} F_{r,df}.$$

In a comparison of current psychometric approaches for analysis of longitudinal data, (Blanchin et al., 2011) compare different methods, including fitting a mixed effects model directly onto the scores as well as using a Rasch model extension. The results show that the CTT and IRT methods have comparable Type I errors and power overall.(Blanchin et al., 2011) In the last part of this dissertation, we apply a multivariate cluster-robust variance estimator to this test.

2.2 Nonparametric Solutions Using the Rasch Model

This section is mostly an overview of the existing methods which extend the Rasch model to a longitudinal setting. For reasons of parameter identifiability, and because of its properties, the Rasch model will be the only IRT model of interest from now on.

The Rasch model assumes a two-dimensional structure between persons and items where each person answers each item exactly once. There are many instances where one would like to measure the evolution of a trait in a longitudinal study. However, this means repeated observations, which violates the independence assumption of the Rasch model since two observations from the same person

are usually not independent: with reason, we can assume that personal traits are consistent and carried throughout an individual's responses over a given time period. Current solutions for this are to ignore the correlation, add a third dimension to the systematic component of the model, or parameterize items and subjects according to time. We argue that item parameters should be fixed over subjects and time, and that therefore only subject parameters should vary over time. This also allows us to measure personal changes over time, which is often the main goal of longitudinal studies. The focus throughout the published literature seems largely on the estimation of subject parameters. Item parameter analysis is often overlooked, even though it occurs first in the estimation process (Blanchin et al., 2011).

The most common solutions for applying the Rasch model to longitudinal studies are to ignore the potential dependence between observations and proceed with the usual estimation methods after some data manipulation. These solutions are called "anchoring", "stacking", and the Mallinson Approach (Mallinson, 2011).

Anchoring

Anchoring is the process of estimating item difficulties with a single time point and using those estimates for person parameter estimation on all time points. Theoretically, if item levels have been

fixed as with this method, subjects' change over time should be reflected in their ability estimates. One potential problem with this is if there is a dramatic time effect. The item range might be appropriate for the first time point but will be off-center by the end of the study, creating floor and ceiling effects. If there are only two time points, there is no way to choose a middle ground. Additionally, this method greatly reduces the number of observed responses. Data in longitudinal health studies on rare diseases may already be scarce; in that case, anchoring is probably not the method of choice.

Stacking

Stacking ignores possible dependence between observations coming from one subject. Each subject-item value is the sum of the observed responses. Estimation is then done in the usual manner, treating the observations as fully independent. Assuming positive correlations within subject response vectors, this will result in underestimation of the variance - which translates to over-restrictive confidence intervals.

Mallinson Approach

In the Mallinson approach, a random sample of patients is gathered at each time point so that each subject is used only once for across time points in the estimation of the items. This removes potential local dependence within subject responses. Once items are

calibrated in this way, subject parameters can be estimated throughout the time points. This approach seems impractical because it requires a large number of subjects: If there are two time points, twice as many patients are required as in a single time-point study. If the longitudinal study is done over say, seven time points, then seven times as many patients will be required.

Comments

Contrary to longitudinal models discussed later in this literature review,anchoring, stacking and the Mallinson approach all use fixed effects on subject parameters. This allows for estimation of actual subject abilities rather than an overall population mean and variance. The models considered below all use a normal distribution on latent population ability, as it reduces the number of parameters to a single mean and variance for the whole population.

2.3 Extended Rasch Models

2.3.1 Discrete Mixture Distributions

For clustered data, mixture distributions are considered. Often called mixture Rasch models, these are not to be confused with GLMMs. Here it is assumed that there are latent subpopulations, and within each, the simplest form of the Rasch model holds for the measure in question. Ability parameters are estimated sepa-

rately within each class. In addition, one must estimate or guess the number of classes, or subpopulations, as well as the probabilities of membership to each of the classes. However, it is assumed that every person in a subpopulation will have the same ability estimate, at least reducing the number of subject ability parameters. The overall probability of a response vector is:

$$P(\mathbf{x}) = \sum_{c=1}^{C} \pi_c P(\mathbf{x}|c, \beta_c, \boldsymbol{\delta}) = \sum_{c=1}^{C} \pi_c \prod_{i=1}^{L} \frac{e^{(x_i\beta_c - \delta_i)}}{1 + e^{(\beta_c - \delta_i)}} \tag{2.2}$$

Since class membership and frequency are unknown, the EM algorithm is used to solve the estimation problem (Fischer & Molenaar, 1995). Within each Maximization step, Newton-Raphson procedure is used. More details can be found in the computing methods section. (Willse, 2011) studies JML methods for a multinomial-response Mixture Rasch Model using the EM algorithm in this way .

As with normal Rasch analysis, once a model has been specified, there can be multiple choices for a parameter estimate. The longitudinal Rasch model parameterizes subject abilities according to time. In the estimation process, one must choose whether to place a random distribution on those values (in which case MML is used) or to keep the values fixed and link time effects using some sort of latent linear model (in which case CML is used). When a random effect is assumed on subjects, the estimate is the mean of the pos-

terior distribution (EAP) which does not seem ideal if the goal is to measure any change in the sample group or in individuals: the EAP is a shrinkage estimator used in conjunction with the latent random effect and assigns every subject the same estimate.

2.3.2 Rasch Poisson Count model

The Rasch Poisson Counts Model (Jansen, 1997) application to longitudinal studies: consider repeated observations on the same item, or multiple items of the same difficulty - each item is a column, each subject is a row, and each cell is the count of (un)successful responses. The probability of a response matrix for a test with n time-points or repeated measurements is then:

$$\mathbb{P}(X_{ij} = x_{ij}) = \frac{e^{-n\lambda_{ij}} \left(n\lambda_{ij}\right)^{x_{ij}}}{x_{ij}!}, \qquad \lambda_{ij} > 0,$$

where

$$\lambda_{ij} = \theta_j \delta_i,$$

Both parameter vectors are constrained to positive numbers since the rate of a Poisson RV is positive. Item difficulties (δ_i's) are assumed to be fixed and constrained to sum to 1 to avoid indeterminacy. Subject abilities (θ_j's) are given a Gamma distribution as it is conjugate to Poisson random variables. Introducing a random effect

on subject abilities allows the model to account for overdispersion, since the unconditional variance of the raw scores will be greater than their means. Concerns regarding this model include

1. Poisson distribution is supposed to have infinite domain. Is this approach justifiable for when the number of repeated observations is small?

2. Poisson distribution is the limiting distribution of a binomial with large N and small p, as $N \to \infty$. It is not representative of the underlying distribution when p is not close to 0 or 1, or when N is small.

3. Poisson distribution can be used for longitudinal settings when the items are dichotomous but this might not extend well to items with categorical answers. Andrich's approach of treating each threshold as a empirically independent binary random variables in a conditioned space and taking values in $\{0, 1\}$ could be extended to a longitudinal setting. However in order to obtain any sort of sensible estimate for the item thresholds, the assumption that item threshold distances are constant throughout items and subjects would have be to made, and that is a strong assumption.

2.3.3 Time as a Third Dimension - Multi-Facet Rasch Model

Time is incorporated into the model structure as its own separate effect. This is similar to incorporating a school or grader effect. In these models, we assume that the time, school or grader effect is constant across subjects and items. See (Farindon, 2007) for more details.

$$\mathbb{P}(Y_{ijt} = y|\theta_j; \delta_i; \lambda_t) = \frac{e^{y(\theta_j - \delta_i + \lambda_t)}}{1 + e^{(\theta_j - \delta_i + \lambda_t)}} \qquad y \in \{0, 1\} \tag{2.3}$$

This model makes the rather stringent assumption that time effects for each subject and item are the same. λ_t is the average time effect on the model across all items, all subjects. This seems inappropriate for any longitudinal study where subjects do not evolve in a parallel manner and does not lend well to longitudinal studies, since a primary interest in these cases is to use calibrated tests to measure individual changes. Rather than estimating an average time effect, a model which parameterizes time into subject abilities might be preferable.

2.3.4 Andersen's Dichotomous Longitudinal Rasch Model

Andersen's Longitudinal Rasch model transforms item difficulties and subject abilities from a single value into vectors indexed through time. Suppose we want to measure over T time points. Each subject

is the following ability vector: $(\theta_j)_{t,t=1,\ldots,T}$.The probability of success for item i, subject j at time t is then:

$$\mathbb{P}(Y_{ij}^{(t)} = y^{(t)}|\theta_j;\delta_i) = \frac{e^{y^{(t)}(\theta_j^{(t)}-\delta_i)}}{1+e^{(\theta_j^{(t)}-\delta_i)}} \tag{2.4}$$

Marginal Maximum Likelihood

A multivariate normal distribution is usually assigned to subject abilities. The marginal likelihood for this model is

$$L(\delta,\mu,\mathbf{\Sigma}|\mathbf{y}) = \prod_{j=1}^{N}\int_{\mathbb{R}^T}\prod_{t=1}^{T}\prod_{i=1}^{L}G(\theta/\mu,\mathbf{\Sigma})\mathbf{d}\theta \tag{2.5}$$

where $G(./\mu,\Sigma)$ is the multivariate distribution function with mean vector $\mu = (\mu_1,\ldots,\mu_T)\prime$ and covariance matrix Σ. Usually we take $\mu_1 = 0$.this implies that $\hat{\mu}_t$ represents the deviation from 0, ie the differences in time $\hat{d}_{1t}, \quad t = 1,\ldots T$.

2.4 Hierarchical Logistic Test Model

Let X_{vit} denote response given by subject v, on item i, at time t. The Linear Logistic Test Model considers a RM where items are allowed to vary over time, while subject abilities are considered constant over time. This does not seem like a good representation of the data, as items are inherently static, while individuals may fluctuate over time.

Item parameters are sometimes also modeled as linear combinations of both fixed and random parameters. A third dimension is added and interpreted as a school effect. Let p_{ijm} be the probability that person j from school m answers item i correctly. Then the systematic component of a GLM with k predictor variables is

$$\eta_{ijm} = \beta_{0jm} + \beta_{1jm}X_{1ijm} + \beta_{0jm}X_{2ijm} + \cdots + \beta_{0jm}X_{(k-1)ijm}$$

where $i = 1, \ldots, k-1, \quad j = 1, \ldots, n, \quad m = 1, \ldots, r$ and X_{qijm} is the qth dummy variable for the corresponding item, student and school (or other third dimension)(Kamata, 2001). This articles uses set-to-0 identifiability constraint (rather than sum-to-0) so that the first item is the "referencing" item.

$$\begin{cases} \beta_{0jm} = \gamma_{00m} + u_{0jm} \\ \beta_{1jm} = \gamma_{10m} \\ \vdots \\ \beta_{(k-1)jm} = \gamma_{(k-1)0m} \end{cases}, \qquad u_{0jm} \sim N(r_{00m}, \tau_{\{\gamma\}})$$

where γ_{00m} is the overall effect of the referencing item in school m. To test variation across school, one could model

$$\begin{cases} \gamma_{00m} = \pi_{000} + r_{00m} \\ \gamma_{10m} = \pi_{100} \\ \vdots \\ \gamma_{(k-1)0m} = \pi_{(k-1)00} \end{cases}, \qquad r_{00m} \sim N(0, \tau_\pi)$$

Here the π's are fixed. r_{00m} is the random effect associated with school m and can be interpreted as average ability of students in school m. A small variance τ_m would of course imply that there is not great variation between schools. u_{0jm} is a person- and school-specific ability and indicates how much that student deviates from the average ability of students in school m. Note that this can be extended to a latent regression model with person characteristic variables (Olsbjerg & Christensen, 2015).

2.5 Concluding Thoughts

Psychometric analyses of longitudinal data were assessed by studying research papers which applied different methods and empirically compared their statistical properties. Some researchers use random effects generalized linear models, but because interest in these experiments is usually to measure individual progressions though time, other options are usually considered as well. Antoine Barbieri

published a number of articles comparing different ways to model longitudinal data, and had some pertinent concluding thoughts regarding CTT methods. In particular, he highlights that ordinal data is bounded and not always symmetrical- while normal distribution (assumed in a linear mixed model) assumes a continuous, unbounded symmetric relationship. As we will see in simulation output, floor and ceiling effects can heavily bias the item difficulty estimates; confirming a concern mentioned by (Barbieri, Tami, et al., 2017; Barbieri, Peyhardi, et al., 2017; Barbieri et al., 2015). On the other hand, he notes that IRT models are used in development and validation of questionnaires but rarely for longitudinal analysis of HRQoL in clinical trials (Anota et al., 2014).

This literature review encompasses the different approaches taken in CTT and IRT. As the reader may have noted, almost all use a random effect, which means that any determined covariance structure will not be robust to misspecification and will affect estimation of item parameters.

We now turn to statistical approaches for correlated binary sequences. Generalizes linear models clearly overlap here and will be discussed quite briefly in the next chapter, as we have spend time on more specified versions of them here.

Chapter 3

Models for Correlated Binary Data

Dichotomous test data is nothing other than sequences of potentially correlated binary responses. While the popularization of IRT bridged the gap between psychometrics and probability theory of generalized linear models, not all statistical options have been explored. In this chapter we focus on statistical approaches to correlated binary data. We look at varying degrees of parameterization and ask ourselves when modeling complicated underlying structures with unverifiable assumptions is worth the complication and the computing effort, especially if resulting estimates are not robust to misspecified covariance structures. Sensible interpretation of the results is a key component of the analysis as these are applied in social sciences and used to make decisive statements to back research

hypotheses.

These the

3.1 Generalized Linear Mixed Models

Generalized linear models are a common way to analyze data which does not have a continuous distribution with additive error. Instead of modeling the outcome of each observation, we model a function of the parameters. The model consists of three components:

1. The *random component* encompasses all of the variation of the model. It describes the distribution of the observed data.

2. The *systematic component* is a linear function of possibly unknown parameters/predictors. In mixed GLMs, these variables can be fixed or random.

3. The *link function* establishes a relationship between the random and systematic component. Instead of modeling observations directly like in linear models, a function (the link function) of the mean is assumed to be a linear combination of the predictors. The most common links for the binomial distribution are *logit* and *probit*.

While hierarchical and random effects generalized mixed models have already been somewhat discussed, the computing methods behind the models are rarely discussed and even less understood

by psychometricians. The binomial distribution has the following likelihood score function:

$$\frac{\partial l}{\partial \beta} = X^T W \begin{bmatrix} (y_1 - \mu_1) g'(\mu_1) \\ \vdots \\ (y_n - \mu_n) g'(\mu_n) \end{bmatrix} = 0.$$

where

$$W = \begin{bmatrix} m_1 & (1-\mu_1) & & \\ & & \ddots & \\ & & & \mu_n(1-\mu_n) \end{bmatrix}.$$

For logit link, $g\prime(\mu_i) = \frac{1}{\mu_i(1-\mu_i)}$.

We briefly come back to the Rasch model. Assuming a crossed, balanced design with fixed effects (δ for item level and β for person ability) leads to a normal GLM. The score function takes the form of Equation 1.7. Having considered the issues with this estimation method and the complexity of the CML, we turn to a normal distribution on the subject ability parameter. This is logical for data with large samples, especially if marginal techniques are used, since at first subject abilities are considered nuisance parameters.

3.1.1 Probit-Normal Model

The probit link is the most common alternative to the logit link. They have been compared and have been shown to give comparable values, especially around a probability of one half, since both functions are locally linear. While the logit link arises naturally from exponential families and the log odds ratio is an interpretable function of the parameter of interest p, a probit link does not lend to such a nice interpretation. We discuss this more in the context of a model for rat litter survival data comparing a control and treatment group as analyzed by McCulloch, Searle, & Neuhaus, 2001. The structure is a GLMM with the probit link and a normal distribution on a latent trait. Only a sum of indicators, $W_i, i = 1, 2$ is observed: the survival of rats in each litter. A continuous distribution is assumed on the probit of the survival probability, with litter as the random effect. Survival rate is calculated in both control and treatment groups by taking the ratio of survival from day 21 of the treatment to day 4. McCulloch introduces the data in the following way:

$$Y_{ij} = \mu_i + u_{ij}; \qquad W_{ijk} = \begin{cases} 1, & w.p. \quad \mathbb{P}(Y_{ij} < 0) \\ 0, & w.p \quad 1 - \mathbb{P}(Y_{ij} < 0) \end{cases}, \qquad u_{ij} \sim N(0, \tau^2)$$

for $i = 1, 2$ *(group)*, $j = 1, \cdots, 16$ *(litter)*, $k = 1, \cdots n_j$ *(rat)*, where

$$\Phi^{-1}(p_i|u_{ij}) = \mu_i + u_{ij} \quad \Leftrightarrow \quad p_i|u_{ij} = \mathbb{P}(Z < \mu_i + u_{ij}).$$

does not depend on j, so the W_{ijk}'s are independent within each group $i = 1, 2$. Therefore, we have $Y_{ij} \overset{indep}{\sim} N(\mu_i, \tau^2)$, $W_{ijk} \overset{indep}{\sim} Ber(p_{ij})$ and $W_i := \sum_{j=1}^{16} \sum_{k=1}^{n_j} W_{ijk} \sim Bin(\sum n_j, p_i)$, where W_i represents the number of rats who survived in group $i, i = 1, 2$.

$$\mathbb{E}(Y_{ij}|W_{ij} = w) = \int_{-\infty}^{\infty} y f(y|w) dy$$

Let Ψ be the conditional density, as follows:

$$\Psi(y; w, n, \mu, \tau) = \frac{\Phi(y\tau + \mu)^w (1 - \Phi(y\tau + \mu))^{(n-w)} \phi(y)}{\int_{-\infty}^{\infty} \Phi(y\tau + \mu)^w (1 - \Phi(y\tau + \mu))^{(n-w)} \phi(y) dy}$$

And so, we get the conditional expectation of the complete data given observed values W:

$$\mathbb{E}(Y|W) = \int_{-\infty}^{\infty} \alpha \Psi(\alpha; w, n, \mu, \tau) d\alpha$$

$$\mathbb{E}(Y^2|W) = \int_{-\infty}^{\infty} \alpha^2 \Psi(\alpha; w, n, \mu, \tau) d\alpha$$

From these, we obtain estimates the mean and variance of the unobserved continuous random variable.

3.1.2 Computing Methods

The unobserved nature of general linear models and Generalized mixed linear models makes estimation procedures often quite computationally heavy, and some iterative algorithm is required to converge to a solution. We studied each of these methods and included some computational details in the appendix, section 9.5.1. However the next chapter looks at a CRVE which avoids latent specification of a random effect, so we will not be using these computational tools.

GLMMs are not robust to misspecification of the random effect

Many researchers caution using random effects models without careful consideration first: "Introducing random effects into the models affects estimation of the fixed effects when the distributions of the random effects are misspecified" (Hubbard et al., 2010), introduces bias into the fixed effect parameter (Greene, 2002), has interpretability issues and "leads to nonsensical inference " with "misleadingly arrow confidence intervals, large t-statistics and low p-values" (Freedman, 2006; Hubbard et al., 2010; Cameron & Miller, 2015). An important observation which was also noted in the simulation is that the need to control such within-cluster correlation only increases with the number of observations within a cluster.

When the covariance structure between clustered observations is an estimation hurdle rather than the parameter of interest, a population average model which is robust to variance misspecifications is therefore a sensible alternative to mixed effects models.

3.2 Beta-Binomial Distributions

The beta-binomial distribution is a hierarchical GLM often used to model overdispersion in binomial experiments. The probability of success for each cluster is drawn from a beta distribution, and the number of successes per cluster, denoted Y_g, is such that

$$Y_g|p_g \sim Bin(n_g, p_g)$$

A common parametrization for the beta distribution on the cluster probabilities uses two shape parameters $\alpha, \beta > 0$. Throughout the rest of this document, the mean parameter p and "correlation" parameter (as we will call it) γ will be used, where

$$\mathbb{E}(p_g) = \frac{\alpha}{\alpha+\beta} := p, \qquad \text{Var}(p_g) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} := \gamma p(1-p) \tag{3.1}$$

The unconditional mean of the binary sums Y_g is unchanged due to the tower property, while the variance is at least as great as the variance of a binomial model:

$$\mathbb{E}(Y_g) = n_g p, \qquad \mathrm{Var}(Y_g) = p(1-p)\big[(1-\gamma)n_g + \gamma n_g^2\big]$$

Simulations in later chapters will use the $\gamma \in [0, 1]$ as a tuning parameter for the amount of variance in each cluster. The variance of Y_g is equal to that of a binomial when $\gamma = 0$ and is of order n_g^2 when $\gamma = 1$.

This model has many advantages. First, it places the random effect directly on the probability of success rather than through a link function. This allows for more direct interpretation.

Second, the beta distribution can have a skewed, bimodal or bell-shaped curve depending on the value of its parameters. It is therefore a great candidate for simulating data with clustered binary values. For a discussion on the relation between the underlying probability and correlated binary data, see section 3.3.

Third, this model limits the covariance between any two observations to the non-negative range. Compared to some latent probit models, this means that the correlation bounds cannot be violated, since this only occurs when the random variables are negatively correlated.

Finally, the beta-binomial distribution is a way to generate clustered binary data with a constant intra-cluster covariance between two observations. In many situations, researchers have reason to

expect some sort of decaying or local covariance structure, like an autoregressive process. Modeling with a beta-binomial model introduces a stronger correlation than a time series process would, and therefore any estimator robust to a beta-binomial clustered structure would also be robust to weaker structures.

3.3 Generalized Estimation Equations

When the research goal is to estimate the level of items on a test, the dependence within longitudinal person vectors is not of interest; however it can greatly affect the standard error of the item estimates and should therefore be taken into account. GLMMs model overdispersion with random effects, adding a latent distribution on each subject. However, as seen in the previous sections, latent covariance structures have complicated estimates which are not robust to misspecification (Hubbard et al., 2010), and there is no way around them. Estimating equations provide a more straightforward alternative by avoiding specification of the dependence structure. (Liang & Zeger, 1986) describe estimating equations and provide consistent robust variance estimator.

Let $R(\delta)$ be a $n \times n$ symmetric correlation matrix, and let δ be and $s \times 1$ vector which fully characterizes $R(\delta)$.

Define

$$V_i = D_i^{\frac{1}{2}} R(\delta) D_i^{\frac{1}{2}} / \phi$$

The generalized estimating equations are

$$\sum_{i}^{K} D_i^T V_i^{-1} S_i = 0,$$

where $D_i = d\{a_i'(\theta)\}/d\beta, S_i = Y_i - \mu_i$

Essentially, GEE's use likelihood equations from an independent model for dependent data. Suppose we have many clusters (ex: schools) independent from each other but with a dependent inner-cluster structure. Let G be the number of clusters, and suppose we are interested in comparing a control and treatment group. the GEE equations (3.3) are used to obtain $\hat{p_1}, \widehat{p_0}$ and the cluster robust estimator, first used by Shah, Holt and Folsom (1977), is a function of within-cluster correlation and uses observed residuals $\hat{u}_g$:

$$\widehat{V} = (X^T X)^{-1} \sum_{g=1}^{G} X_g^T \hat{u}_g \hat{u}_g^T X_g (X^T X)^{-1} \tag{3.2}$$

where g is the number of clusters.

Let Y_{jg} be the number of successes in cluster $g, g = 1, \cdots, G_j$ of group $j, j = 0, 1$. Then

$$Y_{j.} = \sum_{g=1}^{G_j} Y_{jg} \sim Binomial(n_j, p_j), \qquad j = 1, 2.$$

And therefore, the Liang and Zeger estimators for this model are:

$$\hat{p_j} = \frac{Y_{j.}}{n_j}, \qquad \widehat{V}_j = \frac{1}{n_j} \sum (Y_{jg} - n_{jg}\hat{p_j})^2 \tag{3.3}$$

For a mixed GLM for binary data with a logit link, the ML equations give unbiased estimating equations:

$$X'y = X'\mathbb{E}(y)$$

The large sample variance of the estimator has been shown to be consistent (Fischer & Molenaar, 1995).

Discussion about correlation in binary random variables Potentially correlated binary sequences have such broad applications that it is worth taking the time to investigate what a dependent binary sequence actually looks like. Bernoulli random variables can be strongly positively dependent and still have a relatively low correlation coefficient ("ChagantyGEEefficiency", n.d.). For example, reliability in CTT is measured directly via correlation of binary or categorical variables. Repeated measures from longitudinal studies also pose the problem of potentially dependent binary sequences, and methods like stacking in IRT disregard these without investigating the consequences. For continuous observations, correlated data is easy to visualize. Dependence between binary sequences, however, looks much different. It is important to note that to model

dependence between a sequence of Bernoulli random variables, some sort of function is placed on the mean parameter p. The probability of success for sequential outcomes or outcomes in the same cluster is then determined by a probability structure. For example, correlation can be modeled as a Markov chain, where the previous success or failure affects the probability of the current one. It can also be thought of as originating from an auto-regressive time series process, which would alter the probability of success over time with a geometrically decaying covariance structure. Finally, the beta-binomial model introduces positive covariance between observations by drawing probability values from a beta distribution. All three of these methods model the probability of a success in three different ways so that it is a function of time, state space or simply has a beta distribution; then observations are drawn with the probability of success in that cluster/state/at that time point.

Simulating dependent sequences of Bernoulli variables using these methods does not exhaust the list of all covariance structures social scientists may encounter in practice. If a patient is very low on the latent scale that is being measured by a test, then it might be more realistic to say that he *will* fail all items whose difficulty level is below a certain threshold. Markov chains with absorbing states might be a way to closely replicate this sort of scenario.

If items on a test are positively correlated, then we would expect similar answers on these items. In other words, given that one

item was answered successfully, we would expect a successful answer on the other items with a higher probability than if the first item was an observed "failure". Therefore, highly correlated items would yield constant sequences of mostly successes (or mostly failures). But how is such a positively correlated sequence distinguished from independent items which all have a high (or low) probability of being endorsed, especially when only a few observations are recorded? This problem extends to items with small numbers of ordinal categories as well and is inevitably tied to a downfall of Cronbach's alpha, which is that a test is rarely deemed unreliable. Indeed (Maul, 2017) show that a test composed of nonsensical lorem ipsum or even blank items is deemed reliable due to the positive correlation of the items. This discussion is kept in mind when looking at simulation results in the coming chapters which use a beta-binomial model and briefly Markov chains to generate the data.

