1-1-2014

Estimation and Inference of the Three-Level Intraclass Correlation Coefficient

Matthew David Davis
University of Pennsylvania, Mat.D.Davis@gmail.com

Estimation and Inference of the Three-Level Intraclass Correlation Coefficient

Abstract
Since the early 1900's, the intraclass correlation coefficient (ICC) has been used to quantify the level of agreement among different assessments on the same object. By comparing the level of variability that exists within subjects to the overall error, a measure of the agreement among the different assessments can be calculated. Historically, this has been performed using subject as the only random effect. However, there are many cases where other nested effects, such as site, should be controlled for when calculating the ICC to determine the chance corrected agreement adjusted for other nested factors. We will present a unified framework to estimate both the two-level and three-level ICC for both binomial and multinomial outcomes. In addition, the corresponding standard errors and confidence intervals for both ICC measurements will be displayed. Finally, an example of the effect that controlling for site can have on ICC measures will be presented for subjects nested within genotyping plates comparing genetically determined race to patient reported race.

In addition, when determining agreement on a multinomial response, the question of homogeneity of agreement of individual categories within the multinomial response is raised. One such scenario is the GO project at the University of Pennsylvania where subjects ages 8-21 were asked to rate a series of actors' faces as happy, sad, angry, fearful or neutral. Methods exist to quantify overall agreement among the five responses, but only if the ICCs for each item-wise response are homogeneous. We will present a method to determine homogeneity of ICCs of the item-wise responses across a multinomial outcome and provide simulation results that demonstrate strong control of the type I error rate. This method will subsequently be extended to verify the assumptions of homogeneity of ICCs in the multinomial nested-level model to determine if the overall nested-level ICC is sufficient to describe the nested-level agreement.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Epidemiology & Biostatistics

First Advisor
Warren B. Bilker

Second Advisor
J. Richard Landis

Keywords
Agreement, Correlation, Homogeneity, ICC, Nested-level, Reliability

ESTIMATION AND INFERENCE OF THE THREE-LEVEL INTRACLASS CORRELATION COEFFICIENT

Matthew Davis

A DISSERTATION

in

Epidemiology and Biostatistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2014

Supervisor of Dissertation	Co-Supervisor of Dissertation
Warren B. Bilker	J. Richard Landis
Professor of Biostatistics	Professor of Biostatistics

Graduate Group Chairperson

John H. Holmes, Associate Professor of Medical Informatics in Epidemiology

Dissertation Committee

Sharon Xiangwen Xie, Associate Professor of Biostatistics

Robert DeRubeis, Professor of Psychology

ESTIMATION AND INFERENCE OF THE THREE-LEVEL INTRACLASS CORRELATION COEFFICIENT

ACKNOWLEDGEMENT

First, I would like to thank God for providing me the strength, fortitude and perseverance to undertake this endeavor and see it to completion.

> Worthy are you, our Lord and God, to receive glory and honor and power,
> for you created all things, and by your will they existed and were created.
> -Revelation 4:11

I would like to thank Dr. Warren Bilker for his unending support of my dissertation. The time, energy and care Dr. Bilker put towards this dissertation went beyond what was necessary, and this work would not have been possible without his guidance. I would also like to thank Dr. J. Richard Landis for overseeing our work, from providing the original dissertation research question to providing careful input and oversight throughout the life of the project. I am grateful to my committee members, Dr. Sharon Xie and Dr. Robert DeRubeis, for providing helpful feedback on this research. In addition, I am grateful to Carla Hultman, Pat Spann, Marissa Fox and Catherine Vallejo for their invaluable organizational support. I would also like to thank Dr. Benjamin French for overseeing my master's thesis and for his guidance through that process. Finally, I would like to acknowledge all of the faculty and staff of the Department of Biostatistics and Epidemiology for their unending support of my work at the University of Pennsylvania.

Of course, none of this work would have been possible without the loving support of my wife, Stephanie. She is my rock and my foundation. She provided support to me through the most difficult times and rejoiced with me in the most exciting times, and I look forward to endeavoring with her on the rest of the adventures that life has to offer. I am grateful for my son, Mason, who has been a source of laughter and

love through this dissertation and understanding during those times he sacrificed so I could finish this work. I would like to thank my family for providing constant love and support, in addition to housing and food in the early years, as well as my friends for their love and understanding through this dissertation. I would like to thank my fellow students, particularly those in my cohort, for sharing knowledge, discussing statistical theory and finding ways to make it through difficult times. Lastly, I would like to thank my employer, Theorem Clinical Research, for the flexibility to attend classes and complete my dissertation, and specifically acknowledge Jeffrey Joseph for his mentorship and guidance both in statistics and my career.

ABSTRACT

ESTIMATION AND INFERENCE OF THE THREE-LEVEL INTRACLASS CORRELATION COEFFICIENT

Matthew Davis

Warren B. Bilker

J. Richard Landis

Since the early 1900s, the intraclass correlation coefficient (ICC) has been used to quantify the level of agreement among different assessments on the same object. By comparing the level of variability that exists within subjects to the overall error, a measure of the agreement among the different assessments can be calculated. Historically, this has been performed using subject as the only random effect. However, there are many cases where other nested effects, such as site, should be controlled for when calculating the ICC to determine the chance corrected agreement adjusted for other nested factors. We will present a unified framework to estimate both the two-level and three-level ICC for both binomial and multinomial outcomes. In addition, the corresponding standard errors and confidence intervals for both ICC measurements will be displayed. Finally, an example of the effect that controlling for site can have on ICC measures will be presented for subjects nested within genotyping plates comparing genetically determined race to patient reported race.

In addition, when determining agreement on a multinomial response, the question of homogeneity of agreement of individual categories within the multinomial response is raised. One such scenario is the GO project at the University of Pennsylvania where subjects ages 8–21 were asked to rate a series of actors' faces as happy, sad,

angry, fearful or neutral. Methods exist to quantify overall agreement among the five responses, but only if the ICCs for each item-wise response are homogeneous. We will present a method to determine homogeneity of ICCs of the item-wise responses across a multinomial outcome and provide simulation results that demonstrate strong control of the type I error rate. This method will subsequently be extended to verify the assumptions of homogeneity of ICCs in the multinomial nested-level model to determine if the overall nested-level ICC is sufficient to describe the nested-level agreement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

Introduction

"It is by universal misunderstanding that all agree. For if, by ill luck, people understood each other, they would never agree." Little did Charles Baudelaire know that his penned words 150 years prior would be a fitting description for statistical studies on methods of agreement. Given multiple ratings on the same object, it is a result of naivety that one would think that all raters would agree in their interpretation of the object. In addition, according to Mr. Baudelaire, even if the raters were lucky enough to fully understand one another's way of thinking, they still would not agree on the individual assessments on the objects. As a result, it is necessary to study statistical measures of agreement to better quantify how well independent raters agree when assessing the same object. This dissertation reviews the scope of available published work on measures of agreement and will add to these measures in two areas. First, a test for homogeneity of intraclass correlation coefficients (ICCs) will be derived across separate responses within a multinomial outcome. Second, the concept of a nested-level of agreement will be examined, and methods for estimating and providing inference on the nested-level agreement will be presented for both binary and multinomial outcomes.

1.1. Introduction to Measures of Agreement

A number of books have recently been written on measures of agreement that provide excellent summaries of the scope of literature published to date on the topic. *Measures of Interobserver Agreement and Reliability*[48] by Shoukri et. al. provides an overall summary of agreement methods for continuous scale measurement, population coef-

ficient of variation, dichotomous outcomes and multiple raters and categories. The summary of methods of agreement for kappa statistics and the intraclass correlation coefficients are of particular interest and provide an important summary of available methods that are directly applicable to this research. While this dissertation focuses mainly on the methods summarized by Shoukri, the following references are provided to more completely describe the current status of the methods of measures of agreement. *Analyzing Rater Agreement: Manifest Variable Methods*[49] by Von Eye and Mun provides a framework to assess rater agreement based on log-linear models. In *Statistical Tools for Measuring Rater Agreement*, Lin et. al.[36] examine methods of rater agreement using the concordance correlation coefficient (CCC) as a basis. In this book, agreement methods for both continuous and categorical data are developed and corresponding power and sample size methods are presented. Lastly, Broemeling provides a Bayesian description of measures of agreement in *Bayesian methods for measures of agreement*[7] focusing both categorical and continuous outcomes.

While this list of books is by no means exhaustive, it provides a good description of the current landscape of research as it relates to measures of agreement. This dissertation focuses solely on measures of agreement as it pertains to categorical outcomes, both for binary and multinomial responses. As a result, I will first outline the building blocks for agreement for categorical outcomes by describing both the kappa statistic and the intraclass correlation coefficient. I will then describe model-based assessments of the ICC using the beta-binomial and multinomial-Dirichlet distributions. Thirdly, I will describe a prior method to determine homogeneity of ICCs among categorical responses for a multinomial outcome, and will finally conclude with a description of the current work completed describing the analysis of nested-level agreement for binomial responses.

1.1.1. Kappa Statistic

The kappa statistic was originally proposed by Cohen (1960)[13] as a chance corrected measure of agreement between two raters and is calculated as

$$\hat{\kappa} = \frac{P_o - P_e}{1 - P_e} \tag{1.1}$$

where P_o is the observed proportion of agreement between the two raters and P_e is the expected measure of agreement by chance. $\hat{\kappa}$ has limits $[\frac{-P_e}{1-P_e}, 1]$ depending on the observed level of agreement. Regarding estimation of a standard error of the kappa statistic, Fleiss, Nee and Landis (1979)[23] wrote "Many human endeavors have been cursed with repeated failures before final success. The scaling of Mount Everest is one example. The discovery of the Northwest Passage is a second. The derivation of a correct standard error for kappa is a third!" A closed-form solution for the exact variance of $\hat{\kappa}$ has not yet been discovered, however an asymptotic variance can be found in Fleiss et. al. (1979)[23]. While $\hat{\kappa}$ is commonly used to quantify measures of agreement among raters, it is only applicable in situations where there are only two raters and a binomial response, necessitating further methods that can handle more diverse cases.

1.1.2. Weighted Kappa Statistic

The weighted kappa statistic was developed 8 years after the original kappa statistic by Cohen (1968)[14], which allows for a measure of agreement for a multinomial outcome based on a set of weights. For k possible outcomes, a $k \times k$ contingency table can be constructed for each possible combination of ratings for two ratings on the same object, and let i and j index the cell for responses i from rater 1 and j from rater 2 $(i, j = 1...k)$. Let v_{ij} be the weight associated with cell (i, j), p_{oij} be the

observed probability of response for cell (i, j) and p_{eij} be the expected probability of response for cell (i, j). Then the weighted kappa statistic can be calculated as

$$\kappa_w = 1 - \frac{\sum v_{ij} p_{oij}}{\sum v_{ij} p_{eij}} \tag{1.2}$$

The weighted kappa statistics allows for researchers to specify weights for the analysis giving stronger weights towards specific levels of agreement, allowing for customizable measures of agreement for a given response. Interestingly, using the weights $v_{ij} = (i-j)^2$, Fleiss and Cohen (1973)[21] proved that the resulting weighted kappa statistic is equivalent to the intraclass correlation coefficient, drawing a direct comparison between the two measures of agreement. Krippendorff (1970)[30] showed a similar result. The remainder of measures of agreement to be covered will focus on the ICC, however the concept of chance corrected agreement will be important to developing an adjusted nested-level ICC estimate.

1.1.3. Intraclass Correlation Coefficient

The intraclass correlation coefficient was first introduced by J. Arthur Harris in 1913 [25] as a measurement of agreement for multiple ratings on the same object. Since its inception, the volume of literature describing and implementing the ICC has grown exponentially. Originally intended for continuous outcomes, the ICC was expanded to describe rater agreement for categorical data as well by Landis et. al. (1977) [33, 34], Fleiss and Cohen (1973)[21] and Krippendorff (1970)[30]. In addition, at this time rules of thumb for interpretation of the ICC were given by Landis and Koch (1977a)[33] that assisted in quantifying the ICC.

The introduction of this method of interpretation provided a common benchmark for researchers to measure their level of agreement against and further promoted the use

Table 1.1: Interpretation of ICC Measures from Landis and Koch (1977a)

Kappa Statistic	Strength of Agreement
<0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost Perfect

of the ICC as a measure of agreement for categorical outcomes. By 1999, Ridout et. al. [44] documented and compared 20 distinct methods for estimating the ICC for binomial data. For the purposes of this research, we will focus on two methods of calculating and providing inference on the ICC, the components of variance model introduced by Landis and Koch (1977c) [34] and the beta-binomial estimate of the ICC introduced by Crowder (1978, 1979) [15, 16].

Components of Variance Model

ANOVA methods of determining the ICC were documented by numerous researchers including Anderson and Bancroft [1], Scheffé [45] and Searle [46], however flexible models to handle varying number of raters per object did not arise until Landis and Koch [34] presented the variance components approach for estimating the ICC. According to their approach, the binomial response y_{ij} for object i and rater j can be modeled by

$$y_{ij} = \mu + s_i + e_{ij} \tag{1.3}$$

where μ is the overall probability of response $y_{ij} = 1$, s_i are normally distributed errors with mean 0 and variance σ_s^2 and e_{ij} are normally distributed errors with mean 0 and variance σ_e^2. The overall variance in the model can be computed as $\sigma_s^2 + \sigma_e^2$ with

the between-subject variance categorized as σ_s^2. The corresponding ICC is calculated as

$$\rho = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2} \tag{1.4}$$

as the ratio of the variance attributed to the between-subject error and the total variance. Therefore, larger ICC values would indicate that the overall variability is dominated by the between-subject error and not the within-subject error attributed to multiple ratings on an object, indicating that the raters in the model exhibit a high level of agreement on the objects they are rating. Landis and Koch extended this model to account for multinomial data and provided an asymptotic standard error calculation that involved the use of complex matrix calculations as described in Koch et. al. (1977) [29]. Further improvements on the standard error calculations were made, such as the development of more computationally simple variance for the ICC published by Mak (1988)[39] that relies on fewer assumptions than the Landis and Koch calculations. However, as these improvements are not needed for the research presented in this dissertation, the methods will not be described in detail here and are summarized nicely by Shoukri [48].

Beta-Binomial Model

Crowder (1977)[15] proposed the beta-binomial distribution as an ANOVA method to model overdispersed binomial data. If y is distributed according to the beta-binomial distribution,

$$P(Y = y) = \binom{n}{y} B(y + \alpha, n - y + \beta)/B(\alpha, \beta) \tag{1.5}$$

where B(x) is the beta function of x, n is the number of trials in the sample, y is sum of the responses in the trial and α and β are the parameters of the model to be fit. Using the binomial distribution and letting $\pi = P(y = 1)$ for a one-sample trial,

$E(y)$=$n\pi$ and $Var(y)$=$n\pi(1-\pi)$. However, as these data are overdispersed, there is an additional overdispersion parameter added to the variance calculation describing the overdispersion such that $Var(y)$=$n\pi(1-\pi)(1-(n-1)\rho)$. In trial design, this overdispersion parameter is commonly known as a design effect, or DEFF. By setting the moments of the beta-binomial distribution and the overdispersed binomial together, it is shown that the ICC can be derived as a function of the parameters of the beta-binomial model $\rho = (\alpha + \beta + 1)^{-1}$.

Given the identities presented regarding π and ρ, the beta-binomial distribution can then be completely specified by the probability of response and the ICC as demonstrated by Crowder [16]. Therefore, the resultant likelihood can be maximized over π and ρ to obtain maximum likelihood (ML) estimates of the parameters. This is an important discovery as the use of an ML estimate for the ICC contains important properties. First, the resultant ML estimator for the ICC is a consistent estimate. Second, using the Fisher Information matrix, the second derivative of the log-likelihood can be used to derive an efficient estimate of the variance for the estimate of the ICC $\hat{\rho}$ using the methodology described by Casella and Berger [8]. In fact, the "asymptotically fully efficient" variance of this estimator was used by Ridout et. al. [44] as the reference by which the efficiency of other estimates of the ICC were measured against.

These desirable properties of the ML estimates of the ICC make the beta-binomial model an excellent choice to expand to attempt to estimate and provide inference on the nested-level ICC. In subsequent chapters, the beta-binomial distribution will be expanded to incorporate multiple levels of ICCs and will form the basis of further exploration into the nested-level ICC. However, the beta-binomial distribution is only flexible enough to model binomial data. In order to have a model-based ICC estimate for multinomial outcomes, the more flexible multinomial analog, the multinomial-

Dirichlet distribution, should be considered and expanded for exploration into the nested-level ICC for multinomial data.

Multinomial-Dirichlet Model

When collecting a response that has multiple outcomes, it is generally of interest to quantify the level of agreement among multiple raters on the multinomial response as a whole. However too often estimation of agreement on the multinomial response as a whole is sacrificed for assessing agreement on each item-wise response. For example, in asking a subject to quantify their race as White, Black, Hispanic or Other, researchers typically look at the level of agreement among raters on each item-wise response such as "White or non-White" or "Black or non-Black". Under certain conditions, such as non-homogeneity of ICCs among the item-wise responses, analyzing each item-wise response has its place, however analyzing data in this manner neglects the relationship that the item-wise responses have given they were all asked in the same question and are therefore correlated. To investigate this overall ICC among all responses, the multinomial-Dirichlet distribution (MDD) can be used to provide estimation and inference of the ICC for a multinomial outcome as proposed by Lui et. al. (1999)[37] and Chen et. al. [9, 10] to do so.

Let y_h be the total number of positive responses for category h of a multinomial response, and let the vector of all responses for a given set of multinomial data be $\mathbf{y}$ with n_h categories per response. In addition, let $\mathbf{Z} = (z_1, z_2 \dots z_{n_h})$ be the vector of parameters that describe the MDD. Then MDD can be written as

$$P(\mathbf{y} = \mathbf{y}|\mathbf{Z}) = \frac{N!}{\prod_{a=1}^{n_h} y_a!} \frac{\Gamma\left(\sum_{a=1}^{n_h} z_a\right)}{\Gamma\left(N + \sum_{a=1}^{n_h} z_a\right)} \prod_{a=1}^{n_h} \frac{\Gamma(y_a + z_a)}{\Gamma(z_a)} \tag{1.6}$$

Chen [10] and Lui et. al. [37] demonstrate that the MDD models the pooled subject-level correlation $\rho = \left(\sum_{a=1}^{n_h} z_a + 1\right)^{-1}$ and item-wise response rate $\pi_i = \frac{z_i}{\sum_{a=1}^{n_h} z_a}$ for category i.

Bartfay et. al. (2000) [2] demonstrated the effect of collapsing multinomial data when assessing agreement. They examined the use of the MDD in modeling the overall ICC for a multinomial response, determining that there is a significant gain in efficiency and reduction in the width of the confidence interval surrounding the overall ICC estimate when analyzing agreement for the multinomial response and opposed to each item-wise response. This gain in efficiency is due to accounting for the non-independent relationships of each item-wise response when assessing agreement. This is of particular interest and appropriate where the ICCs are equivalent for each item-wise response as there is no difference in the measures of agreement, and the pooled ICC is sufficient to model the level of agreement for any individual response. Therefore, attempts should be made where appropriate to quantify the measure of agreement among multinomial responses as overall pooled ICCs. However, in situations where these assumptions are violated and there is heterogeneity among item-wise responses, the pooled ICC should not be considered adequate to model the data and the loss of efficiency from analyzing each item-wise response should be considered an acceptable trade-off for the flexibility to individually model each item-wise response. There has been some work done on determining whether homogeneity of item-wise ICCs exists for a multinomial response. These existing methods will be summarized (see Homogeneity of ICCs below) and extended (see Chapter 2).

Nested-Level ICC

Generally when studying measures of agreement among raters, the only factors taken into account are objects being rated, the rating scale and the raters themselves. Sit-

uations may arise, however, where there are additional factors that need to be taken into account. Landis et. al. (2011)[31] provide one such example. Westlund and Kurland (1953)[50] published the results of a study where two independent neurologists from Winnipeg and New Orleans classified subjects from their own patients, then each other's patients, on the certainty of Multiple Sclerosis (MS) diagnosis using the following ordinal measurement: (1) Certain MS, (2) Probable MS, (3) Possible MS, (4) Doubtful, Unlikely or Definitely Not MS. These results had been previously analyzed for rater agreement by Landis et. al. (1977a)[33], however were reanalyzed in Landis et. al. (2011)[31] to determine if there is a nested-level factor that helps explain part of the measures of agreement. In this study, the patients at each site are the objects being rated. Each patient is nested within one and only one site. Consider the level of agreement that exists within a nested-level by combining all ratings for all patients within a nested level and assessing the level of agreement for all ratings combined. At first this may seem to be a futile exercise as reasonable individuals may not expect there to be any agreement among seemingly independent subjects within a site. However, this may not be the case. Consider the extreme example where all certain MS subjects were located in Winnipeg and all doubtful MS subjects were located in New Orleans. In that case, the measure of agreement within each nested-level may actually be significant and of interest, causing researchers to wonder whether the observed agreement among raters in this case is valid or simply due to the clustering of patients within the sites.

Landis et. al. (2013)[31] set out to answer the question of how to quantify this nested level of agreement using the random effects model as a framework, formulating estimates for the object-level and nested-level ICCs using the variance components from a three-level random effects model. They continue the research to lay out a framework for an estimate of the variance of the nested-level ICC using the delta method and

the variance/covariance matrix of the mean square error estimates, however could not provide a closed form solution of the variance/covariance estimates and therefore could not specify the variance estimate for the nested-level ICC. In addition, in the presence of nested-level agreement, the corresponding object-level agreement could potentially be inflated, and more investigation should be conducted on the effect this could have on apparent object-level agreement.

1.1.4. Homogeneity of ICCs

Measuring agreement on a multinomial response requires more work and more assumptions than assessing agreement on the binomial counterpart. First, there is only one ICC associated with a binary outcome, whereas there are k potential ICCs for a multinomial response with k outcomes that could be derived by dichotomizing each of the item-wise assessments of the multinomial response. Second, in order to accurately describe the measure of agreement on the multinomial response as a whole, it is helpful (yet not necessarily imperative) that the item-wise agreement measures for each dichotomized response are equivalent. It may ofter occur that raters agree more strongly on certain items than they do on others.

Landis et. al. (1977c)[34] provide an example of estimating ICCs for psychiatric diagnoses for six raters on one of five response categories: depression, personality disorder, schizophrenia, neurosis or other diagnosis. The range of resultant ICCs was 0.254–0.575 with raters most often agreeing on "other diagnosis" and least often agreeing on "depression" or "personality disorder". An overall ICC was provided of 0.440, describing the overall agreement across all diagnoses, however it is difficult to determine if this is an accurate summary of agreement across all responses, or if the summary of agreement for each item-wise response would best describe the data. Therefore an investigation into the homogeneity of item-wise ICCs would be prudent

to assess whether the overall measure of agreement is the best fit for the data.

The question of whether to summarize agreement on the response as a whole or by each item-wise response is important. Bartfay et. al. [2] demonstrated the gain in efficiency that can occur by combining responses where appropriate. However, in order to combine responses (without any a priori hypotheses regarding overall agreement), homogeneity of item-wise ICCs should be demonstrated, otherwise important differences in agreement on item-wise responses could be lost.

Chen et. al. [9, 10] described the assessment of overall agreement for a trinomial response using the trinomial-Dirichlet distribution, but also provided a framework to determine whether homogeneity of ICCs exists across the item-wise responses. Chen then introduced the double beta-binomial model which is a distribution comprised of the product of two beta-binomial distributions. Under the condition of homogeneity of item-wise ICCs, Chen showed that the double beta-binomial distribution devolves into the trinomial-Dirichlet distribution. As the two distributions are nested, this allows for the use of the likelihood-ratio test to test whether the assumption of homogeneity of item-wise ICCs is valid.

