RANK-BASED ANALYSES AND DESIGNS WITH CLUSTERED DATA

By

Shengxin Tu

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Biostatistics

August 9, 2024

Nashville, Tennessee

Approved:

Frank E. Harrell, Ph.D.

Bryan E. Shepherd, Ph.D.

Jonathan S. Schildcrout, Ph.D.

Peter F. Rebeiro, Ph.D.

In loving memory of my dear grandfather, Mr. Tu.

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere gratitude to my Ph.D. advisor, Dr. Bryan E. Shepherd, whose unwavering support, profound wisdom, and invaluable mentorship have illuminated my academic journey. Dr. Bryan E. Shepherd is always encouraging, enthusiastic, optimistic, and willing to listen. He has taught me not only the principles of scholarly thinking, rigorous research, and interdisciplinary collaboration, but also the joy of enjoying life's simplest moments and being grateful for them. Furthermore, I am profoundly thankful to Dr. Chun Li, a professor at the University of Southern California, for his invaluable guidance on my dissertation. Dr. Chun Li has taught me the spirit of perpetual rigor and curiosity. Without the support and guidance of both Dr. Bryan E. Shepherd and Dr. Chun Li, I would not have reached this significant milestone.

I am deeply grateful to my esteemed dissertation committee members, Dr. Frank E. Harrell, Dr. Bryan E. Shepherd, Dr. Jonathan S. Schildcrout, and Dr. Peter F. Rebeiro, whose inspirational suggestions and insightful feedback greatly enhanced my dissertation. I would also like to thank Dr. Donglin Zeng, a professor at the University of Michigan at Ann Arbor, for his essential help regarding asymptotic theory in the Chapter 2 project. Moreover, I appreciate the study investigators who provided data for the example applications in my dissertation. This dissertation was supported in part by funding from the National Institutes of Health (R01AI093234; R01NS113171 for BRIDGE; R01MH113478 for HoPS+; U01DK112271 for the Nigerian uACR study; P30AI110527 and K23AI120875 for the longitudinal biomarker data).

Furthermore, I wish to express my appreciation to the investigators with whom I have had the honor to collaborate, Dr. Jessica Castilho, Dr. Rachael Pellegrino, Dr. Valeria Fink, Dr. Gabriel C. Rozas, and Dr. Claudia Cortes. Through our interdisciplinary collaboration, I have had the opportunity to apply my statistical knowledge to answer questions about the HIV epidemic. Their expertise and scientific passion have expanded the breadth of my research endeavors, empowering me to make meaningful contributions to global public health.

I wish to extend my sincere appreciation to the faculty, staff, and graduate students in the Department of Biostatistics at Vanderbilt University, especially my cohort, Jamie G. Joseph, Caroline Birdrow, Xiangyu Ji, Tianyi Sun, Kaidi Kang, Yan Yan, Aaron Lee, and Justin Jacobs. Being part of this vibrant and supportive community has been an enriching experience that I deeply value. Studying at Vanderbilt University has truly been one of the most rewarding decisions of my life. Additionally, I am immensely thankful to all the friends that I have made during my time at Vanderbilt. Special thanks belong to Siyuan Yu, Tianyi Sun, Wenying Gu, and Jingxuan He, who have always been by my side, especially during challenging times. As we embark on the next chapter of our lives, I am grateful for our friendship and hope it will last forever. I always remember the afternoon when Dr. Robert Greevy called me with an offer from Vanderbilt University, making the moment when the gears of fate began to turn.

My heartfelt thanks to the Tu and Xu families for their strong support and presence throughout the journey of my life. Among them, I hold a special place in my heart for my beloved parents, to whom I owe immeasurable gratitude. Your unconditional love, unwavering support, and consistent encouragement have been the pillars upon which I have built my aspirations and pursued my dreams. Thank you for always believing in me and guiding me with your wisdom and care.

Lastly, I would like to express my deepest gratitude to the person who is indispensable in making all this happen: myself. You deserve sincere commendation for achieving this significant milestone. As you step into the next chapter of your journey, always remember the courage and kindness in your heart. Amidst the ups and downs, stay true to yourself and embrace the journey with resilience and spontaneity, "to see the world as it is, and to love it" by Romain Rolland. As you move forward, do not forget to glance back at the journey you have traversed, where flowers bloom along the way.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

Clustered data are common in biomedical research. Conventional statistical approaches, including intraclass correlation coefficients and Pearson correlations, are frequently used to handle clustered data. The intraclass correlation coefficient (ICC), first introduced by R. A. Fisher (1925), is used to measure the degree of similarity within clusters (Murray et al., 2004; Hedges and Hedberg, 2007). The total, between-, and within-cluster Pearson correlations are used in the analyses of correlation between two variables with clustered data, together providing an enriched perspective of the correlation (Snijders and Bosker, 1999; Ferrari et al., 2005). However, these conventional approaches are sensitive to extreme values and skewness, and depend on the scale of the data. They also are not applicable to ordered categorical data.

In practice, variables of interest often include skewed continuous variables, ordinal variables, or mixtures of the two (e.g. outcomes subject to a detection limit). For example, in an observational study, people living with HIV on antiretroviral therapy had repeated measurements of CD4 and CD8 counts (Castilho et al., 2016). The data of CD4 and CD8 counts are both right-skewed and sometimes transformed prior to analyses; estimates of the ICC and Pearson correlations will vary with the transformation. As another example, in a cluster randomized controlled trial on childhood epilepsy care (Aliyu et al., 2019), the number of seizures from 18 to 24 months is an irregularly distributed count variable. Data of these types may be more appropriately analyzed using rank-based approaches. However, rank-based methods for clustered data are under-developed.

Several studies have proposed rank-based measures to evaluate intraclass similarity (Rothery, 1979; Shirahata, 1981); however, they are probabilities of concordance and do not share the same spirit as Fisher's ICC, which is a correlation measure. In Chapter 2, we define the rank ICC as a natural extension of Fisher's ICC to the rank scale, and describe its corresponding population parameter. The rank ICC is simply interpreted as the rank correlation between a random pair of observations from the same cluster. We also extend the definition when the underlying distribution has more than two hierarchies. We describe estimation and inference procedures, show the asymptotic properties of our estimator, and conduct simulations to evaluate its performance. We also use three real data examples to illustrate our method, including skewed data, count data, and three-level ordered categorical data. Furthermore, an R package, `rankICC`, has been developed and is available on CRAN, implementing our new method.