Part II

Cluster-Robust Variance Estimator

"Of two equivalent theories or explanations, all other things being equal, the simpler one is to be preferred". - William of Ockham

Chapter 4

Cluster-Robust Variance Estimator

Introduction The last chapter concluded with general estimating equations (GEE) as a way to obtain consistent mean estimates while avoiding model over-specification in correlated binary models. In this chapter, we look at a cluster-robust sandwich estimator which uses GEEs to obtain an asymptotically normal test statistic, and apply it to binary response data. Because the within-cluster correlation is unspecified, this estimator is robust to a wide range of covariance structures. The following theory was developed in conjunction with (Carter, Marquis, & Steigerwald, 2020), who develop consistency conditions based on cluster variances. We refer the reader to that paper, which will from now on be referred to as the CMS paper, for proofs developed prior to this dissertation.

The cluster-robust variance estimator (CRVE) is presented in section 1 along with the aforementioned consistency conditions. CMS show that heteroscedasticity across clusters or variation in cluster sizes lead to increased inaccuracy in the normal approximation, and therefore a large number of clusters might not be a good enough criteria for valid inference. They develop an *effective* number of clusters G^* which measures the degrees of freedom in test statistic and calculate a bias as a function of G^*. The GEE variance estimator is known to have downward bias. Section 2 presents this exact bias and the aforementioned consistency conditions are used to develop a conservative multiplicative bound. In Section 3, we discuss a Student' t approximation in small-sample situations with possibly unaccounted variation and derive the degrees of freedom for an approximate chi-square distribution for the variance estimator as proposed by (Satterthwaite, 1946a). The two-sample test statistic and its approximate degrees of freedom for testing treatment effects in clustered groups are given in Section 4. The next chapter will explore the bias adjustments and t-distribution and compare the results with other possible estimators using a simulation study of coverage percentages.

4.1 Cluster-Robust Variance Estimator

Many publications have proposed cluster-robust variance estimators in heteroscedastic experiments. White (1984, Thm 6.3, p.136) establish two results for clusters of equal size: First, that the cluster-robust t statistic has a Gaussian asymptotic null distribution, and second, the variance component is consistently estimated through use of a cluster-robust variance estimator. Consistency of the variance estimator is established by Hansen (2007), but is not robust to cluster heterogeneity. (Carter, Schnepel, & Steigerwald, 2017) use cluster-robust modified variance estimators with a set of relaxed sufficient conditions for asymptotic normality for linear models with continuous outcomes. However, Carter et al. rely on a fourth-order moment condition that binary errors do not generally satisfy. Other papers rely on bounding the largest cluster size ((Djogbenou, MacKinnon, & Nielsen, 2018; Hansen & Lee, 2018)). The proof of the asymptotic normality of the test statistic depends on the variance of $\widehat{V}$ being negligible relative to the size of the true variance V; that is; no cluster is contributing to the estimator in a disproportionately large way. CMS give sufficient conditions for normal approximation which control individual variances. We show through theory and simulation that cluster sizes and differences in cluster variances lead to non-normality of the test statistic, and suggest a reasonable *effective* number of clusters which guarantees the

size of the test to be at most the desired α.

We begin by stating the test statistic of interest:

$$t = \frac{\widehat{p} - p}{\sqrt{\widehat{\widehat{V}}}} \tag{4.1}$$

where p is an unknown proportion common to a clustered population. Observations are comprised of data clustered into G independent groups with unspecified intra-cluster variations which are, for the purpose of this research question, considered a nuisance to the estimation process, since the parameter of interest is the mean. For each cluster g, denote $(Y_g)_{g=1,\ldots,G}$ as the sum of binary observations $Y_{ig} \in \{0,1\}, i = 1, \ldots n_g$ where n_g are the corresponding cluster sizes. A binomial assumption would be erroneous at this point, because of possible correlation within each cluster which affect the cluster variances. Although ignoring it does not bias the estimate of the mean, its variance becomes grossly underestimated, as simulations in the next chapter confirm (Figures 5.11). The estimate of p is the sample average of all binary observations

$$\widehat{p} = \frac{\sum_{g=1}^{G} Y_g}{\sum_{g=1}^{G} n_g} = \frac{Y}{N}, \qquad \text{where} \qquad Y = \sum_{g=1}^{G} Y_g, \qquad N = \sum_{g=1}^{G} n_g.$$

We now give the cluster robust variance estimator (CRVE) which

is an estimate of $\mathrm{V} = \mathrm{Var}(\widehat{p})$:

$$\widehat{V} = \frac{1}{N^2} \sum_{g=1}^{G} (Y - n_g \widehat{p})^2 \tag{4.2}$$

This sandwich estimator measures residuals over clusters rather than each binary observation.

We now list four sufficient conditions for asymptotic normality, derived in CMS.

Condition 4.1.1. *The number of clusters $G_j \to \infty$ for $j = 0$ and 1.*

While this condition is not necessary for consistency of $\widehat{p}$, it is crucial for consistency of the cluster-robust variance estimators in 4.2. Because the cluster-robust variance estimator depends on the outcomes only through the cluster-level sums Y_{jg}, the other conditions restrict the distribution of the Y_{jg}. Condition 2 bounds the kurtosis. Condition 3 bounds the variation in cluster sizes. Condition 4 is a stronger condition which bounds the fourth moment of the cluster sums.

Condition 4.1.2. *The kurtosis of each Y_{jg} is bounded*

$$\mathrm{Var}\left([Y_{jg} - n_{jg} p_j]^2 \right) \leq \kappa \ [\mathrm{Var}(Y_{jg})]^2$$

for all g, $j = 0$ and 1, and a constant κ.

Condition 4.1.3. *The empirical coefficient of variation of the n_g's is negligible as $G \to \infty$*

$$\sum_{g=1}^{G} \left(\frac{n_g}{N} - \frac{1}{G} \right)^2 \to 0.$$

Condition 4.1.4.

$$\sum_{g=1}^{G} \left(\frac{\mathrm{Var}(Y_g)}{N^2 V} - \frac{1}{G} \right)^2 \to 0,$$

If the variances of the Y_g are identical, then this relation holds exactly. When the model is nearly binomial, meaning that the variance of each cluster is close to $n_g p(1-p)$, then Condition 4.1.4 follows immediately from Condition 4.1.3. This might occur in experiments with time measurements where the cluster correlations decay geometrically. Condition 4.1.4 is needed in case the model has a stronger correlation structure.

These conditions are more than sufficient to have asymptotic normality when the population variance is known; the proof of the following lemma is therefore omitted.

Lemma 4.1.1. *If Conditions 4.1.1–4.1.4 are satisfied,*

$$\frac{\widehat{p} - p}{\sqrt{V}} \overset{appr}{\sim} \mathcal{N}(0, 1).$$

The main result of CMS is that the test statistic with the variance estimate is also approximately normally distributed for a relatively large effective number a clusters. The proof to show this convergence in distribution uses Slutsky's theorem and demonstrates consistency of the CRVE by showing that $\frac{\hat{V}}{V} \xrightarrow{\mathbb{P}} 1$. The details of the proof are available in the appendix of CMS. This main result is now stated:

Theorem 4.1.1. *If Conditions 4.1.1–4.1.4 are satisfied then*

$$t = \frac{\widehat{p} - p}{\sqrt{\widehat{\widehat{V}}}} \overset{appr}{\sim} \mathcal{N}(0, 1).$$

Asymptotic normality has been established. What does this mean for a practitioner? The rest of this chapter examines the behavior of the test statistic under different experimental settings and aims to provide helpful guidelines. The GEE cluster-robust estimator has been known to underestimate the variance. The conditions above aim to control the difference in the variances of the Y_g (Condition 4.1.4) and the variation in cluster sizes (Condition 4.1.3), because those are directly related to the their variances. When heterogeneity between groups is significant and the quantities in 4.1.3 and 4.1.4 are relatively large, the test statistic may not follow an approximate normal distribution. In fact, simulations in Chapter 6 reveal a slight bimodal density in situations where one cluster holds more than half of the observations (Figure 5.29). An effective number of clusters is developed as a function of the variation of cluster variances, which

represents a more accurate measure of the degrees of freedom in the experiment.

4.2 Effective number of clusters

The **effective number of clusters** is defined as

$$G^* = G\Big(1 + \frac{1}{G}\sum_{g=1}^{G}\frac{[\mathrm{Var}(Y_g) - \bar{V}]^2}{\bar{V}^2}\Big)^{-1}, \tag{4.3}$$

where $\bar{V} = V \cdot N^2/G$. This quantity attempts to measure cluster homogeneity, in that smaller values indicate variation between cluster variances, and $G^* \to G$ as the clusters become more homogeneous. Note that this is a multiplicative measure, so that $1 \leq G^* \leq G$. As theoretical results backed by simulation results will show, G^* can be used as a measure of the consistency of the variance estimator and consequent asymptotic normality of the test statistic. In fact, $G^* \to \infty$ is sufficient for convergence of the test statistic. Simulations in the next chapter will show that even a G^* which is small compared to G suffices for the test statistic to have at least 94% coverage.

The quantity in (4.3) is based on the unknown cluster variances. This allows for some flexibility in calculating a G^* based on the presumed underlying model. We chose to use the beta-binomial distribution as it provides a matching structure of independent groups while allowing for the cluster variances to have different magnitudes.

This is a common way to account for overdispersion in generalized linear models (see Section 3.2). The number of successes in each cluster can have a variance as small as an independent binomial model, $np(1-p)$, and reach an order of $n^2p(1-p)$ when γ is close to one, which is a larger variance than in many other clustered models (CMS):

$$\text{Var}(Y_g) = (1-\gamma)\cdot n_g p(1-p) + \gamma \cdot n_g^2 p(1-p) \tag{4.4}$$

Plugging $\gamma = 1$ into 4.4 to give the largest possible variance, p cancels out of the equation and the resulting G^* is then only a function of cluster size variances:

$$G^* = G\left[1+\frac{1}{G}\sum_{g=1}^{G}\left(\frac{n_g^2-\nu}{\nu}\right)^2\right]^{-1} \qquad \text{(beta-binomial assumption, } \gamma = 1) \tag{4.5}$$

where $\nu = \frac{\sum n_g^2}{G}$. See section 3.2 for details on parameterization for the beta-binomial distribution.

The value in 4.5 provides a conservative effective number of clusters that can be used in practice. It is a more sensitive measurement of the variation between the cluster variances. An example of a low effective number of clusters is one in which one cluster contains most of the observations. The variance of that cluster is proportionate to its size. In other words, if this cluster offers a poor estimate of the

mean, then it will add significant error to the sample average, because $\widehat{p}$ is calculated by weighing each binary observation equally. It would therefore be preferable to have observations distributed evenly throughout the clusters, so that no one cluster dominates the sample average estimate. This is precisely what G^* is attempting to measure. A low effective number of clusters results in a more significant downward bias, calculated next. In Section 4.4, Satterthwaithe approximation is used to show that under varying cluster sizes and smaller samples, the test statistic is closer to a t-distribution with approximately G^* degrees of freedom.

4.3 Bias Calculation and Bound

The bias $b(\widehat{V}) = \mathbb{E}(\widehat{V}) - V$, stated here, is calculated in detail in the appendix 9.1.

$$b(\widehat{V}) = V \sum_{g=1}^{G} \left(\frac{n_g}{N}\right)^2 - \frac{2}{N^3} \sum_{g=1}^{G} n_g \operatorname{Var}(Y_g) \tag{4.6}$$

For homogeneous clusters the bias simplifies down to $-(\frac{1}{G})V$. This is the same bias that the sample variance s^2 has for σ^2 in a regular identically distributed and independent setting. The form in Equation (4.6) is not very intuitive. We therefore rewrite the bias to show that its magnitude increases directly as a function of cluster size variation by making a comparison to the homogeneous case

where $n_g = N/G$:

$$b(\widehat{V}) = V\left[-\frac{1}{G} + \Gamma^2 - \frac{2}{N^2 V}\sum_g \sigma_g^2\left(\frac{n_g}{N} - \frac{1}{G}\right)\right] \tag{4.7}$$

where Γ^2 is the variation in cluster sizes

$$\Gamma^2 = \sum_g \left[\frac{n_g}{N} - \frac{1}{G}\right]^2 \tag{4.8}$$

The form (4.7) shows that the bias of the CRVE is at best $-1/G$ when $G = G^*$, and is inversely proportional to the effective number of clusters. Figure 4.3 provides a good visualization of this relationship.

A bound is now derived using one of the consistency conditions on the cluster variances.

ℓ^2 bound

The last term in the bias in 4.7 involves fourth moments of the distribution and can be bounded using consistency Condition 4.1.4.

$$\left[\frac{1}{G}\sum_g \sigma_g^2\left(\frac{n_g}{N} - \frac{1}{G}\right)\right]^2 \leq \left[\frac{1}{G}\sum_g \sigma_g^4\right]\left[\frac{1}{G}\sum_g \left(\frac{n_g}{N} - \frac{1}{G}\right)^2\right] \tag{4.9}$$

The first factor in 9.4 can be related to G^* and bounded using beta-binomial moments with $\gamma = 1$. The bias bound, derived as a func-

tion of Γ (4.8), simplifies to

$$b(\widehat{V}) \geq V\left[-\frac{1}{G}-\frac{1}{G^*}+\left(\Gamma-G^{*-\frac{1}{2}}\right)^2\right] \tag{4.10}$$

and is used to obtain a bias-adjusted variance estimate $\widehat{V}^{BC}$:

$$\widehat{V}^{BC}=\widehat{V}\cdot\left[1-\frac{1}{G}-\frac{1}{G^*}+\left(\Gamma-G^{*-\frac{1}{2}}\right)^2\right]^{-1}$$

Naturally, this bias correction is effective when compared to the theoretical variance of a beta-binomial model, as seen in Figure 4.3, since we assumed a beta-binomial fourth moment when using Condition 4. The bias is directly proportional to G^*; erratic cluster variances result in a low effective number of clusters, and therefore a more important bias. Again, using a bound based on the beta-binomial variance with $\gamma = 1$ in a way ensures that we have accounted for the "worst-case" scenario (cluster variances order of n_g^2). In Chapter 6 we explore the possibility of substituting in a smaller γ for a finer bound, based on the discussion of beta-distributions in (3.2).

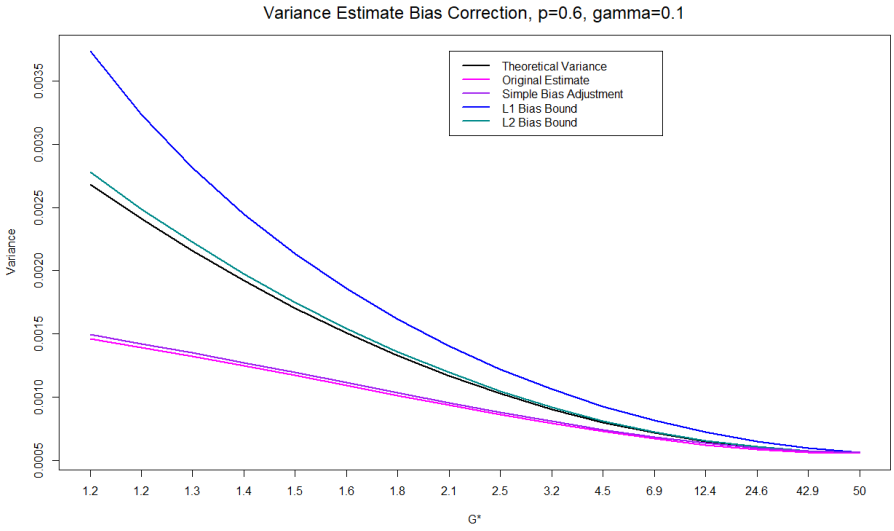


Figure 4.1: Variance Bias Corrections

Figure 4.1 shows the theoretical variance under a beta-binomial model (black line), along with the original GEE estimate and the bias-adjusted ones as a function of the effective number of clusters. The "Simple Bias Adjustment" refers to the standard $\frac{G}{G-1}$ unbiased estimator for homogeneous clusters.

4.4 t Distribution with G^* degrees of freedom

The asymptotic coverage of the test statistic in Eq(4.1) depends heavily on the effective number of clusters. For small values of G^*, the empirical densities of the resulting standardized scores is closer to a Gamma distribution. Empirical densities plotted in Figure 5.29 show a bi-modal density for high values of γ and low values of G^*.

This is likely due to the effect of the very large cluster carrying most of the weight in $\widehat{p}$, and occasionally having a very large residual. This increased non-normality suggests that perhaps a t distribution would be more appropriate.

4.4.1 Satterthwaite Approximation

If we have G clusters, the CRVE can be written as a linear function of terms like

$$U_g = \frac{(Y_g - n_g\widehat{p})^2}{V_g} \approx \chi^2_1$$

where $V_g = \text{Var}(Y_g)$. The exact distribution is complex since the terms are not independent, so we use the approach taken in (Satterthwaite, 1946a), which approximates the distribution by matching moments with those of a χ^2 variable. The degrees of freedom for the chi-square distribution are then

$$2\alpha^* = G\left[1 + \frac{1}{G}\sum_g \frac{(\hat{V}_g - \hat{\bar{V}})^2}{\hat{\bar{V}}^2}\right]^{-1} = G\left[1 + \frac{1}{G}\sum_g \frac{(n_g^2 - v)^2}{v^2}\right]^{-1} = G^*$$

When the number of clusters is undermined by a large difference in cluster variances, the test statistic 4.1 is closer to a t distribution. We reasoned above that $\widehat{V} \approx \chi^2_{G^*}$, so that the degrees of freedom for the t distribution are equal to the effective number of clusters. Simulations will show the effect of this approximation on

coverage percentages. The approximation can also be applied to the bias-corrected estimate $\widehat{V}^{BC}$, however simulations show that for $G^* < 2$, (which is an extremely unbalanced situation), both bias and t-distribution overcompensate and result in oversized confidence intervals.

4.5 Two-sample problem

The CRVE and its adjustments are now applied to a treatment-effect problem. If the control and treatment groups are clustered, the variance of the difference in proportions can be tested using the bias-adjusted cluster-robust estimators. The statistic for testing if two samples have the same population proportions is

$$t = \frac{\widehat{p}_1 - \widehat{p}_0}{\sqrt{\widehat{V}_0 + \widehat{V}_1}}, \tag{4.11}$$

where $\widehat{p}_j$, $\widehat{V}_j$ are the sample proportions and CRVE estimators as in (4.2), respectively, for the control and treatment groups. The bound in (4.6) can be used on each $\widehat{V}_j$ to obtain bias-adjusted estimates.

Asymptotic normality of (4.11) is preserved if the Conditions listed in Section 4.1 are true for both populations. For the degrees of freedom, the magnitude of the variance estimates must be taken into account, as would be done in an unequal variance two-sample t-test.

Unlike in the one-sample problem, the approximate degrees of freedom for comparing two clustered populations are sample-dependent, meaning that they depend on the variance estimate:

$$D = \left[\left(\frac{\widehat{V}_1}{\widehat{V}_0 + \widehat{V}_1} \right)^2 \frac{1}{G_0^*} + \left(\frac{\widehat{V}_1}{\widehat{V}_0 + \widehat{V}_1} \right)^2 \frac{1}{G_1^*} \right]^{-1}, \tag{4.12}$$

Chapter 5

Simulation Results

The experiment described in Chapter 5 involves many variables: number of clusters and effective clusters, cluster size, overall probability of success, and within-cluster variances. The theory described in the previous chapter is asymptotic, and the conditions depend on quantities like the number of clusters growing to infinity. Other than reinforcing the already-known truth that larger samples yield better results, this requirement is somewhat useless to a practitioner. In this chapter we study the behavior of the CRVE under different situations, discuss the effect of the bias adjustment and Satterthwaithe degrees of freedom approximation, and compare the asymptotic coverage percentages to other commonly used estimators, like a GLM or quasibinomial model. Simulations help us check asymptotic normality of the test statistics under conditions spanning the spectrum of possible parameter values. More specifically, we show that the

effective number of clusters is a good measure for cluster homogeneity and that asymptotic convergence breaks down when G^* is very low (less than 2). Our simulations show that the test statistic given in 4.1 is very close to a standard normal distribution for as little as 30 or 50 clusters, depending on the intra-group variance structure. The beta-binomial distribution, described in section 3.2, is used to generate clustered binary values. To check the distribution of the test statistic, its value is repeatedly calculated over thousands of simulations, and 95% coverage percentages are estimated by taking the number of resulting p-values which fall within the standard normal confidence interval:

$$\text{Estimated coverage percentage} = \frac{\#\{\texttt{Z-scores} \in (-1.96, 1.96)\}}{\texttt{nsim}}$$

The exact details of the simulation process are available in the appendix (9.4). CMS show that the CRVE (4.2) has an approximate normal distribution when the effective number of clusters is large; when cluster variances are vary greatly, the test statistic may not have a normal distribution. To demonstrate this, we calculate coverage percentages of the test statistic over a range of cluster distributions. Cluster variances are largely a function of cluster sizes, and the γ parameter from the beta distribution:

For each graph presented in this chapter, the simulation process evaluates coverage percentages over a range of cluster sizes, starting

with one dominant cluster holding over 40% of the responses and gradually homogenizing cluster sizes. Each graph represents a different value of p and γ. We present many situations to see how extreme values of p affect test statistics, and discuss a sensible range of values for γ. The effective number of clusters G^* is used as a measure of heterogeneity. We focus in particular on situations when the effective number of cluster is small. Coverage probabilities are evaluated for the different variance estimators. We then consider other estimators which are used in practice: the regular proportion estimator which assumes independent and identically distributed observations, and will be called *iid* throughout this chapter; the generalized linear model (GLM) using quasi-likelihood to deal with overdispersion, and a Wild bootstrap estimator. The results show that the adjusted CRVE is comparable or superior to the other options for highly heterogeneous cases. It is also the only conservative one, in that it overestimates the variance. We then extend this estimator to the treatment-control problem described in 4.5 and look at some situations where the effective number of clusters in each the populations are highly unequal. In addition, we are aware that the effective number of clusters and bias correction are based on a beta-binomial model with the largest possible variance, and that therefore the adjusted CRVE will estimate data simulated under this model best. We therefore consider a second simulation method with a Markov-chain model that mimics any sort of geometric decay in correlation.

In all simulations, one aspect seems to remain constant: the *iid* assumption is erroneous and sensitive to any amount of overdispersion, with coverage percentage that drop below 50%. The GLM and CRVE estimators are comparable in all situations, and are robust to overdispersion. In general, the bias-adjusted CRVE is a percentage point or two above the quasi-binomial estimate, and both have a slight under-coverage for $G = 50$. This is remedied as the numbers of clusters grows. The use of G^* degrees of freedom in a t- distribution results in conservative coverage for $G^* < 10$ but then drops to meet the other estimators. It therefore has a significant effect when there is a lot of cluster heterogeneity, which was indeed the goal. However, we will discuss whether or not it is overcompensating, and how useful the resulting confidence intervals will be.

5.1 CRVE vs Independence Assumption

The first steps in our simulations were to compare the CRVE to the binomial test statistic which assumes a cluster-free population. This estimator is explicitly stated in the appendix, Section 9.4. These simulations confirm that the effective number of clusters G^* can be used as a measure of cluster heterogeneity and therefore coverage percentage. Simulations span multiple values of p and γ for both the CRVE and the test statistic resulting from an independence

assumption.

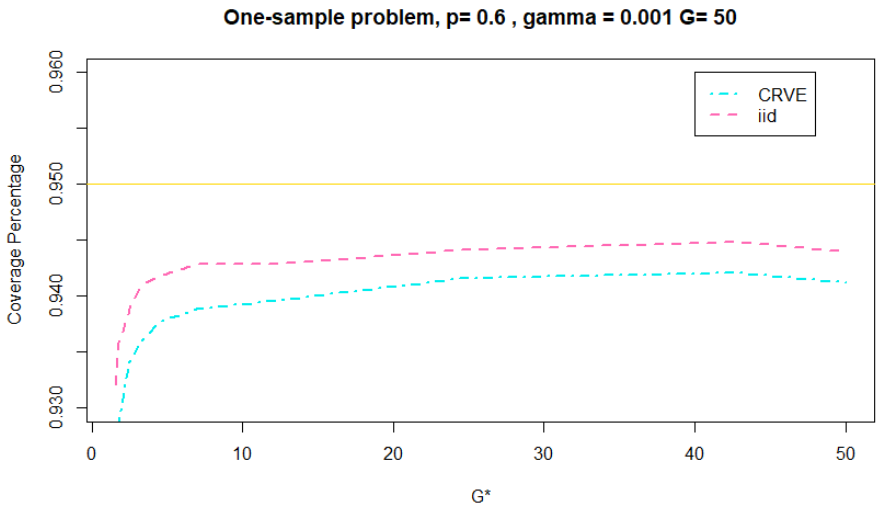


Figure 5.1: Coverage percentages for the CRVE test statistic and "iid" estimator, independent model.

We start with a nearly independent model, with $\gamma = 0.001$, and obtain a somewhat expected result (Figure 5.1). 50 clusters is perhaps too low for normal approximation to be accurate enough, which is why both statistics have a slight under-coverage of about 94%. The "iid" test statistic performs slightly better than the unadjusted CRVE in this situation. However, increasing the value of γ only slightly results in clusters with more unequal variances which are larger than $p(1-p)/n$ and therefore the "iid" estimator performs poorly, and clearly does not follow a normal distribution and underestimates the variances greatly, as can be seen in Figures 5.2, 5.3, and 5.4.