These methods fall short in being able to be widely applied in two areas. First, if there is not homogeneity of ICCs among the item-wise responses, there are 3 potential expressions of the double beta-binomial distribution as the particular expression relies on a specific breakdown of conditional beta-binomial distributions. Chen (1991) [9] analyzes all three distributions and concludes that all three test statistics are greater than the upper 1st percentile of the appropriate chi-square distribution and can therefore reasonably conclude heterogeneity of ICCs. However, no mention is made of how to appropriately control the corresponding type I error rate, and the use of this methodology without such control will lead to unacceptable inflation in the overall type I error. Second, while Chen mentions that these methods can be extended

to the quadrinomial case, there is no explicit mention of the form such a distribution would take nor the proof that the dirichlet-Multinomial distribution would be similarly nested within the quadrinomial-Dirichlet distribution. Therefore, the method should be extended to the more general case where the multinomial-Dirichlet distribution is nested within a multiple beta-binomial distribution to test for homogeneity of item-wise ICCs, and more consideration should be given to the process and proof of control of the type I error rate with simulations to support such findings.

1.2. Estimation and Inference of the Three-Level Intraclass Correlation Coefficient

Chapter 2 of this dissertation will present a test of homogeneity of item-wise intraclass correlation coefficients for multinomial data. First, the methods originally derived by Chen et. al. [9, 10] will be presented for the trinomial case and extended to any number of responses. Second, recommendations for controlling the overall type I error rate will be presented and simulations provided to show the strong control of the type I error rate when testing for homogeneity of ICCs for multinomial data. Finally, the test will be applied to two separate studies concerning cervical cancer diagnoses and facial recognition to assess whether homogeneity of ICCs exist in either case.

Chapter 3 will provide a framework based on the beta-binomial distribution that allows for estimation and inference on the nested-level ICC for binary responses. A likelihood framework will be developed based on estimates of the probability of positive response and object-level ICCs. Using maximum likelihood techniques, a formula for the variance of the nested-level ICC along with a corresponding confidence interval will be presented. Then, a nested-level adjusted object-level ICC will be derived that provides a measure of agreement adjusted for the nested-level

agreement. A simulation study will then be performed to demonstrate the bias of the nested-level ICC estimate and corresponding coverage of the confidence interval. Finally, we illustrate the impact of the differential prevalence of the response attribute across object-level clusters on estimates of nested-level agreement by examining agreement between self-reported race/ethnicity of 3,546 study participants and genetically-inferred race/ethnicity assessed across 47 genotyping plates within a GWAS.

Chapter 4 will combine the results from chapters 2 and 3 and discuss a method to derive the nested-level ICC for multinomial data. The multinomial-Dirichlet distribution will be modified, similarly to the beta-binomial distribution as shown in Chapter 3, to account for nested-level data, however this method is proved to be valid in one of two ways. First, this method can be used if there is demonstrated homogeneity of object and nested-level ICCs, and therefore the methods derived in chapter 2 will be used to test homogeneity among object-level ICCs and extended to test homogeneity of nested-level ICCs. Second, the method is found to be valid when there are a large number of objects on average per nested-level. The methods presented in this chapter are applied to the study examining agreement between self-reported race/ethnicity across plates within a GWAS presented in Chapter 3, providing an overall estimate of the nested-level ICC and testing whether homogeneity of ICCs exists across each item-wise response.

CHAPTER 2

A Test of Homogeneity of Dependent Intraclass Correlation Coefficients for Multinomial Data

2.1. Introduction

Whether considering if a second opinion is needed or looking at the reliability of a result, the question of agreement among multiple ratings on the same object has attracted interest since J. Arthur Harris' seminal paper on the intraclass correlation coefficient (ICC) in 1913 [25]. Most often, the discussion centers around results that have continuous outcomes to ensure continuity across multiple ratings. However, in the biological and clinical setting, the categorical outcome is often of more interest than the continuous outcome. While methods such as the ANOVA based intraclass correlation coefficient and the concordance correlation coefficient have spanned the chasm between continuous and categorical outcomes when answering the questions of rater agreement, to truly understand the levels of agreement in the categorical setting, a qualitative-specific framework is needed.

The question of the agreement among multiple raters on a binomial outcome has been well-documented, starting with Cohen's kappa statistic [13] and branching out to a number of methods, many of which are summarized and critiqued by Ridout et. al. [44]. There has been a larger focus on analyzing the agreement among raters on binary outcomes at the expense of developing more robust theory on analyzing multinomial outcomes. Some methods exist that are appropriate to assess agreement for multinomial data. The ANOVA method proposed by Landis and Koch [34] is easily extendable to multinomial data. Fleiss and Cohen both suggested kappa statistics

appropriate for multinomial data [14, 22]. The concordance correlation coefficient has also been extended to multinomial data [28]. There are other methods that have focused on this area, but the development of likelihood-based methods is of particular interest due to the desirable properties of maximum likelihood estimators.

The multinomial-Dirichlet distribution has classically been used to model overdispersed multinomial data, and estimates from this model can be obtained to make inference on the corresponding intraclass correlation coefficient, assuming that the ICC is constant across each response. Chen et. al. [9] presented the Dirichlet-trinomial model to make inference on the proportion and ICC of a three-level multinomial outcome, as well as the more flexible double beta-binomial model. In their work, the double-beta binomial model was used to assess the goodness of fit regarding the trinomial-Dirichlet model to determine if the assumption of heterogeneity of ICCs across responses was valid. Furthermore, Bartfay and Donner described the gain in efficiency when using all possible outcomes to make inference on a homogeneous ICC for multinomial outcomes compared to modeling and making inference on the level of agreement of each outcome separately [2]. However in both cases, the conversation is mostly restricted to the three-outcome case. In addition, the goodness-of-fit test presented by Chen [9] contains a potential flaw due to the fact that given the same set of data, there are three separate expressions for the double beta-binomial distribution that could likely lead to different conclusions for the goodness-of-fit test depending on the decomposition. This paper will provide a generalization of the double beta-binomial distribution to the multiple beta-binomial distribution and demonstrate how the goodness of fit test for the homogeneity of ICCs across all possible responses originally presented by Chen [9] can be extended to multinomial data with any number of outcomes. In addition, particular attention will be paid to the question of how to handle various decompositions of the multiple beta-binomial distribution when test-

ing against the multinomial-Dirichlet distribution. Simulation studies are presented to demonstrate the control of the goodness of fit test over the type I error rate and power under various assumptions. Finally, two examples are provided on assessing homogeneity of ICCs, and recommendations are provided on how to analyze the data if the assumption of the homogeneity of ICCs across responses is violated.

2.2. Notation and Motivation

2.2.1. Notation

Let y_{hij} be a binary outcome (0 or 1) for the j^{th} rater ($j = 1, ..., n_{i\cdot}$) on the i^{th} object ($i = 1, ..., n_{\cdot\cdot}$) for the h^{th} response ($h = 1, ..., n_h$) where n_h is the number of outcomes of the response of interest, and let $\mathbf{y}$ be the vector of all responses. Let $x_{hi} = \sum_{j=1}^{n_{i\cdot}} y_{hij}$ be the total number of positive responses for object i for response h, let $\mathbf{x}$ be the vector of all such responses and let $\mathbf{x_i}$ be the vector of responses for object i. Let π_h be the proportion of objects with the trait being assessed such that $P(y_{hij} = 1) = \pi_h$. y_{hij} is assumed to follow a multinomial distribution where $E(y_{hij}) = \pi_h$ and $Var(y_{hij}) = \pi_h(1 - \pi_h)$. Let ρ_h be the object-level intraclass correlation coefficient for response h and let $\rho_{\cdot}$ be the overall object-level intraclass correlation.

2.3. Distributions for Overdispersed Multinomial Data

2.3.1. Beta-Binomial Distribution

The beta-binomial distribution can be specified as $\binom{n}{k} B(k + \alpha, n - k + \beta)/B(\alpha, \beta)$ where $B(x)$ is the beta function of x, n is the number of ratings in the sample, k is number of positive responses in the trial and α and β are the parameters of the model to be fit. If $y \sim$ Beta-binomial (α, β), then E(y) $= n\alpha/(\alpha + \beta)$

and $\text{Var(y)} = [n\alpha\beta(\alpha+\beta+n)] / \left[(\alpha+\beta)^2(\alpha+\beta+1)\right]$. As has been previously demonstrated, $E(x_{hi})=n_{i\cdot}\pi_h$ and $Var(x_{hi}) = n_{i\cdot}\pi_h(1-\pi_h)(1+(n_{i\cdot}-1)\rho_h)$, which means that $\pi_h=\alpha/(\alpha+\beta)$ and $\rho_h = (\alpha+\beta+1)^{-1}$. This leads to the solution $\rho_h = \pi_h/(\pi_h+\alpha)$, implying $\alpha = \pi_h(1-\rho_h)/\rho_h$ and $\beta = (1-\pi_h)(1-\rho_h)/\rho_h$. The details of the maximum likelihood estimates of these parameters have been documented elsewhere and will not be discussed further [15, 16, 39].

2.3.2. Dirichlet-Trinomial and Double Beta-Binomial Models

Chen et. al. [9] developed the Dirichlet-trinomial model to model a trinomial outcome with an overdispersion of variance. Specifically in his example, Chen modeled observations with three potential outcomes: x_{ij}, y_{ij} and z_{ij}. Define $n_{ij} = x_{ij}+y_{ij}+z_{ij}$. Then, the Dirichlet-trinomial can be defined as a Dirichlet-multinomial model with only three outcomes:

$$P(x_{ij}, y_{ij}, z_{ij}) = \frac{n_{ij}!\Gamma(\alpha_i+\beta_i+\gamma_i)\Gamma(x_{ij}+\alpha_i)\Gamma(y_{ij}+\beta_i)\Gamma(z_{ij}+\gamma_{ij})}{x_{ij}!y_{ij}!z_{ij}!\Gamma(n_{ij}+\alpha_i+\beta_i+\gamma_i)\Gamma(\alpha_i)\Gamma(\beta_i)\Gamma(\gamma_i)} \tag{2.1}$$

However, this distribution assumes that the ICCs among different responses are equivalent. Therefore, Chen broadened the distribution using the double beta-binomial distribution, which is a joint distribution for the responses that allow for separate ICCs for each response category. The double beta-binomial model can be written as the product of two conditional beta-binomial distributions:

$$\begin{aligned} P(x_{ij}, y_{ij}, z_{ij}) = & \frac{n_{ij}!\Gamma(\alpha_i+\beta_i)\Gamma(x_{ij}+\alpha_i)\Gamma(n_{ij}-x_{ij}+\beta_i)}{x_{ij}!(n_{ij}-x_{ij})!\Gamma(n_{ij}+\alpha_i+\beta_i)\Gamma(\alpha_i)\Gamma(\beta_i)} \\ & \times \frac{(n_{ij}-x_{ij})!\Gamma(\gamma_i+\delta_i)\Gamma(\gamma_i+y_{ij})\Gamma(n_{ij}-x_{ij}-y_{ij}+\delta_i)}{y_{ij}!(n_{ij}-x_{ij}-y_{ij})!\Gamma(n_{ij}-x_{ij}+\gamma_i+\delta_i)\Gamma(\gamma_i)\Gamma(\delta_i)} \end{aligned} \tag{2.2}$$

Chen determined that the Dirichlet-trinomial model is a special case of the double

beta-binomial model, implying that the likelihood-ratio test can be used to test the homogeneity of ICCs across responses within an object.

2.3.3. Multinomial-Dirichlet Distribution

For a given set of multinomial data $\mathbf{y}$ with n_h categories per response, a Dirichlet distribution can be assumed as the prior distribution for the probability of response for each category and a multinomial likelihood for the response vector. By invoking Bayes' rule, one obtains the multinomial-Dirichlet distribution (MDD). Let $\mathbf{M} = (m_1, m_2 \dots m_k)$ be the vector of parameters that describe the MDD. Then the MDD can be written as

$$P\left(\mathbf{X_i} = \mathbf{x_i} | \mathbf{M}\right) = \frac{N!}{\prod_{f=1}^{n_h} x_{fi}!} \frac{\Gamma\left(\sum_{f=1}^{n_h} m_f\right)}{\Gamma\left(N + \sum_{f=1}^{n_h} m_f\right)} \prod_{f=1}^{n_h} \frac{\Gamma\left(x_{fi} + m_f\right)}{\Gamma\left(m_f\right)} \tag{2.3}$$

Using the fact that the MDD models the overall ICC, $\rho_{\cdot} = \left(\sum_{f=1}^{n_h} m_f + 1\right)^{-1}$, and probability of response, $\pi_h = m_h / \left(\sum_{f=1}^{n_h} m_f\right)$ [10], it can be shown that $m_h = \pi_h \left(1 - \rho_{\cdot}\right) / \rho_{\cdot} \; \forall h$. The likelihood for $n_{\cdot\cdot}$ observations can be written as

$$\begin{aligned} L\left(\mathbf{M} | \mathbf{X} = \mathbf{x}\right) = & \prod_{a=i}^{n_{\cdot\cdot}} \left[\frac{n_{a\cdot}!}{\prod_{q=1}^{n_h} x_{qa}!} \left[\prod_{d=1}^{n_{a\cdot}} \left(d + \frac{1 - \rho_{\cdot}}{\rho_{\cdot}} - 1 \right) \right]^{-1} \right. \\ & \left. \times \left[\prod_{b=1}^{n_h} \prod_{c=1}^{x_{ab}} \left(c + \frac{1 - \rho_{\cdot}}{\rho_{\cdot}} \pi_b - 1 \right) \right] \right] \end{aligned} \tag{2.4}$$

This likelihood can be directly maximized to obtain MLE's of each π_i and $\rho_{\cdot}$. In addition, the standard error of each can be found by inverting the negative of the information matrix appropriately for each parameter, the details of which can be found elsewhere [42].

2.3.4. Multiple Beta-Binomial Distribution

It is well known that a common distribution to describe the presence of overdispersed binomial responses is the beta-binomial distribution [16, 44]. When a response vector has more than two responses, to maintain the concept of overdispersion, one can use the multinomial-Dirichlet distribution to capture the overdispersion [9]. However, this distribution makes the strong assumption that $\rho_h = \rho . \forall h$, which is unreasonable in many situations as raters on the same object may agree for certain responses and disagree for others. Therefore, other considerations need to be made to allow for the flexibility of separate ρ_h responses for different categories.

Originally, when considering the multinomial-Dirichlet distribution, the joint distribution of responses for object i, $P(\mathbf{X_i} = \mathbf{x_i}|\mathbf{M})$, is modeled. However, using the definition of conditional probability, this probability can be rewritten as

$$\begin{aligned} P(\mathbf{X_i} = \mathbf{x_i}|\mathbf{M}) &= P(X_{2i} = x_{2i}, \ldots X_{hi} = x_{hi}|\mathbf{M}, X_{1i} = x_{1i})\, P(X_{1i} = x_{1i}|\mathbf{M}) \\ &= P(X_{3i} = x_{3i}, \ldots X_{hi} = x_{hi}|\mathbf{M}, X_{1i} = x_{1i}, X_{2i} = x_{2i}) \times \\ &\quad P(X_{2i} = x_{2i}|\mathbf{M}, X_{1i} = x_{1i})\, P(X_{1i} = x_{1i}|\mathbf{M}) \ldots \end{aligned}$$

Therefore, the probability model can be written as a product of successive conditional beta-binomial distributions. In order to make the model more flexible, the restriction based on the parameters of the MDD can be removed and each conditional beta-binomial distribution can be modeled with its own set of parameters α and β. Define $\sum_{i=m}^{n} z_i = 0$ where $n < m$ and let $\mathbf{A} = (\alpha_1, \alpha_2, ...\alpha_{n_h - 1})$ and $\mathbf{B} = (\beta_1, \beta_2, ...\beta_{n_h - 1})$ be the vectors of parameters that describe each conditional beta-binomial distribution. Then, the multiple beta-binomial distribution (MBBD) for the vector of responses

for object i can be written as

$$P\left(\mathbf{X_i} = \mathbf{x_i}|\mathbf{A},\mathbf{B}\right) = \prod_{f=1}^{n_h-1} \binom{N - \sum_{g=1}^{f-1} x_{gi}}{x_{fi}} \times \frac{\Gamma\left(x_{fi}+\alpha_f\right)\Gamma\left(N - \sum_{g=1}^{f} x_{gi} + \beta_f\right)\Gamma\left(\alpha_f+\beta_f\right)}{\Gamma\left(N - \sum_{g=1}^{f-1} x_{gi} + \alpha_f + \beta_f\right)\Gamma\left(\alpha_f\right)\Gamma\left(\beta_f\right)} \tag{2.5}$$

Unlike the standard beta-binomial distribution, however, the conditional beta-binomial distribution does not have a direct parametrization that links it to the unconditional probability of response and corresponding ICC. Instead, each conditional beta-binomial distribution models the response probability and level of agreement given the predecessors it is conditional upon have already occurred. Assuming that conditioning occurs in order of response such that

$$\begin{aligned} P\left(\mathbf{X_i} = \mathbf{x_i}|\mathbf{A},\mathbf{B}\right) =& P(X_{1i} = x_{1i}|\mathbf{A},\mathbf{B})P(X_{2i} = x_{2i}...X_{n_hi} = x_{n_hi}|\mathbf{A},\mathbf{B},X_{1i} = x_{1i}) \\ =& P(X_{1i} = x_{1i}|\mathbf{A},\mathbf{B})(X_{2i} = x_{2i}|\mathbf{A},\mathbf{B},X_{1i} = x_{1i})... \\ & P(X_{n_hi} = x_{n_hi}|\mathbf{A},\mathbf{B},X_{1i} = x_{1i}...X_{(n_h-1)i} = x_{(n_h-1)i}) \end{aligned}$$

then the conditional beta-binomial distribution can instead be written in terms of the probability of response h, $\pi_{h|1,2,...,h-1}$, and conditional level of agreement of response

h, $\rho_{h|1,2,...h-1}$, conditional on all previous responses.

$$P\left(\mathbf{X_i} = \mathbf{x_i}|\mathbf{A},\mathbf{B}\right) = \prod_{f=1}^{n_h-1}\left[\prod_{a=1}^{x_{fi}}\left(a + \frac{\left(1-\rho_{f|1...f-1}\right)\pi_{f|1...f-1}}{\rho_{f|1...f-1}} - 1\right)\times \right. \tag{2.6}$$

$$\prod_{a=1}^{N_{i\cdot}-\sum_{j=1}^{f}x_{ji}}\left(a + \frac{\left(1-\rho_{f|1...f-1}\right)\left(1-\pi_{f|1...f-1}\right)}{\rho_{f|1...f-1}} - 1\right)\times$$

$$\left.\prod_{a=1}^{N_{i\cdot}-\sum_{j=1}^{f-1}x_{ji}}\left(a + \frac{1-\rho_{f|1...f-1}}{\rho_{f|1...f-1}} - 1\right)^{-1}\right]$$

It can be shown that the MDD is a special case of the MBBD. With n_h possible outcomes of the response vector of interest, the MDD has n_h parameters that define the distribution. Call these parameters $m_1, m_2, ..., m_{n_h}$. In contrast, the MBBD would have $2(n_h - 1)$ parameters that define the distribution. Assume for the MBBM that each outcome x_{hi} has two parameters that comprise its conditional beta-binomial distribution, a_h and b_h. Under the conditions $a_h = m_h$ and $b_h = m_{h+1} + m_{h+2} + ...m_{n_h} \forall h$, the MBBD devolves into the MDD. The proof is provided in Appendix A.1.

2.4. Testing for Homogeneity of Dependent Intraclass Correlation Coefficients

2.4.1. Estimation of Parameters

As previously mentioned, the MDD can be written in terms of n_h parameters $\pi_1, \pi_2, ...\pi_{n_h-1}, \rho_\cdot$ (since the probabilities of response are constrained by the equality $\sum_{i=1}^{n_h}\pi_i = 1$). The corresponding MBBD has similar constraints and can be written in terms of parameters $\pi_1, \pi_{2|1}, ...\pi_{n_h-1|1,2...,n_h-2}, \rho_1, ...\rho_{n_h-1|1,2...,n_h-2}$. One will notice that $\pi_{n_h|1,2...,n_h-1}$ and $\rho_{n_h|1,2...,n_h-1}$ are not accounted for in the MBBD, but this is expected as the conditional beta-binomial distribution for the ${n_h}^{th}$ response is trivial

conditional on all other possible responses.

Because the primary focus of these methods is on determining differences among the ICCs, each of the probabilities will be obtained prior to estimating the parameters of the final model using the moment estimator $\pi_h = \left(\sum_{q=1}^{n_{\cdot\cdot}} \sum_{p=1}^{n_{i\cdot}} y_{hqp}\right) / \left(\sum_{q=1}^{n_{\cdot\cdot}} n_{q\cdot}\right)$. Given the proportion π_h within each model, the ICC can subsequently be determined. For the MDD, given the assumptions outlined earlier, the conditional ICC can be completely determined by the parameters $m_1 ... m_k$ in the MDD. Recall within each conditional beta-binomial distribution, the ICC is specified as $\frac{1}{a_h+b_h+1} = \frac{1}{\sum_{i=h}^{n_h} m_i+1}$, so no further estimation of the ICC is required. However, within the MBBD, for each conditional beta-binomial distribution, the ICC for each conditional distribution needs to be estimated maximizing the respective likelihood. Estimation of the ICC in this fashion has been documented elsewhere and will not be discussed further [16].

2.4.2. Testing Homogeneity of Intraclass Correlation Coefficients

Given that the MDD is a special case of the MBBD and the fact that the proportion parameters are the same between the two models, the MDD is a nested model within the MBBD under the constraint that $\rho_1 = \rho_2 = ... = \rho_{\cdot}$. If this is true, the likelihood under the MBBD is equivalent to the likelihood under the MDD, and different otherwise. Given the nested likelihoods, one can test the hypothesis that $\rho_1 = \rho_2 = ... = \rho_{\cdot}$ using a likelihood ratio test.

Let L_{MDD} be the likelihood of the parameters given the data assuming the MDD, and let L_{MBBD} be the likelihood of the parameters given the MBBD. Then, the test statistic $\psi = 2log\frac{L_{MBBD}}{L_{MDD}}$ follows a chi-square distribution with $n_h - 2$ degrees of freedom

$\left(\chi^2_{n_h-2}\right)$[8]. Thus, the following test of hypotheses can occur:

$$\begin{cases} H_0: & \rho_1 = \rho_2 = ... = \rho_. \\ H_A: & \rho_1 \neq \rho_. \text{ or } \rho_2 \neq \rho_. \text{ or } ...\rho_{n_h} \neq \rho_. \end{cases}$$

However, this likelihood-ratio test is not as straightforward as it may appear. Recall that the MBBD is a decomposition of the joint distribution of all possible responses of the outcome of interest. Under the null hypothesis laid out above, different decompositions of the joint distribution can be obtained using various orderings of conditional beta-binomial distributions. Therefore, the following hold true for object i:

$$\begin{aligned} P\left(\mathbf{X_i} = \mathbf{x_i}|\mathbf{A},\mathbf{B}\right) =& P(X_{1i} = x_{1i}|\mathbf{A},\mathbf{B})(X_{2i} = x_{2i}|\mathbf{A},\mathbf{B},X_{1i} = x_{1i})... \\ & P(X_{n_hi} = x_{n_hi}|\mathbf{A},\mathbf{B},X_{1i} = x_{1i}...X_{(n_h-1)i} = x_{(n_h-1)i}) \\ =& P(X_{n_hi} = x_{n_hi}|\mathbf{A},\mathbf{B})(X_{(n_h-1)i} = x_{(n_h-1)i}|\mathbf{A},\mathbf{B},X_{n_hi} = x_{n_hi})... \\ & P(X_{1i} = x_{1i}|\mathbf{A},\mathbf{B},X_{n_hi} = x_{n_hi}...X_{2i} = x_{2i}) \end{aligned}$$

Of course these are only two examples, and for each distribution there are $n_h!/2$ unique decompositions of the joint likelihood since the last two beta-binomial distributions in the decomposition are interchangeable as the two decompositions will result in equivalent likelihoods. The test statistic from any one of the possible decompositions can be used to reject the null hypothesis that the ICC for all responses are equal. However, performing all possible tests and observing whether any test indicates that there is enough information to reject the null hypothesis of equivalent ICCs will inflate the type I error rate as the issue of multiple comparisons arises. Therefore, multiple comparison methods must be employed to ensure that the type I error rate is controlled.