In the analyses of correlation between two variables with cluster data, current rank-based measures are only for the total correlation (Rosner and Glynn, 2017; Shih and Fay, 2017; Hunsberger et al., 2022). There

is a need to develop rank-based between- and within-cluster correlations. In Chapter 3, we define population parameters for the between- and within-cluster Spearman rank correlations. The definitions are natural extensions of the Pearson between- and within-cluster correlations to the rank scale. We show that the total Spearman rank correlation approximates a weighted sum of the between- and within-cluster Spearman rank correlations, where the weights are functions of rank ICCs of the two random variables. We also discuss the equivalence between the within-cluster Spearman rank correlation and the covariate-adjusted partial Spearman rank correlation. Furthermore, we describe estimation and inference for the three Spearman rank correlations, and conduct simulations to evaluate the performance of our estimators. We also illustrate our method with data from a longitudinal biomarker study and a clustered randomized trial. A developed R package, `rankCorr`, is accessible on CRAN.

In sample size calculations for cluster randomized controlled trials (RCTs), a design effect based on the ICC is commonly used to inflate the sample size of an adequately powered individual RCT (Campbell and Walters, 2014; Rutterford et al., 2015). However, this design effect was derived for comparisons of means and may not apply to skewed or ordinal data. In addition, as mentioned previously, there are limitations to using the ICC for handling skewed or ordinal data. There are two rank-based sample size calculation approaches proposed as alternatives for calculating sample sizes in cluster RCTs (Kim et al., 2005; Rosner and Glynn, 2011). However, the two approaches are complex, lack closed forms, involve numerous calculations, and rely on additional assumptions. In Chapter 4, we introduce a design effect that incorporates the rank ICC, and propose new sample size calculations for cluster RCTs with skewed or ordinal outcomes using this new design effect. Our calculations involve inflating the sample size for an adequately powered individual RCT for an ordinal outcome with the new design effect. For continuous outcomes, our calculations set the number of distinct ordinal levels to the sample size. Our calculations are unified and simple. Furthermore, we show that with continuous data, our calculations closely approximate more complicated calculations based on clustered Wilcoxon rank-sum tests. We conduct simulations to evaluate our calculations' performance and illustrate their use in the design of two cluster RCTs, one with a skewed continuous outcome and a non-inferiority trial with an irregularly distributed count outcome.

CHAPTER 2

Rank Intraclass Correlation

This chapter is from Rank Intraclass Correlation for Clustered Data published in *Statistics in Medicine* and has been reproduced with the permission of the publisher and my co-authors Chun Li, Donglin Zeng, and Bryan E. Shepherd.

2.1 Introduction

With clustered data, observations in the same cluster are often more similar to each other than to those from other clusters. The degree of similarity is frequently measured by the intraclass correlation coefficient (ICC). R. A. Fisher first introduced the ICC to assess familial resemblance of a trait between siblings (Fisher, 1925). The ICC has since been used in various disciplines including epidemiology, genetics, and psychology. It also has been employed in clinical trial design (Murray et al., 2004; Hedges and Hedberg, 2007). Fisher's ICC measures the correlation between a random pair of observations from a random cluster. When the cluster size is infinite, Fisher's ICC is equal to the variance of cluster means divided by the total variance (Harris, 1913). Because of this, the ICC has also been estimated with random effects models, in which it is estimated as the proportion of total variance attributable to the clusters (Shrout and Fleiss, 1979; Donner, 1986).

The ICC is fundamental to the analysis of clustered data. However, similar to Pearson's correlation, it is sensitive to extreme values and skewed distributions, and it depends on the scale of the data. When a variable is transformed to a different scale, the ICC may change. For some non-Gaussian distributions, the ICC might be estimated using generalized linear random effects models. In this case, the ICC is defined on the link function transformed scale and it may be sensitive to the non-normality of random effects or the method used to derive the within-cluster variance (Nakagawa et al., 2017). The ICC is also not applicable to ordered categorical data. For ordered categorical data, ordinal regression models with random effects may be used to estimate variance components, but the total variance is undefined unless numbers are assigned to levels of the ordinal response (Hallgren, 2012; Denham, 2016).

Several studies have proposed nonparametric measures to evaluate intraclass similarity based on the notion of concordance. One measure is the probability that a random observation from a cluster does not fall between a random pair of observations from a different cluster (Rothery, 1979). Another measure is the probability that a random pair of observations from a cluster does not fall between two random observations each from a different cluster (Shirahata, 1981). Shirahata (1982) performed comparisons between the two measures and a modification of Kendall's measure of dependence (Shirahata, 1982). All three measures

are rank-based; however, they are probabilities of concordance and do not share the same spirit as Fisher's ICC, which is a correlation measure. Methods to estimate the ICC for categorical data have been developed (Chakraborty et al., 2021), but they ignore the order information when applied to ordered categorical data.

In this Chapter, we define the rank ICC as a natural extension of Fisher's ICC to the rank scale. We provide its population parameter and extend it when the underlying distribution has more than two hierarchies. Our estimator of the rank ICC is insensitive to extreme values and skewed distributions, and does not depend on the scale of the data. It can be used for ordered categorical variables. We also show that our estimator is consistent and asymptotically normal. We have developed an R package, `rankICC`, available on CRAN, which implements our new method. The R script for the three application examples and simulations is on our Github page, https://github.com/shengxintu/rankICC.

2.2 Population Parameters

2.2.1 Two Hierarchies

Consider a two-level hierarchical distribution. A random variable from the distribution is denoted as X_{ij}, where i represents the cluster it belongs to and j is the index within cluster i. Fisher defined the ICC as the correlation between a random pair from the same cluster; that is, $\rho_I = corr(X_{ij}, X_{ij'})$, where $j \neq j'$, indicating that two different observations are drawn from cluster i. For a continuous hierarchical distribution, the ICC has also been expressed as the ratio of the between-cluster variance to the total variance (Fieller and Smith, 1951); $\rho_{Ir} = \sigma_b^2/(\sigma_b^2 + \sigma_w^2)$, where σ_b^2 is the between-cluster variance (i.e., the variance of cluster means), and σ_w^2 is the within-cluster variance (i.e., the mean of within-cluster variances). These two definitions are equivalent only when cluster sizes are infinite. In general, the relationship between these two definitions is

$$\begin{aligned}
\rho_I &= cov(X_{ij}, X_{ij'}) \Big/ \sqrt{var(X_{ij})var(X_{ij'})} \\
&= \{cov[E(X_{ij}|\mu_i), E(X_{ij'}|\mu_i)] + E[cov(X_{ij}, X_{ij'}|\mu_i)]\} \Big/ (\sigma_b^2 + \sigma_w^2) \\
&= \{cov(\mu_i, \mu_i) + E[cov(X_{ij}, X_{ij'}|\mu_i)]\} \Big/ (\sigma_b^2 + \sigma_w^2) \\
&= \rho_{Ir} + E[cov(X_{ij}, X_{ij'}|\mu_i)] \Big/ (\sigma_b^2 + \sigma_w^2), \qquad (2.1)
\end{aligned}$$

where μ_i is a random variable representing the mean of cluster i. If cluster sizes are finite, $\rho_{Ir} > \rho_I$ because $E[cov(X_{ij}, X_{ij'}|\mu_i)]$ in (2.1) is negative. With equal cluster sizes of m, the value of ρ_I is constrained between $-1/(m-1)$ and 1 (Fisher, 1925). Note that ρ_I can be negative when cluster sizes are finite, whereas ρ_{Ir} is always non-negative. While ρ_I is a correlation measure, ρ_{Ir} is a measure of the fraction of total variance attributable to cluster means. Hence, ρ_I is a more general measure of the intraclass correlation.