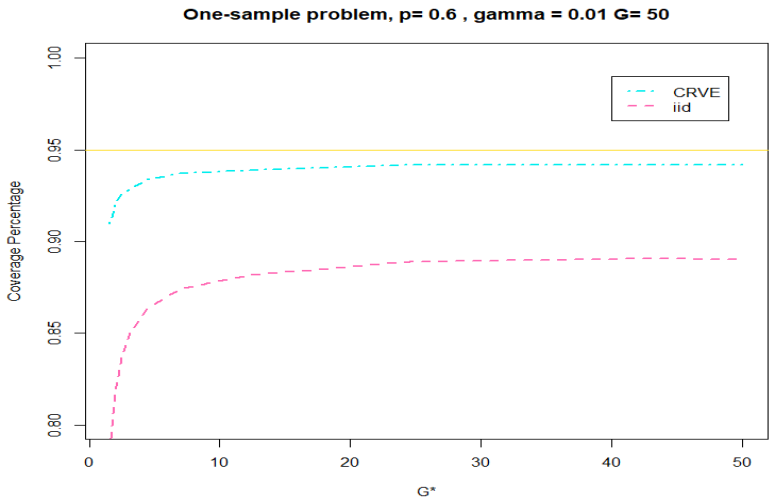


Figure 5.2: Coverage percentages for the CRVE test statistic and "iid" estimator, slight overdispersion.

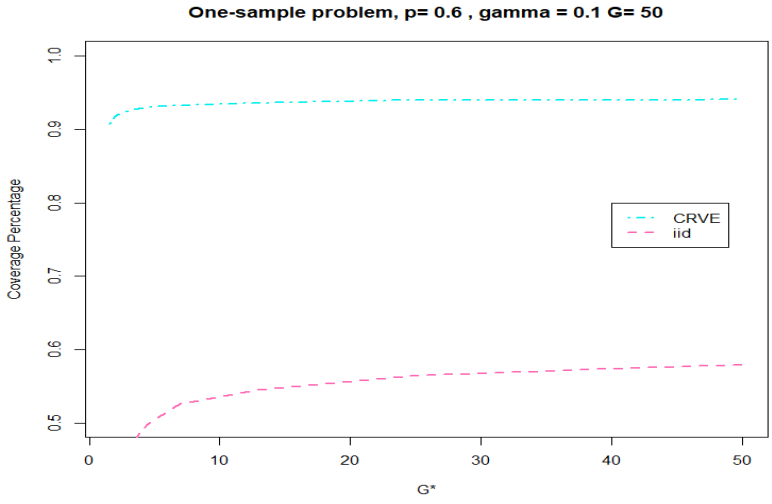


Figure 5.3: Coverage percentages for the CRVE test statistic and "iid" estimator, overdispersion.

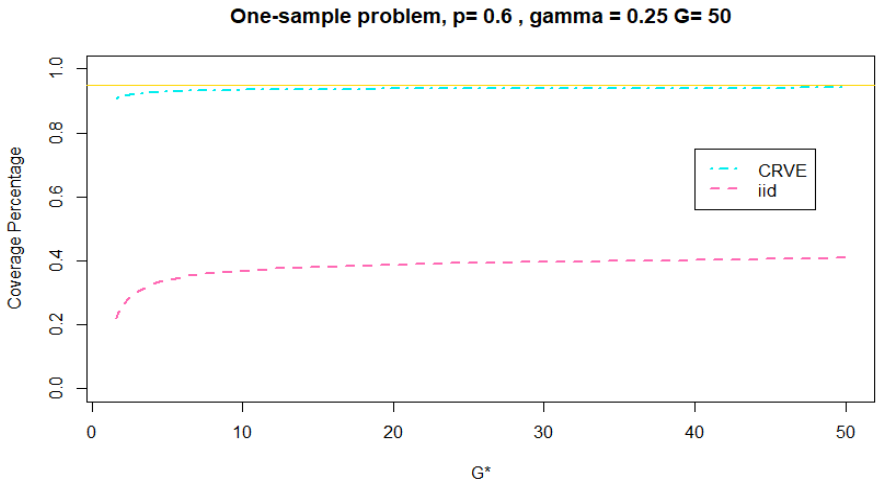


Figure 5.4: Coverage percentages for the CRVE test statistic and "iid" estimator, heavy overdispersion.

These figures show that the test statistic (4.1) is indeed robust to variation in cluster variances, and accounts for any overdispersion. Assuming independence in these situations is obviously a serious mistake which will lead to invalid inference. The CRVE is known to have a slight downward bias, as was observed in these graphs. In the next section we therefore inspect the effect of the bias adjustment and compare z-scores with t_{G^*}-scores.

5.2 CRVE Adjustments: Bias Bound and t_{G^*} Distribution

We now use simulations to compare the original CRVE with its bias-adjusted estimator, and apply the Satterthwaite approxima-

tion to the degrees of freedom for both of the resulting test statistics (G^* degrees of freedom). Our simulations revealed that while both the bias correction and the use of G^* degrees of freedom in a t-distribution helped individually, together they might overcompensate in extremely unbalanced cases. This is quite a low effective number of clusters and would occur in an extreme situation with very unbalanced groups. We look at simulation results which demonstrate this occurrence. The adjustments are evaluated at different values of p and γ to bring out any patterns or issues in cases of low p or large γ.

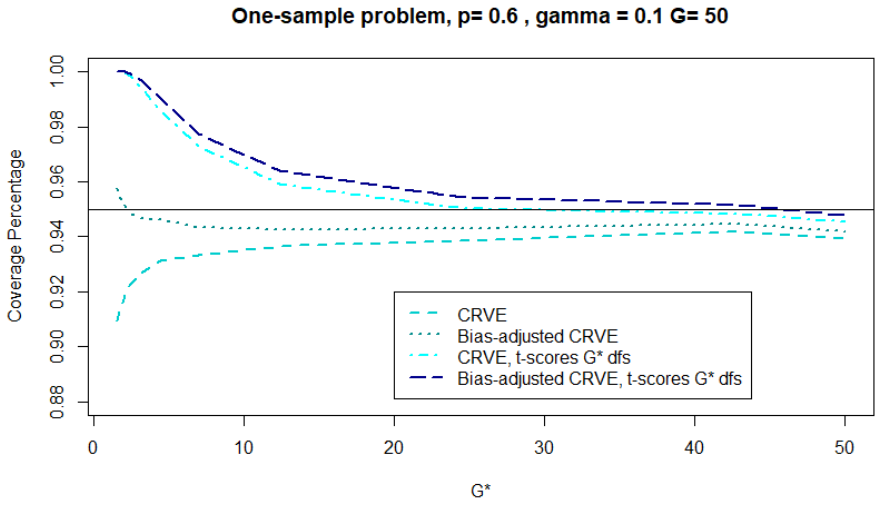


Figure 5.5: Bias and t-df adjustments on the CRVE

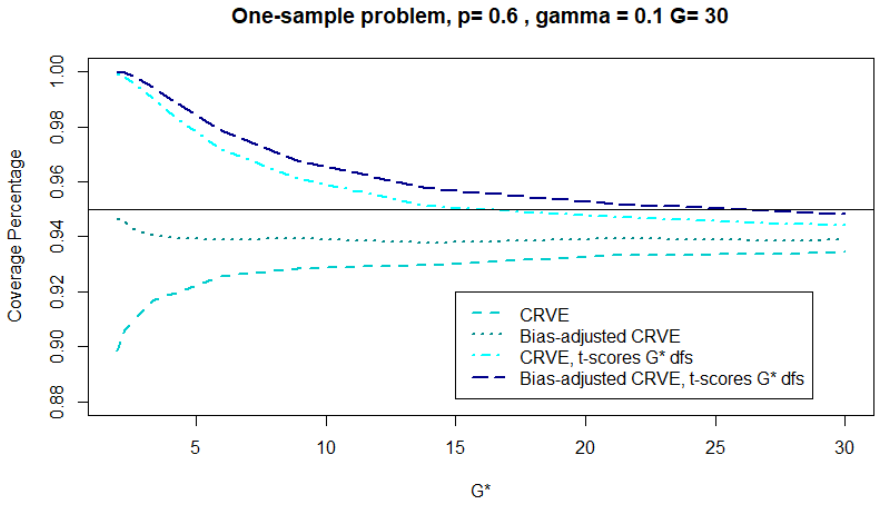


Figure 5.6: Bias and t-df adjustments on the CRVE for 30 clusters.

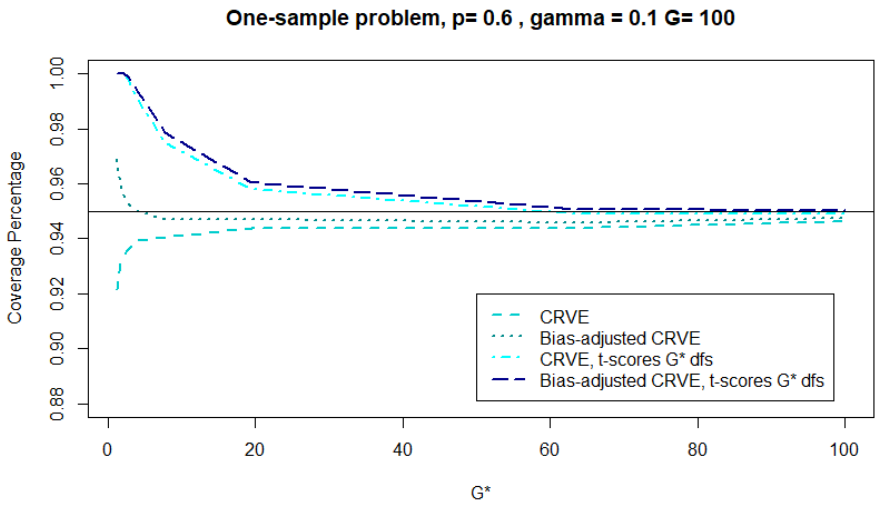


Figure 5.7: Bias and t-df adjustments on the CRVE for 100 clusters.

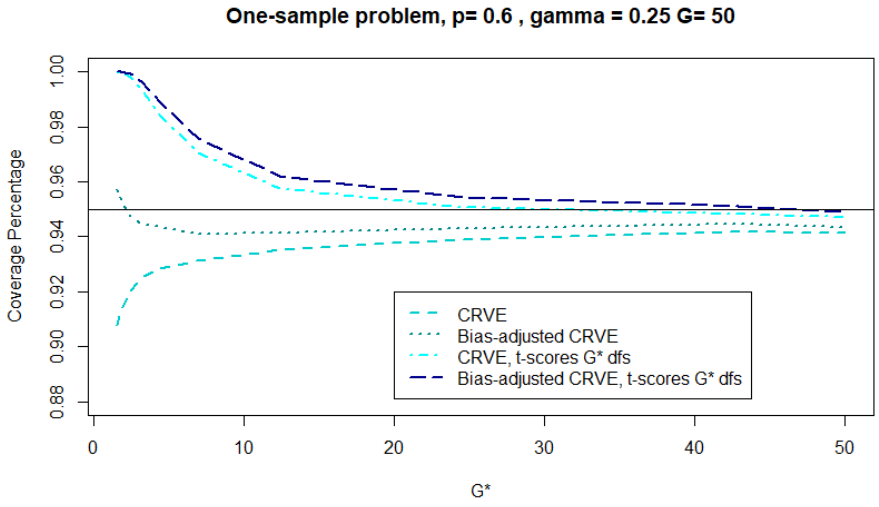


Figure 5.8: Bias and t-df adjustments on the CRVE for 50 clusters with widespread distribution of cluster means.

These plots were simulated over three different probabilities and values of γ, but they all consistently reveal the same result: Although the bias adjustment estimates the theoretical variance quite accurately, the bias-corrected CRVE still results in a slight under-coverage ($> 94\%$) for small numbers of clusters. Even with cluster homogeneity at $G = 50$, the coverage percentage is slightly below the desired 95%. We remind the reader that the distribution is asymptotic and depends on the effective number of clusters getting large, and the effective number of clusters is at most G. As the number of clusters grows, the coverage percentage approaches its theoretical size. We ran a simulation with 100 clusters and the resulting coverage percentages at $G^* = 100$ were within .2% of 95 for

all of the adjusted test statistics (Figure 5.7).

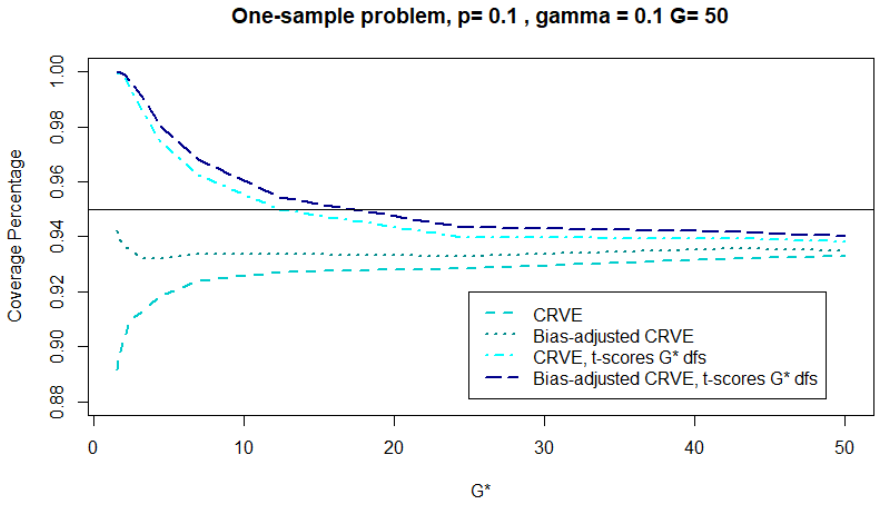


Figure 5.9: Bias and t-df adjustments on the CRVE with some cluster means set to zero.

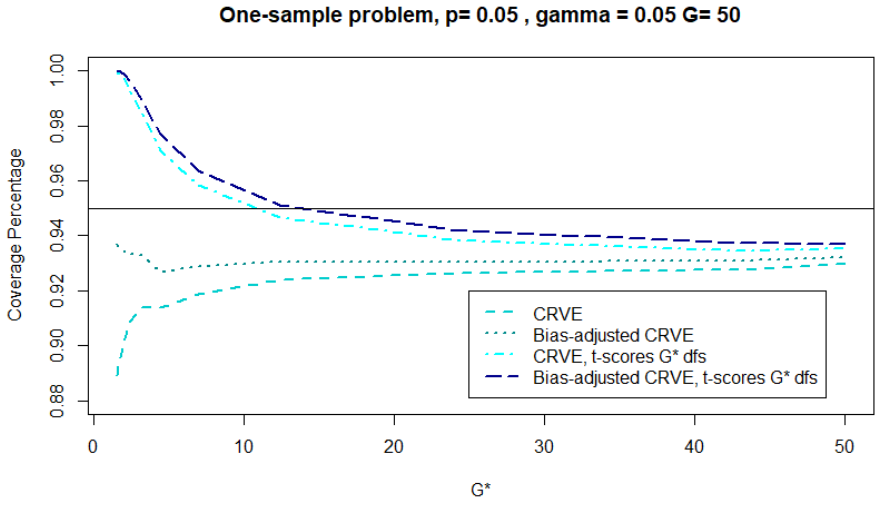


Figure 5.10: Bias and t-df adjustments on the CRVE with very low probabilities of success for all cluster means.

The last simulations in this section (Figures 5.9 and 5.10) test the CRVE for very small values of p. The CRVE seems to be more sensitive to floor and ceiling values of p than to correlation, but only slightly. Coverage percentages in Figure 5.10 still remain above 92% when the overall probability of success is $p = 0.05$. In these situations, even a small value of γ can result in cluster means set to 0.

We conclude from this part of the simulation study that bias adjustment improves coverage percentages for unbalanced clusters. In very unbalanced experiments, a researcher who wishes to be conservative might consider t scores with G^* degrees of freedom.

5.3 CRVE and Other Methods

This section compares the bias-adjusted CRVE with Z and t_{g^*} scores with a quasi-binomial generalized linear model and a wild bootstrap. The simulation method remains the same; details on other estimators can be found in the appendix, Section 9.4. For a probability of $p = 0.6$ and slight overdispersion of $\gamma = 0.1$, we can see that for effective number of clusters less than twenty, the quasibinomial estimator fails to capture the underlying variance in the model and has slightly small confidence intervals (Figure 5.11). The wild bootstrap seems to match the bias-adjusted CRVE quite closely, but has a slight downward bias for the homogeneous case of 50 observations

per each of the 50 clusters.

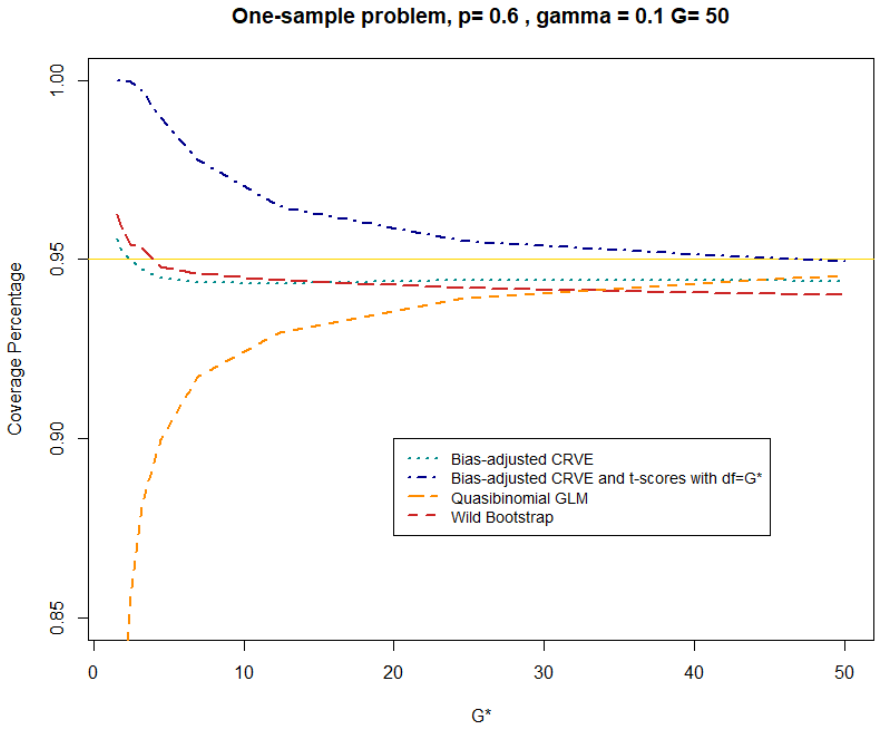


Figure 5.11: Comparing cluster-robust test statistics, slight overdispersion

Unlike the binomial model, all of these methods are clearly robust to overdispersion, as can be seen in Figure 5.12 where γ is set to 0.25. The CRVE with a t distribution and G^* degrees of freedom is the only one to yield a conservative test statistic with confidence intervals that are at least as large as the desired coverage percentage for all levels of cluster heterogeneity. However, some investigation into the width of the resulting confidence intervals for about $G^* < 10$ remains to be done.

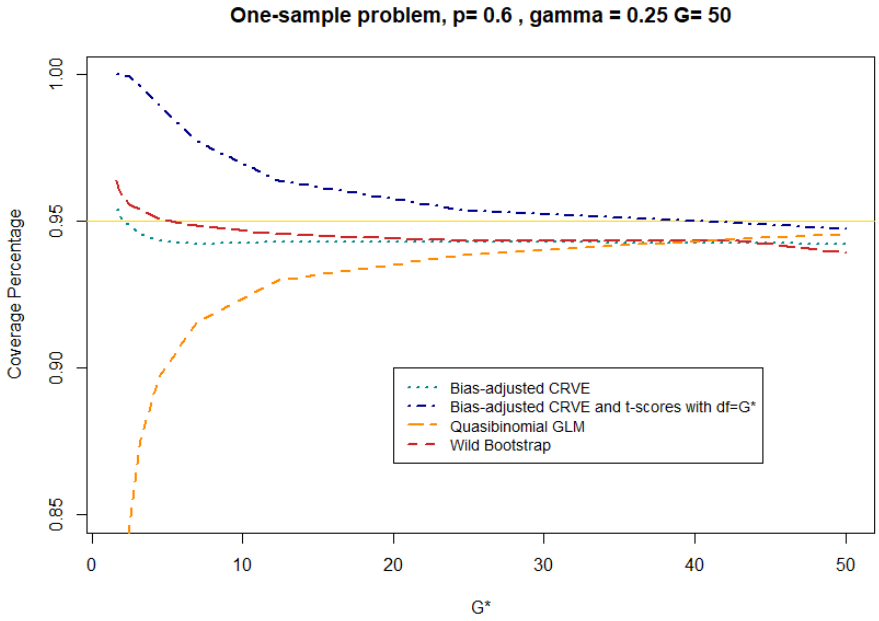


Figure 5.12

The next few graphs give coverage percentages at an underlying probability of success of 10%. We vary the value of γ but keep it rather small to avoid too many cluster means set to 0.

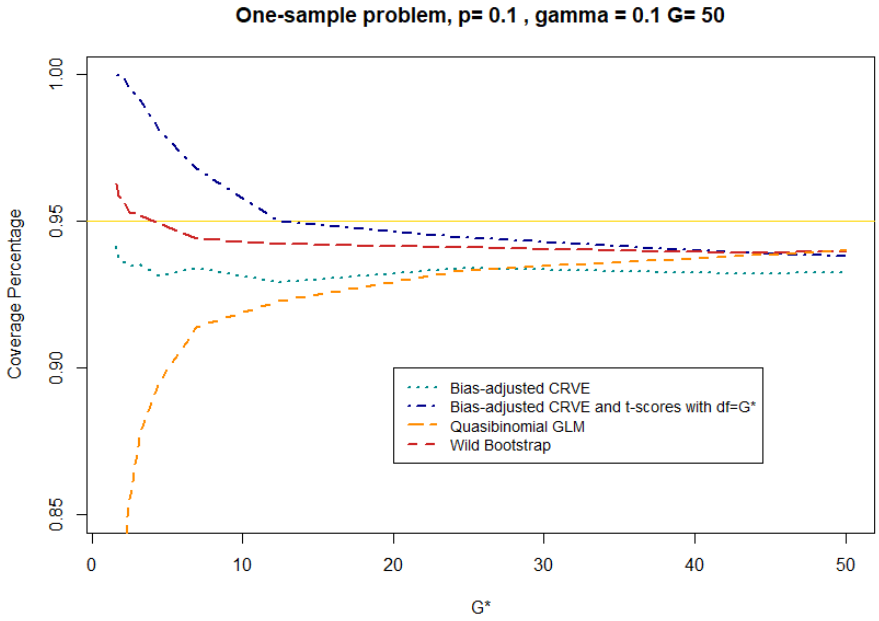


Figure 5.13: Comparing cluster-robust test statistics, slight overdispersion with a probability of success of 10%.

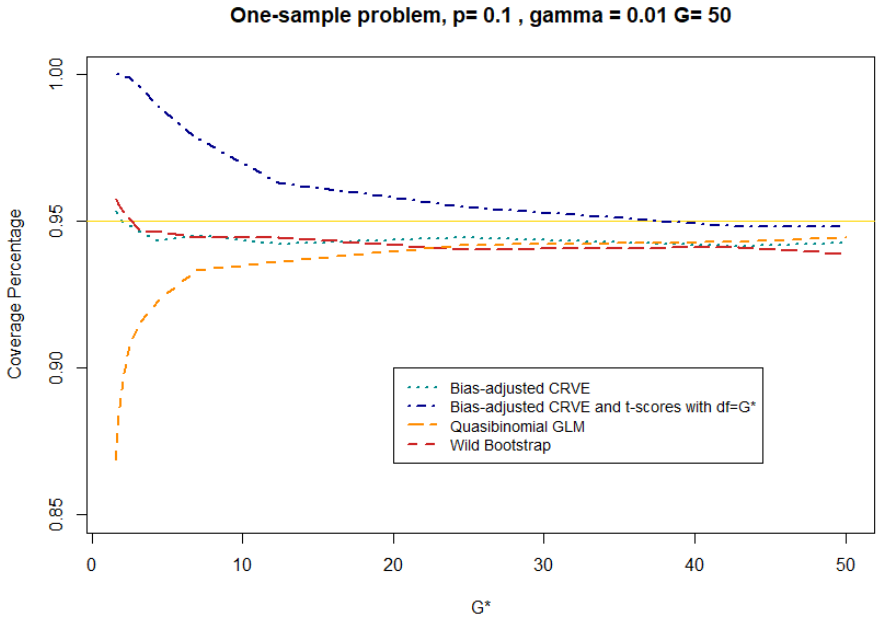


Figure 5.14: Comparing cluster-robust test statistics, slight overdispersion with a probability of success of 10%.

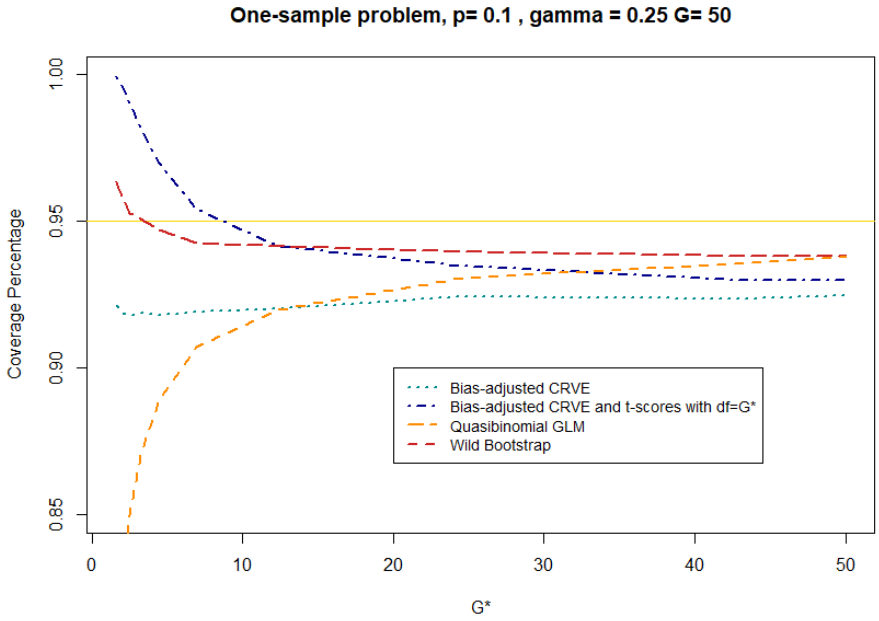


Figure 5.15: Comparing cluster-robust test statistics, heavy overdispersion with a probability of success of 10%.