2.4.3. Multiple Comparisons Considerations

Given n_h possible outcomes to the response of interest, there are $n_h!/2$ possible unique decompositions of the MBBD. Let L_b be the likelihood of the parameters given the data under the b^{th} decomposition of the MBBD ($b = 1...n_h!/2$). Let p_b be the p-value associated with the likelihood ratio test comparing the b^{th} decomposition of the MBBD to the MDD according to the $\chi^2_{n_h-2}$ distribution. Finally, the ordered p-values will be denoted as $p_{(1)}...p_{(n_h!/2)}$ where $p_{(1)} \leq p_{(2)}... \leq p_{(n_h!/2)}$.

In practice, there is no true decomposition of the likelihood as the decomposition is arbitrary, necessitating that all possible decompositions are considered. As each of the likelihood ratio test statistics are based on the same data, have the same reference null-hypotheses, and in some cases use some of the same parameters, each of the statistics are positively correlated. Unfortunately, the research on the joint distribution of correlated chi-square variables has yet to reveal a closed-form solution of the joint distributions in many situations [12], which leaves little room to either attempt to estimate the correlation among the test statistics or use that information for multiple comparisons. Thus, one is relegated to using methods based on the ordered p-values to control the type I error rate.

To obtain strong control over the type I error rate, the Bonferroni-Holms method [26] lends itself to a simple solution to control the error rate. However, this method was demonstrated to control the type I error rate when assessing multiple independent test statistics. For the purposes of this test where the aim is to test the homogeneity of ICCs among all responses, only one of the test statistics is required to be significant at the α level of interest in order to reject the null hypothesis. After performing all $n_h!/2$ possible likelihood-ratio tests, the concern is not which test rejects the null hypothesis,

only that one of the tests rejects the null hypothesis. Therefore, using the Bonferroni-Holms procedure in this case is equivalent to observing only whether $p_{(1)} \leq 2\alpha/n_h!$. However, due to the high correlation among each of the test statistics, the Bonferroni-Holms procedure will actually prove to be too conservative in its control of the type I error rate, leading to an overall loss of power of the test [24].

In contrast, alternative methods serve to provide weak control over the type I error rate by controlling the false-discovery rate (FDR). The Benjamini-Hochberg method [3] has been widely used as a step-down procedure that provides control over the FDR, but has been criticized in its use for not providing strong control over the type I error rate. To define this procedure, let $z = max\left(g : p_{(g)} \leq 2g\alpha/n_h!\right)$ if such a g exists, otherwise let $z = 0$. If $z > 0$, the null hypothesis of homogeneity of ICCs is rejected in favor of the alternative that the ICCs are not all equivalent across responses. This procedure has two benefits over the Bonferroni-Holms procedure. First, it uses all available likelihood-ratio tests to compare against the null hypothesis. Second, Benjamini and Yekutieli [4] showed that the FDR is well controlled in the case of comparing positively correlated test-statistics, which lends credence to the results.

2.4.4. Test Conclusions

This likelihood-ratio test is intended to test multinomial responses for homogeneity of ICCs across each potential outcome. However, the proposed MBBD does not provide an avenue to simultaneously estimate the ICC of each response. Many methods suggest dichotomizing the multinomial response into binary yes/no results for each possible response, then assess the ICC for each dichotomized response [34, 19, 2]. One commonly used method to assess the ICC for each dichotomized result is the beta-binomial model[15, 16, 39, 44], which has been indirectly utilized in the MBBD. In each decomposition of the MBBD, the first term in the decomposition will re-

sult in an ICC estimate equivalent to the beta-binomial ICC. Therefore, each of the ICCs are estimated among all permutations of the MBBD by maximizing the first dichotomized beta-binomial distribution with respect to the ICC of the outcome of interest. Therefore, it would appear to be most appropriate, in the case that the hypothesis of homogeneity of ICCs across all responses is rejected, to estimate the ICC and corresponding standard error of each dichotomized outcome using the likelihood-based method of maximizing the dichotomized beta-binomial distribution. If not rejected, the methods presented by Lui et. al. [37] can be employed to estimate the common ICC among all outcomes.

2.5. Simulations

2.5.1. Simulation Methodology

To test the overall type I error rate of the test of homogeneity of ICCs, as well as to demonstrate the power of the procedure, a simulation study was carried out. All data were generated under the MBBD, and in the case of the null model, the assumptions of the MBBD which equate to the MDD were implemented as described in section 3.4. Recall that the MBBD is a product of conditional beta-binomial distributions such that if

$$X_i \sim MBBD\left(\pi_1, \pi_{2|1}...\pi_{n_h-1|1,2...n_h-2}, \rho_1, \rho_{2|1}...\rho_{n_h-1|1,2...n_h-2}, N, N-x_{1i}, ...N-\sum_{j=1}^{n_h-2} x_{ji}\right)$$

and $BB(\pi_h, \rho_h, N)$ denotes the beta-binomial distribution with probability of success π_h and ICC ρ_h with N responses, then

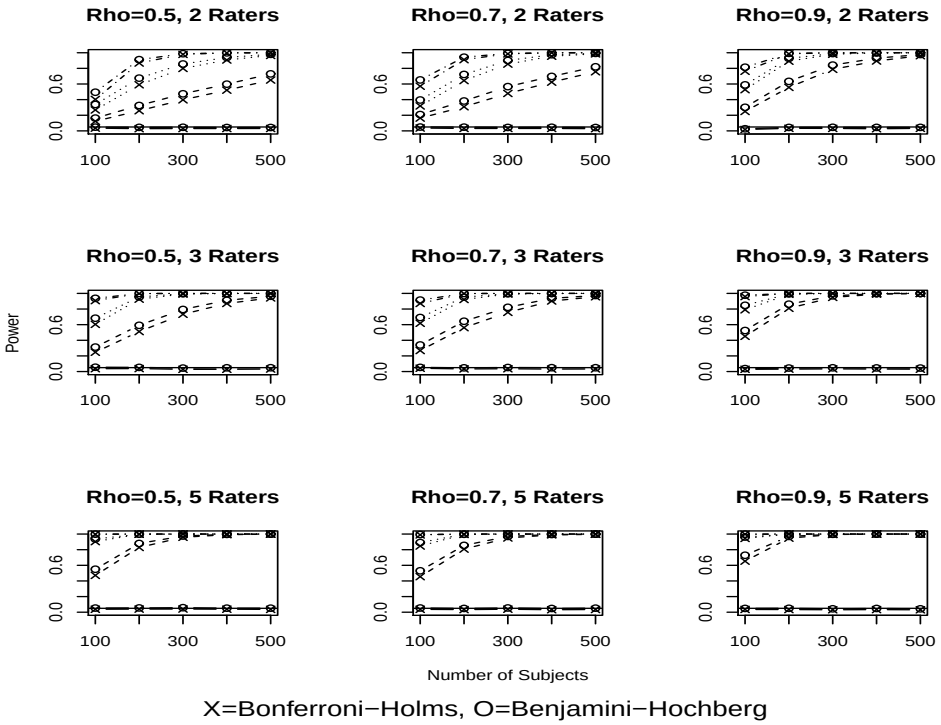
$$P(\mathbf{X_i} = \mathbf{x_i}) = BB(\pi_1, \rho_1, N) \times BB\left(\pi_{2|1}, \rho_{2|1}, N - x_{1i}\right) \times ... \\ BB\left(\pi_{n_h-1|1,2...n_h-2}, \rho_{n_h-1|1,2...n_h-2}, N - \sum_{j=1}^{n_h-2} x_{ji}\right)$$

Simulating parameters under this distribution involves specifying a number of options governing the simulation including:

1. The number of objects, i
2. The number of raters, j
3. The probability of response for each possible outcome, $\left(\pi_1, \pi_{2|1}..., \pi_{n_h-1|1,2...n_h-2}\right)$
4. The overall ICC under the null hypothesis, $\rho_.$
5. For power studies, the deviation from the ICC under the null hypothesis, $\left(\rho_{d1}, \rho_{d2}..., \rho_{d(n_h-1)}\right)$

Then, to obtain the sampled data, first set a sample of data from the first beta-binomial distribution $BB(\pi_1, \rho_. - \rho_{d1}, j)$. After obtaining the number of positive responses for the first outcome, continue to obtain the number of positive responses for the second outcome by sampling from the second beta-binomial distribution $BB\left(\pi_{2|1}, \rho_{2|1} - \rho_{d2}, j - x_1\right)$, where $\pi_{2|1}$ and $\rho_{2|1}$ are the conditional probability of success and conditional ICC under the MDD assumption as previously described. The deviation from the MDD assumption lies in the specification of ρ_{d2}. Continue this process for all possible outcomes up to $n_h - 1$. The final set of outcomes are specified as $j - \sum_{f=1}^{n_h-1} x_f$. All simulations were performed using the R software pack-

Figure 2.1: Homogeneity of ICC Power Plots



age [43], and sampling from the beta-binomial distribution was performed using the rbetabinom function from the emdbook package written by Bolker [5].

Table 2.1: Power of Homogeneity of ICC Test

Raters	Diff	$\rho = 0.5$				$\rho = 0.7$			
		Number of objects							
		100	200	300	400	100	200	300	400
2	(0,0)	3.5 [4.8]	2.9 [4.5]	3.0 [4.6]	3.0 [4.3]	3.7 [4.9]	3.1 [4.8]	3.1 [4.4]	3.1 [4.6]
	(.2,0)	12.3 [16.4]	25.8 [32.5]	39.7 [47.5]	52.1 [59.8]	16.4 [20.9]	30.9 [38.2]	47.9 [56.7]	62.4 [69.7]
	(.3,0)	26.8 [33.8]	59.0 [67.5]	79.9 [85.6]	90.7 [94.0]	31.9 [39.5]	64.4 [72.2]	85.5 [90.1]	95.5 [97.5]
	(.4,0)	40.2 [49.5]	86.8 [91.2]	97.9 [98.8]	99.8 [99.9]	57.2 [65.2]	91.1 [94.3]	98.6 [99.2]	99.9 [100.0]
	(.2,.2)	7.7 [10.7]	14.1 [18.6]	22.5 [27.7]	31.4 [37.5]	9.5 [12.3]	16.9 [21.5]	26.8 [32.7]	38.3 [45.0]
	(.3,.3)	12.0 [15.9]	33.3 [39.7]	53.5 [60.8]	71.0 [77.0]	15.2 [19.9]	37.7 [44.0]	59.1 [65.5]	76.9 [81.6]
	(.4,.4)	15.1 [18.9]	53.2 [60.1]	82.0 [86.7]	94.7 [96.1]	28.2 [34.5]	63.7 [69.8]	86.3 [89.9]	96.0 [97.3]
3	(0,0)	3.7 [5.5]	3.8 [5.3]	2.8 [4.4]	2.9 [4.8]	4.2 [5.3]	3.4 [4.8]	3.6 [5.2]	3.1 [4.9]
	(.2,0)	24.7 [31.3]	51.2 [59.2]	73.3 [79.4]	87.1 [91.2]	26.9 [34.0]	56.2 [64.4]	75.7 [82.2]	90.3 [93.7]
	(.3,0)	60.4 [68.3]	92.6 [95.5]	99.1 [99.6]	99.9 [99.9]	61.7 [69.2]	92.2 [95.3]	99.1 [99.5]	99.9 [99.9]
	(.4,0)	90.8 [93.7]	99.7 [99.9]	100.0 [100.0]	100.0 [100.0]	87.1 [91.4]	99.8 [99.9]	100.0 [100.0]	100.0 [100.0]
	(.2,.2)	14.8 [19.2]	33.8 [39.7]	53.9 [60.3]	68.1 [73.5]	15.9 [20.4]	35.1 [40.6]	55.4 [60.8]	70.7 [76.1]
	(.3,.3)	37.3 [43.9]	75.3 [80.4]	92.2 [94.4]	98.4 [98.9]	37.1 [43.3]	74.3 [78.9]	92.4 [94.2]	98.3 [98.9]
	(.4,.4)	67.1 [73.1]	97.3 [98.2]	99.9 [100.0]	100.0 [100.0]	64.7 [70.4]	96.0 [97.1]	99.7 [99.8]	100.0 [100.0]
5	(0,0)	3.7 [5.3]	4.0 [5.5]	4.1 [5.8]	3.8 [5.2]	3.7 [5.4]	3.3 [4.8]	3.8 [5.4]	3.2 [4.8]
	(.2,0)	47.1 [55.0]	82.8 [88.0]	95.9 [97.5]	99.2 [99.6]	45.5 [53.0]	80.7 [85.5]	95.0 [97.0]	98.9 [99.4]
	(.3,0)	90.3 [93.7]	99.7 [99.9]	100.0 [100.0]	100.0 [100.0]	84.7 [89.2]	99.5 [99.7]	100.0 [100.0]	100.0 [100.0]
	(.4,0)	99.8 [99.9]	100.0 [100.0]	100.0 [100.0]	100.0 [100.0]	98.9 [99.4]	100.0 [100.0]	100.0 [100.0]	100.0 [100.0]
	(.2,.2)	32.5 [38.4]	65.6 [71.3]	85.2 [88.7]	94.6 [96.2]	31.6 [36.8]	62.1 [67.6]	83.1 [86.7]	93.8 [95.4]
	(.3,.3)	75.0 [79.9]	98.4 [98.9]	100.0 [100.0]	100.0 [100.0]	66.9 [72.5]	96.1 [97.1]	99.6 [99.8]	100.0 [100.0]
	(.4,.4)	98.2 [99.0]	100.0 [100.0]	100.0 [100.0]	100.0 [100.0]	94.2 [95.6]	100.0 [100.0]	100.0 [100.0]	100.0 [100.0]

Note: Power calculations are represented as Bonferroni-Holms [Benjamini-Hochberg]

2.5.2. Simulation Results

The results of a series of simulations performed based on the methodology set forth in section 2.5.1 are provided in both figure 2.1 and table 2.1. The simulations results provided in figure 2.1 examine only the effect of a difference in ρ on one of the four possible outcomes, where the difference between the ICC of the response in question and the rest of the responses is programmed to be either 0, .2, .3 or .4. Power calculations are displayed for both Bonferroni-Holms and Benjamini-Hochberg corrections for multiple comparisons. In all cases, as expected, the power from the Benjamini-Hochberg correction was greater than that of the Bonferroni-Holms. Examining the type I error rate for these tests leads to two conclusions. First, the Bonferroni-Holms correction is a conservative correction with type I error rates ranging from 0.028 to 0.042 percent for an $\alpha = 0.05$ test, which also shows that the FWER is strongly controlled for this test under these conditions. Secondly, the Benjamini-Hochberg correction yields a type I error rate closer to the nominal 0.05 level with a range of 0.043 to 0.058, however more powerful than the expected error-rate.

In general, the power of the test for homogeneity of ICCs is greater when either the number of objects or raters is increased. In addition, with all other parameters equivalent, a higher ICC for the majority of tests results in greater power. For example, a result with one outcome with an ICC of 0.5 and the rest 0.7 will have less power than another situation where one outcome has an ICC of 0.7 and the others 0.9. Therefore, both the difference of the discrepant ICC and the magnitude of the ICCs must be taken into consideration when considering power analyses for this test. Finally, a close analysis of table 2.1 reveals that greater power is observed with only one discrepant ICC instead of two, all other parameters being equal. Consider the case of 5 raters and 200 subjects where the majority $\rho = 0.5$. The case with only

one discrepant outcome yields 82.8% power using the Bonferroni-Holms correction [88.0% using Benjamini-Hochberg], however two discrepant outcomes results in only 65.6% [71.3%] power. Therefore investigators should consider the number of expected discrepant outcomes to properly power this test understanding that fewer discrepant results will result in greater power.

2.6. Applications

2.6.1. Cervical Cancer Diagnoses

As originally reported by Holmquist, McMahan and Williams [27], ratings for the classification of carcinoma in situ of the uterine cervix from seven pathologists were recorded. Each rater gave their interpretation of 118 slides and rated each slide as one of the following five ordinal categories: negative, atypical squamous hyperplasia, carcinoma in situ, squamous carcinoma with early stromal invasion, and invasive carcinoma. Both Landis and Koch [32] and Landis et. al. [31] investigated these data to assess rater agreement and to test for potential rater bias. These data will now be assessed to determine whether the levels of agreement within each response are equivalent and can be jointly modeled, or whether an assumption of homogeneity of ICCs for each response is violated.

2.6.2. Face Recognition

The GO Project at the University of Pennsylvania enrolled subjects ages 8-21 who entered the emergency room at Children's Hospital of Philadelphia for any reason. Demographic information was collected for each subject. In addition, the subjects' medical health was rated on a scale of 0 to 4 (0 being the healthiest) and their cognitive ability was rated as 'typically developing', 'other psychiatric', 'sub psychotic'

or 'psychosis spectrum'. Each subject was shown 3D images of 40 faces and asked to specify an emotion on each face from happy, sad, anger, fear or neutral. This study has the primary goal of establishing a cohort to follow to better understand the factors leading to psychosis. However, we wish to examine the level of agreement of face recognition among healthy, developed subjects. Therefore, the ratings of a subset of 73 subjects 18 or older who are typically developing with medical rating 0 will be analyzed to measure the level of agreement of facial recognition among healthy volunteers.

2.6.3. Application Results

Table 2.2 shows the proportion, ICC and corresponding ICC standard error for each of the 5 responses from the two scenarios described above. In addition, the overall ICC based on the multinomial-Dirichlet distribution is displayed along with its corresponding standard error. Figure 2.2 displays the $-log_{10}$ p-value for each of the 60 possible decompositions of the MBBM from largest p-value to smallest. In addition, the corresponding Benjamini-Hochberg boundary for significance is displayed for each decomposition as a reference. In both cases the hypothesis of homogeneity of ICC across responses is violated. Table 2.2 shows a diversity of ICCs for both scenarios, but particularly in the cervical cancer diagnoses data with a range in ICC of .147 for atypical hyperplasia to .546 for invasive cancer. As expected, the level of agreement among healthy subjects for the facial recognition data is relatively homogeneous, except for their assessment of happy faces where there is markedly better agreement among subjects. There is strong evidence in both cases that the assumption of homogeneity is violated, and the graph of the distribution of p-values accurately portrays a much stronger signal for the cervical diagnoses data (note that 5 p-values were calculated as zero and a value of 20 was imputed for display purposes). As a result, the

overall ICC from the multinomial-Dirichlet distribution does not accurately portray the true level of agreement among responses due to the lack of homogeneity, and researchers should consider the levels of agreement among the 5 item-wise responses as an alternative. This may not be a surprising result for the cervical cancer diagnosis data, but is unexpected for the facial recognition data. One would posit that typically developing, healthy 18–21 year old subjects would be able to agree on the emotions displayed on 40 artificial faces, but clearly this is not the case. Therefore, caution should be taken when using these facial recognition results as outcomes for research as even healthy subjects do not agree what emotions are being shown.

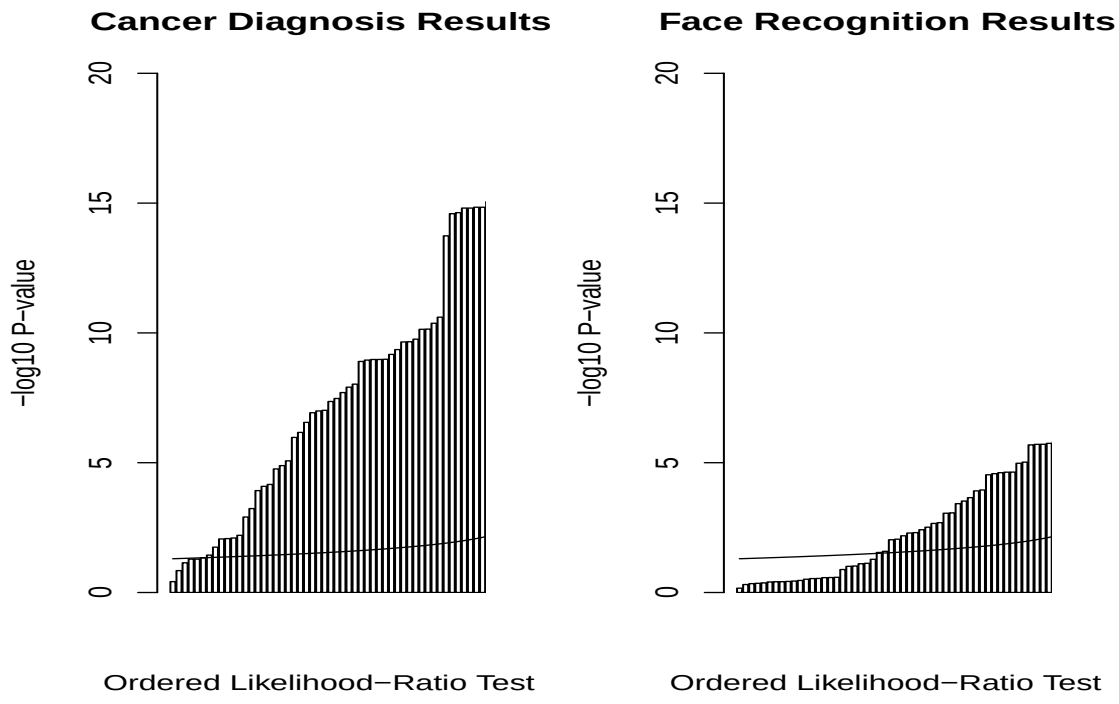
Table 2.2: Application Results

Cervical Cancer Results				Face Recognition Results			
Category	$\hat{\pi}$	$\hat{\rho}$	SE($\hat{\rho}$)	Category	$\hat{\pi}$	$\hat{\rho}$	SE($\hat{\rho}$)
Negative	.281	.518	.046	Happy	.207	.861	.043
Atyp. Hyperplasia	.254	.147	.033	Sad	.194	.573	.060
Ca in Situ	.364	.377	.042	Anger	.171	.660	.065
Ca w/ Early Invas.	.074	.184	.054	Fear	.198	.638	.062
Invasive Cancer	.027	.546	.137	Neutral	.230	.593	.060
Overall ICC		.332	.028			.637	.031

2.7. Conclusion

Comparing the level of agreement among various outcomes of a multinomial response is important when testing reliability and reproducibility of an assessment. Ideally there would be no difference in the level of agreement among any of the outcomes, however practically it may not be the case. In order to appropriately test this phenomenon taking the dependency of the data into account, we have presented a likelihood-based approach to test the homogeneity of ICCs across multiple outcomes in a multinomial response. This test was demonstrated to have an appropriate, yet conservative, type I error rate when using the Bonferroni-Holms correction and a close to nominal, albeit

Figure 2.2: Application Results Distribution of -log10 P-values



The Benjamini-Hochberg boundary for significance is displayed for each test as a reference

slightly liberal, type I error rate when using Benjamini-Hochberg correction. This test is applicable to any number of potential multinomial outcomes, however simulations were completed on only the case of a four-outcome response. The test was observed to have increased power when the number of subjects, number of raters or majority $\rho_{\cdot}$ was increased. Finally, in addition to testing the homogeneity of ICCs, this test can also be used to test the appropriateness of modeling the data with a common ICC, which can result in a more accurate estimate of the overall ICC while reducing the corresponding standard error if the method is deemed appropriate.