The rank ICC, to be defined below, is the rank-based version of Fisher's ICC, similar to Spearman's

rank correlation which is the rank-based version of Pearson's correlation (Kruskal, 1958). The relationship between the population parameters of Fisher's ICC and the rank ICC is identical to the relationship between those of Pearson's correlation and Spearman's rank correlation. The population parameters of Fisher's ICC and Pearson's correlation are correlations on the original scale of the variables, while the population parameter of Spearman's rank correlation is the grade correlation (i.e., the correlation between CDFs) for continuous variables (Kruskal, 1958), or more generally, the correlation of the population versions of midranks or ridits (Bross, 1958; Kendall, 1970).

Let F be the CDF of the two-level hierarchical distribution. Let $F(x-) = \lim_{t \uparrow x} F(t)$ and $F^*(x) = \{F(x) + F(x-)\}/2$. The population version of the rank ICC is defined as

$$\gamma_I = corr[F^*(X_{ij}), F^*(X_{ij'})], \tag{2.2}$$

where $(X_{ij}, X_{ij'})$ is a random pair drawn from a random cluster and $j \neq j'$. If X is continuous, $\gamma_I = 12cov[F(X_{ij}), F(X_{ij'})]$, because $F^*(X) = F(X) \sim \text{Unif}(0,1)$ and its variance is $1/12$. If X has a discrete or mixture distribution, $F^*(X_{ij})$ corresponds to the population version of ridits (Bross, 1958). The rank ICC γ_I given in (2.2) is therefore Fisher's ICC on the cumulative probability scale. The rank ICC has the same boundaries as Fisher's ICC and can be negative with finite cluster sizes.

2.2.2 Multiple Hierarchies

We extend the definition of the rank ICC to multiple hierarchies. For ease of understanding, we begin with three hierarchies. Starting from the innermost level, the three levels are named level 1, level 2, and level 3. One example is a population of schools, in which there are different classrooms and different students within each classroom. Here level 1 is the student, level 2 is the classroom, and level 3 is the school. Correlation may exist within both level-2 and level-3 units. A random variable drawn from a three-level hierarchical distribution is denoted as X_{ijk}, where i, j, and k are indices for levels 3, 2, and 1, respectively. Let F be the CDF of the three-level hierarchical distribution and $F^*(x) = \{F(x) + F(x-)\}/2$. The rank ICC at level 2, denoted as γ_{I2}, measures the correlation between a random pair of level-1 observations from the same level-2 unit. It is defined as

$$\gamma_{I2} = corr[F^*(X_{ijk}), F^*(X_{ijk'})], \tag{2.3}$$

where $k \neq k'$. The rank ICC at level 3, denoted as γ_{I3}, measures the correlation between a random pair of level-1 observations from the same level-3 unit but different level-2 units. It is defined as

$$\gamma_{I3} = corr[F^*(X_{ijk}), F^*(X_{ij'l})], \tag{2.4}$$

where $j \neq j'$ but k and l can be equal or different. At level 3, there are two potential sources of within-cluster correlation: one due to different level-2 units within the same level-3 unit and the other due to different level-1 units within the same level-2 unit. Our definition of γ_{I3} captures the former; the latter has already been captured by γ_{I2}. If we were to ignore the second level and consider the rank correlation between two random level-1 units from the same level-3 unit irrespective of their level-2 information, the resulting definition would reflect both sources of correlation, which is not ideal; it could be quite different from γ_{I3} with a small number of level-2 units within each level-3 unit. Our rank ICC definitions given by (2.3) and (2.4) have comparable interpretations to previously proposed definitions of ICC for 3 hierarchies on the original scale (Siddiqui et al., 1996).

The general definition of the rank ICC for a multiple-level hierarchical distribution can be similarly defined. Let Q be the number of hierarchies and $X_{I_Q I_{Q-1}...I_1}$ denote a random variable from a Q-level hierarchical distribution, where I_Q, I_{Q-1}, ..., I_1 are indices for levels Q, $Q-1$, ..., 1, respectively. The CDF of the Q-level hierarchical distribution is denoted as F, and $F^*(x) = \{F(x) + F(x-)\}/2$. The rank ICC at level j ($j \in \{2,3,...,Q\}$) measures the correlation between level-1 observations from the same level-j unit and different level-$(j-1)$ unit:

$$\gamma_{Ij} = corr[F^*(X_{I_Q I_{Q-1}...I_j I_{j-1}...I_1}), F^*(X_{I_Q I_{Q-1}...I_j I'_{j-1}...I'_1})], \tag{2.5}$$

where $I_{j-1} \neq I'_{j-1}$, and for $l < j-1$, I_l and I'_l can be the same or different.

2.3 Estimation and Inference

2.3.1 Estimation

Since the rank ICC can be viewed as a function of the underlying distribution $\gamma_I(F)$, then our estimator of γ_I is $\hat{\gamma}_I = \gamma_I(\hat{F})$. Given two-level data $\{x_{ij} : i = 1,2,...,n, j = 1,2,...,k_i\}$ with a total number of observations of $N = \sum_{i=1}^{n} k_i$, a nonparametric estimator of the CDF is $\hat{F}(x) = \sum_{i=1}^{n}\sum_{j=1}^{k_i} w_{ij} I(x_{ij} \leq x)$, where w_{ij} is the weight of observation x_{ij} and $\sum_{i=1}^{n}\sum_{j=1}^{k_i} w_{ij} = 1$. The weight w_{ij} depends on how we believe the data reflect the composition of the underlying hierarchical distribution; for example, $w_{ij} = 1/(nk_i)$ corresponds to equal weights for clusters and $w_{ij} = 1/N$ corresponds to equal weights for observations. Other weighting options will be described later in this section. The weight of cluster i is denoted as $w_{i\cdot} = \sum_{j=1}^{k_i} w_{ij}$. Similarly, we estimate $\hat{F}(x-) = \sum_{i=1}^{n}\sum_{j=1}^{k_i} w_{ij} I(x_{ij} < x)$, and define $\hat{F}^*(x) = \{\hat{F}(x) + \hat{F}(x-)\}/2$. Then our estimator of γ_I is $\hat{\gamma}_I = corr\{\hat{F}^*(X_{ij}), \hat{F}^*(X_{ij'})\}$, where $(X_{ij}, X_{ij'})$ is a random pair drawn from a random cluster and $j \neq j'$.