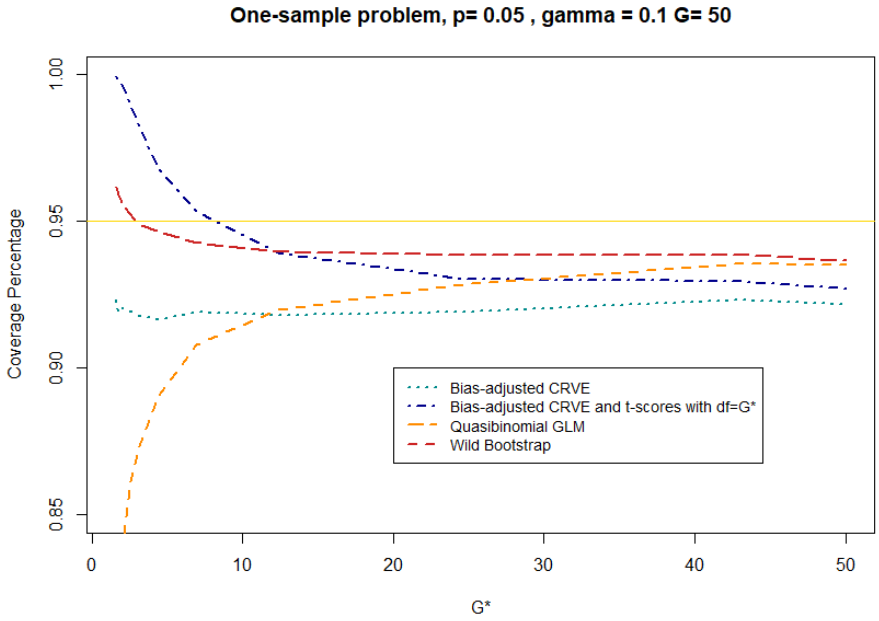


Figure 5.16: Comparing cluster-robust test statistics for $p = 0.05$.

We see that in cases of floor and ceiling probabilities of success and larger overdispersion (Figures 5.15, 5.16, and 5.17), the bootstrap and quasibinomial compensate slightly better than the CRVE statistics.

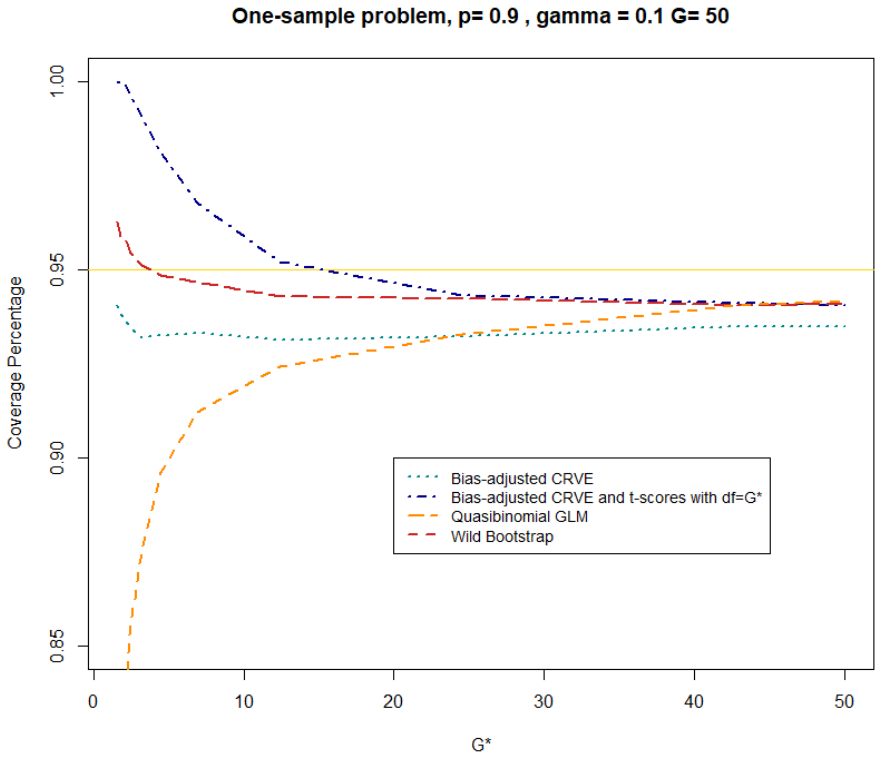


Figure 5.17: Comparing cluster-robust test statistics for $p = 9$.

5.4 Comparing Two Samples

We then performed the same simulation study on the difference of two proportions. In this case, the test statistic was the one given in (4.11).These simulations show that the difference in proportion is better captured by the CRVE statistic than by quasibinomial estimator. Because the bootstrap estimator was quite time-costly, we look at it separately in the next section.

Somewhat surprisingly, we see in these next few graphs that the

CRVE test statistic for difference in proportions performs better than in the one sample problem for very small values of p, and is superior to the quasibinomial GLM in the treatment-control problem.

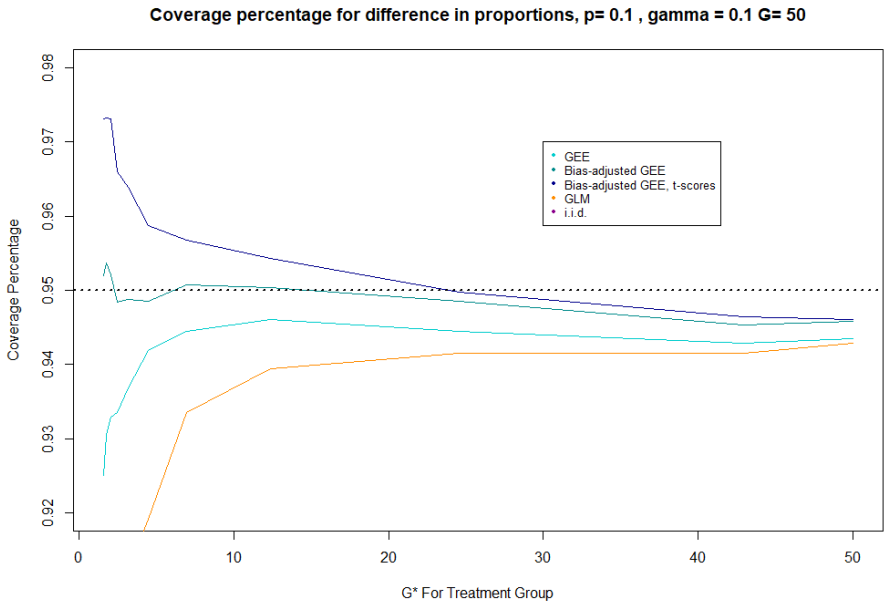


Figure 5.18: Treatment-control experiment with $p = 0.1, \gamma = 0.1$.

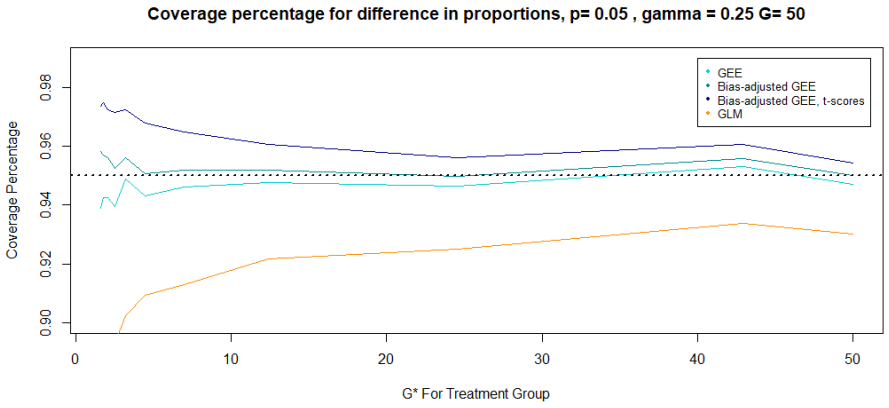


Figure 5.19: Treatment-control experiment with $p = 0.1, \gamma = 0.25$.

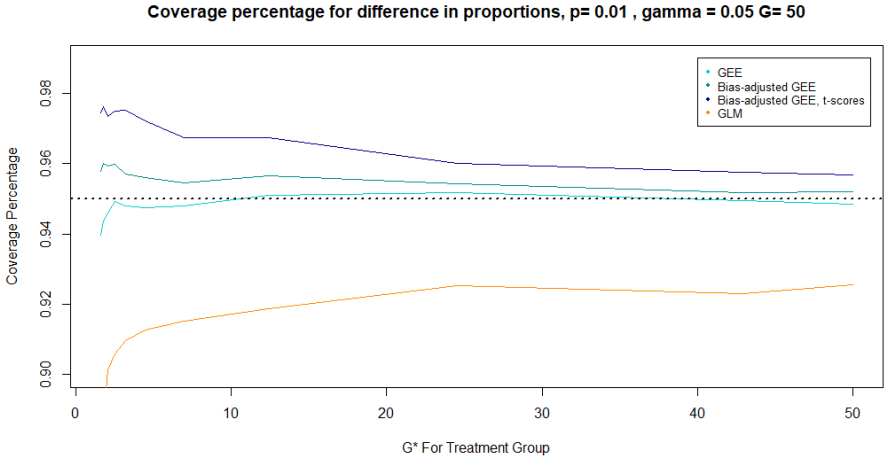


Figure 5.20: Very small values of p will cause the GLM to break down, while the CRVE for difference in proportions is robust.

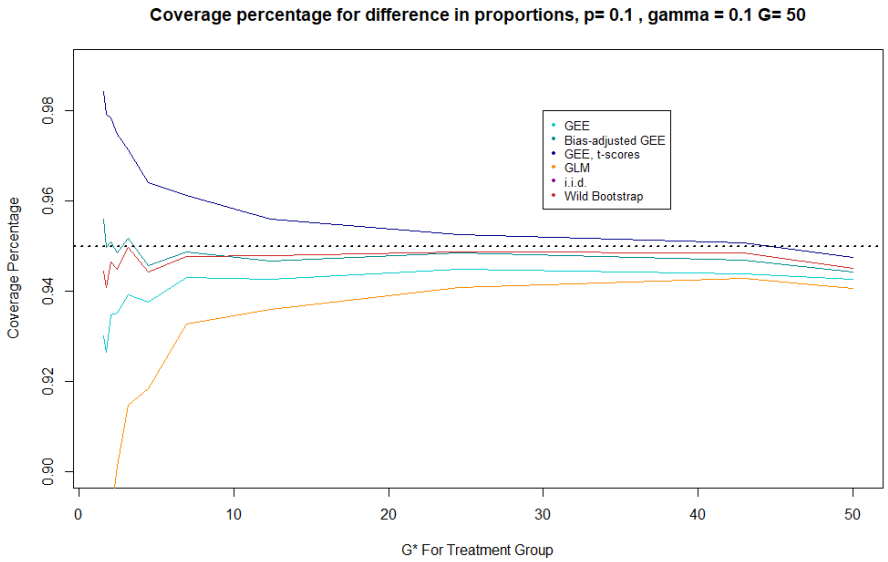


Figure 5.21: Coverage percentages for cluster-robust statistics for difference in proportions, $p = 0.1$.

5.4.1 Bootstrap Estimator

Because we are already running a high number of simulations, computing a bootstrap estimate at each point resulted in much longer computing times, so the comparisons were done separately. We are using the wild bootstrap, described in the Appendix. This estimator has been known to perform well in cases where the correlation is high, so we show two simulation results, both at critically low values of p, one with almost independent observations, and the other with $\gamma = \frac{1}{2}$. We note that the correlation bounds imposed by the parameters of the binomial distribution are not be violated here since the beta-binomial model only introduces positive correlation

between observations from the same group. The GEE estimator is shown to be robust to extreme cases. Consider $p = 0.1, \gamma = 0.5$, as for the coverage percentages shown in Figure 5.22. Such a large variance in the beta part of the model generates over a quarter of all cluster probabilities as zero. While the quasibinomial loses a few percentage points, the CRVE remains right around 95% coverage with the wild bootstrap.

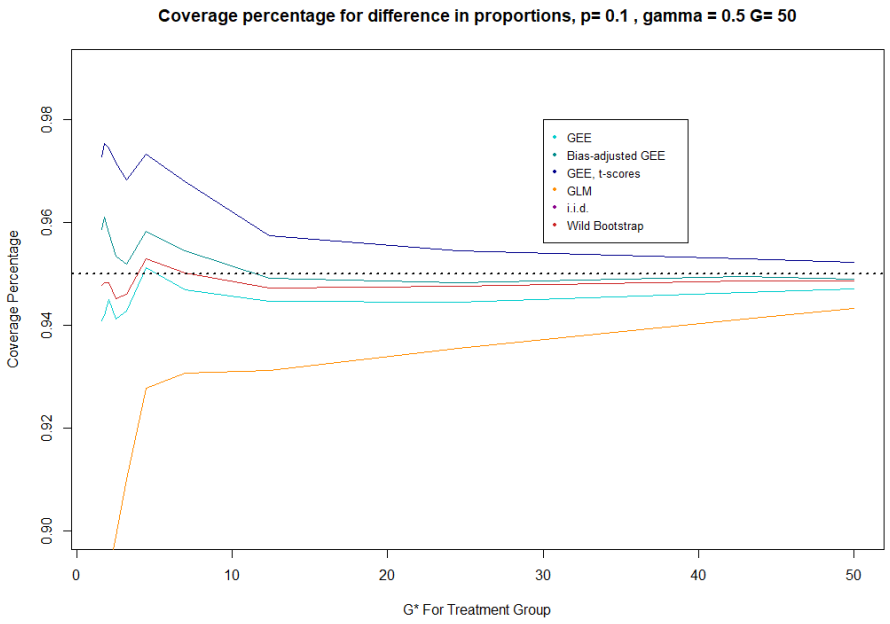


Figure 5.22: Comparison of GEE estimators with wild bootstrap estimator with $p = .1, \gamma = 0.5$

The bootstrap estimator, while effective, can yield values outside the range of the pseudo-binomials when a cluster residual is large. For example, consider a cluster with 50 observations which had $Y_g =$

28 successes, but the overall estimate was $\widehat{p} = 0.8$. Then the residual for this cluster is $r_g = -12$ so that $Y_{g(boot)}$ is equally likely to be equal to 28 or 52. Y_g only has 50 observations, so the range of the bootstrapped observations is larger than that of the distribution it is replicating.

5.5 Markov Chain Model Simulation

The calculations for the bias and effective number of clusters are functions of unknown cluster variances, for which we substituted a beta-binomial variance with the largest magnitude possible. Because the beta-binomial model is so flexible, this substitution should preserve the robust attributes of the estimator. We decided to verify this by running the coverage percentage simulations with another underlying model. These models will consider different correlation structures.

A two-state Markov Chain model is such a way to create a correlated set of binary values. Although the overall probability of a success (which is represented by the stationary distribution of the Markov chain) is fixed, it is assumed that the previous observation affects the current probability of a success. Suppose $(\boldsymbol{X})_n$, $X_i \in \{0, 1\}$ represents a path on a two-state Markov chain with

transition probability matrix

$$p_{i,j} = \begin{bmatrix} 1-\gamma p & \gamma p \\ \gamma(1-p) & 1-\gamma(1-p) \end{bmatrix} \tag{5.1}$$

as first introduced by (Sponsler, 1957). In this equation, p is the parameter of interest and γ is an arbitrary value such that the correlation between sequential observations is

$$\rho = 1 - \gamma$$

This parameter plays an important role in the rate of convergence of the test statistic. For example, a value of $\gamma > 1$ would imply a negative correlation. If such an anticorrelation factor is significant, binary vectors will have strong alternating tendencies (1010101...). A value of γ close to 0 will lead to positive correlation between observations, increasing cluster variances. The relationship between p and γ is also intricate and discussed in detail in (?, ?), who state admissible bounds for the correlation as a function of p. These are stated later and briefly discussed. The n^{th} step transition probability can be written as

$$p_{i,j}^{(n)} = \begin{bmatrix} (1-p) & p \\ (1-p) & p \end{bmatrix} + \rho^n \begin{bmatrix} p & -p \\ -(1-p) & (1-p) \end{bmatrix} \tag{5.2}$$

The correlation disappears as $n \to \infty$ for $|\rho| < 1$. In other

words, observations far apart are almost independent. The n^{th} step transition matrix in (5.2) converges to its stationary distribution, which by design is the desired probability of success:

$$\begin{bmatrix} (1-p) & p \\ (1-p) & p \end{bmatrix}$$

Variance in the model The variance of the pseudo-binomial $Y = \sum_{i=1}^{n} X_i$, calculated in (Sponsler, 1957) using recurrence times, is

$$\text{Var}(Y) = np(1-p)\left[\frac{1+\rho}{1-\rho}\right] \tag{5.3}$$

which for $\rho = 0$ is the variance of a binomial random variable. We note that the variance explodes as $\rho \to 1$. Simulations show that coverage percentages drop as p and γ approach 1. This is because it is possible for the variance of this model to be greater than $n^2p(1-p)$, when

$$\rho > \frac{n-1}{n+1}.$$

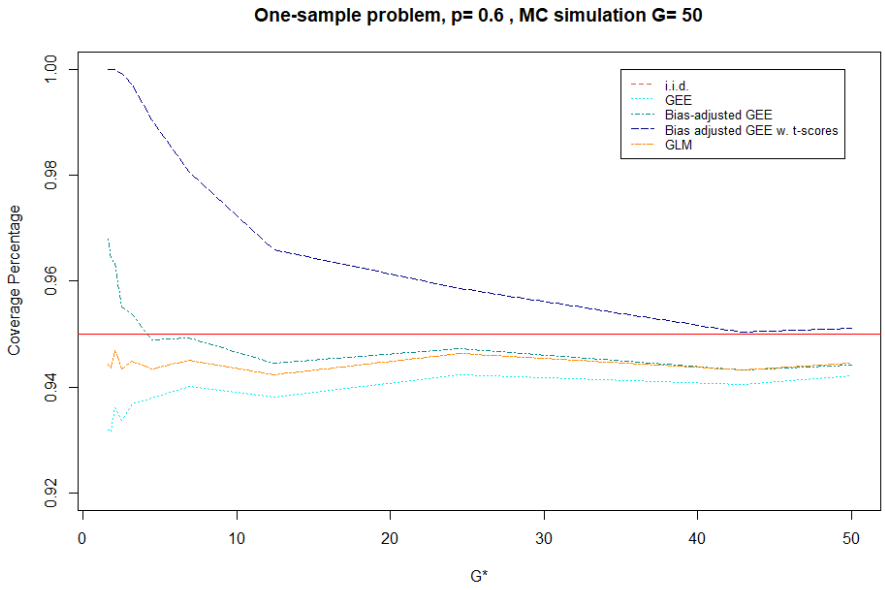


Figure 5.23: Coverage Percentages for Markov chain two-state model with $p = 0.6$ and $\rho = 0.3$.

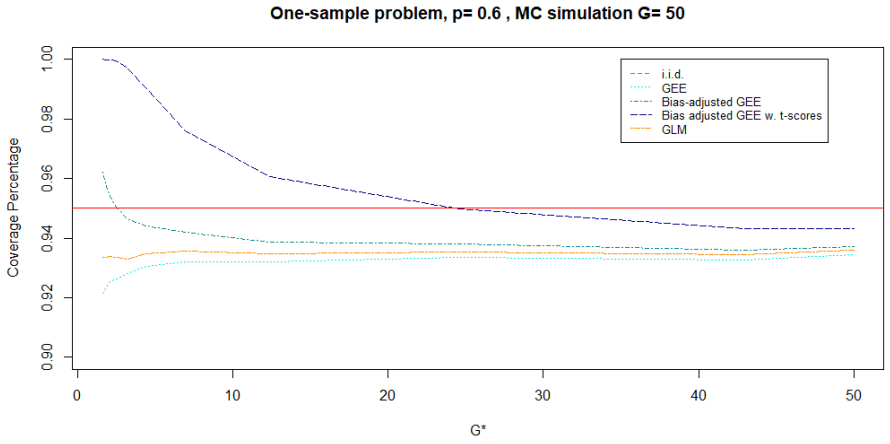


Figure 5.24: Coverage Percentages for Markov chain two-state model with $p = 0.6$ and $\rho = 0.7$.

Importance of the initial distribution. Although the stationary distribution is independent of the initial distribution, cluster sizes in this particular application are often not large enough for this independence to take effect; so for chains of length 50, the initial distribution can greatly influence the path taken and therefore the sample proportions. The initial distribution for the simulations was therefore chosen to be the underlying distribution of the chain:

$$p_0 = (1 - p, \quad p)$$

Simulation results show how much all of the sample proportions can be affected by the initial distribution. The probability of success and the correlation coefficient can interact in a way that yields a Markov chain which will stay in the state it was initially placed in. This occurs near the boundaries of p. We demonstrate with the following example: If $p = .9$ and the correlation is very high, $\rho = .9$, then while the stationary distribution is [.1 .9], the one-step transition probability matrix is

$$\begin{bmatrix} .91 & .09 \\ .01 & .99 \end{bmatrix}$$

If the initial distribution places the chain in state 0, then it will most likely stay there, and the estimate from this cluster will grossly underestimate p (Figure 5.26). This obviously creates a bad esti-

mate. If p is the probability of correctly answering an item on a test, we argue that this example will most likely not occur in practice. In terms of test measurement theory, it would imply first that an individual's ability is almost outside the range of the test, meaning that this is a bad test for this particular student, and second, that this individual gets "stuck" in a state and never changes their answer. Using the stationary distribution as the initial distribution balances out this issue (Figure 5.26) whereas

Figure 5.25 is the result of the initial distribution being equal to the stationary distribution.

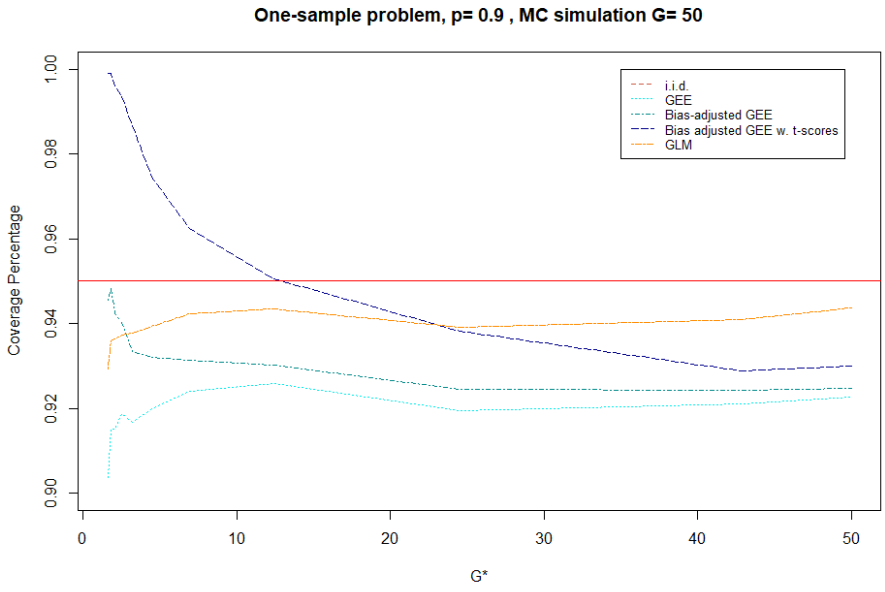


Figure 5.25: Coverage Percentages for Markov chain two-state model with $p = 0.9$ and $\rho = 0.1$. The initial distribution here is $p_0 = (1-p, p)$.

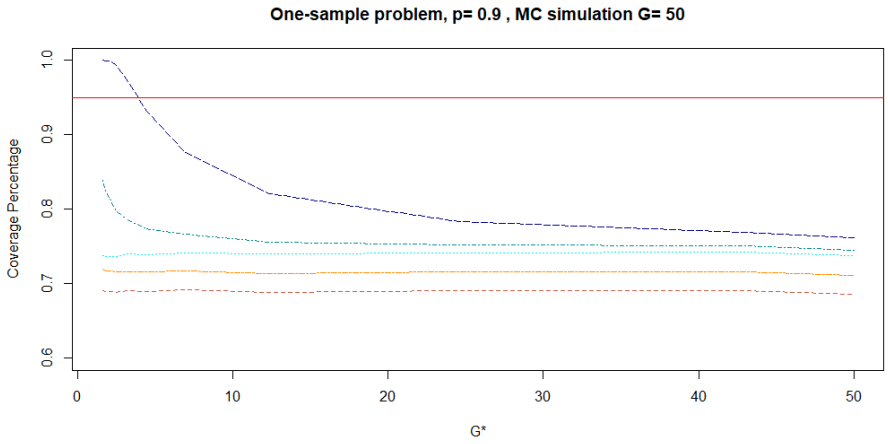


Figure 5.26

Coverage Percentages for the Markov chain two-state model with $p = 0.9$ and $\rho = 0.9$ are shown in 5.26. The large value of the correlation coefficient causes binary vectors to repeat their initial value and get stuck in a loop (111..., or 000...). In this case we see that all estimators are affected; there is no robust coverage percentage to this sorts of patters with such severe correlation .

5.6 Empirical Densities

The empirical density of the variances estimator and the resulting test statistic were also collected and revealed a right-skewed variance estimator for low effective number of clusters. The corresponding t-statistic shows a bimodal density in these same cases. Otherwise the empirical densities of the test statistic appear to match a normal distribution.