CHAPTER 3

Estimation and Inference of the Three-Level Intraclass Correlation Coefficient for Binomial Data

3.1. Introduction

Many different methods have been developed to assess inter-rater agreement on repeated measures on the same object. In studying the measures of agreement and reliability of measuring a particular object multiple times, the intraclass correlation coefficient (ICC) has been proposed as a measure of agreement. Given normal, continuous data, the ICC can be relatively easily calculated using the one-way random effect model $y_{ijk} = \mu + s_{ij} + \epsilon_{ijk}$ where $\epsilon_{ijk} \sim N(0, \sigma_e^2)$ and $s_{ij} \sim N(0, \sigma_s^2)$ for the k^{th} rating for the j^{th} object in the i^{th} nested-level [34]. The ICC can then be calculated from the model as $\sigma_s^2 / (\sigma_s^2 + \sigma_e^2)$. [21]. For the continuous case, the normality assumption can generally be satisfied and few issues arise in inference. Inference procedures are also straightforward given normal continuous data as presented in Searle [47].

Typically, only the object-level agreement is considered. However, this object-level agreement can be nested within other levels of agreement, artificially inflating the observed object-level agreement if the nested-level agreement is not taken into account. The correlation that can be found among ratings within a site, for example, can artificially inflate the object-level ICC. As shown by Landis et. al. [31], the random effects model for object and nested-level effects can be written as $y_{ijk} = \mu + c_i + s_{ij} + \epsilon_{ijk}$ where $\epsilon_{ijk} \sim N(0, \sigma_e^2)$, $s_{ij} \sim N(0, \sigma_s^2)$ and $c_i \sim N(0, \sigma_c^2)$. Then, the nested-level ICC is calculated as $\sigma_c^2 / (\sigma_c^2 + \sigma_s^2 + \sigma_e^2)$; whereas the object-level ICC is calculated as $(\sigma_c^2 + \sigma_s^2) / (\sigma_c^2 + \sigma_s^2 + \sigma_e^2)$. While this method can estimate the ICC among all

observations within a nested-level, there is currently no method to determine the corresponding standard error of the nested-level agreement.

Given dichotomous binary responses, the normality assumption is not verified and issues can arise in calculating confidence intervals and standard errors. Landis and Koch [34] showed the consistency of the point estimate of the ICC in the dichotomous case using the one-way random effects model, however challenging issues still arose with deriving a simplified version of the linearized Taylor-series based variance estimator for the ICC. Koch et. al. [29] developed a general method of estimation for repeated measures of categorical data that allows for asymptotic estimation of the standard error of the ICC estimate, but this method requires ≥ 5 observations per cluster to be valid and requires expressing the ICC estimator as a compounded function of the underlying multinomial proportions, leading to a series of matrix products to formulate the variance estimators. As an improvement, Mak [39] developed an "exact asymptotic" variance of the ICC with dichotomous outcomes using the one-way random effects ANOVA model. This model provides more accurate standard errors than using methods that assume normality. However, none of these models work optimally on binary data and are not sufficient to estimate the standard error of the nested-level ICC.

Beyond these ANOVA methods, Ridout et. al. [44] compared 20 different estimates of the ICC for binary response data, each with their own efficiencies and drawbacks. Of these methods, one method that performed particularly well was using the Beta-Binomial model to estimate the ICC [15, 16]. Following these favorable comparisons, this paper will extend the estimation of the ICC using the Beta-Binomial model to determine the level of agreement between different objects within the same nested-level and calculate the corresponding standard error. Then, we will demonstrate that a

level of agreement existing among different objects within the same nested level artificially inflates the estimate of the object-level agreement and will provide a nested-level adjusted object-level correlation that will better reflect the true level of agreement among objects. Finally, we will apply this method to the actual race/ethnicity classification data arising from a genome-wide association study (GWAS) in which serious identity misalignment was discovered. Comparison of self-reported race/ethnicity and genetically-inferred race/ethnicity, separately within each of 47 genotyping plates, led to isolation of several genotyping plates with substantial misalignment of study subject identity with mis-matched genotyping plate results.

3.2. Notation and Motivation

3.2.1. Motivating Example: Investigating Identity Misalignment within a GWAS

A mini-GWAS (50K single-nucleotide polymorphisms - SNPs) conducted within a multipurpose cohort study of renal and cardiovascular outcomes led to the troubling discovery that intentionally duplicate genotyping results were paired with totally different subject IDs. Fortunately, within the same clinical research network, a full-scale GWAS (1 million SNPs) was conducted shortly thereafter, and the "fingerprinting" step was used to correctly realign nearly 4% of the subject IDs to their correct genotyping results. Since each study participant was classified by self-reported race/ethnicity as 1) Non-Hispanic White; 2) Non-Hispanic Black; 3) Hispanic; and 4) Other, the cross-classification of genetically-inferred race/ethnicity with self-reported race summarized in Table 3.1 illustrates the impact of the re-alignment of subject IDs quite strikingly among the NH-White and NH-Black discordant cells. In particular, prior to re-alignment of IDs, there were 55 misclassifications (left panel), but after re-alignment of IDs there were 0 (right panel), with the estimator of simple kappa

Table 3.1: Agreement between Self-Reported and Genetically-Inferred Ethnicity

Original PID (Kappa=0.921)

Self-reported Race/ Ethnicity	Genetically-inferred Race/Ethnicity				Total
	NH-White	NH-Black	Hispanic	Asian/ Other	
NH-White	**1,474 (96.91)**	26 (1.71)	8 (0.53)	13 (0.85)	**1,521 (42.89)**
NH-Black	29 (1.94)	**1,447 (96.98)**	3 (0.20)	13 (0.87)	**1,492 (42.08)**
Hispanic	16 (3.99)	11 (2.74)	**359 (89.53)**	15 (3.74)	**401 (11.31)**
Asian/ Other	31 (23.48)	6 (4.55)	3 (2.27)	**92 (69.70)**	**132 (3.72)**
Total	**1,550 (43.71)**	**1,490 (42.02)**	**373 (10.52)**	**133 (3.75)**	**3,546**

Re-aligned PID (Kappa=0.951)

Self-reported Race/ Ethnicity	Genetically-inferred Race/Ethnicity				Total
	NH-White	NH-Black	Hispanic	Asian/ Other	
NH-White	**1,505 (98.75)**	0 (0.00)	7 (0.46)	12 (0.79)	**1,524 (42.98)**
NH-Black	0 (0.00)	**1,478 (99.13)**	1 (0.07)	12 (0.80)	**1,491 (42.98)**
Hispanic	15 (3.75)	8 (2.00)	**361 (90.25)**	16 (4.00)	**400 (11.28)**
Asian/ Other	30 (22.90)	4 (3.05)	4 (3.05)	**93 (70.99)**	**131 (3.69)**
Total	**1,550 (43.71)**	**1,490 (42.02)**	**373 (10.52)**	**133 (3.75)**	**3,546**

Note: Cell proportions are displayed as row percentages to illustrate the accuracy of the genetically-inferred race/ethnicity within each self-reported category

increasing from 0.92 to 0.95.

On further investigation, among the final set of 3,546 study participants with resolved data from both GWAS studies, it was discovered that the biospecimens from the Hispanic study participants were heavily clustered on 5 of the 47 genotyping plates, as illustrated in Figure 3.1. This differential prevalence of self-reported Hispanic ethnicity across the 47 genotyping plates, together with the race-ethnicity agreement between self-report and genetically inferred race/ethnicity will be used as a motivating example to illustrate the impact of multi-stage clustering on ICC measures of agreement. Given that 400/3,546 = 11.3% of the study participants self-reported Hispanic ethnicity, the observed distribution of subjects across genotyping plates in

Figure 3.1: Distribution of Self-Reported Hispanics by Plate in a GWAS

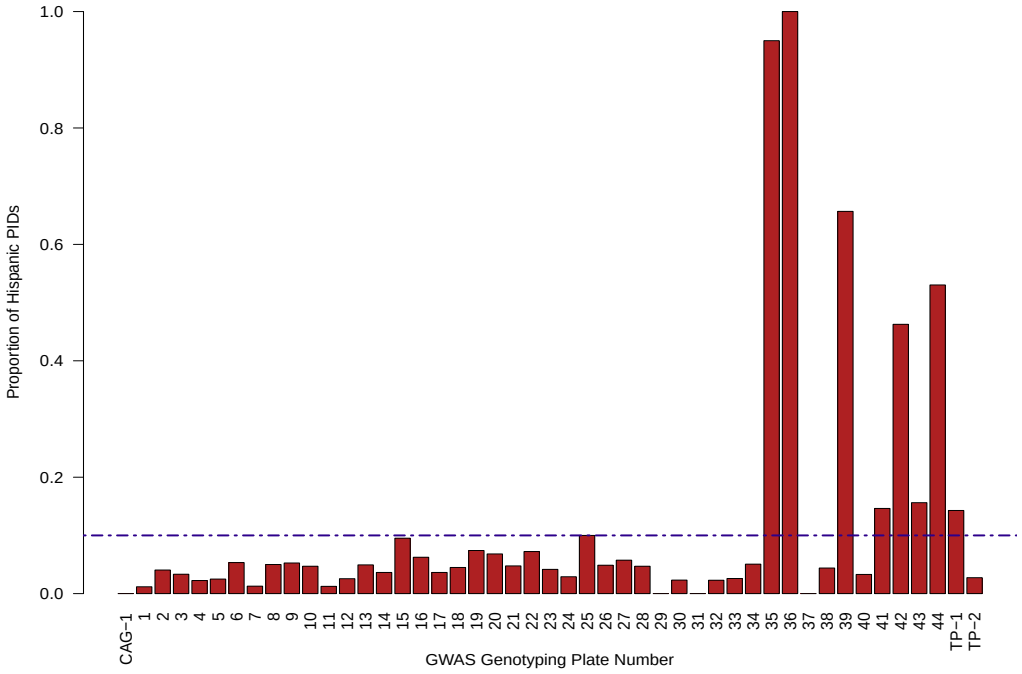


Figure 3.1 illustrates the differential prevalence, with most plates having far less than 11%, and 5 plates with more than 50% prevalence. With the discrepant number of Hispanics present per plate, there is an issue that could potentially arise when assessing the level of agreement between the self-reported race and the genetically-determined race. Particularly, when looking at the level of agreement of responses among Hispanics, there is a possibility that there is a large level of agreement simply due to the distribution of Hispanics across plates which could artificially inflate the apparent subject-level race agreement.

Table 3.2: Levels of Agreement among Race Responses in a GWAS

Alignment	Response Category	Avg. Prop. ($\hat{\pi}$)	2-level[a] Subject-ICCs ($\hat{\rho}$)	3-level VC Model[b] Plate-ICCs ($\hat{\rho}_c$)	3-level VC Model[b] Subject-ICCs ($\hat{\rho}$)
Original PIDs	1:NH-Whites	0.433	0.929	0.063	0.929
	2:NH-Blacks	0.420	0.949	0.041	0.949
	3:Hispanics	0.109	0.919	0.427	0.919
	4:Others	0.037	0.683	0.000	0.683
Re-aligned PIDs	1:NH-Whites	0.433	0.963	0.063	0.963
	2:NH-Blacks	0.420	0.986	0.041	0.986
	3:Hispanics	0.109	0.926	0.428	0.926
	4:Others	0.037	0.693	0.000	0.693

[a] 2-level Variance Components Model [Landis and Koch [34]]: 1) Subjects; 2) Race Classifier (Subjects)

[b] 3-level Variance Components Model [Choi and Landis [11]]: 1) Plates; 2) Subjects (Plates); 3) Race Classifier (Subjects)

To focus particular attention on the impact of object-level clustering on ICC measures of agreement, four separate category-specific binary ICCs were estimated for each race/ethnicity category in Table 3.2. The level of agreement among subjects was obtained via the 2-level variance components model, which is asymptotically equivalent to using the beta-binomial model to assess agreement [15]. The level of agreement among responses within the same genotyping plate was obtained using the methods described in Landis et. al. [31]. These methods are used to determine the subject-level agreement ρ and the plate-level agreement ρ_c. As can easily be seen, the level of agreement among subjects is very high, while the level of agreement among results within the same plate is low. However, looking at the results from Hispanic subjects, we see that the intraclass correlation coefficient for results within a plate is 0.427, which corresponds to a moderate level of agreement for responses that should be uncorrelated, according to the criteria set out by Landis and Koch [33] . However, this method for assessing the ICC of responses within the same plate has a drawback. The method of deriving the standard error of the ICC estimate specified in [31] required estimating the variance/covariance matrix of the expected ANOVA mean squares estimates. This research did not provide formulas for the corresponding variance/covariance matrix, leaving a need for an explicit formula for the standard error of the nested-level ICC.

In this paper, we set out to find methods that will serve to appropriately determine the level of agreement among different objects within the same nested-level and find the standard error and confidence interval for this measure of agreement. In addition, we will examine the effect that this level of agreement can have on the object-level agreement and demonstrate cases where this nested-level agreement can positively bias object-level agreement in a way that overstates the true level of agreement of

raters on the same object.

3.2.2. Notation

Understanding the roles of "raters", "objects" and "nested-levels" are crucial to understanding the methods presented in this paper. In this context of rater-agreement, "objects" will refer to the item being rated and "raters" will refer to the process assigning the ratings. A "nested-level" refers to a grouping that could be applied to all "objects" such that each "object" has an identity in one and only one "nested-level". For example, in the motivating GWAS example, the "object" being rated is the race of each subject and the "raters" refer to either the subjects' self-assessment of race or the genetically determined race. The "nested-level" is considered to be the genotyping plate as each subject was assigned to one and only one genotyping plate.

Let y_{ijk} be a binary outcome (0 or 1) for the k^{th} rater ($k = 1, ..., n_{ij\cdot}$) on the j^{th} object ($j = 1, ..., n_{i\cdot\cdot}$) in the i^{th} nested-level ($i = 1, ..., n_{\cdot\cdot\cdot}$), and let $\mathbf{y}$ be the vector of all responses. y_{ijk} is assumed to follow a binomial distribution where $E(y_{ijk}) = \pi$ and $Var(y_{ijk}) = \pi(1-\pi)$. Let π be the proportion of objects with the trait being assessed such that $P(y_{ijk} = 1) = \pi$. Let ρ be the object-level intraclass correlation coefficient and ζ be the nested-level intraclass correlation coefficient. Given $n_{ij\cdot}$ ratings per object, the sum of all responses for an object can be written as $x_{ij} = \sum_{k=1}^{n_{ij\cdot}} y_{ijk}$, while the sum of all responses within a nested-level can be written as $x_{i\cdot} = \sum_{j=1}^{n_{i\cdot\cdot}} \sum_{k=1}^{n_{ij\cdot}} y_{ijk}$. Let $\mathbf{x_{ij}}$ and $\mathbf{x_{i\cdot}}$ be vectors that contain the sum of responses for all object or nested-levels for the i^{th} nested-level for the j^{th} object. Let $m_i = \left(\sum_{j=1}^{n_{i\cdot\cdot}} \binom{n_{ij\cdot}}{2} \right) / \binom{\sum_{j=1}^{n_{i\cdot\cdot}} n_{ij\cdot}}{2}$, which can be interpreted as the proportion of area of the upper diagonal of the correlation matrix contributed to by the object-level ICC contributes towards. Let $d_i = m_i\rho + (1-m_i)\zeta$, which is the weighted average of all pair-wise correlations within nested-level i.

3.3. Obtaining the Variance of $x_{i\cdot}$

3.3.1. Estimation of the Variance of x_{ij}

In general, when multiple raters assess the same object, the results are correlated together. When studying the reliability of the raters' assessments, that correlation is of high interest. For the typical object level ICC estimation, the correlation ρ among ratings within each object is assumed to be constant while ratings on different objects are considered to be independent. Given $n_{ij\cdot}$ ratings per object, the vector of all responses for an object will have an $n_{ij\cdot}$ x $n_{ij\cdot}$ dimension correlation matrix in the form of

$$\Sigma_{ij\cdot} = \begin{bmatrix} 1 & \rho & \dots & \dots & \rho \\ \rho & \ddots & \dots & \dots & \vdots \\ \vdots & \dots & 1 & \dots & \vdots \\ \vdots & \dots & \dots & \ddots & \rho \\ \rho & \dots & \dots & \rho & 1 \end{bmatrix} \tag{3.1}$$

Then a consistent estimate of the proportion π is

$$\hat{\pi} = \frac{\sum_{i=1}^{n_{\dots}} \sum_{j=1}^{n_{i\cdot\cdot}} \sum_{k=1}^{n_{ij\cdot}} y_{ijk}}{\sum_{i=1}^{n_{\dots}} \sum_{j=1}^{n_{i\cdot\cdot}} n_{ij\cdot}} \tag{3.2}$$

As the theory is developed further, it is important that any ICC estimate allows for varying number of replicate observations per object. Restricting analysis only to objects with a certain number of ratings is unrealistic, so the theory must be kept robust to account for these cases.

The variance of x_{ij} can be found directly. Consider p and q to be the position of the

response in the covariance matrix. Given the correlation matrix for multiple ratings on the same object, the variance for the total number of responses within an object can be written as

$$\begin{aligned} Var(x_{ij}) &= \sum_{k=1}^{n_{ij\cdot}} Var(y_{ijk}) + 2\sum_{p<q} Cov(y_{ijp}, y_{ijq}) \\ &= n_{ij\cdot}\pi(1-\pi) + 2\binom{n_{ij\cdot}}{2} Cov(y_{ijp}, y_{ijq}) \\ &= n_{ij\cdot}\pi(1-\pi) + n_{ij\cdot}(n_{ij\cdot}-1) Cov(y_{ijp}, y_{ijq}) \\ &= n_{ij\cdot}\pi(1-\pi) + n_{ij\cdot}(n_{ij\cdot}-1)\rho\pi(1-\pi) \\ &= n_{ij\cdot}\pi(1-\pi)\left[1 + (n_{ij\cdot}-1)\rho\right] \end{aligned} \tag{3.3}$$

As a result, it is clear that an over dispersion parameter exists that inflates the variance of correlated binomial data more than independent binomial data. However, the correlation due to this over dispersion increases in the presence of a nested level of correlation.

3.3.2. Estimation of the Variance of $x_{i\cdot}$.

When a nested-level of correlation exists beyond object-level correlation, the correlation matrix of x_{ij} no longer contains all of the information regarding $x_{i\cdot}$. Therefore, the entire $x_{i\cdot}$ needs to be considered as the cluster of interest rather than the object level x_{ij}. The correlation matrix for the vector of responses within a nested-level contains two parameters, the object-level ICC, ρ, and the nested-level ICC, ζ. This correlation matrix can be expressed by a combination of object and nested-level correlations. Let $\mathbf{1}_{i\cdot\cdot}$ be a $\sum_{j=1}^{n_{i\cdot\cdot}} n_{ij\cdot} \times \sum_{j=1}^{n_{i\cdot\cdot}} n_{ij\cdot}$ matrix with all matrix elements equal to 1, and $\mathbf{1}_{ij\cdot}$ be a $n_{ij\cdot} \times n_{ij\cdot}$ matrix with all matrix elements equal to 1. Then the

correlation matrix for nested-level i can be written as

$$\Sigma_{i\cdot\cdot} = DIAG(\Sigma_{i1\cdot} - \zeta \mathbf{1}_{i1\cdot}, \Sigma_{i2\cdot} - \zeta \mathbf{1}_{i2\cdot}, ..., \Sigma_{in_{i\cdot\cdot}} - \zeta \mathbf{1}_{in_{i\cdot\cdot}}) + \zeta \mathbf{1}_{i\cdot\cdot} \tag{3.4}$$

Then we can find the nested-level variance in a similar fashion as the object-level variance. Recall that m_i is the proportion of area of the upper diagonal of the correlation matrix that the object-level ICC contributes towards. Then:

$$Var\left(x_{i\cdot}\right) = \left(\sum_{j=1}^{n_{i\cdot\cdot}} n_{ij\cdot}\right) \pi(1-\pi) \left[1 + \left(\sum_{j=1}^{n_{i\cdot\cdot}} n_{ij\cdot} - 1\right) \left[m_i \rho + (1 - m_i)\, \zeta\right]\right] \tag{3.5}$$

The details of this derivation can be found in Appendix A.2. This variance looks similar to the variance of x_{ij} with two exceptions:

1. The number of responses for x_{ij} is the number of responses per object while the number of responses for $x_{i\cdot}$ is the number of responses within the entire cluster

2. The object-level ICC in x_{ij} is replaced by a mixture of the object and nested-level ICCs proportional to the area of the covariance matrix occupied by ρ.

This necessitates that a method should be developed such that this nested-level over dispersion parameter can be accounted for. The beta-binomial distribution will allow for the accurate modeling of the first two moments of the correlated binomial data.

3.4. Current ICC Methods

3.4.1. Object-Level ICC Estimation Framework

The beta-binomial distribution can be specified as $\binom{n}{k} B(k+\alpha, n-k+\beta)/B(\alpha,\beta)$ where B(x) is the beta function of x, n is the number of trials in the sample, k

is sum of the responses in the trial and α and β are the parameters of the model to be fit. If y $\sim$ Beta-binomial (α, β), then E(y) $= n\alpha/(\alpha+\beta)$ and Var(y) $= [n\alpha\beta(\alpha+\beta+n)]/\left[(\alpha+\beta)^2(\alpha+\beta+1)\right]$. As demonstrated earlier, $E(x_{ij})$=$n_{ij.}\pi$, which means that π=$\alpha/(\alpha+\beta)$. It follows that $1-\pi$=$\beta/(\alpha+\beta)$ as described by Crowder[16]. As a result, the variance can then be written as

$$\begin{aligned} Var(x_{ij}) &= \frac{n_{ij.}\pi(1-\pi)[\alpha+\beta+n_{ij.}]}{\alpha+\beta+1} \\ &= n_{ij.}\pi(1-\pi)\left[1+\frac{(n_{ij.}-1)}{\alpha+\beta+1}\right] \\ &= n_{ij.}\pi(1-\pi)\left[1+\frac{(n_{ij.}-1)}{\frac{\alpha}{\pi}+1}\right] \\ &= n_{ij.}\pi(1-\pi)\left[1+\frac{\pi(n_{ij.}-1)}{\pi+\alpha}\right] \end{aligned} \tag{3.6}$$

Recall from earlier that $Var(x_{ij})$=$n_{ij.}\pi(1-\pi)[1+(n_{ij.}-1)\rho]$ and π=$\alpha/(\alpha+\beta)$. This leads to the solution $\rho = \pi/(\pi+\alpha)$. Then $\alpha = \pi(1-\rho)/\rho$ and
$\beta = (1-\pi)(1-\rho)/\rho$.
Now instead of optimizing the beta-binomial distribution over α and β, the likelihood can be maximized over π and ρ. Then the log likelihood can be expressed as

$$\begin{aligned} LogL(\pi,\rho|\mathbf{X_{ij}}=\mathbf{x_{ij}}) = \sum_{i=1}^{n_{...}}\sum_{j=1}^{n_{i..}}\Bigg[& log\binom{n_{ij.}}{x_{ij}} + \sum_{a=0}^{x_{ij}-1} log\left(x_{ij}+\pi\frac{1-\rho}{\rho}-1-a\right) + \\ & \sum_{b=0}^{n_{ij.}-x_{ij}-1} log\left(n_{ij.}-x_{ij}+(1-\pi)\frac{1-\rho}{\rho}-1-b\right) - \\ & \sum_{c=0}^{n_{ij.}-1} log\left(n_{ij.}+\frac{1-\rho}{\rho}-1-c\right)\Bigg] \end{aligned} \tag{3.7}$$

The object-level beta-binomial distribution has been demonstrated to consistently estimate object-level ICC [15, 44] and will not be discussed further.