Because the rank ICC measures the correlation of a random pair from the same cluster, we could consider Monte Carlo estimation. That is, we first randomly select clusters with replacement and then randomly draw

pairs of observations from the selected clusters. Then γ_I could be estimated as the sample correlation of $\hat{F}^*(x)$ between the sampled pairs of observations. As the number of sampled pairs increases, the estimate of this approach will converge to a limit, which is our estimator:

$$\hat{\gamma}_I = \frac{\sum_{i=1}^{n} w_{i\cdot} \sum_{1\le j<j'\le k_i} \frac{2}{k_i(k_i-1)} [\hat{F}^*(x_{ij}) - \bar{\hat{F}}^*][\hat{F}^*(x_{ij'}) - \bar{\hat{F}}^*]}{\sum_{i=1}^{n}\sum_{j=1}^{k_i} w_{ij}[\hat{F}^*(x_{ij}) - \bar{\hat{F}}^*]^2}, \tag{2.6}$$

where $\bar{\hat{F}}^* = \sum_{i=1}^{n}\sum_{j=1}^{k_i} w_{ij}\hat{F}^*(x_{ij})$, and $k_i(k_i-1)/2$ is the number of possible unordered pairs in cluster i.

The estimator $\hat{\gamma}_I$ given by (2.6) is consistent for γ_I and is asymptotically normal. The proof of consistency and asymptotic normality and the variance estimation of $\hat{\gamma}_I$ are in Section 2.7. The results allow us to compute standard errors (SEs) of $\hat{\gamma}_I$ and to construct confidence intervals (CIs) for γ_I.

The selection of weights, w_{ij}, warrants additional discussion. For populations with finite and unequal cluster sizes, if there is ambiguity in the relative contributions of clusters in a hierarchical distribution, then the rank ICC can have some ambiguity. One could assume all clusters have an equal contribution regardless of their cluster sizes, in which case it would be sensible to set $w_{ij} = 1/(nk_i)$. Or one could assume the relative contributions are proportional to cluster sizes, in which case it would be sensible to set $w_{ij} = 1/N$. The choice of w_{ij} should be driven by subject matter knowledge. For example, if one is measuring the repeatability of an assay by collecting specimens (one per person) and measuring them multiple, unequal numbers of times, then it seems sensible to assume the clusters (people) contribute equally in the population. In contrast, if one is interested in the correlation of a trait between individuals within the same family, then it may (or may not) be sensible to assume each family contributes proportionally to the family size. These two weighting approaches have been applied to estimating the ICC on the original scale under variable cluster sizes (Karlin et al., 1981). For perfectly balanced data, $\hat{\gamma}_I$ is the same regardless of the weighting approach used.

However, in practice, there is often uncertainty in how we should assume clusters contribute to the underlying distribution and we may want to consider different weighting schemes. In fact, there may be bias-variance considerations that might suggest using weights that do not exactly match the true cluster contributions. For example, consider a population with equal cluster contribution. When γ_I is close to zero, observations in the same cluster are almost independent, so treating all observations equally regardless of the cluster size can be more efficient than weighting observations inversely proportional to the size of their cluster. In contrast, when γ_I is close to one, observations in the same clusters are almost redundant, favoring equal weight per cluster. But whether γ_I is close to zero or one is often unknown before analysis. Therefore, one might use an iterative procedure to identify a more efficient weighting scheme. One approach is to use a linear combination of the two weights above, where the combination depends on the value of γ_I; that is,

$w_{ij}(\gamma_I) = (1-\gamma_I)/N + \gamma_I/(nk_i)$. We call this the combination approach. Another approach is to compute the effective sample size (ESS) (Kish, 1965) for the clusters (i.e., $n_i^{(e)} = k_i/(1+k_i\gamma_I)$) and weight clusters and their observations in clusters accordingly; that is, $w_{i.}(\gamma_I) = n_i^{(e)}/\sum_{j=1}^{n} n_j^{(e)}$ and $w_{ij}(\gamma_I) = w_{i.}(\gamma_I)/k_i$. We call this second approach the ESS approach. These approaches require a working value of γ_I. We implement iterative procedures in which we (a) start with an initial value of γ_I, (b) update the weights, and (c) compute a new estimate of γ_I. We repeat steps (b) and (c) multiple times until the estimate of γ_I converges. Our simulations suggest that the choice of the initial value has no effect on the final estimate.

With three or more hierarchies, the estimation of the rank ICC is similar to that described above for two hierarchies. Given three-level nested data $\{x_{ijk}; i = 1,2,...,n, j = 1,2,...,n_i, k = 1,2,...,m_{ij}\}$, the nonparametric estimator for the CDF is $\hat{F}(x) = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\sum_{k=1}^{m_{ij}} w_{ijk} I(x_{ijk} \le x)$, where $\sum_{i=1}^{n}\sum_{j=1}^{n_i}\sum_{k=1}^{m_{ij}} w_{ijk} = 1$. Similarly, $\hat{F}(x-) = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\sum_{k=1}^{m_{ij}} w_{ijk} I(x_{ijk} < x)$. Let $\hat{F}^*(x) = \{\hat{F}(x) + \hat{F}(x-)\}/2$. The general form of the estimator of γ_{I2} is

$$\hat{\gamma}_{I2} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n_i} w_{ij.} \sum_{1\le k<k'\le m_{ij}} \frac{2}{m_{ij}(m_{ij}-1)} [\hat{F}^*(x_{ijk}) - \bar{\hat{F}}^*][\hat{F}^*(x_{ijk'}) - \bar{\hat{F}}^*]}{\sum_{i=1}^{n}\sum_{j=1}^{n_i}\sum_{k=1}^{m_{ij}} w_{ijk}[\hat{F}^*(x_{ijk}) - \bar{\hat{F}}^*]^2}, \tag{2.7}$$

where $w_{ij.} = \sum_{k=1}^{m_{ij}} w_{ijk}$ and $\bar{\hat{F}}^* = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\sum_{k=1}^{m_{ij}} w_{ijk}\hat{F}^*(x_{ijk})$. The general form of the estimator of γ_{I3} is