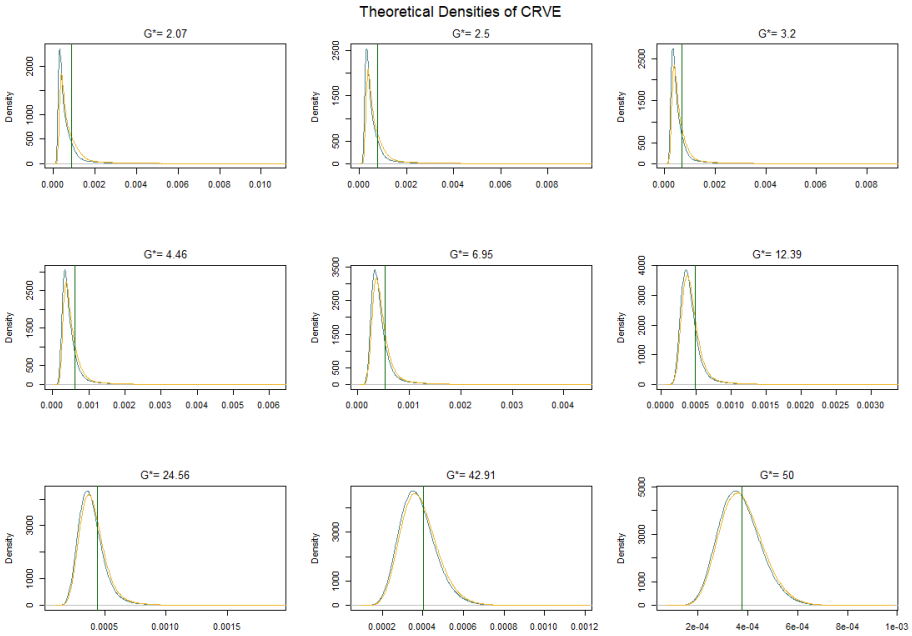


Figure 5.27: Empirical density of the variance estimator for $p = .2$ and $\gamma = .1$ over the distribution of cluster sizes given in Table 9.1

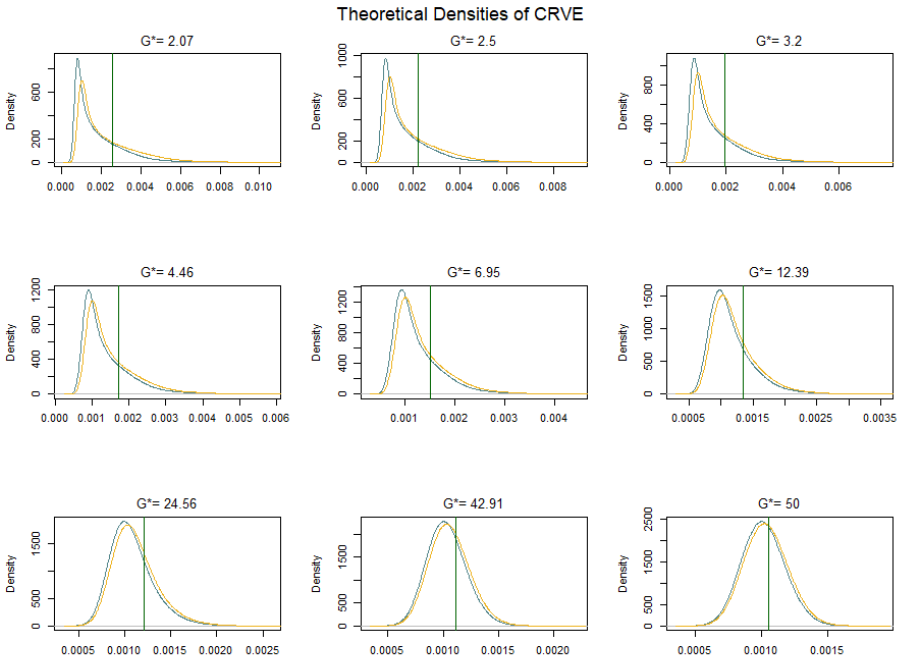


Figure 5.28: Empirical Densities of variance estimator for $p = .6$ and $\gamma = .2$

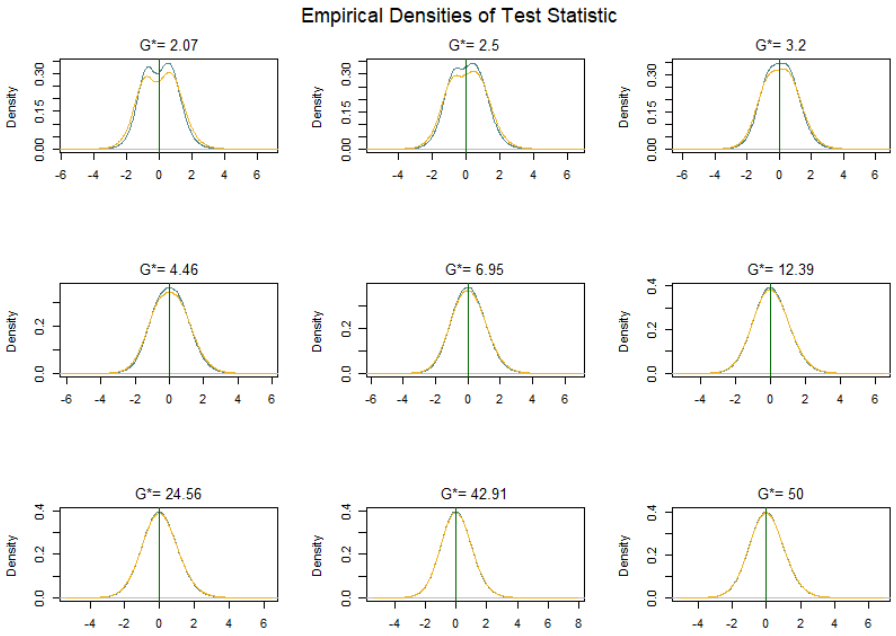


Figure 5.29: Empirical Densities of test statistic for $p = .6$ and $\gamma = .5$

5.7 Discussion

These simulations enabled us to confirm many hypotheses as well as highlight interesting aspects of the CRVE. GEE Empirical Densities show a right-skewed distribution of the CRVE. Our guess is that the estimator slightly overestimates a little bit most of the time, and on occasion will underestimate greatly due to one large cluster with a marginal probability that is substantially different from the mean. The empirical densities of the test statistic look symmetric and almost normal for homogeneous clusters, but develop a bimodal tendency as cluster heterogeneity increases.

We also investigated the effect of the bias bound correction and the Satterthwaite approximation to the degrees of freedom. The bias adjustment improves the estimator marginally, but the variance is still underestimated for small number of effective clusters. Using a t-distribution with G^* degrees of freedom provides the only conservative interval out of all of the options. These adjustments seem to be invariant to the value of the beta distribution parameters p and γ.

We compared the bias-adjusted CRVE with other variance estimators commonly applied to this sort of data, and concluded, with no surprise, that assuming an independent and identically distributed distribution is definitely not the answer. The variance of $\widehat{p}$ is grossly underestimated and gets worse as probabilities near 0 and

1 or as the correlation parameter γ increases. The quasibinomial estimator, which should model correlation, behaved in a somewhat parallel manner to the CRVE. We attempted a wild bootstrap estimator which seemed to be on par with the CRVE, only much more computationally intensive. All estimators appear to be affected by low probabilities. the GLM tends to be more robust to large differences between cluster variances. The iid estimator is highly incompetent for any model which has correlation within the clusters. The bootstrap and GEE estimator are robust in the same ways.

Our conclusion from these simulations is that the bias-adjusted CRVE is comparable to the bootstrap, yet much simpler computationally than both the quasi-binomial and the bootstrap methods. Any inference assuming independence within clusters will be near useless. The CRVE seemed to be more robust to other cluster dependence structures than the GLM. The bias bound correction rectified the underestimation of the CRVE but seems to over-correct for effective numbers of clusters less than 2. These are extreme situations, in which researchers should perhaps be aware that they are using conservative confidence intervals. Any concern we had about the slight under-coverage (94%) was remedied when we increased the total number of clusters G to 100, where we observed 95% coverage for homogeneous clusters, even with large covariance present in the model. We therefore conclude that this bias-adjusted CRVE offers a

compelling alternative to variance estimates from generalized mixed effects models and should be considered when the primary research interest is the mean parameter in an unevenly clustered population.

Chapter 6

Multivariate Cluster-Robust Variance Estimator

Conditions of consistency and asymptotic normality are now extended to a multivariate experiment where multiple probabilities are of interest across clustered populations. This natural extension from a one-dimensional problem to L dimensions mainly involves work around the covariance estimates. A chi-square test is developed to test for equal probabilities, and more generally for a set of contrasts of elements of the probability vector. We begin by setting up the experiment in question.

6.1 Multivariate Clustered Binomial Experiments

A researcher may find himself in a situation where more than one proportion is of interest across a clustered population. We assume the same cluster structure as for the CRVE, with any within-cluster dependence unspecified. Suppose we are interested in a vector of probabilities:

$$\boldsymbol{p} = \big(p_1, \ldots, p_L\big)$$

A single observation from a cluster is now a vector of binary values of length L. As in the univariate case, we collect n_g of these observations from each independent cluster and write the sums as

$$\boldsymbol{Y}_g = (Y_{g1}, \ldots, Y_{gL})$$

where Y_{gl} is equal to a sum of binary values which may not be independent or identically distributed; we therefore do not make the assumptions that these have a marginal binomial distribution. In the one-dimensional problem we argued that consistency of the variance estimator was dependent on a reasonable level of cluster size homogeneity, captured by the effective number of clusters G^* (4.3). This notion extends naturally to the multivariate case because of the aforementioned structure of the data, in which the cluster sizes are unchanged: a vector of binary values is collected at each point

rather than a single value, but the number of points inside each cluster remains the same. This will later be applied to longitudinal test data, in which we consider the clusters to be individuals. Each subject answers a full test, repeatedly over time. The size of each cluster is how many times each subject has taken the test. Note that this structure allows for differently sized-clusters. Let $n = \sum_{g=1}^{G} n_g$, where $(n_1, \ldots, n_G)$ are the cluster sizes. Each probability is estimated using its overall sample proportion:

$$\widehat{p}_i = \sum_{g=1}^{G} Y_{gi}/n, \qquad \widehat{\boldsymbol{p}} = \frac{1}{n}\boldsymbol{Y}\mathbb{1}_G$$

Each individual probability estimate can be shown to have the same properties as in Chapter 4. That is, marginally,

$$\frac{\widehat{p}_i - p}{\sqrt{\widehat{V}_i}} \overset{appr}{\sim} \mathcal{N}(0,1), \qquad \text{for reasonable } G^*$$

We are still using cluster-robust estimators, so that the covariance between two binary observations within any cluster is not specified. However, the covariance of any two sample proportions $\widehat{p}_i$ and $\widehat{p}_j$ should be modeled, since this is a covariance across clusters that can be specified in the same way as the cluster-robust variance estimator.

Covariance Estimate

Although each cluster is independent, the within-cluster correlation structure is not specified. As a result, the estimates $\widehat{p}_1, \ldots, \widehat{p}_L$

may not be independent. Any covariance between $\widehat{p}_i$ and $\widehat{p}_j$ would come from dependence between two observations within the same cluster, Y_{gi} and Y_{gh}. Since we will apply this to a situation in which clusters are subjects, it makes sense to think that two observations taken from the same person might be correlated.

We denote the potential covariance between two probability estimates as

$$\gamma_{ij} = \text{Cov}(\widehat{p}_i, \widehat{p}_j)$$

for $i \neq j$. Because clusters are independent, any covariance will come from intra-cluster associations:

$$\gamma_{ij} = \frac{1}{n^2} \sum_{g}^{G} \gamma_{gij} \qquad \Big(\text{Cov}(Y_{gi}, Y_{g'j}) = 0 \quad \forall i, j, \quad g \neq g' \Big)$$

where γ_{gij} is the unspecified covariance for an individual's responses to items i and j. Similarly to the cluster-robust variance estimate, we estimate cluster-wide covariances (rather than on the binary level). The extension of the cluster-robust variance estimator is then given by the matrix $\hat{\Sigma}$:

$$\hat{\Sigma} = \frac{1}{n^2} \sum_{g=1}^{G} \big(\boldsymbol{Y_g} - n_g \widehat{\boldsymbol{p}} \big) \big(\boldsymbol{Y}_g - n_g \widehat{\boldsymbol{p}} \big)^\top \tag{6.1}$$

As for the cluster-robust variance estimator, the covariance estima-

tor

$$\hat{\gamma}_{ij} = \frac{1}{n^2} \sum_{g=1}^{G} (Y_{gj} - n_g \widehat{p}_j)(Y_{gi} - n_g \widehat{p}_i) \tag{6.2}$$

is unbiased estimator for homogeneous clusters with the simple correction $\frac{G}{G-1}\hat{\gamma}_{ij}$. For bias calculations of the covariance estimate, see appendix section 9.2.3. If the underlying covariance matrix is diagonal, then $\gamma_{ij} = 0$ and the multiplicative bias of the covariance estimate disappears.

6.2 Asymptotic Chi-Square Test

In this chapter we show that the variance-covariance estimator $\hat{\boldsymbol{V}}$ can be used to obtain a chi-squared test under similar conditions to those stated by the CMS paper. The following lemma is an established multivariate result.

Lemma 6.2.1. *Suppose that for N large,*

$$\widehat{\boldsymbol{p}} - \boldsymbol{p} \overset{appr}{\sim} N(\boldsymbol{0}, \Sigma)$$

Then

$$\big(\widehat{\boldsymbol{p}} - \boldsymbol{p}\big)^\top \Sigma^{-1} \big(\widehat{\boldsymbol{p}} - \boldsymbol{p}\big) \rightsquigarrow \chi^2_L.$$

The main result of this chapter is the fact that the same result applies when the cluster-robust estimate of the covariance matrix of $\boldsymbol{Y}$ is used. This estimator will be labeled $\hat{\Sigma}$ and is the L-dimensional version of the CRVE defined in Eq.4.2.

This result is shown easily enough when the variance-covariance matrix is known to be diagonal as the quadratic form can be written as a sum of independent terms:

$$\begin{aligned}
\widehat{\boldsymbol{p}}^{\top}\Sigma^{-1}\widehat{\boldsymbol{p}} &= \widehat{\boldsymbol{p}}^{\top}\left(\hat{\Sigma}^{-1}\hat{\Sigma}\right)\Sigma^{-1}\widehat{\boldsymbol{p}} \\
&= \widehat{\boldsymbol{p}}^{\top}\begin{bmatrix} 1/\widehat{V}_1 & 0 & \dots \\ 0 & 1/\widehat{V}_2 & \dots \\ 0 & \dots & 1/\widehat{V}_k \end{bmatrix}\begin{bmatrix} \widehat{V}_1/V_1 & 0 & \dots \\ 0 & \widehat{V}_2/V_2 & \dots \\ 0 & \dots & \widehat{V}_k/V_k \end{bmatrix}\widehat{\boldsymbol{p}} \\
&= \sum_{j=1}^{L-1}\left(\frac{\widehat{p}_j^2}{\widehat{V}_j}\right)\cdot\left(\frac{\widehat{V}_j}{V_j}\right)
\end{aligned} \tag{6.3}$$

Here we can use the univariate proof. We have established that $\frac{\widehat{V}_j}{V_j} \xrightarrow{P} 1$ and therefore $\sqrt{\frac{\widehat{p}_j^2}{\widehat{V}_j}} = \frac{\widehat{p}_j}{\sqrt{\widehat{V}_j}}$ has an asymptotic standard normal distribution. Therefore $\frac{\widehat{p}_j^2}{\widehat{V}_j} \rightsquigarrow \chi_1^2$ from the univariate proof and a diagonal matrix implies that the $\widehat{p}_k$'s are independent, so that

$$\sum_{j=1}^{L-1}\frac{\widehat{p}_j^2}{\widehat{V}_j} \rightsquigarrow \chi_{L-1}^2$$

By an application of Slutsky's theorem we therefore obtain the following lemma

Lemma 6.2.2. *For a set of estimates $\widehat{\boldsymbol{p}}$ that are known to be independent,*

$$\left(\widehat{\boldsymbol{p}}-\boldsymbol{p}\right)^{\top}\hat{\Sigma}^{-1}\left(\widehat{\boldsymbol{p}}-\boldsymbol{p}\right)\rightsquigarrow\chi_{L}^{2}.$$

Without loss of generality, we can assume that our covariance matrix is diagonal, as long as covariances are still estimated. Indeed, in many situations, the elements of (6.9) may not be independent. For any set of such estimates, we argue that an equivalent set of *independent* estimates can be generated by a rotation of the data. Therefore, the theory for underlying diagonal covariance matrices can be extended to any covariance matrix due to the equivariant property of the variance estimator. We now state the main result.

Theorem 6.2.1 (Asymptotic Chi-Square Test)**.** *Suppose the entries of $\widehat{\boldsymbol{p}}$ are all independent but this is unknown to the researcher. That is, each covariance term $\gamma_{ij}=0$ but is estimated using Eq.(6.2). Then*

$$\left(\widehat{\boldsymbol{p}}-\boldsymbol{p}\right)^{\top}\hat{\Sigma}^{-1}\left(\widehat{\boldsymbol{p}}-\boldsymbol{p}\right)\rightsquigarrow\chi_{L}^{2}. \tag{6.4}$$

We prove the chi-square approximation result for a more general setting in the following manner: Let $z = \Sigma^{-1/2}\big(\widehat{\boldsymbol{p}} - \boldsymbol{p}\big)$. $\widehat{\boldsymbol{p}}$ is an unbiased estimator so that for large sample sizes, z has an approximate multivariate standard normal distribution. That is, the covariance matrix of z is diagonal and close to the identity matrix.

Per Lemma 6.2.1, $z^\top z$ follows an approximate chi-square distribution. We will use K degrees of freedom from here on out for the sake of generality. The quadratic term with the estimated covariance matrix can be expressed in the following way:

$$\big(\widehat{\boldsymbol{p}} - \boldsymbol{p}\big)^\top \hat{\Sigma}^{-1}\big(\widehat{\boldsymbol{p}} - \boldsymbol{p}\big) = z^\top z + z^\top B z, \tag{6.5}$$

where the matrix B is equal to $\mathbf{I} - \Sigma^{1/2}\hat{\Sigma}^{-1}\Sigma^{1/2}$. Therefore it is clear that convergence of (6.5) to a chi-square distribution depends on $z^\top B z$ becoming negligible in probability. To prove this, we will show that both the expectation and variance of this term converge to zero in probability.

By a geometric series argument (see appendix, Section 9.2.2), it is sufficient to show that

$$z^\mathsf{T}\Sigma^{-1/2}\left(\mathbf{I} - \Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2}\right)\Sigma^{-1/2}z \xrightarrow{\mathbb{P}} 0$$

for convergence of $z^\top B z$ to zero. Since the z_i $(i = 1, \ldots, K)$ are

uncorrelated with mean 0 and variance 1,

$$\mathbb{E}\big(z^\top Bz\big) = \sum_i b_{ii} z_i^2 = \operatorname{tr}(B) = K - \operatorname{tr}(\hat{\Sigma}\Sigma^{-1})$$

and

$$\operatorname{Var}(z^\top Bz) = 2\operatorname{tr}(B^2) = 2\big(K - 2\operatorname{tr}(\hat{\Sigma}\Sigma^{-1}) + \operatorname{tr}(\hat{\Sigma}^2\Sigma^{-2})\big)$$

Suppose the underlying matrix is diagonal, but we estimate covariances anyway. That is, $\Sigma = \operatorname{Diag}(V_j)_{j=1,\dots K}$ and

$$\hat{\Sigma} = \begin{pmatrix} \widehat{V}_1 & \hat{\gamma}_{12} & \dots \\ \hat{\gamma}_{12} & \widehat{V}_2 & \dots \\ \dots & \hat{\gamma_{jk}} & \widehat{V}_K \end{pmatrix}$$

Then $\operatorname{tr}(\hat{\Sigma}\Sigma^{-1}) = \sum_{j=1}^{K} \frac{\widehat{V}_j}{V_j}$ and

$$\operatorname{tr}(\hat{\Sigma}^2\Sigma^{-2}) = \sum_{j=1}^{K} \Big(\frac{\widehat{V}_j}{V_j}\Big)^2 + \sum_{j=1}^{K} \Big(\frac{\sum_{i\neq j} \hat{\gamma}_{ij}^2}{V_j^2}\Big)$$

The last term involving the correlation can be rearranged to give

$$\sum_{j=1}^{K} \sum_{i>j} \Big(\frac{\hat{\gamma}_{ij}}{V_i V_j}\Big)^2 \big[V_i^2 + V_j^2\big] \tag{6.6}$$

where we know V_i^2 and V_j^2 to be $O_P(1)$. Plugging everything back into the expectation and variance of the quadratic term then shows

that

$$\mathbb{E}\big(z^\top Bz\big) = K - \sum_{j=1}^{K} \frac{\widehat{V}_j}{V_j} \qquad \xrightarrow{\mathbb{P}} \quad 0,$$

and

$$\mathrm{Var}(z^\top Bz) = 2\Big(\underbrace{K - 2\sum_{j=1}^{K} \frac{\widehat{V}_j}{V_j} + \sum_{j=1}^{K} \Big(\frac{\widehat{V}_j}{V_j}\Big)^2}_{\xrightarrow{\mathbb{P}} 0} + \sum_{j=1}^{K}\sum_{i>j} \Big(\frac{\hat{\gamma}_{ij}}{V_i V_j}\Big)^2 \big[\underbrace{V_i^2 + V_j^2}_{\xrightarrow{\mathbb{P}} 1} \big] \Big)$$

The implied convergences above result from the already-proven fact that $\frac{\widehat{V}_j}{V_j} \xrightarrow{\mathbb{P}} 1$ for each j. Sufficient conditions for convergence should therefore ensure that (6.6) becomes negligible in probability. In order to show this, we inspect the covariance estimates.

All that remains is to show that (6.6) is negligible in probability for large samples. Define

$$\begin{aligned} \tilde{\gamma}_A &= \frac{1}{n^2}\sum_{g=1}^{G} (Y_{gj} - n_g p_j)(Y_{gi} - n_g p_i) \quad =: \quad \frac{1}{n^2}\sum_{g=1}^{G} \tilde{\gamma}_{Ag} \\ \tilde{\gamma}_B &= \frac{1}{n^2}\sum_{g=1}^{G} n_g^2 (\widehat{p}_i - p_i)(\widehat{p}_j - p_j) \end{aligned} \tag{6.7}$$

Using Cauchy-Schwarz inequality on each term $\tilde{\gamma}_{Ag}$, the first term can be bounded by the fourth central moments of the individual

subject-item responses. Assuming $\gamma_{gij} = 0$,

$$\mathrm{Var}(\tilde{\gamma}_A / V_i V_j) \leq \frac{1}{n^4} \sum_{g=1}^{G} \frac{\left[\mathbb{E}(Y_{gi} - n_g p_i)^4\right]^{\frac{1}{2}}}{V_i} \cdot \frac{\left[\mathbb{E}(Y_{gj} - n_g p_j)^4\right]^{\frac{1}{2}}}{V_j}$$

Each of these summands is bounded and goes to zero by Conditions 1, 2 and 4. For details, see the proof of Lemma 1 in CMS. Similarly, the second term $\tilde{\gamma}_B$ is bounded by the fourth central moments of $\widehat{p}_i$ and $\widehat{p}_j$:

$$\mathrm{Var}(\tilde{\gamma}_B / V_i V_j) \leq \frac{\left[\mathbb{E}(\widehat{p}_i - p_i)^4\right]^{\frac{1}{2}}}{V_i} \cdot \frac{\left[\mathbb{E}(\widehat{p}_j - p_j)^4\right]^{\frac{1}{2}}}{V_j} \cdot \sum_{g=1}^{G} \left(\frac{n_g}{n}\right)^4 \quad (6.8)$$

Because we have already shown asymptotic normality of $\left(\widehat{p}_i - p_i\right)/\sqrt{V_i}$ for each category or item i, each expectation in the bound of (6.8) is like the second moment of a chi-square random variable with one degree of freedom, which is equal to 3. Therefore, (6.8) goes to zero with Condition 4.1.3 and $\tilde{\gamma}_A / V_i V_j$ and $\tilde{\gamma}_B / V_i V_j$ are negligible in probability as $G \to \infty$. Their cross product is therefore also negligible, and the desired result is proven. ■

6.2.1 Contrasts

Suppose we are interested in an arbitrary contrast α, which is taken to be a linear combination of the elements of $\boldsymbol{p}$ depending on the hypothesis of interest. For example, one might wish to test if all items have the same approximate difficulty level. Such a question could be answered with a chi-square test and a contrast of the form

$$\alpha = \begin{bmatrix} p_1 - p_2 & \dots & p_1 - p_i & \dots & p_1 - p_L \end{bmatrix}. \tag{6.9}$$

We will denote $\widehat{\boldsymbol{V}}_{\hat{\alpha}}$ as the cluster-robust variance-covariance estimator of the contrasts estimate $\hat{\alpha}$, given by the matrix $\boldsymbol{A}$:

$$\widehat{\boldsymbol{V}}_{\hat{\alpha}} = \frac{1}{n^2}(W - T\hat{\alpha})^\top (W - T\hat{\alpha}) = \frac{1}{n^2} A^\top \hat{\Sigma} A \tag{6.10}$$

In the case of (6.9),

$$\boldsymbol{A}_{L\times(L-1)} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ -1 & 0 & \dots & \dots \\ 0 & -1 & 0 & \dots \end{pmatrix} \tag{6.11}$$

As we have already established a chi-square distribution under the aforementioned conditions, 6.4 can be naturally extended to the contrasted data so that

Lemma 6.2.3.

$$\hat{\alpha}^{\top}\widehat{\boldsymbol{V}}^{-1}\hat{\alpha} \overset{appr}{\sim} \chi^2_{L-1}.$$

The next chapter extends beta-binomial simulations from the CRVE to test for a hypothesis of equal probabilities, as with the contrast matrix 6.11. Coverage percentages confirm the asymptotic distribution of this test statistic for Eq. (6.5).