3.4.2. ICC Estimation within Nested-Levels

Landis et. al. [31] described a method to obtain the object-level ICC. A similar framework was described where the following marginal and pairwise probabilities were defined:

1) $Pr(Y_{ijk} = 1) = \pi$

2) $Pr(Y_{ijk} = 1, Y_{ij'k'} = 1) = \delta_c$ if j$\neq$ j', $\forall$ i

3) $Pr(Y_{ijk} = 1, Y_{ijk'} = 1) = \delta_s$, for the ij^{th} object and k $\neq k'$

Then the category specific variance components model can be described as $y_{ijk} = \pi + c_i + s_{ij} + r_{ijk}$ where

- $c_{i..}$ are independent nested-level effects with variance component σ_c^2, indexed by i=1,. . .,$n_{...}$;
- $s_{ij.}$ are object effects (nested within nested-levels) with variance component σ_s^2, indexed by j=1,. . .,$n_{i..}$;
- r_{ijk} are rater effects (nested within objects) with variance component σ_r^2, indexed by k=1,. . .,$n_{ij.}$.

Then it follows that the components of variance can be written as

$$\begin{aligned} \sigma_c^2 &= \rho_c \pi(1-\pi) \\ \sigma_s^2 &= (\rho - \rho_c)\pi(1-\pi) \\ \sigma_r^2 &= (1-\rho)\pi(1-\pi) \end{aligned}$$

Then the nested-level ICC is derived as

$$\begin{aligned}\rho_c &= \frac{cov(Y_{ijk}, Y_{ij'k'})}{\sqrt{var(Y_{ijk})var(Y_{ij'k'})}} \\ &= \frac{\delta_c - \pi^2}{\sqrt{\pi(1-\pi)}\sqrt{\pi(1-\pi)}} = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_s^2 + \sigma_r^2}.\end{aligned}$$

and the object-level ICC is derived as

$$\begin{aligned}\rho &= \frac{cov(Y_{ijk}, Y_{ijk'})}{\sqrt{var(Y_{ijk})var(Y_{ijk'})}} \\ &= \frac{\delta_s - \pi^2}{\sqrt{\pi(1-\pi)}\sqrt{\pi(1-\pi)}} = \frac{\sigma_c^2 + \sigma_s^2}{\sigma_c^2 + \sigma_s^2 + \sigma_r^2}\end{aligned}$$

While ρ proves to be a consistent estimate of the object-level ICC [44], ρ_c proves to consistently estimate ζ. However, this problem needs to be addressed in a different fashion in order to both estimate ζ and obtain an accurate standard error estimate for ζ. The beta-binomial distribution can be modified to estimate ζ as well as obtain a standard error estimate for ζ.

3.5. The Nested-Level ICC

3.5.1. Nested-Level ICC Likelihood Framework

To accurately estimate ζ, the variance derived in section 3.2 can be used with the beta-binomial distribution to obtain appropriate estimation of ζ. To estimate the object-level ICC, the beta-binomial distribution was specified with $\alpha = \pi(1-\rho)/\rho$ and $\beta = (1-\pi)(1-\rho)/\rho$. Using the variance described in section 3.2, however, α and β are presented as $\alpha = \pi[1-(m_i\rho + (1-m_i)\zeta)] / [(m_i\rho + (1-m_i)\zeta)] = \pi(1-d_i)/d_i$ and $\beta = (1-\pi)[1-(m_i\rho + (1-m_i)\zeta)] / [(m_i\rho + (1-m_i)\zeta)] = (1-\pi)(1-d_i)/d_i$.

Then the updated log-likelihood can be written as

$$LogL(\pi,\zeta,\rho|\mathbf{X_{i\cdot}}=\mathbf{x_{i\cdot}}) = \sum_{a=1}^{n_{\cdots}}\left[log\binom{n_{a\cdot\cdot}}{x_{a\cdot}} + \sum_{b=0}^{x_{a\cdot}} log\left(b+\pi\frac{1-d_a}{d_a}-1\right) + \sum_{c=0}^{n_{a\cdot\cdot}-x_{a\cdot}} log\left(c+(1-\pi)\frac{1-d_a}{d_a}-1\right) - \sum_{e=0}^{n_{a\cdot\cdot}} log\left(e+\frac{1-d_a}{d_a}-1\right)\right] \tag{3.8}$$

However, direct maximization of this likelihood does not yield unique solutions for ρ or ζ. Note that this likelihood is the same as the two-level likelihood where ρ is replaced by $m_i\rho+(1-m_i)\,\zeta$. Just as ρ can be consistently estimated using the object-level model, $m_i\rho+(1-m_i)\,\zeta$ can also be consistently estimated. However, given $n_{i\cdot\cdot}$, $x_{i\cdot}$ and m_i, there is not enough information to estimate both ζ and ρ strictly using the nested-level likelihood. Therefore, alternative methods should be used for accurate estimation of ζ.

3.5.2. Estimation of ζ

As described in section 3.4.1, the beta-binomial distribution can consistently estimate both the object-level ICC ($\hat{\rho}$) and its corresponding standard error. Generally, when analyzing data on agreement, the true proportion π needs to be estimated, but inference on this parameter is not usually of interest. Therefore, π will be estimated as described in section 3.3.1 and will be considered known throughout estimation of ρ and ζ.

In order to estimate ζ, the following steps should be followed:

1. Since ρ can be estimated from the object-level data, along with its corresponding standard error, we can take advantage of the consistency of the estimate in

estimation ζ. Therefore, the first step in this process should be to estimate ρ given the data.

2. Given that we now have an estimate of $\rho\,(\hat{\rho})$, we can know maximize the likelihood (3.8) using $\hat{\rho}$ as the consistent estimate of ρ. Direct maximization of the resultant profile-likelihood will yield a consistent estimate of ζ and an asymptotic standard error estimate.

Given these steps, we now seek to maximize the profile-likelihood using standard maximum likelihood techniques. First, the gradient vector can be derived as

$$\begin{aligned}\frac{\delta LogL(\zeta|\mathbf{X_{i\cdot}}=\mathbf{x_{i\cdot}},\hat{\rho},\hat{\pi})}{\partial\zeta} =& \sum_{a=1}^{n_{\cdots}}\left[\sum_{b=0}^{x_{a\cdot}}\frac{-\hat{\pi}\,(1-m_a)}{d_a^2\left(b+\hat{\pi}\frac{1-d_a}{d_a}-1\right)}+\right.\\ & \sum_{c=0}^{n_{a\cdot\cdot}-x_{a\cdot}}\frac{-(1-\hat{\pi})\,(1-m_a)}{d_a^2\left(c+(1-\hat{\pi})\frac{1-d_a}{d_a}-1\right)}+\\ & \left.\sum_{e=0}^{n_{a\cdot\cdot}}\frac{1-m_a}{d_a^2\left(e+\frac{1-d_a}{d_a}-1\right)}\right]\end{aligned} \tag{3.9}$$

While an explicit solution for $\frac{\delta LogL(\pi,\zeta,\rho|\mathbf{X_{i\cdot}}=\mathbf{x_{i\cdot}})}{\delta\zeta}=0$ is not apparent, this gradient can be numerically solved to obtain $\hat{\zeta}$, the maximum likelihood estimate of ζ.

Next, the second derivative with respect to ζ can be calculated using the following equations:

$$\begin{aligned}\frac{\partial^2 LogL(\zeta|\mathbf{X_{i\cdot}}=\mathbf{x_{i\cdot}},\hat{\rho},\hat{\pi})}{\partial\zeta^2} =& \frac{-2\frac{\partial LogL(\zeta|\mathbf{X_{i\cdot}}=\mathbf{x_{i\cdot}},\hat{\rho},\hat{\pi})}{\partial\zeta}}{d_i^3}-\sum_{a=1}^{n_{\cdots}}\left[\sum_{b=1}^{x_{a\cdot}}\frac{(1-m_a)^2\hat{\pi}^2}{d_a^4\left(b+\hat{\pi}\frac{1-d_a}{d_a}-1\right)^2}\right.\\ &\left.+\sum_{c=1}^{n_{a\cdot\cdot}-x_{a\cdot}}\frac{(1-m_a)^2(1-\hat{\pi})^2}{d_a^4\left(c+(1-\hat{\pi})\frac{1-d_a}{d_a}-1\right)^2}-\sum_{e=1}^{n_{a\cdot\cdot}}\frac{(1-m_a)^2}{d_a^4\left(e+\frac{1-d_a}{d_a}-1\right)^2}\right]\end{aligned} \tag{3.10}$$

It is easy to infer that when the second derivative is evaluated at $\hat{\zeta}$, $\delta LogL(\zeta|\mathbf{X_{i\cdot}} = \mathbf{x_{i\cdot}}, \hat{\rho}, \hat{\pi})/\delta\zeta = 0$. The second derivative can then be inverted and evaluated and the parameters of interest, resulting in the standard error

$$SE\left(\hat{\zeta}\right) = \left[\sum_{a=1}^{n_{\cdots}} \left[\frac{d_a^4}{(1-m_a)^2}\left[\sum_{b=0}^{x_{a\cdot}} \frac{\hat{\pi}^2}{\left(b+\hat{\pi}\frac{1-d_a}{d_a}-1\right)^2} + \sum_{c=0}^{n_{a\cdot\cdot}-x_{a\cdot}} \frac{(1-\hat{\pi})^2}{\left(c+(1-\hat{\pi})\frac{1-d_a}{d_a}-1\right)^2} - \sum_{e=0}^{n_{a\cdot\cdot}} \frac{1}{\left(e+\frac{1-d_a}{d_a}-1\right)^2}\right]\right]\right]^{-\frac{1}{2}} \tag{3.11}$$

3.5.3. Asymptotic Properties of $\hat{\zeta}$

Recall that $Var(x_{i\cdot}) = \left(\sum_{j=1}^{n_{i\cdot\cdot}} n_{ij\cdot}\right)\pi(1-\pi)\left[1+\left(\left(\sum_{j=1}^{n_{1\cdot\cdot}} n_{ij\cdot}\right)-1\right)[m_i\rho+(1-m_i)\zeta]\right]$. m_i can be rewritten as

$$m_i = \frac{\sum_{j=1}^{n_{i\cdot\cdot}} n_{ij\cdot}(n_{ij\cdot}-1)}{\left(\sum_{j=1}^{n_{i\cdot\cdot}} n_{ij\cdot}\right)\left(\sum_{j=1}^{n_{i\cdot\cdot}} n_{ij\cdot}-1\right)} \tag{3.12}$$

It is easy to show that as $n_{i\cdot\cdot} \to \infty, m_i \to 0$. Therefore, as $n_{i\cdot\cdot} \to \infty, Var(x_{i\cdot}) \to \left(\sum_{j=1}^{n_{i\cdot\cdot}} n_{ij\cdot}\right)\pi(1-\pi)\left[1+\left(\left(\sum_{j=1}^{n_{1\cdot\cdot}} n_{ij\cdot}\right)-1\right)\zeta\right]$.

In this case, the problem devolves to a two-level maximum likelihood ICC estimation, where $\hat{\zeta}$ proves to be a consistent estimate of ζ. In some highly controlled situations, such as a well-controlled clinical trial, it may be possible to hold the number of objects per nested-level and the number of ratings per object constant, in which case m_i can

be written as:

$$m_i = \frac{n_{i..}\left(n_{ij.}\right)\left(n_{ij.}-1\right)}{\left(n_{i..}n_{ij.}\right)\left(n_{i..}n_{ij.}-1\right)} = \frac{\left(n_{ij.}-1\right)}{\left(n_{i..}n_{ij.}-1\right)} \tag{3.13}$$

As $n_{ij.} \to \infty$, m_i converges to $1/n_{i..}$ and $\hat{\zeta}$ is not a maximum likelihood estimator of ζ. However, as $n_{i..} \to \infty$, $m_i \to 0$ and $\hat{\zeta}$ becomes the maximum likelihood estimate of ζ and possesses all of the properties thereof. Then, due to the efficient asymptotic properties of maximum likelihood estimators, $\sqrt{n_{...}}\left(\hat{\zeta}-\zeta\right) \to N\left[0, SE\left(\hat{\zeta}\right)^2\right]$ [8]. Therefore, the asymptotic normality of $\hat{\zeta}$, a Wald 95% confidence interval around $\hat{\zeta}$ can be derived using the 1-α^{th} quantile of the normal distribution $Z_{1-\alpha}$ for a given type I error rate α, resulting in the 95% confidence interval $\left[\hat{\zeta} - Z_{1-\alpha}SE\left(\hat{\zeta}\right), \hat{\zeta} + Z_{1-\alpha}SE\left(\hat{\zeta}\right)\right]$. In addition, a Z-statistic from the standard normal distribution can be found testing $\hat{\zeta}$ against an alternative value ζ_a as $Z = \left(\hat{\zeta}-\zeta_a\right)/SE(\hat{\zeta})$. Appropriate hypothesis testing can then be performed by comparing the Z-statistic to the standard normal distribution.

3.5.4. Nested-Level Adjusted Object-Level ICC

Thus far, there has been no consideration towards the effect that ρ has on ζ and vice-versa. In estimating ζ, π and ρ were estimated first based on the object-level data, then used to estimate ζ. Therefore, the estimate of ζ had no effect on the estimation of ρ. However, in practice, a nested-level ICC could have some significant influence over the object-level ICC. If there is a large nested-level ICC effect, then there will be an artificially large object-level ICC effect. Consider the extreme case where there was perfect agreement on every rating within each site in a multi-center trial (for example, each rating within a site was either positive or negative). Then, regardless

of how well raters within a site could actually rate the outcome of interest, there would be perfect object-level agreement due to the fact that there was perfect site-level agreement. Therefore, a correction to the object-level ICC needs to be made. Consider the definition of object-level ICC as defined by Landis, et. al. [31].

$$\zeta = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_s^2 + \sigma_r^2}$$
$$\rho = \frac{\sigma_c^2 + \sigma_s^2}{\sigma_c^2 + \sigma_s^2 + \sigma_r^2}$$

As previously mentioned, ζ can have the effect of artificially inflating ρ. To demonstrate that effect, we want to investigate the minimum ρ that can exist for a given ζ. Then, using the variance components definitions of the object-level and nested-level ICCs, we can derive

$$\begin{aligned}
\rho =& \frac{\sigma_c^2 + \sigma_s^2}{\sigma_c^2 + \sigma_s^2 + \sigma_r^2} \\
=& \frac{\sigma_s^2}{\sigma_c^2 + \sigma_s^2 + \sigma_r^2} + \frac{\sigma_c^2}{\sigma_c^2 + \sigma_s^2 + \sigma_r^2} \\
=& \frac{\sigma_s^2}{\sigma_c^2 + \sigma_s^2 + \sigma_r^2} + \zeta \\
\rho \geq& \zeta
\end{aligned}$$

Thus, for a given ζ, the corresponding object-level ρ has the range $[\zeta, 1]$. Therefore, in order to adjust ρ to obtain a range of $[0, 1]$, we can compute the nested-ICC adjusted object-level ICC as

$$\rho^* = \frac{\rho - \zeta}{1 - \zeta} \tag{3.14}$$

In the presence of nested-level ICC, ρ^* should be a better representation of the true level of agreement among objects since it adjusts for the effect of ζ that exists from

observations that should be uncorrelated. Let Σ^* be the variance/covariance matrix for $\hat{\pi}, \hat{\rho}$ and $\hat{\zeta}$. Then the variance of the estimate of ρ^* $\left(\hat{\rho^*}\right)$ can be estimated using the delta method.

$$Var\left(\hat{\rho^*}\right) = \left(\frac{1}{1-\hat{\zeta}}, \frac{\hat{\rho}-1}{\left(1-\hat{\zeta}\right)^2}\right) \Sigma^* \left(\frac{1}{1-\hat{\zeta}}, \frac{\hat{\rho}-1}{\left(1-\hat{\zeta}\right)^2}\right)' \tag{3.15}$$

3.6. Simulations

In order to accurately estimate ζ, we first estimated π and ρ from object-level data, then used $\hat{\pi}$ and $\hat{\rho}$ to estimate ζ. Ideally, the data would be simulated in a similar fashion, where object-level data were simulated from a beta-binomial distribution with a given π and ρ, and were then used to generate nested-level data. Methods currently exist to generate random observations from a beta-binomial distribution [51]. However, this method assumes a constant ρ among all observations, which is clearly not the case when presented with nested categorical data. As a a result, object-level data cannot be generated first as objects are no longer independent in the presence of nested-level correlations.

Instead, a method needs to be used that has the capability to generate correlated binomial outcomes with the specified nested-level correlation matrix that will provide object-level estimates. As a result, we chose to simulate data from a multivariate normal distribution and dichotomize the outcome vectors according to the method of Emrich and Piedmonte [20]. While these random observations are not generated from the same distribution as assumed by the theory thus far, the inference on the correlation parameters should provide similar results. For a given number of nested-level clusters ($n_{\cdots}$), simulate the number of responses ($x_{i\cdot}$) per rater within a cluster

for a given π, m_i, ρ and ζ, then proceed to estimate π, ρ and ζ as previously outlined. As documented by Demirtas et. al. [18], this method does not perform as well on inference about the correlation as the methods of Poisson sums [41] or archemedian copulas [35]. However, the method of Emrich and Piedmonte allows for the specification of a nested-level type correlation matrix, suiting our needs, while the other two methods are restricted to assuming a common correlation among all observations and are not suited for nested-level analysis. Therefore, the simulations will show that the nested-level maximum profile-likelihood method works well despite the method of simulation. After performing these simulations, we found that this method appears to be asymptotically unbiased for almost all results. We see that the coverage of the confidence interval is closest to 95% when π=0.5 as opposed to π=0.3, which is expected since 0.5 is further from the boundary of π than 0.3. In addition, we found the initially surprising result that, when holding the number of objects per nested-level constant, increasing the number of raters per object actually hurts the performance of both the nested-level ICC estimator and the corresponding confidence interval. However, the theory supports this finding as described earlier.

Table 2 shows a larger range of simulations carried out in this manner. This simulation method works well when the correlations and prevalence are not near the [0,1] boundary. The simulation appears to perform better as either ζ or $\pi \to 0.5$. In addition, the simulation confirms that the method works best as $m_i \to 0$ as previously discussed.

Figure 3.2: Simulation Results for $\zeta = 0.5$, $\rho = 0.7$, $\pi = 0.5$

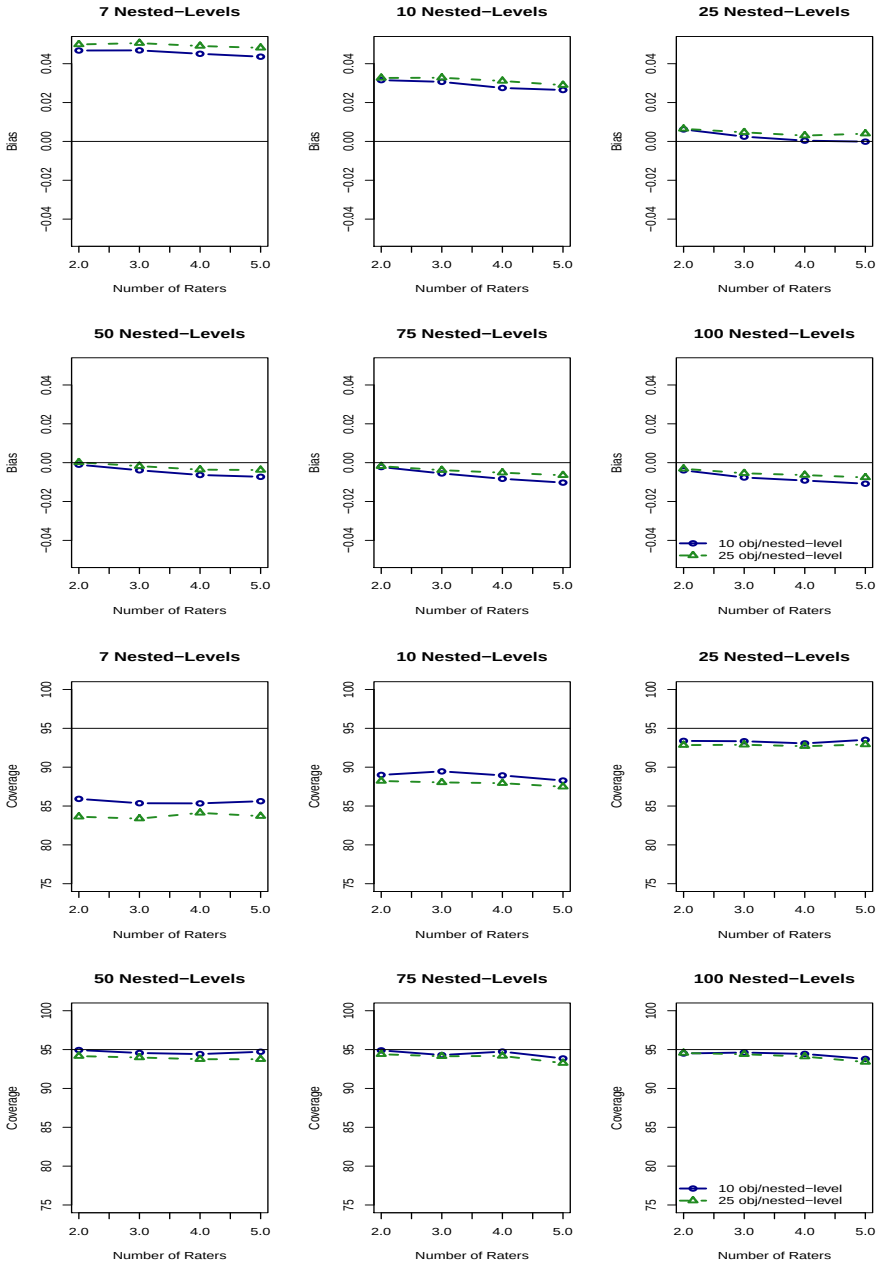


Table 3.3: Simulation Results for $\pi = 0.3$

		ρ	.5			.7		
		ζ	.1	.3	.5	.1	.3	.5
$n_{...}$	$n_{i..}$	$n_{ij.}$						
	10	2	.00(91.2)	-.01(91.0)	-.02(92.3)	.00(91.2)	-.01(90.9)	-.01(91.8)
25		5	.00(90.6)	-.01(90.1)	-.02(91.7)	.00(90.4)	.00(89.8)	.00(91.0)
	25	2	.00(89.6)	-.01(89.1)	-.02(90.1)	.00(89.7)	-.01(89.2)	-.01(89.6)
		5	-.01(88.7)	-.02(88.4)	-.02(89.5)	.00(88.3)	-.01(88.6)	-.01(89.2)
	10	2	.00(93.2)	-.01(92.6)	-.01(93.0)	.00(92.7)	.00(92.0)	.00(93.2)
50		5	.00(93.2)	-.01(92.7)	-.01(93.3)	.00(92.7)	.01(92.0)	.00(93.1)
	25	2	.00(92.2)	-.01(90.9)	-.01(91.5)	.00(92.0)	-.01(91.0)	.00(91.6)
		5	.00(91.6)	-.01(91.1)	-.01(91.0)	.00(91.6)	.00(91.2)	.00(91.6)
	10	2	.00(92.9)	-.01(92.7)	-.01(93.6)	.00(93.0)	.00(92.4)	.00(93.3)
75		5	.00(93.0)	.00(92.2)	-.01(94.0)	.00(92.4)	.01(91.8)	.01(92.7)
	25	2	.00(92.5)	.00(91.1)	-.01(91.8)	.00(92.5)	.00(91.2)	.00(92.1)
		5	.00(92.7)	.00(90.4)	-.01(91.5)	.00(92.3)	.00(91.1)	.00(91.1)
	10	2	.00(94.0)	.00(93.1)	.00(93.9)	.00(93.7)	.00(93.5)	.00(93.5)
100		5	.00(93.9)	.00(94.2)	.00(93.8)	.01(93.1)	.01(92.6)	.01(92.7)
	25	2	.00(93.6)	-.01(91.6)	.00(92.6)	.00(93.3)	.00(92.0)	.00(92.0)
		5	.00(92.7)	-.01(91.7)	.00(92.2)	.00(92.7)	.00(92.1)	.00(91.9)