$$\hat{\gamma}_{I3} = \frac{\sum_{i=1}^{n} w_{i..} \sum_{1\le j<j'\le n_i}\sum_{k=1}^{m_{ij}}\sum_{l=1}^{m_{ij'}} \frac{1}{c_i} [\hat{F}^*(x_{ijk}) - \bar{\hat{F}}^*][\hat{F}^*(x_{ij'l}) - \bar{\hat{F}}^*]}{\sum_{i=1}^{n}\sum_{j=1}^{n_i}\sum_{k=1}^{m_{ij}} w_{ijk}[\hat{F}^*(x_{ijk}) - \bar{\hat{F}}^*]^2}, \tag{2.8}$$

where $w_{i..} = \sum_{j=1}^{n_i}\sum_{k=1}^{m_{ij}} w_{ijk}$, and c_i is the total number of possible unordered pairs in a level-3 unit; $c_i = \{(\sum_{j=1}^{n_i} m_{ij})^2 - (\sum_{j=1}^{n_i} m_{ij}^2)\}/2$. We show the asymptotic normality and consistency of $\hat{\gamma}_{I2}$ and $\hat{\gamma}_{I3}$ in Section 2.7. There are several options for w_{ijk} with three-level data, such as assigning equal weights to all level-1 units (i.e., $w_{ijk} = 1/(\sum_{i=1}^{n}\sum_{j=1}^{n_i} m_{ij})$), assigning equal weights to all level-2 units (i.e., $w_{ijk} = 1/(m_{ij}\sum_{i=1}^{n} n_i)$), or assigning equal weights to all level-3 units (i.e., $w_{ijk} = 1/(nn_i m_{ij})$).

2.3.2 Inference

The distribution of $\hat{\gamma}_I$ can be approximated using asymptotics. The asymptotic standard error (SE) of $\hat{\gamma}_I$, presented in Section 2.7, can be used to construct confidence intervals for γ_I under normality. Because γ_I is bounded, one might also consider estimating the large sample distribution of the Fisher transformed value (i.e., $\log\{(1+\hat{\gamma}_I)/(1-\hat{\gamma}_I)\}/2$) by the delta method to obtain confidence intervals (Fisher, 1915).

An alternative approach for estimating the distribution of $\hat{\gamma}_I$ is bootstrapping. There are two general ways to implement bootstrapping in clustered data; the cluster bootstrap and the two-stage bootstrap (Davison and Hinkley, 1997; Field and Welsh, 2007). In the cluster bootstrap, clusters are randomly selected with replacement. The two-stage bootstrap has an extra step, where in the selected clusters the observations are

randomly drawn with replacement. In our setting, this intracluster sampling in the two-stage bootstrap can cause positive bias in estimating γ_I, because the same observation may be sampled twice in a two-stage bootstrap sample, thus inflating the estimated ICC, particularly in settings with smaller cluster sizes. Hence, we recommend using the cluster bootstrap for bootstrapping.

The standard errors of $\hat{\gamma}_{I2}$ and $\hat{\gamma}_{I3}$ with three hierarchies can be similarly computed. We have derived analytic formulas for asymptotic SEs of $\hat{\gamma}_{I2}$ and $\hat{\gamma}_{I3}$ given by (2.7) and (2.8) in Section 2.7. In addition, one could bootstrap. Considering computational efficiency and the bias caused by intracluster sampling with replacement, we suggest a one-stage bootstrap for bootstrapping with three hierarchies; i.e., only sampling level-3 units with replacement.

2.4 Simulations

A simple additive model was used to generate two-level data: $X_{ij} = U_i + R_{ij}$, where $U_i \overset{i.i.d}{\sim} N(1,1)$ and $R_{ij} \overset{i.i.d}{\sim} N(0,(1-\rho)/\rho)$ with ρ varying in $[0,1]$. Let Y_{ij} be the observation of the jth individual in the ith cluster, where $i = 1,2,....,n$; $j = 1,2,...,k_i$; and k_i is the cluster size of the ith cluster. We considered three scenarios: (I) $Y_{ij} = X_{ij}$; (II) $Y_{ij} = \exp(X_{ij})$; (III) $Y_{ij} = U_i^{'} + R_{ij}$, where $U_i^{'}$'s are i.i.d. following a log-normal distribution such that $var(U_i') = 1$ and $\log(U_i') \overset{i.i.d}{\sim} N(1, \log(1/2 + \sqrt{\exp(-2)+1/4}))$. In Scenarios I and II, since X_{ij} is normally distributed, the rank ICC is $\gamma_I = 6\arcsin(\rho/2)/\pi$ (Pearson, 1907). The rank ICC is identical in Scenarios I and II while Fisher's ICC, ρ_I, is sensitive to skewness and depends on the scale of interest (Figure 2.1). When the variable of interest is normal (Scenario I), γ_I is close to ρ_I. In Scenario III, Y_{ij} is not normally distributed so we empirically computed γ_I by generating a million clusters each with 2 observations, and then computing Spearman's rank correlation.

We first evaluated the performance of our estimator of γ_I for two-level data. The simulations were conducted at different sample sizes n = 25, 50, 100, 200, 500, and 1000 with an equal cluster size (k_i=30). Furthermore, we also performed simulations with various configurations of cluster size at $n = 200$: $k_i = 2$; $k_i = 30$; k_i uniformly ranging from 2 to 50; and $k_i = 2$ for half of the clusters and $k_i = 30$ for the other half. Unless stated otherwise, for estimation, we assigned equal weights to clusters (i.e., $w_{ij} = 1/(nk_i)$), which corresponds with the underlying equal cluster contribution in the simulated hierarchical distribution. We computed 95% confidence intervals for γ_I using the asymptotic SE and bootstrapping.

The bias of our estimator and the coverage of 95% CIs based on the asymptotic SE under the different scenarios described above are shown in Figures 2.2 and 2.3. In summary, our estimator of γ_I had very low bias and good coverage with modest numbers of clusters across all scenarios we considered. It was also robust to the skewed data in Scenarios II and III. Although our estimator had slightly negative bias with a small number of clusters, this bias decreased as the number of clusters increased. Confidence intervals for γ_I