Chapter 7

Simulation for Coverage Percentages of the Chi-Square Tests

7.1 Comparing with Other Estimators

Simulations for this chapter were run in a similar manner to the CRVE. To test coverage percentages of our chi-square statistic, we simulated data under the null hypothesis that all items had the same difficulty, as per the contrast defined in 6.9. Results were similar to the univariate CRVE coverage percentages, with a possibly faster convergence (under the null hypothesis, each question had the same difficulty, essentially multiplying the sample size by the number of probabilities being estimated across the clustered popu-

lations). Graphs were plotted as a function G^* or $\log G^*$ to better inspect ranges of fast convergence. The MVCRVE chi-square tests had 95% coverage regardless of the probability values under the null hypothesis and were robust to large values of the correlation tuning parameter for the beta distribution, γ. To test for robustness to dependent observations, time trends were added within clusters so that the probability of success changed according to a simple linear function across the clusters. While the MVCRVE was robust to these time trends, the estimator modeling every observation as independent was greatly affected and had much lower coverage percentages. We tested both chi-square and F statistics and noticed a slight increase of about one percentage point for the F-statistics. Overall, the simulations confirm that the multivariate cluster-robust variance estimator yields a quadratic term which can be well approximated by a chi-square distribution with as little as 10 clusters, even in an unbalanced cluster distribution 7.2.

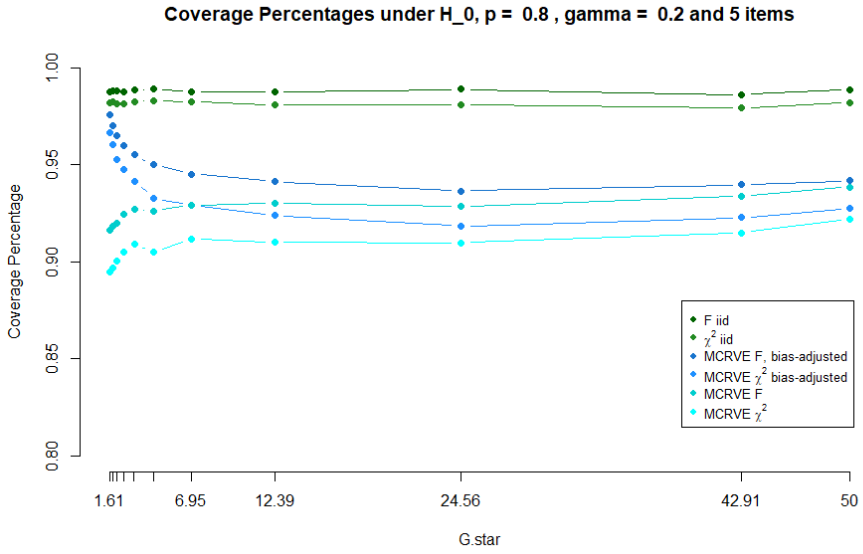


Figure 7.1: Coverage Percentages under null hypothesis and independent model

The "iid" test statistic. Suppose all clustering was disregarded and each binary observation was treated as independent. Suppose further that we treat every binary outcome within a vector as also independent.That is, $Y_{gli} \in \{0, 1\}$, and $Y_{gli} \perp Y_{g'l'i'}$, for all values of g', l', and i'. In a longitudinal testing study where clusters are individuals, this would imply that any person's answer to two questions on the same testing instance would not be correlated. The covariance matrix is then diagonal and we only estimate variances $\widehat{V}_i, \quad i = 1, \ldots, L$. This estimator was plotted agai nst the MVCRVE under both an independent and identically distributed setting (Figure 7.1) Such an assumption is not robust to time effects, as can be seen in Figure 7.2.

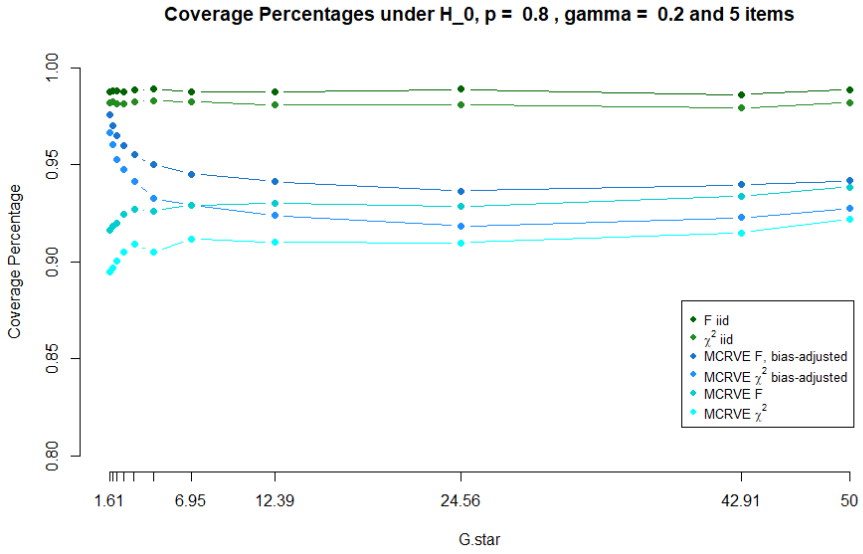


Figure 7.2: Coverage percentages for MVCRVE when each cluster's probability fluctuates slightly and randomly over time.

As with the univariate case, there is still a slight under-coverage even for equally-sized clusters. However, this gap is remedied as the number of clusters, or subjects, increases. For G=100 clusters with an average of 20 observations each (for a total of 2,000 binary observations), the coverage percentage at $G^* = G$ is 0.9497. It is already a widely accepted result that ignoring potential correlation results in underestimating the variance, and this fact is confirmed in these simulations, where we see coverage percentages drop significantly, even for large numbers of clusters.

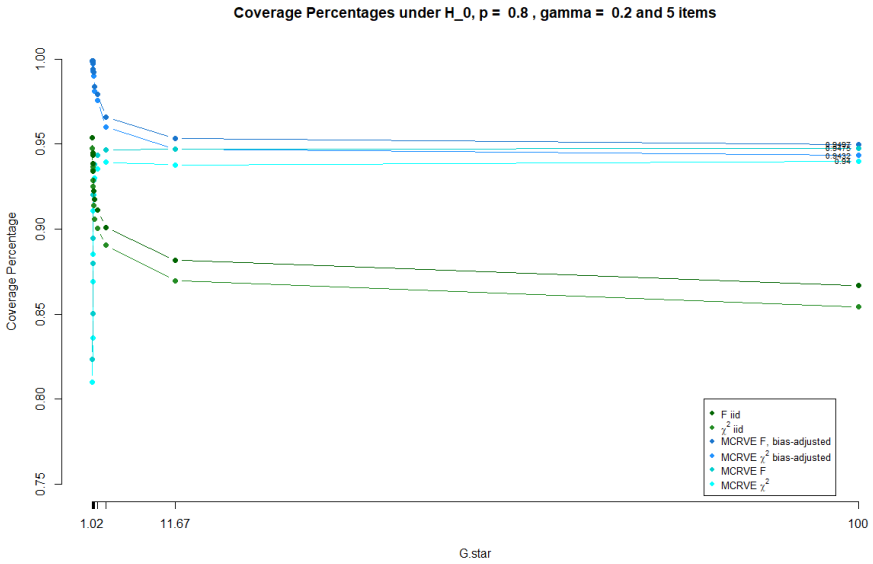


Figure 7.3: Coverage Percentages get closer to 95% for the MVCRVE statistics as the number of clusters increases.

7.2 Effect of Cluster Sizes on the binomial model estimate

It was noted in part 1 that increasing cluster sizes only made any measurement error from unmodeled correlations worse (Barbieri et al., 2015). That is, increasing the total number of observations (rather than the effective number of clusters) actually makes the iid. estimator worse, as seen in Figures 7.2 and 7.2, which both have 400 total observations. Having 20 clusters with 20 observations each resulted in coverage percentages of between 85% and 90% for the statistic assuming independence, while 10 clusters of 40 observations

yielded a much worse coverage percentages (lessened by at least 15%).

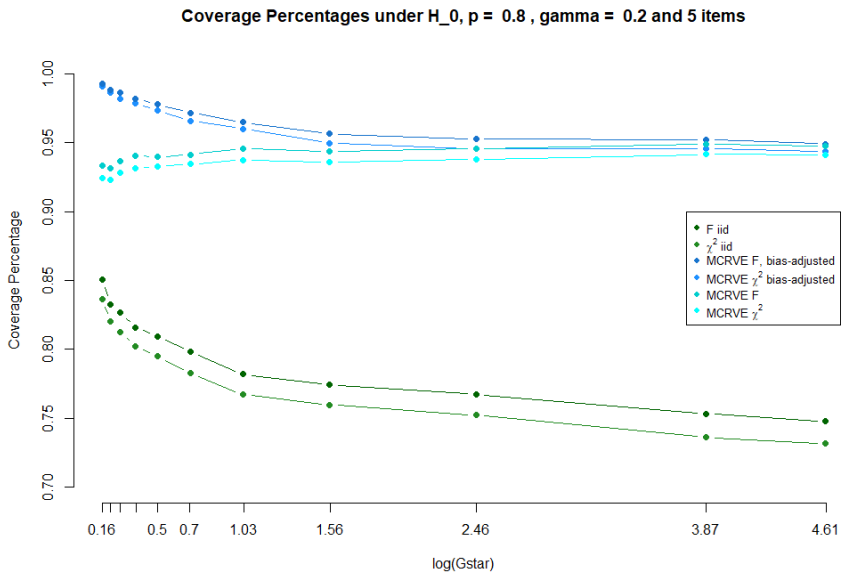


Figure 7.4: Coverage Percentages with 400 observations, 10 clusters

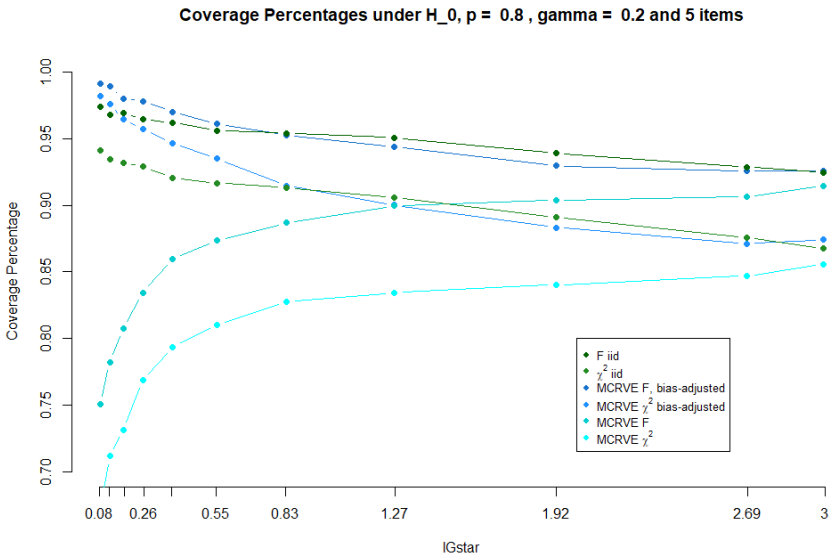


Figure 7.5: Coverage Percentages with 400 observations, 20 clusters

Part III

Application of Cluster-Robust Variance Estimators to Longitudinal Test Data

Chapter 8

Cluster-Robust Variance Estimators for Psychometric Analysis of Longitudinal Datasets

Longitudinal data collected from a group of individuals being administered the same test over a period of time matches a clustered population, where individuals are independent, but their repeated answers to the same items may be correlated. The current practices for repeated measures data in classical test theory and IRT models are discussed discussed in Chapter 2. The cluster-robust variance estimator measures the variance of $\widehat{p}$ whether it be univariate or multivariate quantity, which means it inherently belongs to classi-

cal test analysis. A cluster-robust variance estimator was derived by (Feddag, Grama, & Mesbah, 2003) for the item parameters of the Rasch model and will be briefly discussed in the next section, partially to illustrate why we chose not to derive the CRVE for the item difficulty logistic parameters δ_i.

8.1 Cluster-Robust Variance Estimators and Classical Test Theory

8.1.1 Testing for Parallel Tests

We present a simple hypothesis test of the time effects using the cluster-robust variance estimator, which can be used to check for parallel tests in CTT, a common test assumption. As defined in Chapter 1, parallel tests have questions of identical difficulty and therefore the sum score has a binomial distribution (under the null hypothesis that the tests are parallel. We now set up this test. Suppose data is collected from a longitudinal trial with G individuals who are administered the same test repeatedly over a period of time. It is often the case in longitudinal studies, that patients have uneven number of visits, because, for instance, some patients drop out. Let $\boldsymbol{Y}$ be the data matrix described in 6.1, of dimension $G \times L$ where G is the number of subjects in the experiment, and L is the number of questions on the test. Under the null hypothesis, each binary value has the same probability of success, however we may

still have correlation within subject responses to the same questions over time. We therefore should not assume a binomial distribution on cluster sums, and proceed with the sample proportion estimate and cluster-robust variance estimates 6.10. Let α and $\boldsymbol{A}$ be the quantities defined in Section 6.2.1. Then

$$\hat{\alpha}^{\top}\widehat{\boldsymbol{V}}^{-1}\hat{\alpha} \overset{appr}{\sim} \chi^2_{L-1}$$

The strength of the bias-adjusted CRVE can be useful in these situations where it is possible to have a very large number of patients at the baseline visit, with a few patients with that come back repeatedly. Therefore the clusters might be quite unbalanced and the effective number of clusters G^* can provide a good measure of how stable their test statistic and resulting inference will be.

8.1.2 Testing for a Time Effect

Suppose a test is repeatedly administered to the same group of patients over the course of a longitudinal study, as with the experiment described in Section 2.1. In this case, we can also apply the MVCRVE to test for an overall time effect on the overall probability of success. Under the null hypothesis, the "level" of the test, to speak in CTT terms, should remain stable throughout time.

8.2 Cluster-Robust Variance Estimators and the Rasch Model

As discussed in Part I of this dissertation, the family of Rasch models has desirable estimation properties which make it one of the most commonly used models for psychometric evaluation and test item calibration. Non-parametric approaches to repeated measures using the Rasch model essentially ignore any correlation (stacking), greatly reduce the sample size to have only independent observations (anchoring), or perform data manipulation to reduce correlation between repeated observations (Mallinson approach). Parametric approaches are mostly mixed effects models with a hierarchical structure and usually a normal distribution on the subjects. Most of these models do not belong to the family of Rasch models, and therefore their estimates do not have the same sufficiency or consistency properties or may not be as robust under model specification.

We apply the MCRVE to longitudinal data and estimate the covariance matrix for the stacked item estimates. The stacking method 2.2 is a way for a practitioner to obtain item estimates using the pairwise CML method used in the Rasch program RUMM. Repeated measures from one individual are treated as independent observations. Here we specify a slightly different structure for the data, allowing for potential time correlation.

8.2.1 GEEs and the Mixed Rasch Model

The application of GEEs to longitudinal data has already been explored in (Feddag et al., 2003). The process is quite intricate due to the fact that a hierarchical logit-normal model does not have closed forms for the first and second moments; these must be estimated using numerical integration. Since one of our goals is to avoid modeling a random latent effect, we refer the reader to the cited text for details on this estimation process.

8.2.2 Estimates of p vs. estimates of δ

The variance estimator discussed in Chapter 4 is robust precisely because it avoids specifying too many assumptions. Many publications compare CTT estimates to IRT estimates (Petrillo et al., 2015; Blanchin et al., 2011; Magno, 2009; Hambleton & Jones, 1993; Barbieri, Peyhardi, et al., 2017). However, no publication was found that drew the mathematical connection between a CTT estimate and the item difficulty of an IRT estimate. This seems natural, since the models are fundamentally different. However we can still compare them for a subject with "average" difficulty. CTT only works if items are centered on population anyway. By the Delta method,

$$g(\widehat{p}) - g(p) \rightsquigarrow \mathcal{N}(0, g'(p)^2 V)$$

Therefore we have

$$\log\Big(\frac{\widehat{p}}{1-\widehat{p}}\Big) - \delta \overset{appr}{\sim} \mathcal{N}\Big((0, \frac{1}{\big(p(1-p)\big)^2}V\Big)$$

In other words, the CRVE test statistic will provide valid inference if the items are on target with the population of interest, because in the case the average value of the ability parameter will be zero.

8.2.3 Testing the relationship between bias and range of item parameters

A quick preliminary simulation was run to see how much the range of the deltas would affect the bias of the CRVE. For each case, 10,000 sets of G cluster means using beta-binomial are generated, sample means and CRVE calculated. Of course, the results depend on the variance of the ability parameters as well - a higher variance will overshadow the effects of the δ parameters.

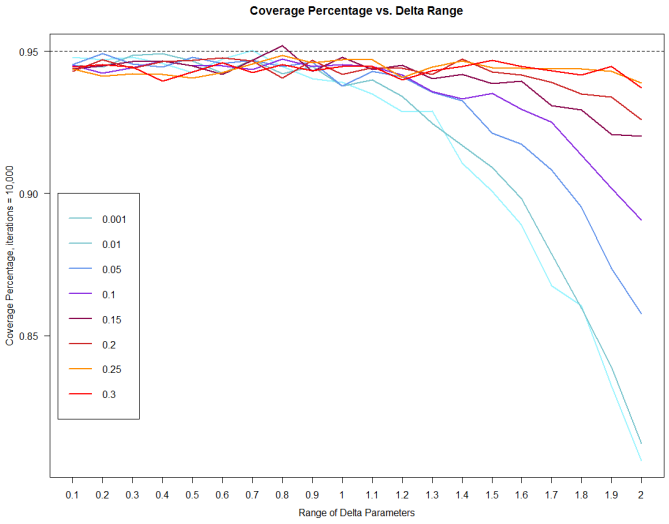


Figure 8.1: Coverage Percentages of the CRVE for sample populations with different variances.

Figure 8.2.3 indicates a positive correlation between the coverage percentage and the variance of the β_g coefficients. Intuitively, this makes sense: the randomness of the ability coefficients cancels out the item levels. Greater values of γ induce a variation that is too large - in this example, $p = 0.75$ and $\gamma \geq 0.4$ (roughly) results in probabilities close to or equal to 1.

Perhaps around $p = \frac{1}{2}$ a greater variation can be beneficial to the Z-estimate, resulting in even better confidence intervals.

8.2.4 Simulations Results of Rasch Estimates

The goal of the set of graphs below is to compare standard errors from the Rasch Anchoring and Stacking methods using CML with

the CRVE. As previously stated, CTT yields unbiased estimators of the probability of an average subject answering a given question - that is, fluctuation in the data due to different individuals being tested is modeled by an additive error to the observed score. These errors are usually assumed to be normally distributed, even though they are clearly not since the data is categorical or binary and there is usually not enough subjects to have a reasonable law of large numbers take place.

That the CTT estimates a biased estimate of the δ parameters when the population is not centered around 0 should therefore be obvious; this is just visualizing the widely accepted result that CTT estimates are sample-dependent. The evidence of the statistical sufficiency of the Rasch CML estimates is obvious in these graphs as well. However the focus should be on the top right graphs, which depict the standard errors, which have been placed on the same scale using the delta method.

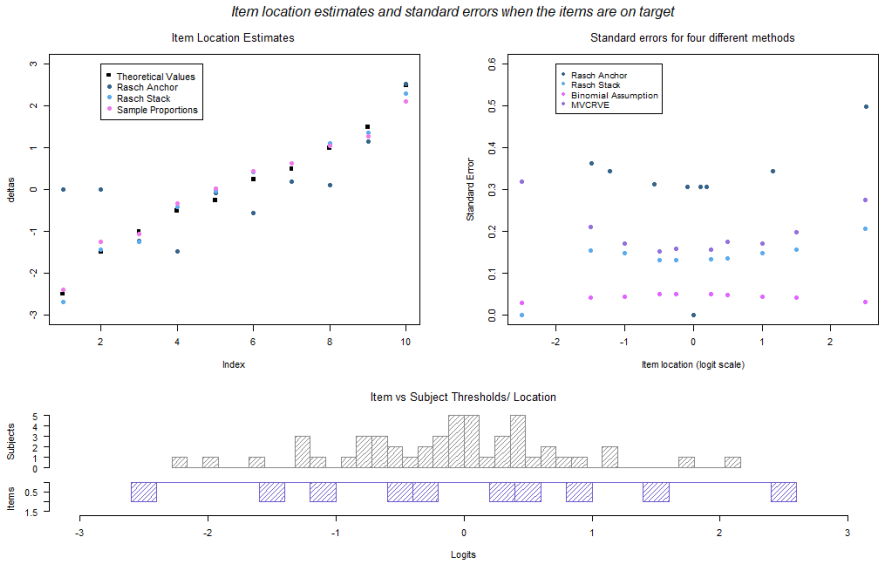


Figure 8.2: Example of a set of items that is centered on the target population

The Rasch anchoring and stacking estimates obtained from conditional maximum likelihood are the only ones to be consistent and robust (in the way that they do not specify a latent distribution on the subjects, therefore it cannot be misspecified). Yet it is known that the anchoring methods yields much larger standard errors due to its sample size reduction, while stacking is known to underestimate the standard errors of the item estimates because it does not account for any potential time dependence. By putting the probability estimates from CTT on the same scale as the Rasch item estimates, we were able to compare them and their standard errors through simulated longitudinal test data. While it may not make much sense to compare these on the edges of the scale, the center

of a scale provides a good opportunity for side-by-side comparison. (Barbieri, Peyhardi, et al., 2017) warns of negative consequences of floor or ceiling effects. When the test is well centered on the sample subject population and has the appropriate spread, however, no question will be too easy or too difficult for *everyone*; that is, no ceiling or floor effects would be observed. In this scenario, each question has a probability of success that is close to one-half. In generalized linear models and the Rasch model, this means that the population and the items are distributed somewhat nicely around 0.

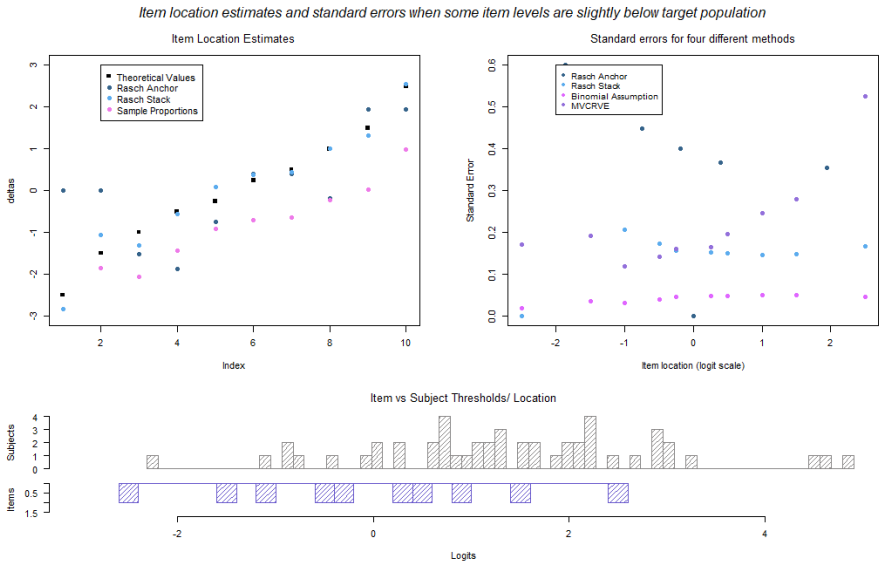


Figure 8.3: Example of a set of items slightly below the target population

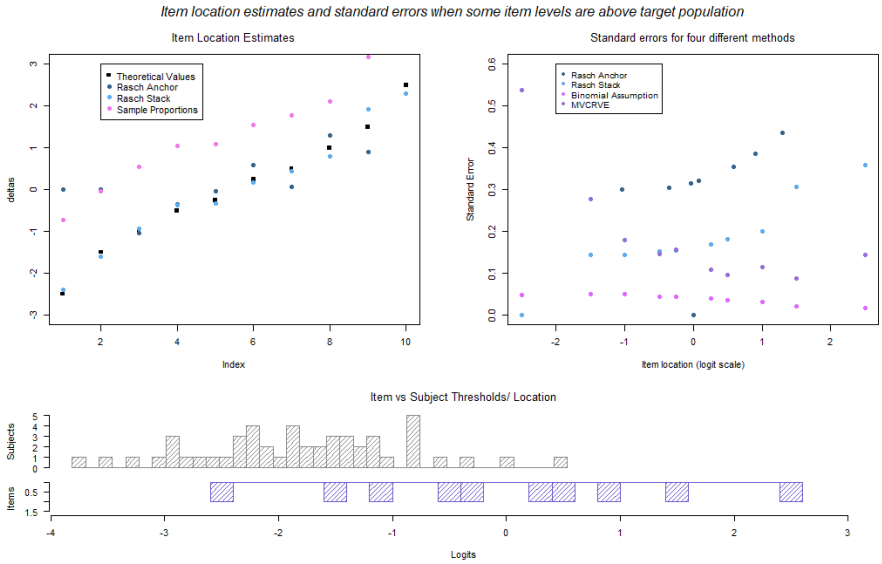


Figure 8.4: Example of a set of items above the target population

Chapter 9

Discussion

This dissertation was a statistical investigation into the potential use of a cluster-robust variance estimator in current test analysis methods and particularly their handling of longitudinal or repeated measures data. We discerned classical test theory, item response theory and Rasch measurement theory as the three branches of psychometric models. After an extensive literature review, we identified limitations of simple methods like CTT as well as potential risks of IRT over-parameterized mixed effects models. As stated in a comparative study, "selection of a psychometric method depends on intended audience and should be justified. For more high-stakes situations, more thorough models like IRT and RMT should be used."(Petrillo et al., 2015). It was concluded that the Rasch model is the only probability model with sufficiency and therefore sample-independent item estimates. Yet important national corpo-

rations like the FDA still use CTT to endorse important questionnaires. The reliability measures calculated to calibrate and validate scales have been criticized extensively and "the issue of test bias is in acute need of scientific scrutiny" when using CTT (Borsboom, 2006). On the other hand, the intricacy of a mixed effect generalized linear model is beyond the mathematical reach of many social scientists and is prone to being blindly interpreted and trusted: " it is essential to understand these differences for researchers deciding which approach to adopt, as well as to know that they each require a complex advanced level of mathematical understanding and unique software. [...] blind application of these methods can result in erroneous conclusions" (Petrillo et al., 2015). Additionally, a covariance structure must be specified when a random effect is placed on the population of interest, and ensuing computing techniques tend to not be robust to structural errors and on occasion even fail to converge. Although numerical methods and fast computing times have made these models more accessible, the tendency in an effort to improve measures is often to add more parameters into the model so that everything in the model is explained, and the issue of parameter identifiability is entirely overlooked.