Note: Results are presented as "Bias(Coverage of 95% Confidence Interval)"

ζ=nested-ICC, ρ=object-ICC, π=prevalence, $n_{...}$=number of nested-effects, $n_{i..}$=number of objects, $n_{ij.}$=number of raters

Table 3.4: Simulation Results for $\pi = 0.5$

		ρ	.5			.7		
		ζ	.1	.3	.5	.1	.3	.5
$n_{...}$	$n_{i..}$	$n_{ij.}$						
	10	2	.00(90.9)	-.01(92.4)	-.01(94.5)	.00(91.1)	-.01(91.9)	-.01(93.4)
25		5	-.01(90.2)	-.01(92.8)	-.01(94.8)	.00(90.1)	.00(91.5)	.00(93.5)
	25	2	.00(90.7)	-.01(91.8)	-.01(93.1)	.00(90.6)	-.01(91.8)	-.01(92.8)
		5	-.01(89.3)	-.01(91.2)	-.01(93.2)	-.01(89.2)	-.01(91.3)	.00(92.9)
	10	2	.00(92.6)	.00(94.5)	.00(95.3)	.00(92.0)	.00(94.3)	.00(94.9)
50		5	.00(93.3)	.00(94.7)	.00(95.9)	.00(92.8)	.01(93.2)	.01(94.7)
	25	2	.00(92.6)	.00(94.0)	.00(94.4)	.00(92.4)	.00(93.6)	.00(94.2)
		5	.00(92.9)	.00(94.5)	.00(95.8)	.00(92.3)	.00(93.6)	.00(93.8)
	10	2	.00(93.3)	.00(94.0)	.00(95.4)	.00(93.1)	.00(94.1)	.00(94.9)
75		5	.00(93.3)	.00(94.3)	.00(95.8)	.00(92.0)	.01(92.7)	.01(93.9)
	25	2	.00(93.2)	.00(93.9)	.00(94.5)	.00(93.0)	.00(94.1)	.00(94.4)
		5	.00(93.8)	.00(93.8)	.00(94.9)	.00(93.2)	.00(93.7)	.01(93.3)
	10	2	.00(94.4)	.00(95.0)	.00(95.9)	.00(93.5)	.00(93.8)	.00(94.5)
100		5	.00(94.3)	.00(95.6)	.00(96.1)	.00(93.3)	.01(93.3)	.01(93.8)
	25	2	.00(94.3)	.00(94.7)	.00(95.0)	.00(93.9)	.00(94.3)	.00(94.6)
		5	.00(93.5)	.00(94.8)	.00(95.1)	.00(93.4)	.00(93.9)	.01(93.4)

Note: Results are presented as "Bias(Coverage of 95% Confidence Interval)"

ζ=nested-ICC, ρ=object-ICC, π=prevalence, $n_{...}$=number of nested-effects, $n_{i..}$=number of objects, $n_{ij.}$=number of raters

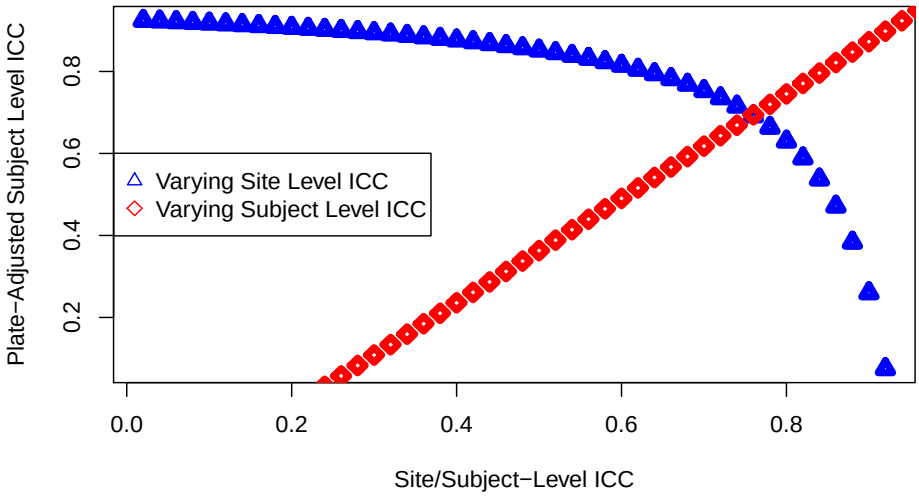
3.7. Nested-Level Agreement within a GWAS

When analyzing the GWAS data presented earlier, it was demonstrated that there was quite a bit of correlation among Hispanic subjects within a genetic plate. However, at the time of the analysis, no methods existed to accurately assess the correlation due to genetic plate nor did a method exist to obtain the standard error surrounding that estimate. Using the methods presented in this paper, this parameter was able to be estimated for all four ethnicities analyzed in the GWAS data (NH-White, NH-Black, Hispanic and Other). These estimates (with their corresponding standard errors and 95% confidence intervals) are displayed in table 3.5.

Table 3.5: Levels of Agreement among Ethnicity Responses in a GWAS Reanalyzed

Alignment	Response Category	Avg. Prop. ($\hat{\pi}$)	2-level Model Subject-ICC (SE) ($\hat{\rho}$)	3-level Model Davis Plate-ICC ($\hat{\zeta}$) (SE)	95% CI	Adj S-ICC
Original PIDs	1:NH-White	0.433	0.929 (0.006)	0.114 (0.025)	[.0650, .1629]	**0.920**
	2:NH-Black	0.420	0.949 (0.005)	0.063 (0.016)	[.0330, .0938]	**0.946**
	3:Hispanic	0.109	0.919 (0.011)	0.215 (0.032)	[.1530, .2769]	**0.897**
	4:Other	0.037	0.682 (0.032)	0.001 (0.003)	[-.0043, .0061]	**0.682**
Re-aligned PIDs	1:NH-White	0.433	0.963 (0.005)	0.114 (0.025)	[.0652, .1635]	**0.958**
	2:NH-Black	0.420	0.986 (0.003)	0.063 (0.016)	[.0330, .0939]	**0.985**
	3:Hispanic	0.109	0.926 (0.010)	0.215 (0.032)	[.1531, .2770]	**0.906**
	4:Other	0.037	0.693 (0.032)	0.001 (0.003)	[-.0043, .0060]	**0.693**

Figure 3.3: Potential Effects of Varying Plate or Subject-Level ICC



From this analysis, the re-aligning of the PIDs resulted in better agreement between self-reported and genetically-inferred ethnicity. Among Hispanics it is clear that there is a nominal level of plate-level agreement among observations on different subjects that should be uncorrelated (ζ=0.215, 95% CI [.1530, .2769] using the original PIDs), significantly different than zero when $\alpha = 0.05$. The resulting plate-adjusted ICC among Hispanics was 0.897, which is only a 0.022 difference from the originally reported subject-level ICC. However, the magnitude of this difference is attributable solely to the high estimate of subject-level agreement among Hispanics. Figure 3.3 shows how the adjusted ICC could be affected by varying the levels of subject or plate-level ICC. As is apparent from equation 3.14, given a constant object-level ICC, the adjusted object-level ICC is a non-linear function of the nested-level ICC with an x-intercept equal to ρ and $\zeta \to 1$ as $\rho \to 0$. On the other hand, given a constant plate-level ICC, the adjusted ICC is a linear function of the object-level ICC with an x-intercept equal to ζ and $\rho \to 1$ as $\zeta \to 1$. Clearly from the potential outcomes of

the adjusted level of agreement, higher levels of nested-level agreement or lower levels of object-level agreement will both reduce the nested-level adjusted object-level ICC.

3.8. Immediate Extension

While this worked focused on only one level of nesting, this method can handle more than one level of nesting. For example, consider another level of nesting creating a first nested-level and second nested-level. Then, three separate correlations would need to be taken into consideration. ρ and ζ would retain their definitions of the object and first nested-level correlations, and a second nested-level correlation ω would be introduced as the measure of agreement among objects within the same second nested-level but in different first nested-levels. In this scenario, similar steps would be followed to first estimate π and ρ, then estimate ζ, and finally estimate ω. Similarly, ζ would be artificially inflated by ω and would have the range $[\omega,1]$. The same correction could be applied to obtain an estimate of ζ adjusted for ω.

3.9. Conclusion

3.9.1. Summary

In this paper, a nested-level profile-likelihood method was presented to estimate the level of agreement that exists within a nested-level factor. In most cases, this correlation parameter should be equal to zero, but in the case that it is not, we have proven that the level of correlation can positively bias the reported object-level ICC. Using profile-likelihood theory (and asymptotic maximum-likelihood theory) we were able to develop a consistent estimate of ζ and provide an α-level Wald type confidence interval around the estimate.

Most importantly, this method demonstrates the necessity of appropriate study plan-

ning when examining levels of agreement. We are accustomed to planning for levels of bias in other studies by stratifying randomizations based on confounding parameters or matching covariates across two different treatment groups, but rarely is the necessary planning applied to studies where levels of agreement are the primary focus to ensure that these nested-level biases do not exist. In the GWAS example presented earlier, perhaps blocking the genotyping plates by patient-reported race would have reduced the plate-level agreement among Hispanics and allowed for appropriate subject-level agreement without adjusting for the plate-level agreement. Clearly this method can adjust object-level agreement for the case where nested-level agreement is non-negligible, however, unless the nested-level agreement is of interest, it is best that this agreement is reduced as much as possible at the planning stage of the study.

3.9.2. Future Work

Currently this work is limited to the case where binomial outcomes are possible. However, even in the case of the GWAS presented above, the data are truly captured in a multinomial fashion (Non-Hispanic Whites, Non-Hispanic Blacks, Hispanics, Others). In order to analyze the data, we dichotomized each response to a binary yes-no answer. Using methods similar to those presented in Bartfay et. al. [2], this method should be able to be extended to capture the nested-level agreement among multinomial responses. In addition, an ICC adjusted for the presence of adjusted nested-level agreement should be derived. Subsequently, there is work to be done to examine the sample-size and power consequences that result in assessing the object-level agreement in the presence of nested-level agreement.

CHAPTER 4

On the Nested-Level Intraclass Correlation Coefficient for Multinomial Data

4.1. Introduction

Nested-level methods have been developed to analyze data in many situations. There is significant potential to increase statistical power when combining results from different nested-levels, however appropriate considerations must be made to ensure that results from various nested-levels are sufficiently homogeneous to be combined. Often times, tests are developed to verify these assumptions, such as the Breslow-Day Test [6] to verify the assumption of homogeneity of odds ratios across independent 2x2 categorical contingency tables to validate the use of the corresponding Cochran-Mantel-Haenszel test. While consideration is often given to this important concept, until recently appropriate focus has not been paid to this phenomenon in the area of rater agreement.

Analyzing agreement among multiple raters on the same object has been generally considered without taking into account the potential effect of nested-level effects that may bias the estimates of object-level agreement. Historically, multiple ratings on the same objects would be analyzed using any of a number of methods to assess the level of agreement among raters, most often using kappa statistics or the intraclass correlation coefficient. Chapter 3 described how to estimate agreement on object-level binary data accounting for a confounding nested-level of agreement using the beta-binomial distribution to model the corresponding ICC. However, many situations arise where raters are asked to provide assessments based on a multinomial scale instead of a

binary scale. Bartfay and Donner [2] and Chen et. al. [9] provided a framework using the multinomial-Dirichlet distribution to model multinomial outcomes and estimate the corresponding pooled ICC. In order to employ their method, however, the assumption of equivalent ICCs across responses must be verified. Chapter 2, as well as Chen et. al.[9], provided likelihood-ratio tests to test the assumption of homogeneity of ICCs across binary responses to verify the assumptions needed for the pooled ICCs. This paper will demonstrate that these frameworks can be further modified to model the nested-level agreement for multinomial outcomes and provide a nested-level adjusted object-level ICC. In addition, under the assumption of either homogeneous object-level ICCs or in the presence of large number of objects per nested-level, a test of homogeneity for nested-level ICCs will be provided. Finally, these methods will be used to reanalyze the GWAS study originally presented in Chapter 3 to provide insight into the measure of nested-level agreement among all responses across races.

4.2. Notation

This methodology generally analyzes repeated ratings on two separate populations referred to as "objects", the items being rated, and "nested-levels", a grouping level that could be applied to all "objects" such that each "object" has a unique identification (and is therefore nested) within a "nested-level". For example, in the case of a diagnostic imaging agent, the "object" would be an image obtained from a patient that would be interpreted multiple times. Each interpreter is hereto referred to as a "rater" who provides a qualitative assessment for each "object". Theses "objects" (images) could then be grouped into the healthcare facilities where they were obtained such that each "object" comes from one and only one "nested-level". These examples are by no means exhaustive and many scenarios could be considered where "objects", "nested-levels" and "raters" take different forms.

Let y_{hijk} be a binary outcome (0 or 1) for the k^{th} rater ($k = 1, ..., n_{ij\cdot}$) on the j^{th} object ($j = 1, ..., n_{i\cdot\cdot}$) in the i^{th} nested-level ($i = 1, ..., n_{\cdot\cdot\cdot}$) on the h^{th} trait ($h = 1, ..., n_h$), and let $\mathbf{y}$ be the vector of all responses. y_{hijk} is assumed to follow a binomial distribution where $E(y_{hijk}) = \pi_h$ and $Var(y_{hijk}) = \pi_h(1-\pi_h)$. Let π_h be the proportion of objects with trait h being assessed such that $P(y_{hijk} = 1) = \pi_h$ and let $\mathbf{p}$ be the $n_h \times 1$ vector of all possible proportions. Let ρ_h be the object-level intraclass correlation coefficient and ζ_h be the nested-level intraclass correlation coefficient for the h^{th} response. Let $\rho_\cdot$ and $\zeta_\cdot$ be the overall object-level and nested-level intraclass correlation coefficients. Given $n_{ij\cdot}$ ratings per object, the sum of all ratings for an object within a given response can be written as $x_{hij} = \sum_{k=1}^{n_{ij\cdot}} y_{hijk}$ where $\mathbf{x_{ij\cdot}}$ is the vector containing all such results for each object, while the sum of all responses for a given outcome within a nested-level can be written as $x_{hi\cdot} = \sum_{j=1}^{n_{i\cdot\cdot}} \sum_{k=1}^{n_{ij\cdot}} y_{hijk}$ with the vector $\mathbf{x_{i\cdot\cdot}}$ containing all such results for each nested-level. Let $m_i = \left(\sum_{j=1}^{n_{i\cdot\cdot}} \binom{n_{ij\cdot}}{2} \right) / \binom{\sum_{j=1}^{n_{i\cdot\cdot}} n_{ij\cdot}}{2}$, which can be interpreted as the proportion of area of the upper diagonal of the correlation matrix contributed to by the object-level ICC. Let $d_i = m_i\rho_\cdot + (1 - m_i)\zeta_\cdot$, which is the weighted average of all pair-wise object-level and nested-level correlations within nested-level i.

4.3. Distributions for Overdispersed Multinomial Data

4.3.1. Multinomial-Dirichlet Distribution: Object-Level Results

For a given set of multinomial data $\mathbf{y}$ with n_h categories per response, a Dirichlet distribution can be assumed as the prior distribution for the probability of response for each category and a multinomial likelihood for the response vector. By invoking Bayes' rule, one obtains the multinomial-Dirichlet distribution (MDD)[38]. Let $\mathbf{Z} = (z_1, z_2 \dots z_{n_h})$ be the vector of parameters that describe the MDD. Then for the j^{th}

object in the i^{th} nested-level, the MDD can be written as

$$P\left(\mathbf{X_{ij\cdot}} = \mathbf{x_{ij\cdot}}|\mathbf{Z}\right) = \frac{N!}{\prod_{a=1}^{n_h} x_{aij}!} \frac{\Gamma\left(\sum_{a=1}^{n_h} z_a\right)}{\Gamma\left(N + \sum_{a=1}^{n_h} z_a\right)} \prod_{a=1}^{n_h} \frac{\Gamma\left(x_{aij} + z_a\right)}{\Gamma\left(z_a\right)} \tag{4.1}$$

Given that Chen [10] demonstrates that the MDD models the pooled object-level correlation $\rho_{\cdot} = \left(\sum_{a=1}^{n_h} z_a + 1\right)^{-1}$ and item-wise response rate $\pi_i = \frac{z_i}{\sum_{a=1}^{n_h} z_a}$, it can be shown that $z_i = \pi_i \frac{1-\rho_{\cdot}}{\rho_{\cdot}} \forall i$ and $\sum_{a=1}^{n_h} z_a = \rho^{-1} - 1$. Using the identity $log\left(\Gamma(z)\right) + log(z) = log\left(\Gamma(z+1)\right)$, the ratio of log-gamma functions can be specified as

$$\frac{log\left(\Gamma(A+B)\right)}{log\left(\Gamma(B)\right)} = \sum_{C=1}^{A-B} log\left(B + C - 1\right) \tag{4.2}$$

Then, the corresponding log-likelihood for object-level results can be written as

$$\begin{aligned} LogL\left(\mathbf{Z}|\mathbf{X_{ij\cdot}} = \mathbf{x_{ij\cdot}}\right) = \sum_{a=1}^{n_{\cdots}} \sum_{b=i}^{n_{a\cdot\cdot}} \Bigg[& log\left(\frac{n_{ab\cdot}!}{\sum_{q=1}^{n_h} x_{qab}!}\right) - \sum_{c=1}^{n_{ab\cdot}} log\left(c + \frac{1-\rho_{\cdot}}{\rho_{\cdot}} - 1\right) \\ & + \sum_{d=1}^{n_h} \sum_{f=1}^{x_{dab}} log\left(f + \frac{1-\rho_{\cdot}}{\rho_{\cdot}} \pi_d - 1\right) \Bigg] \end{aligned} \tag{4.3}$$

This likelihood can be directly maximized to obtain maximum likelihood estimates of each π_i and $\rho_{\cdot}$. In addition, the standard error of each can be found by inverting the negative of the information matrix appropriately for each parameter, the details of which can be found elsewhere [42].

4.3.2. Multinomial-Dirichlet Distribution: Nested-Level Results

Generally, only the object-level of agreement is considered to be important when looking at reliability of a set of data. Researchers are often interested only in how well raters agree when looking at the same object and not the same set of objects

within the same nested-level. However, this second level of information can be important to identify additional levels of bias that may artificially inflate the object-level agreement.

If $\mathbf{y_{ij}}$. were distributed according to the multinomial-Dirichlet distribution, $E(\mathbf{y_{ij}}.) = n_{ij}.\mathbf{p}$ and $Var(\mathbf{y_{ij}}.) = n_{ij}.\left[1 + (n_{ij}. - 1)\,\rho.\right](\mathrm{DIAG}\,(\mathbf{p}) - \mathbf{pp}')$ [40]. Consider the covariance matrix for a set of object-level responses. Given $n_{ij}.$ ratings for the j^{th} object in the i^{th} nested-level, the vector of all responses for the object will have an $n_{ij}.$ x $n_{ij}.$ dimension correlation matrix in the form of

$$\Sigma_{ij}. = \begin{bmatrix} 1 & \rho. & \dots & \dots & \rho. \\ \rho. & \ddots & \dots & \dots & \vdots \\ \vdots & \dots & 1 & \dots & \vdots \\ \vdots & \dots & \dots & \ddots & \rho. \\ \rho. & \dots & \dots & \rho. & 1 \end{bmatrix} \tag{4.4}$$

Ideally, two results from the same nested-level that are not from the same object should be uncorrelated. However, there may be situations where these results are correlated. In this case, the nested-level, rather than the object, should be considered the true cluster as clustering on the object-level does not account for the correlation that exists among separate objects in a nested-level. Let $\mathbf{1}_{i..}$ be a $\sum_{j=1}^{n_{i..}} n_{ij}. \times \sum_{j=1}^{n_{i..}} n_{ij}.$ matrix with all matrix elements equal to 1, and $\mathbf{1}_{ij}.$ be a $n_{ij}. \times n_{ij}.$ matrix with all matrix elements equal to 1. Then the correlation matrix for nested-level i can be written as

$$\Sigma_{i..} = DIAG(\Sigma_{i1\cdot} - \zeta.\mathbf{1}_{i1\cdot}, \Sigma_{i2\cdot} - \zeta.\mathbf{1}_{i2\cdot}, ..., \Sigma_{in_{i..}\cdot} - \zeta.\mathbf{1}_{in_{i..}\cdot}) + \zeta.\mathbf{1}_{i..} \tag{4.5}$$

Under the logic laid out in Chapter 3, the nested-level ICC describes the level of agreement that exists among separate observations within the same nested-level that are generally considered to be independent, such as separate objects or separate subjects. At the nested level, the moments of the multinomial-Dirichlet distribution can be written as $E(y_{i\cdot\cdot}) = n_{i\cdot\cdot}\mathbf{p}$ and

$$\begin{aligned} Var(y_{i\cdot\cdot}) &= \left(\sum_{j=1}^{n_{i\cdot\cdot}} n_{ij\cdot}\right)\left[1 + \left(\left(\sum_{j=1}^{n_{i\cdot\cdot}} n_{ij\cdot}\right) - 1\right)\left(m_i\rho_\cdot + (1-m_i)\,\zeta_\cdot\right)\right]\left(\mathrm{DIAG}\left(\mathbf{p}\right) - \mathbf{p}\mathbf{p}'\right) \\ &= \left(\sum_{j=1}^{n_{i\cdot\cdot}} n_{ij\cdot}\right)\left[1 + \left(\left(\sum_{j=1}^{n_{i\cdot\cdot}} n_{ij\cdot}\right) - 1\right) d_i\right]\left(\mathrm{DIAG}\left(\mathbf{p}\right) - \mathbf{p}\mathbf{p}'\right) \end{aligned}$$

which models the overall nested-level ICC as a linear combination of the object and nested-level ICCs. Therefore, the corresponding log-likelihood can be expressed as:

$$\begin{aligned} LogL\left(\mathbf{Z}|\mathbf{X}_{\mathbf{i}\cdot\cdot} = \mathbf{x}_{\mathbf{i}\cdot\cdot}\right) = \sum_{a=1}^{n_{\cdot\cdot\cdot}} \Bigg[log\left(\frac{n_{a\cdot\cdot}!}{\prod_{q=1}^{n_h} x_{qa\cdot}!}\right) - \left[\sum_{c=1}^{n_{a\cdot\cdot}} log\left(c + \frac{1-d_i}{d_i} - 1\right)\right] + \\ \left[\sum_{e=1}^{n_h}\sum_{f=1}^{x_{ea\cdot}} log\left(f + \frac{1-d_i}{d_i}\pi_e - 1\right)\right]\Bigg] \end{aligned} \tag{4.6}$$

The nested-level ICC describes the level of agreement that exists among separate observations within the same nested-level that are generally considered to be independent, such as separate objects or separate subjects. The pooled ICC under the assumption that the ICC is constant across separate responses can be accurately estimated using the multinomial-Dirichlet distribution. Therefore, in order to estimate the nested-level ICC, the following steps can be used:

1. Estimate $\mathbf{p}$ using the estimate derived from the multinomial distribution
2. Estimate $\rho_\cdot$ using the object-level multinomial-Dirichlet distribution
3. Test whether homogeneity of ICCs exist among object-level responses

4. If homogeneity of object-level ICCs exists, using the estimates from the two steps above and the profile-likelihood for the nested-level data, estimate the overall nested-level ICC

The maximum profile-likelihood estimate for $\zeta_{\cdot}$ can be found by finding $\hat{\zeta}$ that sets the score equation of the profile-likelihood with respect to ζ to zero. The standard error can be calculated using the second derivative of the log-likelihood with respect to ζ, resulting in

$$SE\left(\hat{\zeta}\right) = \left[-1 \cdot \frac{\partial^2 LogL\left(\mathbf{Z}|\mathbf{X_{i\cdot\cdot}} = \mathbf{x_{i\cdot\cdot}}\right)}{\partial \zeta^2}\right]^{-\frac{1}{2}} \tag{4.7}$$

Due to the asymptotics associated with maximum-likelihood estimates, an α level Wald confidence interval can be constructed around the point estimate using $Z_{\alpha/2}$, which is the $(1-\alpha)/2$ quartile of the standard normal distribution. The confidence interval can then be calculated as $\hat{\zeta}_{\cdot} \pm Z_{\alpha/2} \times SE\left(\hat{\zeta}\right)$[8]. The explicit formulas for the estimate and standard error of the overall nested-level ICC can be found in Appendix A.3.