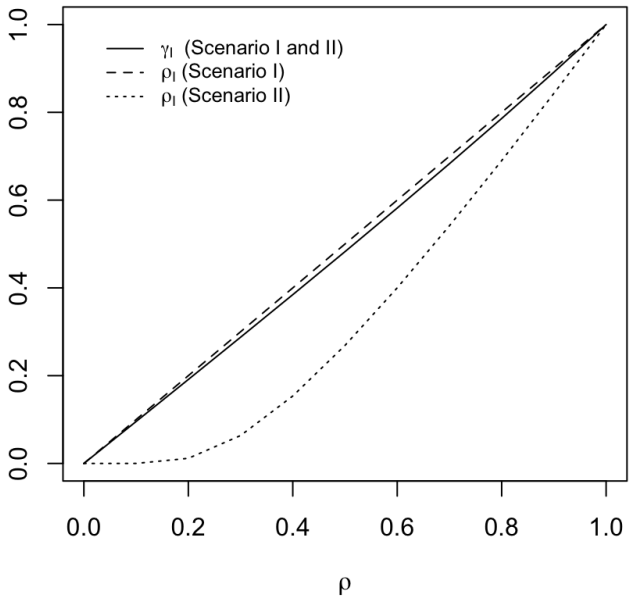
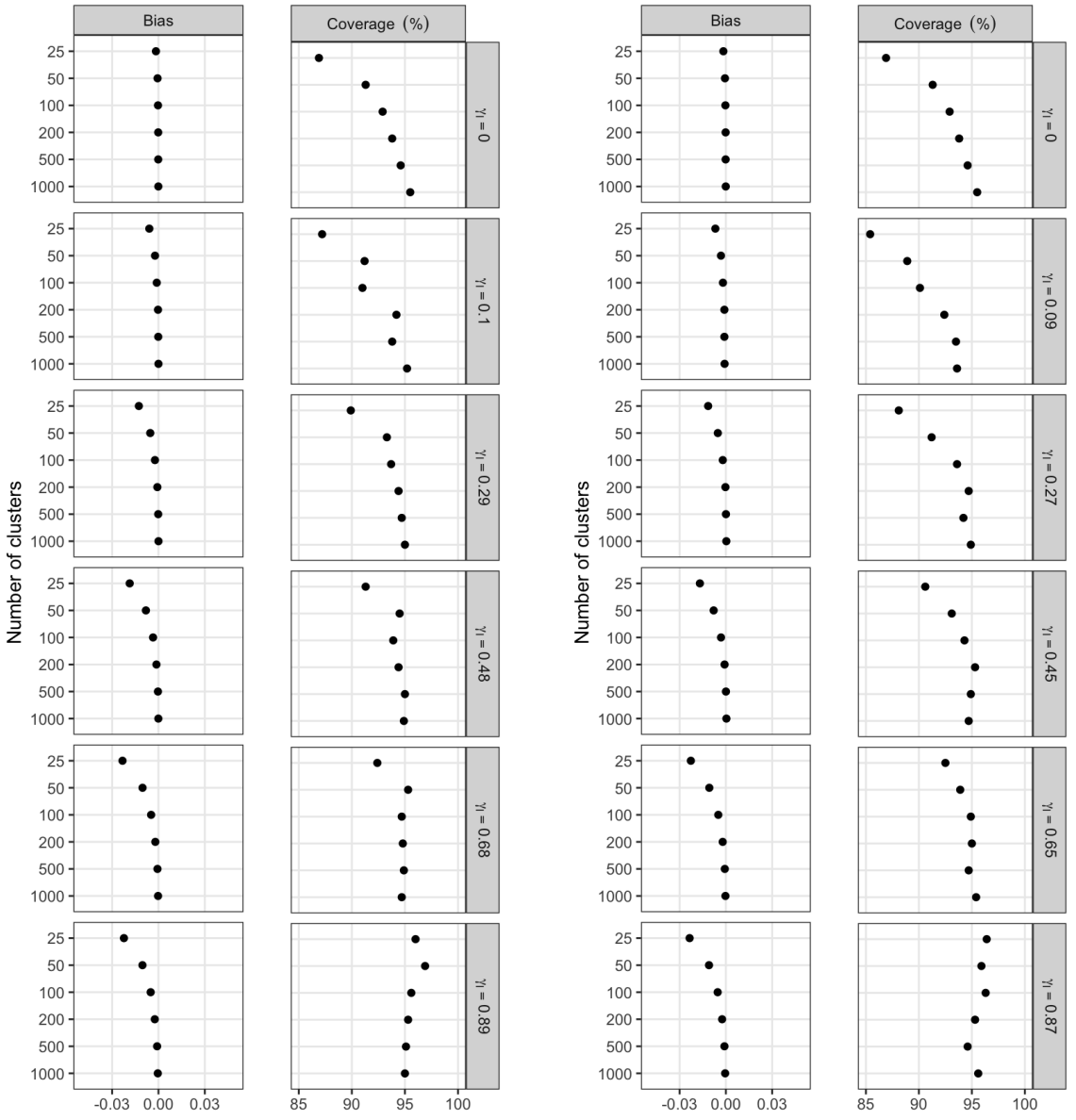


Figure 2.1: Parameters of rank ICC (γ_I) and Fisher's ICC (ρ_I) as a function of the within-cluster correlation (ρ) of X_{ij} under normality (Scenario I) and after exponentiating the data (Scenario II)

based on the asymptotic SE approximately covered at their nominal 0.95 level with ≥ 200 clusters for all true values of γ_I. For smaller values of γ_I, coverage could be low for ≤ 100 clusters. Fisher transformation did not appear to improve coverage (Figure 2.11). The performance of estimators was fairly similar regardless of the size of clusters (Figure 2.3). Additional simulations reported in Tables 2.4 – 2.8 show that confidence intervals based on both the cluster bootstrap SE and percentiles had good coverage.

We then evaluated the performance of our estimator of γ_I when the cluster size is 2 in the population and the rank ICC varies between -1 and 1. Let X_{i1} and X_{i2} be the two observations in cluster i. We generated the two observations in cluster i as follows: $X_{i1} = U_i + R_i$ and $X_{i2} = U_i - R_i$, where $U_i \overset{i.i.d}{\sim} N(1,1)$, $R_i \overset{i.i.d}{\sim} N(0,(1-\rho)/(1+\rho))$, and ρ varies over $[-1,1]$ (we set $var(U_i) = 0$ and $var(R_i) = 20$ when $\rho = -1$). We conducted 1000 simulations at $n = 200$. Our estimator of γ_I had low bias and good coverage (Figure 2.4).