The underlying covariance structure of longitudinal data is not of interest when the goal is item calibration of a rating scale; rather it is a nuisance that should be addressed so that proper inference can be carried out. A population-average estimator robust to var-

ious covariance designs therefore seems like a sensible idea. We elaborated on the cluster-robust variance estimator based on GEEs for sums of dependent binary values and showed that the behavior of the test statistic is largely controlled by imbalance in cluster variances. Simulations using a beta-binomial model and controlling for the effective number of clusters enabled us to visualize coverage percentages as a function of cluster homogeneity. We derived a bias bound based on the CMS conditions which significantly improved coverage percentages for low effective numbers of clusters, and argued that the test statistic might be close to a t-distribution with the effective number of clusters as the degrees of freedom. While both the bias and Satterthwaite approximation improved the downward bias, both seem to be an over-correction and research remains to be done on how to better control the test statistic in extreme situations with very unbalanced clusters. Conditions were developed to ensure the convergence of hypothesis tests for multidimensional binary data. Under some straightforward conditions, we proved convergence of the chi-square statistic and showed its potential use for testing various assumptions of equivalence, independence or parallel tests in CTT. These simulations of longitudinal Rasch model data confirmed the certification that if a test is well-designed, then it will be on target for the population it is attempting to measure, and CTT estimates will perform reasonably well. In other words, if we assume a symmetric population sample, the average of the ability

parameters is close to zero, which means that the probability of success for an average person on any of the items is simply the inverse *logit* the Rasch item level parameter. Consequently, theory and tests derived for the sample proportions CRVEs can be translated to apply to the Rasch item estimates *for a subject whose ability is equal to 0.* In this case, the CTT estimates are almost identical to the Rasch model ones, and it would seem that doing all of the extra work that goes behind conditional maximum likelihood estimation is not worth it. Again, this is only when we have a nice picture with a set of items and population that are evenly stacked. However, as is now widely recognized, sample proportions as item estimates are sample-dependent and suffer a large bias error whenever the test is not "on target" for its population of interest. The cluster-robust variance estimator, which is an estimator of p and not δ, suffers in the same manner. Despite the abundant criticism of CTT, it still presents a simple, interpretable model which at its best offers results comparable to IRT. It seems therefore that with a simple hypothesis test of parallel items or insignificant time effects, reasonable inference is still in reach and may be a suitable option for simple inference. More work remains to be done on a CRVE which keeps its simplicity yet provides a consistent estimator of the variance of stacked item estimates, which would also be independent of ability estimates.

Appendix

9.1 CRVE Bias Calculations

The detailed calculations for the bias in Eq.(4.6) and bias re-write Eq.(4.7) are now given. Recall that $V = \text{Var}(\widehat{p}) = \frac{1}{N^2}\sum_{g=1}^{G}\text{Var}(Y_g)$ and its estimate $\widehat{V}$ is the CRVE from Chapter 4, Equation (4.2).

Calculating the Bias

Let $\sigma_g^2 = \text{Var}(Y_g)$.

$$
\begin{aligned}
b(\hat{V}) &= \mathbb{E}(\widehat{V} - V) \\
&= \mathbb{E}(\frac{1}{N^2}\sum_{g=1}^{G}(Y_g - n_g\widehat{p})^2) - \frac{1}{N^2}\sum_{g=1}^{G}\sigma_g^2 \\
&= \frac{1}{N^2}\sum_{g=1}^{G}\left(\mathbb{E}(Y_g - n_g\widehat{p})^2 - \sigma_g^2\right) \\
&= \frac{1}{N^2}\sum_{g=1}^{G}\left(n_g^2 V + (1 - 2n_g/N)\sigma_g^2 - \sigma_g^2\right) \\
&= \frac{1}{N^2}\sum_{g=1}^{G}\left(n_g^2 V - \frac{2}{N}n_g\sigma_g^2\right) \\
&= \sum_{g=1}^{G}\frac{n_g^2}{N^2}\cdot V - \frac{2}{N^2}\sum_{g=1}^{G}\frac{n_g\sigma_g^2}{N}
\end{aligned}
\tag{9.1}
$$

Rewriting the bias as a function of G^*

For homogeneous clusters, the bias simplifies down to $-(\frac{1}{G})V$. We bound the bias by making a comparison to the homogeneous case where all of the n_g are the same: $n_g = N/G$. These terms in the bias are re-written:

$$\begin{aligned}\sum_g \frac{n_g\sigma_g^2}{N} &= \frac{1}{G}\sum_g \sigma_g^2 + \sum_g \sigma_g^2\left(\frac{n_g}{N} - \frac{1}{G}\right) \\ &= \frac{N^2}{G}V + \frac{1}{N}\sum_g \sigma_g^2(n_g - N/G) \end{aligned} \tag{9.2}$$

and

$$\sum_g \frac{n_g^2}{N^2} = \sum_g \left[\frac{n_g}{N} - \frac{1}{G}\right]^2 + \frac{1}{G}.$$

Plugging this into the last equation of (9.1),

$$\begin{aligned} b(\hat{V}) &= V\left(\sum_g \left[\frac{n_g}{N} - \frac{1}{G}\right]^2 + \frac{1}{G}\right) - \frac{2}{N^2}\left(\frac{N^2}{G}V + \frac{1}{N}\sum_g \sigma_g^2(n_g - N/G)\right) \\ &= V\left[\sum_g \left[\frac{n_g}{N} - \frac{1}{G}\right]^2 + \frac{1}{G} - \frac{2}{G}\right] - \frac{2}{N^2}\sum_g \sigma_g^2\left(\frac{n_g}{N} - \frac{1}{G}\right) \\ &= V\left[-\frac{1}{G} + \Gamma^2 - \frac{2}{N^2V}\sum_g \sigma_g^2\left(\frac{n_g}{N} - \frac{1}{G}\right)\right] \end{aligned} \tag{9.3}$$

where Γ^2 is the variation in cluster sizes:

$$\Gamma^2 = \sum_g \left[\frac{n_g}{N} - \frac{1}{G}\right]^2$$

Calculating a bound

Looking at the last term in the bias, we use the following bound:

$$\left[\frac{1}{G}\sum_g \sigma_g^2\left(\frac{n_g}{N}-\frac{1}{G}\right)\right]^2 \le \left[\frac{1}{G}\sum_g \sigma_g^4\right]\left[\frac{1}{G}\sum_g\left(\frac{n_g}{N}-\frac{1}{G}\right)^2\right] \quad (9.4)$$

The second factor here is known, and it will generally be smaller than the maximum. The first factor is similar to one of our effective number of clusters terms.

$$\frac{1}{G}\sum_g \sigma_g^4 = \frac{n^4}{G^2}V^2 + \frac{1}{G}\sum_g\left(\sigma_g^2 - \frac{n^2}{G}V\right)^2 \quad (9.5)$$

The fourth condition of the CMS paper (4.1.4) and plugging in a beta-binomial model with $\gamma = 1$ (for the largest possible variance) yields the bound

$$\sum_g\left(\frac{\sigma_g^2}{n^2V}-\frac{1}{G}\right)^2 \le \sum_g\left(\frac{n_g^2}{\sum n_g^2}-\frac{1}{G}\right)^2 \quad (9.6)$$

implying

$$\frac{1}{N^2V}\sum_g \sigma_g^2\left(\frac{n_g}{N}-\frac{1}{G}\right) \le \left[1+\frac{1}{G}\sum_g\left(\frac{n_g^2}{\sum n_g^2/G}-1\right)^2\right]^{1/2}\left[\frac{1}{G}\sum_g\left(\frac{n_g}{N}-\frac{1}{G}\right)^2\right]^{1/2}. \quad (9.7)$$

Therefore, the bound on the bias is

$$b(\hat{V}) \ge V\left[1-\frac{1}{G}-2\left[1+\frac{1}{G}\sum_g\left(\frac{n_g^2}{\sum n_g^2/G}-1\right)^2\right]^{1/2}\left[\frac{1}{G}\sum_g\left(\frac{n_g}{N}-\frac{1}{G}\right)^2\right]^{1/2} + \right.$$
$$\left. + \sum_g\left[\frac{n_g}{N}-\frac{1}{G}\right]^2\right] \quad (9.8)$$

Relating this to Γ and our effective number of clusters

$$\frac{G}{G^*} = 1 + \frac{1}{G}\sum_g \left(\frac{n_g^2}{\sum n_g^2/G} - 1\right)^2$$

we obtain the bias bound

$$b(\hat{V}) \geq V\left[1 - \frac{1}{G} - \frac{1}{G^*} + \left(\Gamma - G^{*\,-1/2}\right)^2\right]$$

and use it to obtain a bias-corrected variance estimate:

$$\hat{V}^* = \hat{V} * \frac{1}{b(\hat{V})}$$

9.2 Details on the MVCRVE

9.2.1 Dimensions for Multivariate Extension

Dimensions for mcrve

- $Y = \quad G \times L$
- $p = \quad 1 \times L$
- $T = \quad G \times 1$
- $A = \quad L \times L - 1$
- $\widehat{p} = \frac{1}{n}\mathbb{1}^\top \cdot Y = \quad 1 \times L$
- $C = (Y - T\widehat{p})^\top \quad = \quad L \times G$
- $\hat{\Sigma}_{\widehat{p}} = \frac{1}{n^2} CC^\top \quad = \quad L \times L$
- $\alpha = pA \quad = \quad 1 \times L - 1$
- $\hat{\alpha} = \widehat{p}A \quad = \quad 1 \times L - 1$
- $W = YA \quad = \quad G \times L - 1$
- $\hat{\Sigma}_{\hat{\alpha}} = \frac{1}{n^2}(W - T\hat{\alpha})^\top (W - T\hat{\alpha}) \quad = \quad L - 1 \times L - 1$
- $\hat{\alpha}\hat{\Sigma}_{\hat{\alpha}}^{-1}\hat{\alpha}^\top \quad = \quad 1 \times 1$

Dropping the subscript for the covariance matrix of the contrasted data, we aim to show consistency in the following way: Let $z = \Sigma^{-\frac{1}{2}}\hat{\alpha}^\top$. We know that $z^\top z = \hat{\alpha}\Sigma^{-1}\hat{\alpha}^\top \rightsquigarrow \chi^2_{L-1}$. Therefore,

$$\hat{\alpha}\hat{\Sigma}^{-1}\hat{\alpha}^\top = z^\top \Sigma^{\frac{1}{2}}\hat{\Sigma}^{-1}\Sigma^{\frac{1}{2}} z \tag{9.9}$$

and consistency of the variance covariance estimator is equivalent to

$$\|\Sigma^{\frac{1}{2}}\hat{\Sigma}^{-1}\Sigma^{\frac{1}{2}} - \mathbf{I}\| \to 0$$

One way to do this is to use the decomposition into two matrices, as in the univariate proof of the variance estimator:

The estimate $\hat{\Sigma}$ can be decomposed in the following way. Let $C^\top = Y - T\widehat{p}, \quad n_G^2 = \sum_{g=1}^{G} T_g^2$. Then $\hat{\Sigma} = CC^\top/n_G^2$, and

$$\begin{aligned} C^\top &= (Y - Tp) - (T\widehat{p} - Tp) \\ \Rightarrow \hat{\Sigma} &= \frac{1}{n^2}(Y - Tp)^\top(Y - Tp) + \frac{T^\top T}{n^2}(\widehat{p} - p)^\top(\widehat{p} - p) - \mathcal{R} \end{aligned} \tag{9.10}$$

Let

$$\tilde{\Sigma}_1 = \frac{1}{n^2}(Y - Tp)^\top(Y - Tp)$$

and

$$\tilde{\Sigma}_2 = \frac{T^\top T}{n^2}(\widehat{p} - p)^\top(\widehat{p} - p)$$

1. $\Sigma^{-\frac{1}{2}}\tilde{\Sigma}_1\Sigma^{-\frac{1}{2}} \xrightarrow{\mathbb{P}} \mathbf{I}$

   $$\mathbb{E}(\tilde{\Sigma}_1) = \Sigma,$$

   $$\mathrm{Var}(\tilde{\Sigma}_1) = \frac{1}{n^2}\mathrm{Var}((Y - Tp)^\top(Y - Tp))$$

   Therefore the convergence relies on the fourth moment of the Ys?

2. $\|\Sigma^{-\frac{1}{2}}\tilde{\Sigma}_2\Sigma^{-\frac{1}{2}}\| \xrightarrow{\mathbb{P}} 0$

3. 1. and 2. $\Rightarrow \tilde{\Sigma}_1 + \tilde{\Sigma}_2 \xrightarrow{\mathbb{P}} \Sigma$, provided that we have the multivariate equivalent of the following theorem:

   Lemma 9.2.1. *Assuming that X_{ni} and Y_{ni} are triangular arrays with $i = 1, \ldots, n$ and*

   $$\sum X_{ni}^2 \xrightarrow{\mathbb{P}} c; \qquad \sum Y_{ni}^2 \xrightarrow{\mathbb{P}} 0.$$

   Then

   $$\sum_{i=1}^{n} (X_{ni} + Y_{ni})^2 \xrightarrow{\mathbb{P}} c. \tag{9.11}$$

9.2.2 Geometric Series for Matrices for proof of chi-squared test

$$\mathbf{I} - \mathbf{A} = \mathbf{A} \sum_{k=1}^{\infty} (\mathbf{I} - \mathbf{A})^k$$

$$\implies \mathbf{I} - \mathbf{A}^{-1} = -\sum_{k=1}^{\infty} (\mathbf{I} - \mathbf{A})^k$$

This implies that

$$\left(\mathbf{I} - \Sigma^{1/2} \hat{\Sigma}^{-1} \Sigma^{1/2}\right) = -\sum_{k=1}^{\infty} \left(\mathbf{I} - \Sigma^{-1/2} \hat{\Sigma} \Sigma^{-1/2}\right)^k$$

A geometric sequence argument implies that as long as the first term converges to 0, the entire sequence will converge to 0. In other words,

$$\mathbf{z}^\mathsf{T} \left(\mathbf{I} - \Sigma^{-1/2} \hat{\Sigma} \Sigma^{-1/2}\right) \mathbf{z} \xrightarrow{\mathbb{P}} 0 \tag{9.12}$$

$$\implies \mathbf{z}^\mathsf{T} \sum_{k=1}^{\infty} \left(\mathbf{I} - \Sigma^{-1/2} \hat{\Sigma} \Sigma^{-1/2}\right)^k \mathbf{z} \xrightarrow{\mathbb{P}} 0. \tag{9.13}$$

The proof of convergence of the chi-square test therefore focuses on showing that (9.12) is true. This geometric argument allows us to bypass having to invert $\hat{\Sigma}$ by inverting the known covariance matrix Σ instead.

9.2.3 Bias of the covariance estimate

We calculate the expectation, where the sum is implicitly over g:

$$\sum \mathbb{E}(Y_{gi} - n_g\widehat{p}_i)(Y_{gj} - n_g\widehat{p}_j) = \sum \mathbb{E}(Y_{gi}Y_{gj}) - n_g[\mathbb{E}(\widehat{p}_iY_{gj}) + \mathbb{E}(\widehat{p}_jY_{gi})] + n_g^2\mathbb{E}(\widehat{p}_i\widehat{p}_j)$$
$$= \sum_{g=1}^{G} \gamma_g + \boxed{(\mathbb{E}(\widehat{p}_i\widehat{p}_j) - p_ip_j)\sum n_g^2 + 2p_ip_j\sum n_g^2 - \sum n_g[\mathbb{E}(\widehat{p}_iY_{gj}) + \mathbb{E}(\widehat{p}_jY_{gi})]}$$

The boxed quantity is the bias. It can be simplified. The first term becomes:

$$(\mathbb{E}(\widehat{p}_i\widehat{p}_j) - p_ip_j)\sum n_g^2 = \mathrm{Cov}(\widehat{p}_i, \widehat{p}_j)\sum n_g^2 = \sum \gamma_g \frac{\sum n_g^2}{N^2}$$

For the rest of the bias, note that

$$\mathrm{Cov}(\widehat{p}_i, Y_{gj}) = \mathrm{Cov}(\widehat{p}_j, Y_{gi}) = \frac{1}{N}\sum_{=1}^{G} \mathrm{Cov}(Y_{li}, Y_{gj}) = \frac{1}{N}\mathrm{Cov}(Y_{gi}, Y_{gj})$$

so that

$$\mathbb{E}(\widehat{p}_iY_{gj}) = \mathrm{Cov}(\widehat{p}_i, Y_{gj}) + n_gp_ip_j = \frac{1}{N}\mathrm{Cov}(Y_{gi}Y_{gj}) + n_gp_ip_j \quad (9.14)$$

The final term is symmetric with respect to i and j, so that $\mathbb{E}(\widehat{p}_iY_{gj})$ is the same as (9.14).

Plugging this into the rest of the bias term:

$$2p_ip_j\sum n_g^2 - \sum n_g[\mathbb{E}(\widehat{p}_iY_{gj}) + \mathbb{E}(\widehat{p}_jY_{gi})]$$
$$= 2p_ip_j\sum n_g^2 - 2\sum n_g[\frac{1}{N}\mathrm{Cov}(Y_{gi}Y_{gj}) + n_gp_ip_j]$$
$$= -\frac{2}{N}\sum n_g\gamma_g$$

Therefore, the bias of the covariance estimate $\hat{\gamma} = \frac{1}{N^2}\sum_{g=1}^{G}(Y_{gi} -$

$n_g\widehat{p}_i)(Y_{gj} - n_g\widehat{p}_j)$ is

$$\begin{aligned} b_{\mathrm{Cov}(\widehat{p}_i,\widehat{p}_j)} &= \frac{1}{N^2}\Big(\sum \mathbb{E}(Y_{gi} - n_g\widehat{p}_i)(Y_{gj} - n_g\widehat{p}_j) - \sum \gamma_g\Big) \\ &= \frac{\sum n_g^2}{N^2}\sum \gamma_g - \frac{2}{N}\sum n_g\gamma_g \\ &= \frac{\sum n_g^2}{N^2}\sum \gamma_g\Big(1 - 2N\frac{n_g}{\sum n_g^2}\Big) \end{aligned} \tag{9.15}$$

For homogeneous clusters, this simplifies down to a similar form

$$\begin{aligned} b_{\hat{\gamma}} &= \frac{\sum (N/G)^2}{N^2}\sum \gamma_g - \frac{2}{N}\sum (N/G)\gamma_g \\ &= \frac{-1}{G}\sum_{g=1}^{G}\gamma_g \end{aligned} \tag{9.16}$$

Therefore, $\hat{\gamma} = \frac{1}{N^2}\sum_{g=1}^{G}(Y_{gi} - n_g\widehat{p}_i)(Y_{gj} - n_g\widehat{p}_j)$ becomes an unbiased estimator with the simple multiplicative correction $\frac{G}{G-1}$:

$$\mathbb{E}((\frac{G}{G-1})\hat{\gamma}) = \frac{G}{G-1}\frac{1}{N^2}\Big(\sum_{g=1}^{G}\gamma_g - \frac{1}{G}\sum_{g=1}^{G}\hat{\gamma}_g\Big) = \gamma$$

9.3 Satterthwaite Approximation

(Satterthwaite, 1946b) proposed an approximation of a linear combination of χ^2 random variables.

$$U = \sum_{i=1}^{r}\frac{a_i U_i}{\nu_i}$$

where the U_i are independent χ^2 random variables with ν_i degrees of freedom.

The $\mathbb{E}(U) = \sum_i a_i$.

$$\mathrm{Var}(U) = \sum_i \frac{a_i^2\,\mathrm{Var}(U_i)}{\nu_i^2} = 2\sum_i \frac{a_i^2}{\nu_i}$$

We are hoping that U has approximately a Gamma distribution. We can choose α and β to match the first two moments of this distribution.

$$\begin{aligned}
\alpha_*\beta_* &= \sum_i a_i \\
\alpha_*\beta_*^2 &= \sum_i 2\frac{a_i^2}{\nu_i} \\
\implies \beta_* &= \frac{2\sum_i \frac{a_i^2}{\nu_i}}{\sum_i a_i} \\
\implies \alpha_* &= \frac{\left[\sum_i a_i\right]^2}{2\sum_i \frac{a_i^2}{\nu_i}}
\end{aligned}$$

where typically this is expressed as a re-scaled χ^2 with $2\alpha_*$ degrees of freedom.

9.3.1 Binomial Version

If we have G clusters, the variance estimator can be written as a linear function of terms like

$$U_g = \frac{(Y_g - n_g\hat{p})^2}{V_g} \approx \chi_1^2$$

where $V_g = \text{Var}(Y_g)$. Therefore, in the previous calculation, the $a_g = V_g/n^2$, and we're assuming that $\hat{p}$ is very close to p.

$$\begin{aligned}
\alpha_* &= \frac{1/n^4\left(\sum_g V_g\right)^2}{2\sum_g V_g^2/n^4} \\
&= \frac{G}{2}\left[1 + \frac{1}{G}\sum_{g=1}^{G}\frac{(V_g - \bar{V})^2}{\bar{V}^2}\right]^{-1}
\end{aligned}$$

where the average value of the variances is $\bar{V} = n^2V/G$. This matches with our *Effective Number of Clusters* calculations and indicates that we end up with a χ^2 distribution with approximately G^* degrees of freedom.

If we multiply our estimate by any arbitrary constant (possibly

to fix the bias), then the β changes but not the shape parameter α.

9.3.2 Two Sample Problem

Suppose that we have a control and a treatment group, we will presume that even if the parameter is the same in each model, it is likely that the variances will not be the same due to different cluster sizes.

The variance of our test statistic $\text{Var}(\hat{p}_1 - \hat{p}_0) = V_0 + V_1$. Thus, we need to compute the distribution of the sum of estimators. From the previous section, we argued that $\hat{V}_j$ is approximately a Gamma distributed random variable with shape $\alpha_* = G^*/2$ and $\beta_* = 2V_j/G^*$. The sum of two such random variables will still be approximately Gamma, but with

$$\alpha_{\text{combined}} = \frac{G_0^* G_1^* \left(V_0 + V_1\right)^2}{2V_0^2 G_1^* + 2V_1^2 G_0^*}.$$

This would imply that we can use our unbiased estimates of the variance $\hat{V}_0$ and $\hat{V}_1$ to suggest the degrees of freedom to use in our t critical values

$$\text{df} = \frac{G_0^* G_1^* \left(\hat{V}_0 + \hat{V}_1\right)^2}{\hat{V}_0^2 G_1^* + \hat{V}_1^2 G_0^*} = \left[\left(\frac{\hat{V}_0}{\hat{V}_0 + \hat{V}_1} \right)^2 \frac{1}{G_0^*} + \left(\frac{\hat{V}_1}{\hat{V}_0 + \hat{V}_1} \right)^2 \frac{1}{G_1^*} \right]^{-1}.$$

9.4 Simulation

To simulate clustered data, a beta-binomial model is used: For a fixed number of clusters G, means $p_1, \ldots, p_G$ are independently drawn from a known beta distribution. These values, which we refer to as cluster means, are then used to generate binomial random variables with sizes proportional to each cluster. As previously shown, consistency of the variance estimator depends on cluster homogeneity, which can be quantified by the effective number of clusters G^*. To show this relationship, 95% confidence intervals are estimated over a range of G^* rather than G. Values of G^* are generated by holding the total number of observations in the experiment fixed and gradually assigning more observations to one cluster, until it contains over 40% of the sample size.

In practice, the number of clusters is not always large enough to support asymptotic theory. We chose $G = 50$ with U.S. state data in mind. Assuming 50 observations per cluster in a homogeneous setting, we end up with $N = 2,500$ total binary observations for one population. We later look at two populations and keep the same G and N for each population, so that the total sample size for the two-sample problem is $N = 5,000$.

Coverage percentages are calculated using Z-estimators:

$$Z = \frac{\widehat{p} - p}{\sqrt{V}}$$

The unbiased estimate $\widehat{p} = \sum_{g=1}^{G} Y_g/N$ is used for each Z-score. These variance estimators are compared:

1. *Binomial Distribution* Assuming $i.i.d$ observations, $Y \sim Bin(p, N)$. The known variance estimate of a sample proportion for a binomial experiment is then:

$$\widehat{V}_{i.i.d} = \frac{1}{N}\sum_{g=1}^{G}\sum_{i=1}^{n_g}(Y_{ig} - \widehat{p})^2 = \frac{\widehat{p}(1-\widehat{p})}{N}$$

2. *Quasi-binomial Generalized Linear Model.* In cases of overdispersion, a quasi-binomial model can be used to model the extra-variation by specifying a mean-variance relationship. This is well documented in statistical literature (McCullach & Nelder, 1898)

$$\widehat{V}_{Quasi.Bin.} =$$

3. *Generalized Estimating Equation Sandwich Estimator* ((Liang & Zeger, 1986), (Carter et al., 2020))

$$\widehat{V}_{GEE} = \frac{1}{N^2}\sum_{g=1}^{G}(Y_g - n_g\widehat{p})^2$$

4. *Adjusted GEE Estimator.* $\widehat{V}_{GEE}$ is known to be biased (Rogers & Stoner, 2015; Mancl & A. DeRouen, 2001). The exact bias is calculated in section 3.2. We propose the following adjustment based on a multiplicative ℓ_2 lower bound:

$$\widehat{V}_{GEE^*} = (1/bias_{\ell_2}) * V_{GEE}, \qquad bias_{\ell_2} = 1 - \frac{G + G^*}{GG^*}$$

where G^* is the effective number of clusters derived by Carter and Steigerwald (2018).