If the nested-level ICC for each trait are either assumed or proven to be different, there is currently no distribution that jointly models separate ICCs for each trait. However, in this case, it may be appropriate (as the multiple beta-binomial distribution would imply) that each response could be analyzed separately and distinct estimates, standard errors and confidence intervals for each separate outcome can be obtained by dichotomizing each result and using the methodology described in Chapter 3 to determine the separate measures of agreement for each response.

4.3.3. Multiple Beta-Binomial Distribution

As described in Chapter 2 and by Chen [9], the multiple beta-binomial distribution (MBBD) is a non-unique decomposition of the multinomial-Dirichlet distribution under the assumption that the ICC among all responses are equivalent. If that assumption does not hold, the MBBD can be used to specify the joint distribution of correlated multinomial responses by writing the likelihood as a product of successive conditional beta-binomial distributions. Define $\sum_{i=m}^{n} z_i = 0$ where $n < m$ and let $\mathbf{A} = (\alpha_1, \alpha_2, ...\alpha_{n_h-1})$ and $\mathbf{B} = (\beta_1, \beta_2, ...\beta_{n_h-1})$ be the vectors of parameters that describe each conditional beta-binomial distribution. Then, the multiple beta-binomial distribution (MBBD) can be written as

$$P\left(\mathbf{X_{ij\cdot}} = \mathbf{x_{ij\cdot}} | \mathbf{A}, \mathbf{B}\right) = \prod_{f=1}^{n_h-1} \binom{N - \sum_{g=1}^{f-1} x_{gij}}{x_{fij}} \times \frac{\Gamma\left(x_{fij} + a_i\right) \Gamma\left(N - \sum_{g=1}^{f} x_{gij} + b_i\right) \Gamma\left(a_i + b_i\right)}{\Gamma\left(N - \sum_{g=1}^{f-1} x_{gij} + a_i + b_i\right) \Gamma\left(a_i\right) \Gamma\left(b_i\right)} \tag{4.8}$$

The MBBD can not be used to describe the unconditional ICCs for each response and can only accommodate the conditional responses, and is therefore better used for a goodness of fit test rather than to estimate the ICC for each specific outcome. The nested-level MBBD is analogous to the object-level MBBD described in Chapter 2, however instead of modeling the object-level correlation, it models the linear combination of object and nested-level correlations, d_i, as described in the nested-level MDD. Estimation of this correlation is carried out the same way as its object-level counterpart, however it is performed at the nested-level instead of the object-level. For more details surrounding the MBBD, see Chapter 2 and Chen [9].

4.4. Goodness-of-Fit Testing

4.4.1. Testing Homogeneity of ICCs

As mentioned earlier, the Dirichlet-multinomial distribution is a special case of, and is nested within, the multiple beta-binomial distribution. Therefore, a likelihood ratio test can be used to test the goodness of fit for separate levels of agreement against a pooled level of agreement. Chapter 2 describes the method of testing for homogeneity of the ICC across multinomial responses by comparing the goodness of fit of the multinomial-Dirichlet distribution to the more flexible multiple beta-binomial distribution using the likelihood ratio test comparing the likelihoods of the two models. This test can be extended to test for homogeneity of the nested-level ICC. If the object-level ICC is found to be homogeneous across results, the multiple beta-binomial distribution can be used to demonstrate the goodness-of-fit of both the object-level and nested-level ICC. In the case that homogeneity of ICCs is observed, the MBBD will devolve into the MDD, which can also be used to evaluate the goodness-of-fit of the model. If the homogeneity of the object-level results is not observed, the assumptions for the MBBD are violated and each response should be analyzed separately according to the beta-binomial distribution.

The benefit of using this test in this setting is that the likelihood-ratio test is still valid using profile-likelihoods[17]. Therefore, one can use the methods presented in Chapter 2 to test for homogeneity of the nested-level ICC using the following steps:

1. Estimate $\rho_{\cdot}$ from the object-level data and test for homogeneity of ICCs across multiple responses of the multinomial object-level data

2. If the object-level ICCs are found to be homogeneous, model the profile likeli-

hood based on the ICCs outlined in (1) using $\hat{\rho}_{\cdot}$.

3. Test for homogeneity of nested-level ICCs in a similar fashion as the object-level method

For n_h responses, there are $n_h!/2$ unique decompositions of the MBBD that could potentially model the data, each with its own set of parameters. In order to model the nested-level agreement, each one of these must be considered. In the case that there is homogeneity of ICCs across the object-level responses, the nested-level MBBD can be used to examine the data for potential heterogeneity of nested-level ICCs. The conditional beta-binomial distribution can instead be written in terms of the probability of response h, $\pi_{h|1,2,\ldots,h-1}$, conditional object-level ICC for response h, $\rho_{h|1,2,\ldots h-1}$, the conditional nested-level agreement, $\zeta_{h|1,2,\ldots h-1}$, and $m_{i,h|1,2,\ldots,h-1}$, which is the proportion of the correlation matrix conditional beta-binomial distribution occupied by $\rho_{h|1,2,\ldots h-1}$. Then, let $d_{h|1,2,\ldots,h-1} = m_{i,h|1,2,\ldots,h-1}\rho_{h|1,2,\ldots h-1}+(1-m_{i,h|1,2,\ldots,h-1})\zeta_{h|1,2,\ldots h-1}$. Let $\mathbf{C}$ be the set of all possible conditional probabilities of response and $\mathbf{D}$ be the set of all possible conditional overall nested level ICCs in the form of $d_{h|1,2,\ldots,h-1}$. Then the nested-level MBBD can be written as

$$P\left(\mathbf{X}_{\mathbf{i}\cdot\cdot} = \mathbf{x}_{\mathbf{i}\cdot\cdot}|\mathbf{C},\mathbf{D}\right) = \prod_{f=1}^{n_h-1}\left[\prod_{a=1}^{x_{fi\cdot}}\left(a+\frac{\left(1-d_{f|1\ldots f-1}\right)\pi_{f|1\ldots f-1}}{d_{f|1\ldots f-1}}-1\right)\times \prod_{a=1}^{n_{i\cdot\cdot}-\sum_{g=1}^{f}x_{gi\cdot}}\left(a+\frac{\left(1-d_{f|1\ldots f-1}\right)\left(1-\pi_{f|1\ldots f-1}\right)}{d_{f|1\ldots f-1}}-1\right)\times \prod_{a=1}^{n_{i\cdot\cdot}-\sum_{g=1}^{f-1}x_{gi\cdot}}\left(a+\frac{1-d_{f|1\ldots f-1}}{d_{f|1\ldots f-1}}-1\right)^{-1}\right]$$

In this expression of the model, d_i can be considered the overall measure of agreement among all responses in a nested-level, which is a linear combination of the overall

object-level and nested-level ICCs. Under the profile-likelihood framework, the pooled object-level ICC is considered to be known as it was previously tested and found to be homogeneous. Therefore, testing the goodness of fit of the MDD given parameter d_i is analogous to testing the goodness of fit of the same model for the nested-level ICC. Therefore, this profile-likelihood can then be used to test the assumption of the homogeneity of ICCs across nested-level responses. Under the assumption that $\zeta_1 = \zeta_2 ... = \zeta.$, fewer parameters are needed to model the distribution than if the flexibility were allowed such that at least one $\zeta_i \neq \zeta_j$. For the MBBD, it has been documented that in the object-level case, the conditional object-level ICC can be specified as $\frac{1}{a_h+b_h+1} = \frac{1}{\sum_{i=h}^{n_h} z_i+1}$ for response h conditional on responses 1,2...$n_h - 1$. As a result, in the nested-level case, $d_{h|1,2...h-1}$ can be exactly specified in the same way. Given that the estimates of the object-level ICC are retained from the previous model, the nested-level ICC can be expressed under the homogeneity assumption as

$$\zeta_{h|1,2...h-1} = \frac{d_{h|1,2...h-1} - m_{i,h|1,2,...,h-1}\rho_{h|1,2...h-1}}{1 - m_{i,h|1,2,...,h-1}} \tag{4.9}$$

Thus, the following test of hypotheses can occur:

$$\begin{cases} H_0 : & \zeta_1 = \zeta_2 = ... = \zeta. \\ H_A : & \zeta_1 \neq \zeta. \text{ or } \zeta_2 \neq \zeta. \text{ or } ...\zeta_k \neq \zeta. \end{cases}$$

As object-level homogeneity is assumed, the likelihood under the null model will be the same regardless of the decomposition while the likelihood under the alternative would continue to yield separate likelihoods. Then, each alternative likelihood can be compared to its corresponding null likelihood yielding the test statistic $\psi = 2log\frac{L_{MBBD}}{L_{MDD}}$ which follows a $\chi^2_{n_h-2}$distribution [8, 17]. Given the number of tests

considered, the multiple decompositions of the likelihood can artificially inflate the type I error rate if not appropriately controlled for. Therefore, multiple comparisons methods such as the Bonferroni-Holms[26] or the Benjamini-Hochberg[3] methods can be employed as appropriate depending on whether controlling the family-wise error rate or the false-discovery rate is of more importance. The Benjamini-Hochberg correction is universally more powerful than the Bonferroni-Holms correction, however does not maintain strong control of the family-wise type I error rate. Therefore, the balance between increased power and potentially inflated type I error rate should be considered and the multiple comparison correction should be decided on prior to conducting the test. The effect of these methods on the type I error rate as well as the degree of inflation of the type I error rate due to the multiple tests is described in detail in Chapter 2 and should be taken into consideration when choosing the appropriate method to control the type I error rate for testing the homogeneity of nested-level ICCs.

4.4.2. Asymptotic Considerations

The multinomial-dirichlet distribution can be used to estimate the nested-level agreement by comparing the overdispersion of the variance of the observed data to the expected variance for the multinomial distribution in order to estimate the ICC where d_i, the linear combination of the object and nested-level agreement, represents the total agreement among all responses within a nested level. However, as $m_i \to 0$, as occurs when $n_{i\cdot\cdot} \to \infty$, d_i devolves into simply the nested-level correlation ζ. Therefore, if there is a sufficiently large number of objects per nested-level, or m_i is sufficiently small, then it can be feasible to disregard the effect of heterogeneity among object-level results to provide further inference on the nested-level results. If this assumption can be made, $\hat{\zeta} \approx \hat{d}_i$ and appropriate inference can be made on $\hat{\zeta}$

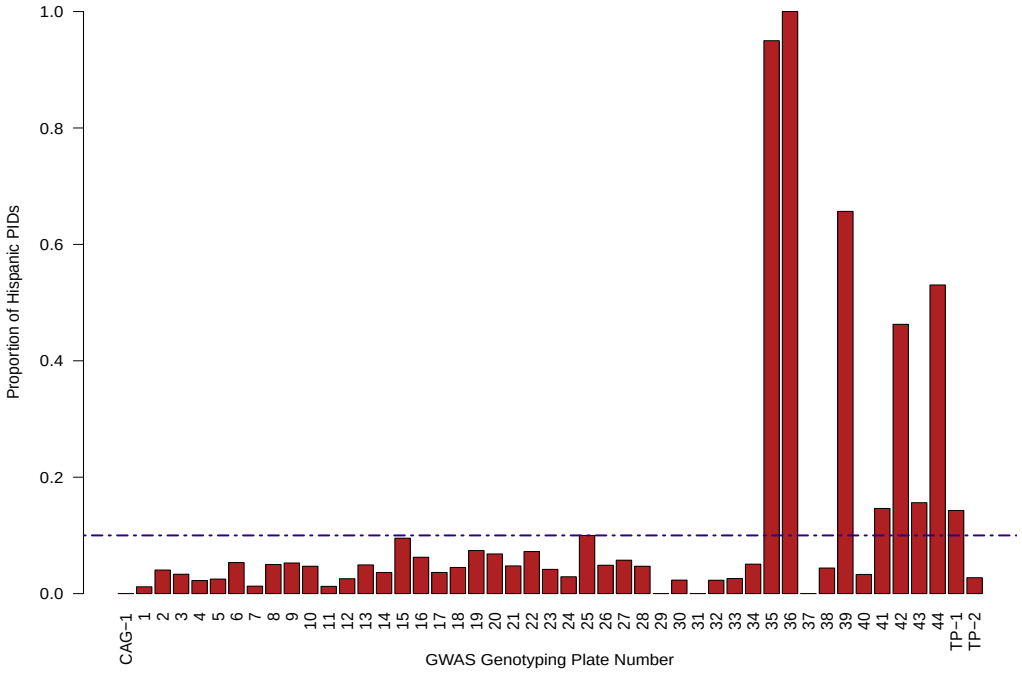
as described in the two-level model presented in Chapter 3. In addition, testing of the homogeneity of the nested-level ICCs can occur as described in Chapter 2 as the likelihood now devolves to a two-level model instead of a three-level model. This is an important benefit as there lies great potential in larger studies where investigation of the nested-level ICC may be of interest for object-level heterogeneity to exist due to the power resulting from the potentially large number of objects in these studies. These asymptotic properties allows for investigation into the nested-level ICCs without consideration of object-level results given a large enough number of objects per nested-level.

4.4.3. Adjusted Object-Level ICC

In the presence of non-zero nested-level ICCs, the unadjusted object-level ICCs can overestimate the measure of agreement that exists among objects. In the binomial case, it was proved in Chapter 3 that for a given object-level correlation ρ_h and nested-level correlation ζ_h, $\rho_h \geq \zeta_h \; \forall \; h$. For the pooled estimate $\zeta_{\cdot}$ to be a valid representation of the nested-level agreement, there must be a demonstration of both object-level and nested-level homogeneity of ICCs among all responses. Therefore, $\rho_1 = \rho_2 = ... = \rho_{\cdot}$ and $\zeta_1 = \zeta_2 = ... = \zeta_{\cdot}$, which therefore necessitates that $\rho_{\cdot} \geq \zeta_{\cdot}$. This reduces the range of the object-level ICC from the expected range of (0,1) to the range ($\rho_{\cdot}$,1). To adjust for the reduced range of the ICC due to the nested-level ICC, the adjusted measure of the object-level ICC, $\rho^*_{\cdot}$, can be derived as

$$\rho^* = \frac{\rho_{\cdot} - \zeta_{\cdot}}{1 - \zeta_{\cdot}} \tag{4.10}$$

Figure 4.1: Distribution of Self-Reported Hispanics by Plate in a GWAS



4.5. Application: "Fingerprinting" within a GWAS

4.5.1. Description

A GWAS conducted within a cohort study led to the troubling discovery that intentionally duplicate genotyping results were paired with totally different subject IDs. Fortunately, within the same clinical research network, a full-scale GWAS (1 million SNPs) was conducted shortly thereafter, and the "fingerprinting" step was used to correctly realign nearly 4% of the subject IDs to their correct genotyping results. Each study participant was classified by self-reported race/ethnicity as

1) Non-Hispanic White; 2) Non-Hispanic Black; 3) Hispanic; and 4) Other. Further analyzing the results, among the final set of 3,546 study participants, it was discovered that the biospecimens from the Hispanic study participants were heavily clustered on

5 of the 47 genotyping plates. Looking at the level of agreement of responses among Hispanics, there is a possibility of a large level of agreement due to the distribution of Hispanics which could artificially inflate the subject-level race agreement.

For the purposes of these data, each subject is considered to be the "object-level" result, and each genotyping plate is considered to be the "nested-level" result as each specimen is nested within each plate. Chapter 3 answered a similar question for response-level results, dichotomizing each race outcome and determining the effect of the nested-level agreement on the corresponding object-level agreement. While those methods were sufficient for the dichotomized responses, it does not analyze the four-part question as a whole and therefore does not use all available information in the analysis for each response. The methods described in this paper serve as an avenue, under the correct circumstances, to either determine the nested-level agreement among all responses or give validity to analyzing each response separately due to the lack of fit of the MDD. To fully analyze the effect of the nested-level correlation on the object-level correlation, the following questions must be answered:

1. Does homogeneity of object-level ICCs exist for these data?
2. Does homogeneity of nested-level ICCs exist for these data?
3. Can all responses be analyzed simultaneously or must each response be analyzed separately?

4.5.2. Results

The methods described thus far will be sufficient to adequately answer all three questions. First, each response will be analyzed separately, and the corresponding object-level and nested-level (in this case, subject-level and plate-level) ICC will be

estimated along with its corresponding standard error. Second, the pooled estimates of the subject and plate-level correlations will be calculated using the MDD. Finally, the MDD and MBBD will be used in combination first to test for the homogeneity of subject-level ICCs and, in the presence of homogeneity of subject-level ICCs, to test for homogeneity of plate-level ICCs.

Table 4.1: Multinomial Object and Nested-Level ICCs for a GWAS

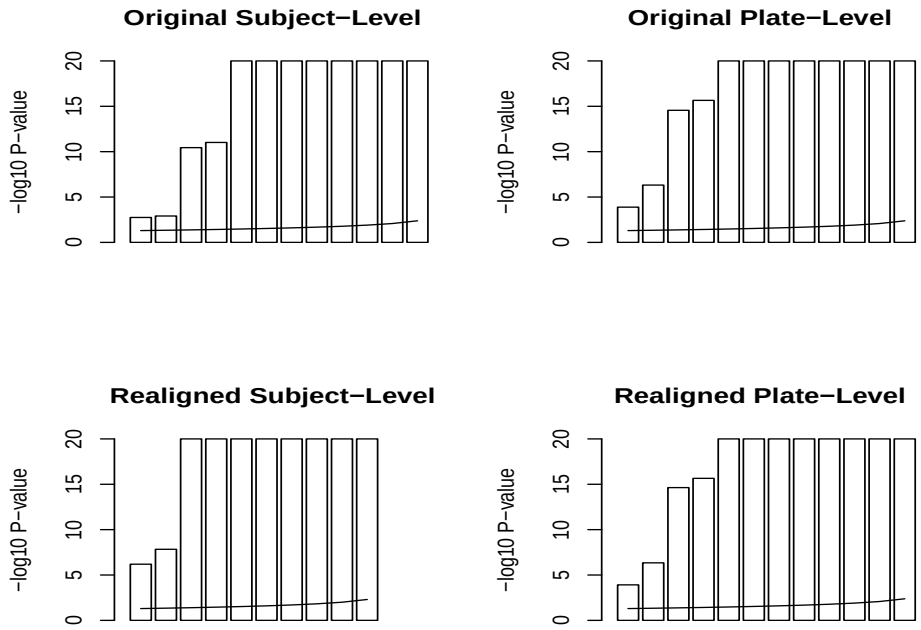
Alignment	Response Category	Avg. Prop. ($\hat{\pi}$)	2-level Model Subject-ICC (SE) ($\hat{\rho}$)	3-level Model Davis Plate-ICC ($\hat{\zeta}$) (SE)	95% CI	Adj S-ICC
Original PIDs	1:NH-White	0.433	0.929 (0.006)	0.114 (0.025)	[.0650, .1629]	**0.920**
	2:NH-Black	0.420	0.949 (0.005)	0.063 (0.016)	[.0330, .0938]	**0.946**
	3:Hispanic	0.109	0.919 (0.011)	0.215 (0.032)	[.1530, .2769]	**0.897**
	4:Other	0.037	0.682 (0.032)	0.001 (0.003)	[-.0043, .0061]	**0.682**
	Overall*		0.921 (0.006)	0.092 (0.011)	[.0705, .1131]	**0.913**
Re-aligned PIDs	1:NH-White	0.433	0.963 (0.005)	0.114 (0.025)	[.0652, .1635]	**0.958**
	2:NH-Black	0.420	0.986 (0.003)	0.063 (0.016)	[.0330, .0939]	**0.985**
	3:Hispanic	0.109	0.926 (0.010)	0.215 (0.032)	[.1531, .2770]	**0.906**
	4:Other	0.037	0.693 (0.032)	0.001 (0.003)	[-.0043, .0060]	**0.693**
	Overall*		0.950 (0.005)	0.092 (0.011)	[.0705, .1132]	**0.945**

*There are heterogeneity of ICCs, therefore the assumptions for the overall ICC estimate to be valid are violated.

The response-level results displayed were previously reported in Chapter 3. The pooled object and nested-level results summarize the measure of agreement among all responses and can be considered assessment-level ICCs, assessing the agreement of the results of the entire four-part question as opposed to the response-level results. In both alignments, the pooled object-level agreement is excellent, while the re-aligned subject ID's prove to have a greater pooled object-level ICC than the original patient ID's. This is expected as one would expect a high level of agreement between patient reported race and genetically-determined race, so appropriately aligning the two responses should result in higher levels of agreement. Interestingly, the nested-level agreement remained remarkably similar between the two alignments, with all point-estimates remaining constant through the alignment. One can argue that this is expected as well as the re-alignment of a small number of 3,546 subjects may result in better object-level agreement, but may not have an effect on the nested-level agreement among only 47 genotyping plates. The estimated pooled nested-level ICC was 0.092 in both alignments, and in both cases the lower-limit of the 95% confidence interval was greater than zero. Therefore, there is a measure of agreement among the plates that is significantly greater than zero and ratings on nested-level results can be categorized as having 'slight agreement' according to Landis and Koch [33]. However, overall the nested-level adjusted object-level ICC did not change much due to the high level of pooled subject-level correlation and the relatively low level of overall nested-level agreement. Therefore, while there was slight agreement among observations within a nested-level where there should be no agreement, the magnitude of object-level agreement as well as the small magnitude of nested-level agreement were enough to only slightly decrease the adjusted object-level ICCs.