We next compared the performance of the four weighting approaches with 1) equal within-cluster variances and equal or unequal cluster sizes, and 2) within-cluster variances varying by cluster size. For 1), we used the same simulations described in the first paragraph of this section. For 2), we supposed the numbers of small clusters of size 2 and large clusters of size 30 are equal in the population, and simulated the data as $X_{ij} = U_i + R_{ij}$, with $R_{ij} \overset{i.i.d}{\sim} N(0, c(1-\rho)/\rho)$, where $c = 0.5$ for small clusters and $c = 1.5$ for large clusters

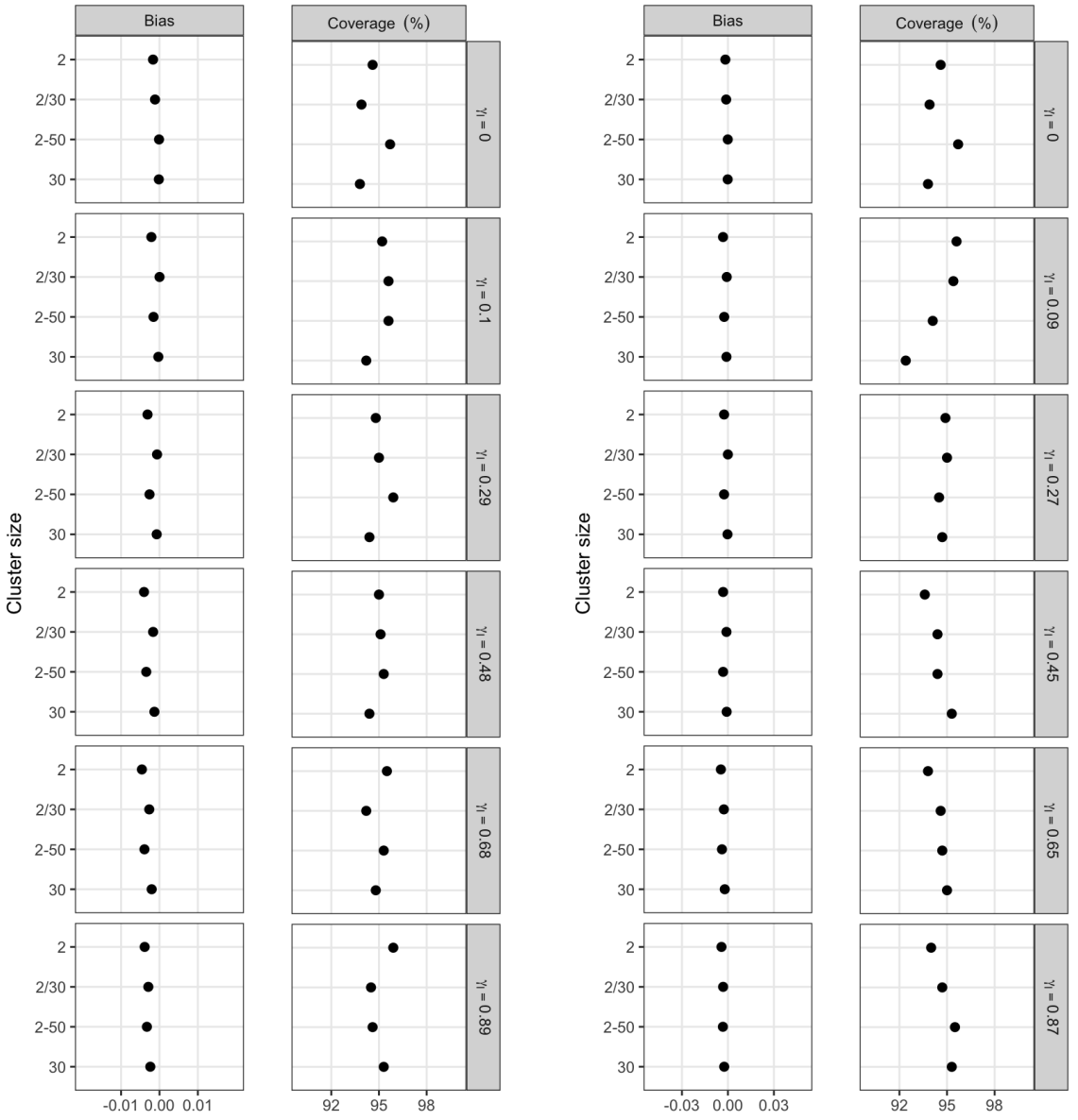


(a) Scenarios I and II

(b) Scenario III

Figure 2.2: Bias and coverage of 95% CIs for $\hat{\gamma}_1$ at different true γ_1 and numbers of clusters under Scenarios I (normality), II (exponentiated outcomes), and III (exponentiated cluster means). The number of observations per cluster was set at 30.

and ρ varying over $[0,1]$. We conducted 1000 simulations at $n = 200$. Results of the two sets of simulations are shown in Figures 2.5 and 2.12. When cluster sizes were equal and within-cluster variances were equal, the estimates of the four weighting approaches were identical. When cluster sizes were unequal and within-cluster variances were equal, the four methods all had low bias and their mean squared errors were dominated by their variances. As hypothesized, assigning equal weights to clusters had the lowest efficiency when the



(a) Scenarios I and II

(b) Scenario II

Figure 2.3: Bias and coverage of 95% CIs for $\hat{\gamma}_I$ at different true γ_I and cluster sizes under Scenarios I, II, and III. The number of clusters was set at 200. "2-50" means the cluster size follows a uniform distribution from 2 to 50, "2/30" means half of the clusters have size 2 and half have 30.

rank ICC was close to zero, because treating large and small clusters equally resulted in lost information, even though the data were simulated in a manner such that equal cluster weighting matched the cluster contribution in the population. In contrast, when within-cluster variances varied by cluster sizes, assigning equal weights to observations contrary to the underlying distribution led to bias. The two iterative approaches had lower mean squared errors than assigning equal weights to clusters or to observations when the rank ICC was

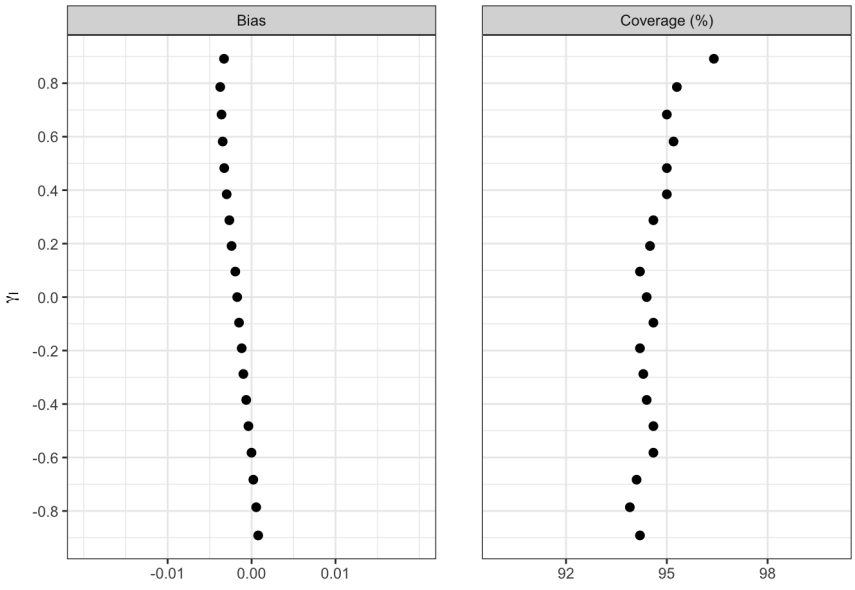


Figure 2.4: Bias and coverage of 95% CIs for $\hat{\gamma}_I$ with cluster sizes of 2 and γ_I varying between -1 and 1. The number of clusters was set at 200.

close to zero.