In our simulation, we generate such Z-scores over varying values of G^* to study the performance of each variance estimator. High coverage percentages indicate that the variance is being overestimated, while under-coverage corresponds to underestimating the variance. The simulation process is now explained step by step for both the one-sample and two-sample problem.

9.4.1 One Proportion

1. Fix number of cluster G, and beta parameters p and γ.
2. Generate vectors of cluster sizes. Start with a leading cluster which makes up over 40% of all the observations. Observations are gradually distributed from the large cluster to all other equally sized clusters, increasing homogeneity until all clusters have the same size (in which case $G^* = G$). G^* is calculated at each point. The distribution of clusters is below. Since $G = 50$ in all simulations, there are 49 small clusters:

Table 9.1: Cluster Size Distribution and Effective Number of Clusters

	Small clusters	Large cluster	G^*
1	40.00	540.00	1.61
2	41.00	491.00	1.80
3	42.00	442.00	2.07
4	43.00	393.00	2.50
5	44.00	344.00	3.20
6	45.00	295.00	4.46
7	46.00	246.00	6.95
8	47.00	197.00	12.39
9	48.00	148.00	24.56
10	49.00	99.00	42.91
11	50.00	50.00	50.00

3. For each row of the table above, `nsim` simulations are run in the following way: for `j in 1 to nsim`:

   (a) Draw G cluster means from the beta distribution.

   $$p_1, \ldots, p_G \sim beta(p, \gamma) \qquad \mathbb{E}(p_g) = p, \qquad \text{Var}(p_g) = p(1-p)\gamma$$

   (b) Draw G binomial random variables with mean and size corresponding to each cluster size:

   $$Y_g | p_g \sim Bin(n_g, p_g)$$

   (c) Calculate $\widehat{p}$ and all variance estimates and corresponding Z-scores. Store Z-scores.

4. Count percentage of resulting Z-scores within (-1.96, 1.96). For one of our estimates, we use a t-distribution with G.star degrees of freedom: `(qt(.025, G.star[i]), qt(.975, G.star[i]))`.

$$\text{Estimated coverage percentage} = \frac{\#\{\texttt{Z-scores} \in (-1.96, 1.96)\}}{\texttt{nsim}}$$

5. Plot the resulting coverage percentages as a function of G^*.

9.4.2 Treatment & Control Problem

Coverage percentages are estimated across different levels of homogeneity, measured by G^*. We are interested in the performance of this estimator when the effective number of clusters varies greatly between the two groups. To that end, the control group is kept homogeneous, while the treatment group follows the same cluster size distribution as for the one-sample problem.

The simulation works as follows:

1. Fix number of clusters for each group G_i and beta distribution parameters p_i and γ_i, where $i = 0$ corresponds to the control group, and $i = 1$ to the treatment group. For the moment, we assume that $G_0 = G_1 = G$, and that the underlying distribution is the same for both control and treatment group, ie $p_0 = p_1, \quad \gamma_0 = \gamma_1$.

2. Cluster size distribution is homogeneous for the control group, with 50 observations per cluster. The treatment group starts with one very large cluster making up over 40% of all the observations and becomes gradually more evenly distributed, until it is homogeneous (in which case $G^* = G$). The effective number of cluster G^* is compared for the control and treatment group below. The size of the large cluster for the treatment group is also given.

Table 9.2: Effective Number of Clusters

Control	Treatment	Large Cluster Size (T)
50.00	1.61	540
50.00	1.80	491
50.00	2.07	442
50.00	2.50	393
50.00	3.20	344
50.00	4.46	295
50.00	6.95	246
50.00	12.39	197
50.00	24.56	148
50.00	42.91	99
50.00	50.00	50

3. For each row of the table above, `nsim` simulations are run in the following way: for `j in 1 to nsim`:

(a) Draw G_i cluster means from the beta distribution for each population $i = 0, 1$.

$$p_{1_i}, \ldots, p_{G_i} \sim beta(p_i, \gamma_i) \qquad \mathbb{E}(p_{g_i}) = p_i, \qquad \text{Var}(p_{g_i}) = p_i(1-p_i)\gamma_i$$

(b) Draw G_i binomial random variables with mean and size corresponding to each cluster size for both control (i=0) and treatment (i=1) groups:

$$Y_{g_i}|p_{g_i} \sim Bin(n_{g_i}, p_{g_i})$$

(c) Calculate $\widehat{p}_i$, all variance estimates separately in both populations, and corresponding Z-scores. The form of the proposed estimator is:

$$Z_j = \frac{\widehat{p}_0 - \widehat{p}_1}{\sqrt{\widehat{V}_0 + \widehat{V}_1}}$$

where $\widehat{V}_i$ is the bias-corrected GEE estimate for population i.

(d) Store Z-scores.

(e) **Wild Bootstrap**. Collect vector of deviations $r_{g_i} = Y_{g_i} - n_{g_i}\widehat{p}$.

i. Generate $K < 2^G$ bootstrap samples of $2G$ observations each by randomly adding or subtracting the residuals from the empirical pooled expectation :

$$\{\boldsymbol{Y}_i^*\}_r, \qquad Y_{g_i}^* = n_{g_i}\widehat{p} \quad \pm \quad r_{g_i}$$

ii. For each $k = 1, \ldots, K$, calculate $(\widehat{p}_1^* - \widehat{p}_0^*)_k$

iii. Calculate 2.5% and 97.5% quantiles of $\{(\widehat{p}_1^* - \widehat{p}_0^*)\}_{(K)}$

4. Count percentage of resulting Z-scores within (-1.96, 1.96). For one of our estimates, we use a t-distribution with G^* degrees of freedom.

$$\text{Estimated coverage percentage} = \frac{\#\{\texttt{Z-scores} \in (-1.96, 1.96)\}}{\texttt{nsim}}$$

5. Count percentage of bootstrap intervals that include 0 in the

interval.

6. Plot the resulting coverage percentages as a function of G_1^*.

9.4.3 Longitudinal Rasch Model

We generate person abilities according to a beta distribution and incorporate difficulty levels using the logistic function. Note that the logistic function returns values roughly in $(-4, 4)$ for realistic values of p, therefore item levels must have a limited range as to not overpower the ability levels.

1. Specify number of subjects (G), items (L), and a mean (p) and variance (γ) for the overall probability of success.

2. δ_i's: subject to constraint $\sum^{L} \delta_i = 0$. Chosen to be spaced symmetrically about zero with unit range. For example, if $L = 5$, $\delta = (-0.4, -0.2, 0, 0.2, 0.4)$. A negative δ corresponds to an item which is easier to endorse than the norm, whereas positive δ values imply more difficult questions.

3. Begin simulation. For each iteration, generate the following:

   - p_g's: Generate subject ability levels using beta distribution with mean and variance as previously specified. Create β_g's by taking the log-odds of the p_g's:

     $$\beta_g = \log\left(\frac{p_g}{1 - p_g}\right)$$

   - For longitudinal data, add time trend to subject parameters (as per the longitudinal dichotomous Rasch model) which is a strictly increasing sequence. Then multiply that sequence by -1, 0 or 1 randomly for each subject. Trend varies for each subject, not for each item
   - Systematic Component: Calculate $\mathbf{X}\theta$ with parameter vec-

tor $\theta = (\beta_1, \cdots, \beta_G, \delta_1, \cdots, \delta_L)$ and design matrix X:

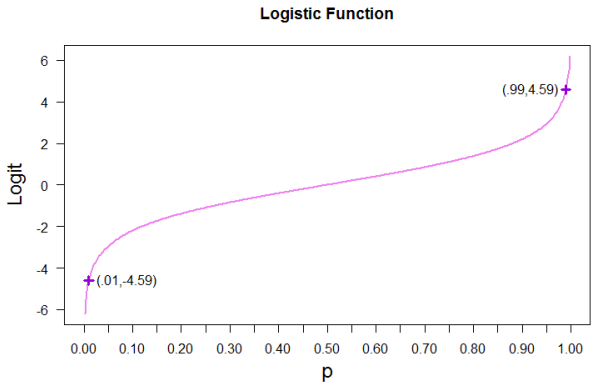
$$\mathbf{X} = \begin{pmatrix} 1 & 0 & \dots & 0 & -1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 & 0 & -1 & \dots & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \ddots & \vdots \\ 1 & 0 & \dots & 0 & 0 & \dots & \dots & -1 \\ 0 & 1 & \dots & 0 & -1 & 0 & \dots & 0 \\ \vdots & & & \vdots & 0 & \ddots & & \vdots \\ 0 & 1 & \dots & 0 & & \dots & \dots & -1 \\ \vdots & & & \vdots & \vdots & & & \vdots \\ 0 & \dots & 0 & 1 & 0 & \dots & 0 & -1 \end{pmatrix}$$

- Take the inverse log-odds to obtain a probability matrix $prob.mat = \frac{e^{\mathbf{X}\theta}}{1+e^{\mathbf{X}\theta}}$
- Y: Generate independent Bernoulli random variables using probability matrix.
- Estimate overall probability of success $\widehat{p}$ and its variance $\widehat{V}$ using cluster-robust estimator.
- Calculate Z statistic.

4. Store all Z-statistics. Calculate percentage which are within the 95% normal confidence interval $[-1.96, 1.96]$.
5. Repeat for different values of G and L.
6. Make a pretty plot

9.4.4 Discussing the *logit* link

As seen in the plot below, taking the log-odds of a percentage point (ie, a value between 0 and 1) returns values mostly within (-4, 4):



Function.png

In practice, it is very unlikely that we observe a value less than .01 or greater than .99. This roughly corresponds to values in the interval (-4.6, 4.6) for the logistic function. Therefore any value of the systematic component outside of this range will return probabilities of 0 or 1.
For the Rasch model, this implies that differences between subject ability and item difficulty should be within that interval. Hence in simulation, both sets of parameters should be calibrated in a way that probabilities very close to 0 or 1 are rare.

9.5 Computational Details

All calculations and simulations were done using R 3.6.0. Base functions were used to calculate the quasi-binomial estimate, as well as to generate time-series data. The following R packages were considered:

1. expm - facilitates matrix algebra
2. eRm - extended Rasch models
3. glmm- for Poisson and Binomial data, with normally distributed random effect and most common link functions used in research. We used the logit link.
4. TAM - Solves estimates for complex logistic structures. Rasch model included.

5. SimMultCor - A package for simulating correlated binary observations by specifying the marginal means and covariance structures.

After reviewing the statistical methods behind the various packages, the eRm package was selected because it was was of the only ones to use conditional maximum likelihood for item estimation in the Rasch model, meaning consistent estimates whose distributions are independent of the ability parameters.

9.5.1 Computing Methods for GLMMs

1. Newton-Raphson- root-finding algorithm which uses Taylor expansion to, in this case, find maximum likelihood estimates. We wish to solve

   $$\frac{\partial f(\theta)}{\theta} = 0,$$

   Expanding about an initial estimate θ_0,

   $$\frac{\partial f(\theta)}{\theta} = f'(\theta) \approx f'(\theta_0) + f''(\theta_0)(\theta - \theta_0)$$

   Set the previous equation to zero to find maximum values. Solving for θ, we get the Newton -Raphson algorithm:

   $$\theta^{(m+1)} = \theta^{(m)} - \frac{f'(\theta^{(m)})}{f''(\theta^{(m)})}$$

2. EM Algorithm-

   For mixed models, the random effect is often considered to be the missing data. Once estimates are given by EM algorithm, the values can be treated as known and fixed values which simplifies the problem.
   EM Iteration Steps for the Logit-Normal Rasch Model
   Recall:

   $$y_{ij}|\beta_j \overset{indep.}{\sim} Bernoulli(\pi_{ij} = \frac{1}{1 + e^{-\delta_i - \beta_j}})$$

   $$\mathbb{E}(y_{ij}|\beta_j) = \pi_{ij}$$

$$g(\pi_{ij}) = \ln \frac{\pi_{ij}}{1-\pi_{ij}} = \delta_i + \beta_j$$

$$\beta_j \sim N(0, \tau^2)$$

(a) Let the complete data be $\mathbf{w}' = (\mathbf{y}', \beta')$

(b) Set m=0. Choose starting values for $\beta^{(0)}$ and $\tau^{(0)}$.

(c) Calculate:

- $\beta^{(m+1)}$ and $\tau^{(m+1)}$ to maximize $\mathbb{E}[\ln f(y|\beta_j, \tau^2)]$

(d) (c) until convergence.

EM Iteration Steps for the Probit-Normal Model
Since this is a nested design, μ_1 and μ_2 are estimated separately, and the variances are not assumed to be equal. Focusing on the treatment group, let $Y = \sum_{j=1}^{n} = Y_{1j}$. The subscript i will be omitted from now on:

(a) Set $\mu^{(0)} = 0$, $\tau^{(0)} = 1$. Set m=0.

(b) E Step - Calculate $\mathbb{E}(Y|W, \mu^{(m)}, \tau^{(m)})$ and $\mathbb{E}(Y^2|W, \mu^{(m)}, \tau^{(m)})$

(c) M Step - Set

$$\mu^{(m+1)} = \mathbb{E}(Y|W, \mu^{(m)}, \tau^{(m)}) \qquad \tau^{(m+1)} = \mathbb{E}(Y^2|W, \mu^{(m)}, \tau^{(m)}) - (\mu^{(m)})^2$$

3. Gaussian Quadrature for GLMM - Hierarchical models with a random latent trait use marginal maximum likelihood to integrate out the possible values of that random effect. Most often the normal distribution is used, resulting in Gaussian expectations as likelihoods. Gauss-Hermite Quadrature is then used to evaluate the marginal likelihood. This method uses cleverly chosen weights w_k and corresponding evaluation points x_k from Hermite polynomials to approximate integrals with infinite bounds. The sum will be exact when the function whose expectation is being calculated (h) can be expressed as a polynomial of degree up to $2d-1$, where d is the number of evaluation points. For MML estimation, "practical experience shows that quadrature with less than 10 points often gives inaccurate answers, while 20 is usually enough for a good approximation"

(McCulloch, 1994; McCulloch et al., 2001). Integrals with respect to the normal density can be approximated as

$$\int_{-\infty}^{\infty} h(x)\frac{e^{-x^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}dx \doteq \sum_{k=1}^{d} h(\sqrt{2}\sigma x_k)w_k/\sqrt{\pi}. \tag{9.17}$$

9.5.2 Correlation bounds for binary random variables

These correlation bounds are a direct consequence of Frechet bounds and are respected by the beta-binomial distribution:

$$\max_{j\neq k}\left\{-\sqrt{\frac{p_j p_k}{q_j q_k}}, -\sqrt{\frac{q_j q_k}{p_j p_k}}\right\} \leq \rho \leq \min_{j\neq k}\left\{\sqrt{\frac{p_j q_k}{q_j p_k}}, \sqrt{\frac{q_j p_k}{p_j q_k}}\right\}$$

These are studied by (Chaganty & Joe, 2004) in the context of GEE's. They show that the correlation matrix derived by Liang and Zeger (Liang & Zeger, 1986) can disregard these bounds when the matrix is misspecified in current GEE software and give a set of simple rules for choosing a correlation matrix. Observations in a beta-binomial model also stay within these bounds since the covariance is always positive.

9.6 Miscellaneous

9.6.1 Useful Inequalities

- **Jensen** $E(g(X)) \geq g(E(X))$ for g convex; reverse if g is concave; both true when $\mathbb{E}|X|$ and $\mathbb{E}|g(X)|$ are finite.
- **Minkowski** for $p \geq 1$,
  $\{\mathbb{E}(|X+Y|^p)\}^{1/p} \leq \{\mathbb{E}(|X|^p)\}^{1/p} + \{\mathbb{E}(|Y|^p)\}^{1/p}$
- Minkowski, $p = 2 : \mathbb{E}(|X+Y|^2) \leq \left(\sqrt{\mathbb{E}(X^2)} + \sqrt{\mathbb{E}(Y^2)}\right)^2$
- **Holder** Let $p, q > 1$ and $p^{-1} + q^{-1} = 1$. Then
  $\mathbb{E}|XY| \leq \{\mathbb{E}(|X|^p)\}^{1/p}\{\mathbb{E}(|Y|^q)\}^{1/q}$. Set $p = q = 2$ to obtain CS:
- **Cauchy-Schwarz** $E(XY)^2 \leq E(X^2)E(Y^2)$

9.6.2 Exponential Parameterization of Outcome Vectors for Longitudinal Data

Suppose we have series of response vectors Y_j all of length L. Following the notation from (Liang & Zeger, 1986), suppose the number of successful answers to one item i for one subject

$$f(y_{it}) = \exp\left[\{y_{it}\theta_{it} - a(\theta_{it}) + b(\theta_{it})\}\phi\right] \tag{9.18}$$

9.6.3 Kronecker Products

The Kronecker product of two matrices $\boldsymbol{A} = \{a_{ij}\}$ and $\boldsymbol{B} = \{b_{ij}\}$ is

$$\boldsymbol{A} \otimes \boldsymbol{B} = \{a_{ij}\boldsymbol{B}\} \tag{9.19}$$

Examples for design matrices:

$$\mathbf{1}_2 \otimes \boldsymbol{I}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\boldsymbol{I}_2 \otimes \mathbf{1}_3 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

References

(n.d.).

Andersen, E. (1971, 01). Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society. Series B*, *32*. doi: 10.1111/j.2517-6161.1970.tb00842.x

Andersen, E. (1973, 01). Conditional inference and models for measuring.

Andrich, D. (1988). *Rasch models for measurement* (Vol. 40). doi: 10.2307/2348242

Anota, A., Barbieri, A., Savina, M., Pam, A., Gourgou, S., Bonnetain, F., & Bascoul-Mollevi, C. (2014, 12). Comparison of three longitudinal analysis models for the health-related quality of life in oncology: A simulation study. *Health and quality of life outcomes*, *12*, 1326. doi: 10.1186/s12955-014-0192-2

Barbieri, A., Anota, A., Conroy, T., Gourgou, S., Juzyna, B., Bonnetain, F., ... Bascoul-Mollevi, C. (2015, 12). Applying the longitudinal model from item response theory to assess the health-related quality of life in the prodige 4/accord 11 randomized trial. *Medical Decision Making*, *36*. doi: 10.1177/0272989X15621883

Barbieri, A., Peyhardi, J., Conroy, T., Gourgou, S., Lavergne, C., & Mollevi, C. (2017, 12). Item response models for the longitudinal analysis of health-related quality of life in cancer clinical trials. *BMC Medical Research Methodology*, *17*. doi: 10.1186/s12874-017-0410-9

Barbieri, A., Tami, M., Bry, X., Azria, D., Gourgou, S., Bascoul-Mollevi, C., & Lavergne, C. (2017, 12). Em algorithm estimation of a structural equation model for the longitudinal study of the quality of life. *Statistics in Medicine*, *37*. doi: 10.1002/sim.7557

Blanchin, M., Hardouin, J.-B., Le Neel, T., Kubis, G., Blanchard, C., Mirallié, E., & Sébille, V. (2011, 04). Comparison of ctt and rasch-based approaches for the analysis of longitudinal patient reported outcomes. *Statistics in medicine*, *30*, 825-38. doi: 10.1002/sim.4153

Borsboom, D. (2006, 09). The attack of the psychometricians. *Psychometrika*, *71*, 425-440. doi: 10.1007/s11336-006-1447-6

Cameron, C., & Miller, D. (2015, 01). A practitioner's guide to cluster-robust inference. *The Journal of Human Resources*, *50*, 317-372.

Carter, A. V., Marquis, S., & Steigerwald, D. (2020). Asymptotic normality of cluster robust variance estimators.

Carter, A. V., Schnepel, K. T., & Steigerwald, D. G. (2017). Asymptotic behavior of a t-test robust to cluster heterogeneity. *The Review of Economics and Statistics*, *99*(4), 698-709. Retrieved from https://doi.org/10.1162/RESTa00639 doi: 10.1162/REST\a\00639

Chaganty, N. R., & Joe, H. (2004). Efficiency of generalized estimating equations for binary responses. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, *66*(4), 851–860. Retrieved from https://doi.org/10.1111/j.1467-9868.2004.05741.x doi: 10.1111/j.1467-9868.2004.05741.x

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models. a generalized linear and nonlinear approach.* doi: 10.1007/978-1-4757-3990-9

Djogbenou, A., MacKinnon, J., & Nielsen, M. (2018). Asymptotic theory and wild bootstrap inference with clustered errors. *manuscript*, *Queens University*. Retrieved from https://doi.org/10.1198/000313006X152207 doi: 10.1198/000313006X152207

Farindon, P. (2007). Rasch models for longitudinal data. In *Multivariate and mixture distribution rasch models* (p. 191-199). Springer, New York, NY. doi: 10.1007/978-0-387-49839-312

Feddag, M.-L., Grama, I., & Mesbah, M. (2003, 01). Generalized estimating equations (gee) for mixed logistic models. *Communications in Statistics-theory and Methods - COMMUN STATIST-THEOR METHOD*, *32*. doi: 10.1081/STA-120018833

Fischer, G., & Molenaar, I. (1995). *Rasch models: Foundations, recent developments, and applications.* doi: 10.1007/978-1-4612-4230-7

Freedman, D. (2006, 02). On the so-called "huber-sandwich estimator" and "robust standard errors". *The American Statistician*, *60*, 299-302. doi: 10.1198/000313006X152207

Greene, W. (2002, 01). The bias of the fixed effects estimator in nonlinear models. *NYU Working Paper*, *EC-02-05*.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*.

Hansen, B., & Lee, S. (2018). Asymptotic theory for clustered samples. *manuscript*, *University of Wisconsin*.

Hubbard, A., Ahern, J., Fleischer, N., Laan, M., Lippman, S., Jewell, N., ... Satariano, W. (2010, 03). To gee or not to gee comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology (Cambridge, Mass.)*, *21*, 467-74. doi: 10.1097/EDE.0b013e3181caeb90

Jansen, M. G. (1997). Applications of rasch's poisson counts model to longitudinal count data. *Applications of latent trait and latent class models in the social sciences*, 380-388.

Kamata, A. (2001, 03). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*, 79 - 93. doi: 10.1111/j.1745-3984.2001.tb01117.x

Liang, K.-Y., & Zeger, S. L. (1986, 04). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13-22. doi: 10.1093/biomet/73.1.13

Linacre, J. (2004, 02). Rasch model estimation: Further topics. *Journal of applied measurement*, *5*, 95-110.

Lord, F., & Novick, M. (1968). *Statistical theories of mental test scores* (1st ed.).

Lord, F. M. (1952). The relation of test score to the trait underlying the test. *Educational Testing Service Research Bulletin Series*, *1952*(2), 517-549. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2333-8504.1952.tb00926.x doi: 10.1002/j.2333-8504.1952.tb00926.x

Magno, C. (2009, 06). Demonstrating the difference between classical test theory and item response theory using derived test data. *CSN: General Cognitive Social Science (Topic)*, *1*.

Mallinson, T. (2011, 01). Rasch analysis of repeated measures. *Rasch Measurement Transactions*, *25*.

Mancl, L., & A. DeRouen, T. (2001, 04). A covariance estimator for gee with improved small-sample properties. *Biometrics*, *57*, 126-34. doi: 10.1111/j.0006-341X.2001.00126.x

Maris, G., & Bechger, T. (2009, 05). On interpreting the model parameters for the three parameter logistic model. *Measurement Interdisciplinary Research and Perspectives*, *7*. doi: 10.1080/15366360903070385

Maul, A. (2017, 08). Rethinking traditional methods of survey validation. *Measurement: Interdisciplinary Research and Perspectives*, 1-19. doi: 10.1080/15366367.2017.1348108

McCullach, P., & Nelder, J. (1898). *Generalized linear models*. Chapman and Hall.

McCulloch, C. (1994, 03). Maximum likelihood variance components estimation for binary data. *Journal of The American Statistical Association - J AMER STATIST ASSN*, *89*, 330-335. doi: 10.1080/01621459.1994.10476474

McCulloch, C., Searle, S., & Neuhaus, J. (2001). *Generalized, linear, and mixed models*. doi: 10.1002/0471722073.scard

Olsbjerg, M., & Christensen, K. (2015, 12). Modeling local dependence in longitudinal irt models. *Behavior research methods*, *47*. doi: 10.3758/s13428-014-0553-0

Petrillo, J., Cano, S., McLeod, L., & Coon, C. (2015, 01). Using classical test theory, item response theory, and rasch measurement theory to evaluate patient-reported outcome measures: A comparison of worked examples. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research*, *18*, 25-34. doi: 10.1016/j.jval.2014.10.005

Rogers, P., & Stoner, J. (2015, 01). Modification of the sandwich estimator in generalized estimating equations with correlated binary outcomes in rare

event and small sample settings. *American journal of applied mathematics and statistics*, *3*, 243-251.

Satterthwaite, F. E. (1946a). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, *2*(6), 110–114. Retrieved from http://www.jstor.org/stable/3002019

Satterthwaite, F. E. (1946b). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, *2*(6), 110–114. Retrieved from http://www.jstor.org/stable/3002019

Sponsler, G. (1957, 04). First-order markov process representation of binary radar data sequences. *Information Theory, IRE Transactions on*, *3*, 56 - 64. doi: 10.1109/TIT.1957.1057395

Sébille, V., Hardouin, J.-B., Le Neel, T., Kubis, G., Boyer, F., Guillemin, F., & Falissard, B. (2010, 03). Methodological issues regarding power of classical test theory (ctt) and item response theory (irt)-based approaches for the comparison of patient-reported outcomes in two groups of patients - a simulation study. *BMC medical research methodology*, *10*, 24. doi: 10.1186/1471-2288-10-24

Willse, J. (2011, 02). Mixture rasch models with joint maximum likelihood estimation. *Educational and Psychological Measurement - EDUC PSYCHOL MEAS*, *71*, 5-19. doi: 10.1177/0013164410387335

Zimmerman, D. (1975, 02). Probability spaces, hilbert spaces, and the axioms of test theory. *Psychometrika*, *40*, 395-412. doi: 10.1007/BF02291765

Zwinderman, A. (1995, 12). Pairwise parameter estimation in rasch models. *Applied Psychological Measurement - APPL PSYCHOL MEAS*, *19*, 369-375. doi: 10.1177/014662169501900406