However, when analyzing the goodness-of-fit of the pooled model, it is clear that the

Figure 4.2: Test of Homogeneity of ICCs for Race Results



Note: The solid line represents the Benjamini-Hochberg criterion for -log10(p-values)

MDD describing the pooled object-level agreement does not accurately fit the data. From a glance at the range of the object-level ICCs, this should come as no surprise. The range of the object level ICCs from the original data is 0.682–0.949 and is 0.693–0.986 among the re-aligned subject ID's. If there was homogeneity of object-level ICCs among the responses, it would be expected that the object-level ICC for each separate race would be roughly equivalent. Due to the large range of object-level ICCs, there does not appear to be homogeneity of ICCs. In both cases, the p-value resulting from each decomposition of the MBBD compared to the MDD was less than 0.0001, indicating that using either the Bonferonni-Holms or Benjamini-Hochberg approach, there is a clear violation of the notion of object-level homogeneity of responses across responses. Therefore, to appropriately determine the object-level agreement in this scenario, each race response should be analyzed separately as opposed to modeling the level of agreement of the question as a whole.

Even though there is not homogeneity among the subject-level ICCs, the asymptotic properties of the MDD can be used to evaluate the homogeneity of nested-level ICCs due to the large number of responses per genotyping plate. On average, there were 150.9 results per plate, resulting in an average m_i of .0067. The overall plate-level ICC d_i is estimated as .098 with standard error .012 and, as m_i is sufficiently small, can be used to estimate the overall plate-level agreement in order to test the homogeneity of nested-level ICCs. The range of nested-level ICCs is 0.001–0.215, so it is expected that there is heterogeneity of nested-level ICCs. Figure 4.2 provides a clear indication that the hypothesis that there is homogeneity among plate-level ICCs should be rejected as the likelihood-ratio tests of all decompositions of the MDD into the MBBD are highly significant. Therefore, there is evidence that the level of agreement among plates is inconsistent across races, signifying that genotyping samples should have been better

stratified among plates in order to achieve balance across plates and reduce the level of bias in analyzing subjects' race. As a result, this provides further evidence that the overall assessment of race should not be analyzed for these data, but rather the dichotomized race assessment for each race in question.

4.6. Conclusion

The presence of nested-level ICCs can artificially inflate the apparent object-level ICCs and should be accounted for when potential nested-level correlations exist. In this paper, the method of modeling pooled nested-level ICCs has been described using the MDD. In addition, in the presence of homogeneity of object-level results, a goodness-of-fit test has been proposed that detects whether the assumption of homogeneity of nested-level ICCs is valid, which indicates whether the pooled nested-level ICC is the appropriate statistic to model the nested-level agreement. These methods were applied to a GWAS study where there was significant nested-level correlation among results that should have been uncorrelated and found that there is strong evidence of heterogeneity of nested-level ICCs among races. However, there is a shortcoming of this approach that mandates that, in the presence of heterogeneity of either the object or nested-level ICC, the level of agreement for each response should be analyzed based on the dichotomized result for each response. Ideally, these results would be modeled using a likelihood-based approach that allows for heterogeneity of ICCs that allows for separate modeling the ICCs for each response. Overall, these methods result in an advance to measure the pooled nested-level ICC for the question as a whole and a goodness-of-fit test to determine if the pooled ICC is an appropriate assessment of nested-level agreement while paving the way for future research to improve on and advance the investigation into measures of agreement.

CHAPTER 5

Discussion

Prior to this dissertation, there was not a comprehensive method to test for homogeneity of ICCs across multiple item-wise responses for a multinomial outcome. In addition, the nested-level ICC for both binomial and multinomial outcomes could neither be accurately estimated nor inference provided on the result. This dissertation has achieved both milestones. Therefore, researchers now have the ability to test the assumption of homogeneity of ICCs across a multinomial response to support summarizing rater agreement by either a pooled ICC or by dichotomized responses. In addition, potential biases due to nested-level agreement can be identified and subsequently corrected to provide unbiased estimates of measures of object-level agreement. Specifically, in this dissertation, we have described and demonstrated the validity of a test for homogeneity of item-wise ICCs across a multinomial response. Although there were a number of potential expressions for the multiple beta-binomial distribution given the number of potential outcomes of the response, recommendations for controlling the type I error rate were presented. Simulations demonstrated not only the strong control of the type I error rate for the test of homogeneity of ICCs across the multinomial response, but also gave some insight into the power of the test under various assumptions for differences in ICC, numbers of subjects and raters, and methods for controlling the type I error rate. As a result, investigators interested in researching the overall measure of agreement for a multinomial response by pooling the ICCs should first test whether homogeneity of item-wise ICCs for the individual responses exist. First, if there is homogeneity among the responses, there is an increase in efficiency to be gained by pooling responses and reporting one overall ICC.

However, if there is heterogeneity among responses, valuable information regarding the differences in measures of agreement for each response will be lost. As a result, in the case of heterogeneity of ICCs, we are recommending at this time that each potential outcome be dichotomized and analyzed separately as there is no likelihood-based framework available to simultaneously estimate the ICCs.

In addition, we have identified the potential issue of a nested-level measure of agreement, and provided frameworks to estimate and provide inference on the nested-level ICC. Using a modification of the beta-binomial distribution for binomial data, we were able to identify the measure of agreement that exists among ratings on separate objects within the same nested-level and provide both variance and confidence interval formulas for the estimate. Simulations verified that the estimation provides unbiased estimates of the nested-level ICC and appropriate coverage of the confidence interval given a large enough sample size. We were also able to prove that the presence of nested-level agreement artificially inflates the apparent object-level agreement, and provided a nested-level adjusted object-level agreement measure to account for this artificial inflation.

In a similar fashion, for multinomial outcomes, the multinomial-Dirichlet distribution was leveraged to estimate and provide inference on the pooled nested-level ICC using the assumption of homogeneity of both object and nested-level ICCs. In order to test this assumption of homogeneity of nested-level ICCs, the multiple beta-binomial distribution was extended to account for nested-levels of agreement and a test for homogeneity of nested-level ICCs was derived. In addition, the asymptotic properties of the model were examined and found that, for a large number of objects per nested-level, that the nested-level ICC can be examined without regard to the object-level ICC. Finally, a nested-level adjusted object-level ICC measure was derived to account for the inflation of the apparent object-level agreement.

Future work should be done investigating a likelihood that can accommodate the data in such a manner to provide simultaneous estimation on separate ICCs for each outcome. This would allow for flexible modeling of ICCs across responses and could be used to conduct pairwise tests of equivalency of item-wise ICCs. In addition, it may lend itself to more flexibility in performing hypothesis tests on the item-wise measures of agreement. Upon its discovery, this model should be used to estimate nested-level ICCs, determine whether homogeneity among the nested-level ICCs exists and provide a nested-level adjusted object-level ICC.

Until this point, few researchers have looked at the potential issue that agreement among a nested-level can cause when estimating agreement. In general, well designed and/or randomized trials will exhibit no nested-level agreement as it is not expected that there would be any type of agreement on ratings on separate objects, however studies that do not pay attention to this point may introduce bias into the results. This emphasizes the point that, if at all possible, attention should be paid to have a random mixture of objects within each nested level to reduce bias in measuring agreement. This methodology also allows for examination into whether nested-level bias exists in previously conducted research studies and to adjust the object-level ICC where appropriate.

APPENDIX A

Technical Arguments

A.1. Chapter 2: Multinomial-Dirichlet Distribution, A Special Case of the Multiple Beta-Binomial Distribution

With n_h response categories for object i, the multiple beta-binomial distribution (MBBD) can be written as

$$P\left(\mathbf{X_i} = \mathbf{x_i}|\mathbf{A}, \mathbf{B}\right) = \prod_{f=1}^{n_h-1} \binom{N - \sum_{g=1}^{f-1} x_{gi}}{x_{fi}} \times \frac{\Gamma\left(x_{fi} + a_f\right) \Gamma\left(N - \sum_{g=1}^{f} x_{gi} + b_f\right) \Gamma\left(a_f + b_f\right)}{\Gamma\left(N - \sum_{g=1}^{f-1} x_{gi} + a_f + b_f\right) \Gamma\left(a_f\right) \Gamma\left(b_f\right)} \tag{A.1}$$

Under the assumption that $a_h = m_h$ and $b_h = m_{h+1} + m_{h+2} + ...m_k \forall h$ where $m_1, m_2, ...m_k$ are the parameters of the multinomial Dirichlet distribution, the MBBD

can be rewritten as

$$
\begin{aligned}
P\left(\mathbf{X_i} = \mathbf{x_i} | \mathbf{A}, \mathbf{B}\right) &= \prod_{f=1}^{n_h - 1} \binom{N - \sum_{g=1}^{f-1} x_{gi}}{x_{fi}} \\
&\times \frac{\Gamma\left(x_{fi} + m_f\right) \Gamma\left(N - \sum_{g=1}^{f} x_{gi} + \sum_{q=f+1}^{n_h} m_q\right) \Gamma\left(\sum_{q=f}^{n_h} m_q\right)}{\Gamma\left(N - \sum_{g=1}^{f-1} x_{gi} + \sum_{q=f}^{n_h} m_q\right) \Gamma\left(m_f\right) \Gamma\left(\sum_{q=f+1}^{n_h} m_q\right)} \\
&= \frac{N!}{\prod_{f=1}^{n_h} x_{fi}!} \\
&\times \prod_{f=1}^{n_h - 1} \frac{\Gamma\left(x_{fi} + m_f\right) \Gamma\left(N - \sum_{g=1}^{f} x_{gi} + \sum_{q=f+1}^{n_h} m_q\right) \Gamma\left(\sum_{q=f}^{n_h} m_q\right)}{\Gamma\left(N - \sum_{g=1}^{f-1} x_{gi} + \sum_{q=f}^{n_h} m_q\right) \Gamma\left(m_f\right) \Gamma\left(\sum_{q=f+1}^{n_h} m_q\right)} \\
&= \frac{N!}{\prod_{f=1}^{n_h} x_{fi}!} \frac{\Gamma\left(\sum_{f=1}^{n_h} m_f\right)}{\Gamma\left(N + \sum_{f=1}^{n_h} m_f\right)} \prod_{f=1}^{n_h} \frac{\Gamma\left(x_{fi} + m_f\right)}{\Gamma\left(m_f\right)}
\end{aligned}
$$

which is simply the multinomial-Dirichlet distribution.

A.2. Chapter 3: Estimation of the Variance of $x_{i\cdot}$

Let y_{ijk} be the binary response 0 or 1 for the k^{th} rater on the j^{th} object in the i^{th} nested-level. Assuming that y_{ijk} follows a binomial distribution $y_{ijk} \sim Bin\left(\pi\right)$, the

variance of $x_{i\cdot} = \sum_{j=1}^{n_{i\cdot\cdot}} \sum_{k=1}^{n_{ij\cdot}} y_{ijk}$ is written as follows:

$$
\begin{aligned}
Var(x_{i\cdot}) &= \sum_{j=1}^{n_{i\cdot\cdot}} \sum_{k=1}^{n_{ij\cdot}} Var(y_{ijk}) + 2 \sum_{m<n} \sum_{p<q} Cov(y_{imp}, y_{inq}) \\
&= \sum_{j=1}^{n_{i\cdot\cdot}} n_{ij\cdot}\pi(1-\pi) + 2 \sum_{j=1}^{n_{i\cdot\cdot}} \binom{n_{ij\cdot}}{2} \rho\pi(1-\pi) + 2\zeta\pi(1-\pi) \\
&\quad \left[\binom{\sum_{j=1}^{n_{i\cdot\cdot}} n_{ij\cdot}}{2} - \sum_{j=1}^{n_{i\cdot\cdot}} \binom{n_{ij\cdot}}{2} \right] \\
&= \sum_{j=1}^{n_{i\cdot\cdot}} n_{ij\cdot}\pi(1-\pi) \left[1 + \frac{2\rho \sum_{j=1}^{n_{i\cdot\cdot}} \binom{n_{ij\cdot}}{2} + 2\zeta \left(\binom{\sum_{j=1}^{n_{i\cdot\cdot}} n_{ij\cdot}}{2} - \sum_{j=1}^{n_{i\cdot\cdot}} \binom{n_{ij\cdot}}{2} \right)}{\sum_{j=1}^{n_{i\cdot\cdot}} n_{ij\cdot}} \right] \\
&= \sum_{j=1}^{n_{i\cdot\cdot}} n_{ij\cdot}\pi(1-\pi) \left[1 + 2 \left(\sum_{j=1}^{n_{i\cdot\cdot}} n_{ij\cdot} - 1 \right) \frac{\rho \sum_{j=1}^{n_{i\cdot\cdot}} \binom{n_{ij\cdot}}{2} + \zeta \left(\binom{\sum_{j=1}^{n_{i\cdot\cdot}} n_{ij\cdot}}{2} - \sum_{j=1}^{n_{i\cdot\cdot}} \binom{n_{ij\cdot}}{2} \right)}{\sum_{j=1}^{n_{i\cdot\cdot}} n_{ij\cdot} \left(\sum_{j=1}^{n_{i\cdot\cdot}} n_{ij\cdot} - 1 \right)} \right] \\
&= \sum_{j=1}^{n_{i\cdot\cdot}} n_{ij\cdot}\pi(1-\pi) \left[1 + \left(\sum_{j=1}^{n_{i\cdot\cdot}} n_{ij\cdot} - 1 \right) \frac{\rho \sum_{j=1}^{n_{i\cdot\cdot}} \binom{n_{ij\cdot}}{2} + \zeta \left(\binom{\sum_{j=1}^{n_{i\cdot\cdot}} n_{ij\cdot}}{2} - \sum_{j=1}^{n_{i\cdot\cdot}} \binom{n_{ij\cdot}}{2} \right)}{\binom{\sum_{j=1}^{n_{i\cdot\cdot}} n_{ij\cdot}}{2}} \right] \\
&= \sum_{j=1}^{n_{i\cdot\cdot}} n_{ij\cdot}\pi(1-\pi) \left[1 + \left(\sum_{j=1}^{n_{i\cdot\cdot}} n_{ij\cdot} - 1 \right) \left(m_i\rho + (1-m_i)\,\zeta \right) \right]
\end{aligned}
$$

A.3. Chapter 4: Estimation and Inference of ζ.

Using the likelihood displayed in equation (5), the properties associated with profile-likelihood can be used to provide a consistent estimate for ζ. and its corresponding standard error. The estimate $\hat{\zeta}$. can be derived by finding the value for ζ that will set

the derivative of the log-likelihood with respect to ζ equal to zero, thereby solving the equation

$$0 = \sum_{a=1}^{n_{\cdots}} \left[\sum_{e=1}^{n_h} \sum_{f=1}^{x_{ea\cdot}} \frac{-(1-m_a)\pi_e}{d_a^2 \left(f + \frac{1-d_a}{d_a}\pi_e - 1 \right)} + \sum_{c=1}^{n_{a\cdot\cdot}} \frac{1-m_a}{d_a^2 \left(c + \frac{1-d_a}{d_a} - 1 \right)} \right]$$

The formula for the variance can be found in section 3.2, however the second derivative can be calculated as follows:

$$\begin{aligned} \frac{\partial^2 LogL(\hat{\mathbf{p}}, \hat{\rho}, \zeta | \mathbf{X_i}_{\cdot\cdot} = \mathbf{x_i}_{\cdot\cdot})}{\partial \zeta^2} = & \sum_{a=1}^{n_{\cdots}} \left[\sum_{e=1}^{n_h} \sum_{f=1}^{x_{ea\cdot}} \left(\frac{-(1-m_a)^2 \pi_e^2}{d_a^4 \left(f + \frac{1-d_a}{d_a}\pi_e - 1 \right)^2} + \frac{2(1-m_a)\pi_e}{d_a^3 \left(f + \frac{1-d_a}{d_a}\pi_e - 1 \right)} \right) \right. \\ & \left. + \sum_{c=1}^{n_{a\cdot\cdot}} \left(\frac{(1-m_a)^2}{d_a^4 \left(c + \frac{1-d_a}{d_a} - 1 \right)^2} - \frac{2(1-m_a)}{d_a^3 \left(f + \frac{1-d_a}{d_a} - 1 \right)} \right) \right] \end{aligned}$$

BIBLIOGRAPHY

[1] R.L. Anderson and T.A. Bancroft. *Statistical Theory in Research*. McGraw Hill, 1952.

[2] E. Bartfay and A. Donner. The effect of collapsing multinomial data when assessing agreement. *International journal of epidemiology*, 29(6):1070–1075, 2000.

[3] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[4] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

[5] Ben Bolker. *emdbook: Ecological Models and Data in R*, 2012. R package version 1.3.2.

[6] NE Breslow and NE Day. *Statistical methods in cancer research*. International Agency for Research on Cancer, 1994.

[7] Lyle D Broemeling. *Bayesian methods for measures of agreement*. CRC Press, 2008.

[8] G. Casella and R.L. Berger. *Statistical inference*, volume 70. Duxbury Press Belmont, CA, 1990.

[9] James J Chen, Ralph L Kodell, Richard B Howe, and David W Gaylor. Analysis of trinomial responses from reproductive and developmental toxicity experiments. *Biometrics*, 47(3):1049–1058, 1991.

[10] James J Chen and Lung-An Li. Dose-response modeling of trinomial responses from developmental experiments. *Statistica Sinica*, 4:265–274, 1994.

[11] J.W. Choi and J.R. Landis. The correlation estimator for unbalanced data. *ASA Proceedings of the Survey Research Methods Section*, pages 858–863, 1998.

[12] Li-Ling Chuang and Yu-Shan Shih. Approximated distributions of the weighted sum of correlated chi-squared random variables. *Journal of Statistical Planning and Inference*, 142(2):457–472, 2012.

[13] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(37):37–46, 1960.

[14] Jacob Cohen. Weighed kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220, 1968.

[15] M.J. Crowder. Beta-binomial anova for proportions. *Applied Statistics*, 27(1):34–37, 1978.

[16] M.J. Crowder. Inference about the intraclass correlation coefficient in the beta-binomial anova for proportions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):230–234, 1979.

[17] Anthony Christopher Davison. *Statistical models*, volume 11. Cambridge University Press, 2003.

[18] H. Demirtas, D. Hedeker, and K. Kapur. A comparative study on most commonly used correlated binary data generation methods. Technical Report 2007-007, University of Illinois at Chicago, Division of Epidemiology and Biostatistics, Department of Psychiatry, December 2007.

[19] Allan Donner and Michael Eliasziw. A heirarchical approach to inferences concerning interobserver agreement for multinomial data. *Statistics in Medicine*, 16(10):1097–1106, 1997.

[20] L.J. Emrich and M.R. Piedmonte. A method for generating high-dimensional multivariate binary variates. *The American Statistician*, 45(4):302–304, 1991.

[21] J.L. Fleiss and J. Cohen. The equivalence of kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33:613–619, 1973.

[22] Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. *Statistical methods for rates and proportions*. John Wiley & Sons, 2013.

[23] Joseph L Fleiss, John C Nee, and J Richard Landis. Large sample variance of kappa in the case of different sets of raters. *Psychological Bulletin*, 86(5):974, 1979.

[24] Wenge Guo. A note on adaptive bonferroni and holm procedures under dependence. *Biometrika*, 96:1012–1018, 2009.

[25] J. Arthur Harris. On the calculation of intra-class and inter-class coefficients of correlation from class moments when the number of possible combinations is large. *Biometrika*, 9(3/4):446–472, 1913.

[26] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[27] N.S. Holmquist, C.A. McMahan, and O.D. Williams. Variability in classification of carcinoma in situ of the uterine cervix. *Archives of Pathology*, 84:334–345, 1967.

[28] Tonya S King and Vernon M Chinchilli. A generalized concordance correlation coefficient for continuous and categorical data. *Statistics in Medicine*, 20(14):2131–2147, 2001.

[29] G.G. Koch, J.R. Landis, J.L. Freeman, D.H. Freeman Jr, and R.G. Lehnen. A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics*, 33(1):133–158, 1977.

[30] K. Krippendorff. K. bivariate agreement coefficients for reliability of data. *Social Methodology*, 1970.

[31] J.R. Landis, T.S. King, J.W. Choi, V.M. Chinchilli, and G.G. Koch. Measures of agreement and concordance with clinical research applications. *Statistics in Biopharmaceutical Research*, 3(2):185–209, 2011.

[32] J.R. Landis and G.G. Koch. An application of hierarchical kappa-type statistics in the assessment of agreement among multiple observers. *Biometrics*, 33(2):363–374, 1977.

[33] J.R. Landis and G.G. Koch. The measurement of observer agreement for categorical data. *biometrics*, 33(1):159–174, 1977.

[34] J.R. Landis and G.G. Koch. A one-way components of variance model for categorical data. *Biometrics*, 33:671–679, 1977.

[35] A.J. Lee. Generating random binary deviates having fixed marginal distributions and specified degrees of association. *The American Statistician*, 47(3):209–215, 1993.

[36] L. Lin, A.S. Hedayat, and W. Wu. *Statistical Tools for Measuring Agreement*. SpringerLink : Bücher. Springer, 2012.

[37] Kung-Jong Lui, William G. Cumberland, Joni A. Mayer, and Laura Eckhardt. Interval estimation for the intraclass correlation in dirichlet-multinomial data. *Psychometrika*, 64(3):355–369, 1999.

[38] Kung-Jong Lui, William G Cumberland, Joni A Mayer, and Laura Eckhardt. Interval estimation for the intraclass correlation in dirichlet-multinomial data. *Psychometrika*, 64(3):355–369, 1999.

[39] T.K. Mak. Analysing intraclass correlation for dichotomous variables. *Applied Statistics*, 37(3):344–352, 1988.

[40] JT Newcomer, NK Neerchal, and JG Morel. Computation of higher order moments from two multinomial overdispersion likelihood models. *Department of Mathematics and Statistics, University of Maryland, Baltimore, USA*, 2008.

[41] C.G. Park, T. Park, and D.W. Shin. A simple method for generating correlated binary variates. *The American Statistician*, 50(4):306–310, 1996.

[42] Sudhir R Paul, Uditha Balasooriya, and Tathagata Banerjee. Fisher information matrix of the dirichlet-multinomial distribution. *Biometrical journal*, 47(2):230–236, 2005.

[43] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

[44] M.S. Ridout, C.G.B. Demetrio, and D. Firth. Estimating intraclass correlation for binary data. *Biometrics*, 55(1):137–148, 1999.

[45] H. Scheffé. *The Analysis of Variance*. John Wiley and Sons Inc., 1959.

[46] S.R. Searle. *Linear Models*. John Wiley and Sons Inc., 1971.

[47] S.R. Searle, G. Casella, and C.E. McCulloch. *Variance components*. Wiley.com, 2009.

[48] Mohamed M Shoukri. *Measures of interobserver agreement and reliability*. CRC Press, 2010.

[49] Alexander Von Eye and Eun Young Mun. *Analyzing rater agreement: Manifest variable methods*. Psychology Press, 2004.

[50] K.B. Westlund and L.T Kurland. Studies on multiple sclerosis in winnipeg, manitoba and new orleans, lousiana i. prevalence; comparison between the patient groups in winnipeg and new orleans. *American Journal of Hygiene*, 57:380–396, 1953.

[51] T.W. Yee. The vgam package. *R News*, 8(2):28–39, 2008.