We also evaluated the performance of our estimator of γ_I for ordered categorical variables. We simulated data of 3-level, 5-level, and 10-level ordered categorical variables by discretizing X_{ij} in Scenario I with cut-offs at quantiles (i.e., using the 1/3 and 2/3 quantiles for 3 levels; the 0.2, 0.4, 0.6, 0.8 quantiles for 5 levels; and the 0.1, 0.2, ..., 0.8, 0.9 quantiles for 10 levels). Similar to Scenario III, we empirically computed γ_I for the ordered categorical variables (Figure 2.6). The rank ICCs of the 5-level and 10-level variables are close to the rank ICC of the continuous variable, while the rank ICC of the 3-level variable is slightly smaller. We conducted simulations for 3-level and 10-level ordered categorical variables at different sample sizes $n = 25$, 50, 100, 200, 500, and 1000 with an equal cluster size (k_i=30). Our estimator of γ_I of the ordered categorical variables generally had low bias and good coverage (Figure 2.13).

We also investigated the performance of our estimators of γ_{I2} and γ_{I3} for data with three hierarchies. Let X_{ijk} be the observation of the kth level-1 unit in the jth level-2 unit and the ith level-3 unit, where $i = 1,2,...,n$, $j = 1,2,...,n_i$; $k = 1,2,...,m_{ij}$. We generated three-level data as follows: $X_{ijk} = U_i + V_{ij} + R_{ijk}$, where $U_i \overset{i.i.d}{\sim} N(1, 20\rho_{I3}), V_{ij} \overset{i.i.d}{\sim} N(0, 20(\rho_{I2} - \rho_{I3})), R_{ijk} \overset{i.i.d}{\sim} N(0, 20(1-\rho_{I2}))$, and $(\rho_{I2}, \rho_{I3}) \in \{(0,0), (0.25, 0.20), (0.55, 0.20), (0.85, 0.20), (0.55, 0.5), (0.85, 0.5), (0.85, 0.8)\}$. Because of normality, the true rank ICCs are $\gamma_{I2} = 6\arcsin(\rho_{I2}/2)/\pi$ and $\gamma_{I3} = 6\arcsin(\rho_{I3}/2)/\pi$. We conducted 1000 simulations for different sample sizes $n = 25$, 50, 100, 200, 500, and 1000 under equal cluster sizes (i.e., $n_i = 15$ and $m_{ij} = 2$). Moreover, we also performed simulations under various patterns of cluster sizes: $(n_i, m_{ij}) \in \{(15,2), (2,15), (4,2), (2-$

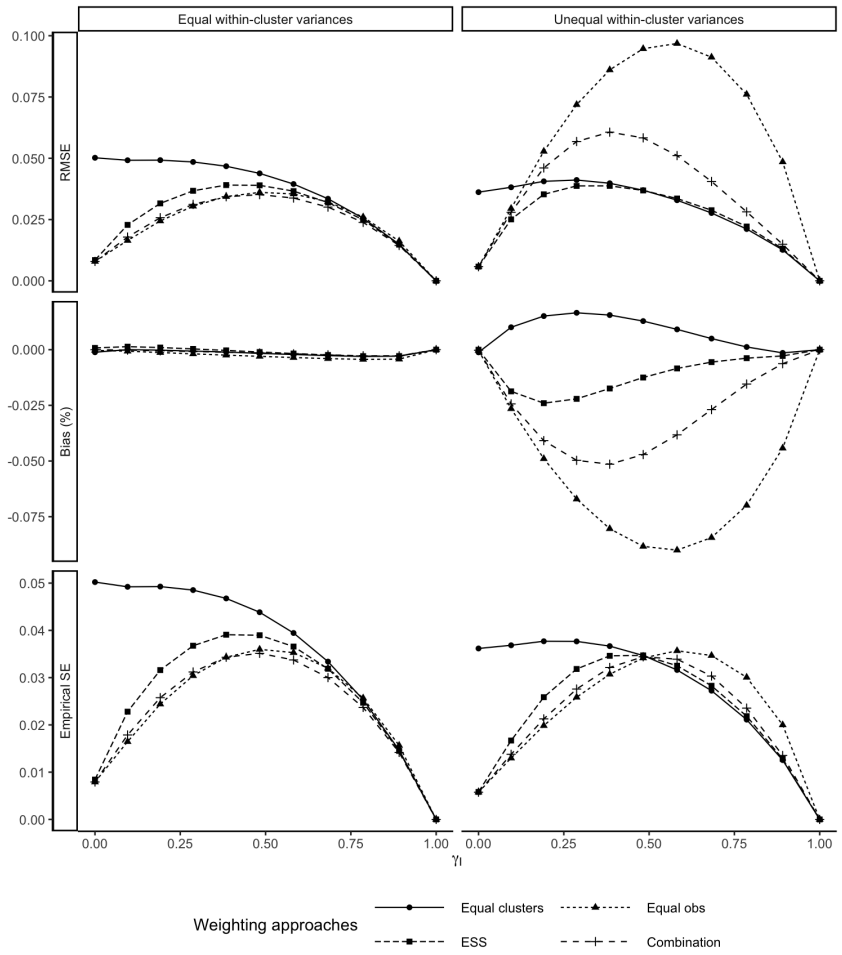


Figure 2.5: Root mean squared error (RMSE), bias, and empirical SE of estimates obtained by the four weighting approaches for our estimator of γ_l. "Equal clusters" refers to assigning equal weights to clusters, "Equal obs" refers to assigning equal weights to observations, "ESS" refers to the iterative weighting approach based on the effective sample size, and "Combination" refers to the iterative weighting approach based on the linear combination of equal weights for clusters and equal weights for observations. We set the tolerance of the two iterative approaches to 0.00001.

15, 2), (2/15, 2), (2, 2-15), (2, 2-15), (2/15, 2/15)}, where "2-15" means the cluster size follows a uniform distribution from 2 to 15, "2/15" means half of the clusters have size 2 and a half have 15. The results for $n_i = 15$ and $m_{ij} = 2$ are shown in Figure 2.7, and the other results are in Figures 2.13 and 2.14 and Tables 2.9 and 2.10. Our estimators of γ_{l2} and γ_{l3} had very low bias and good coverage in all cases we considered. The asymptotic SE and the one-stage bootstrap had good performance in constructing confidence intervals, and the former was computationally efficient.

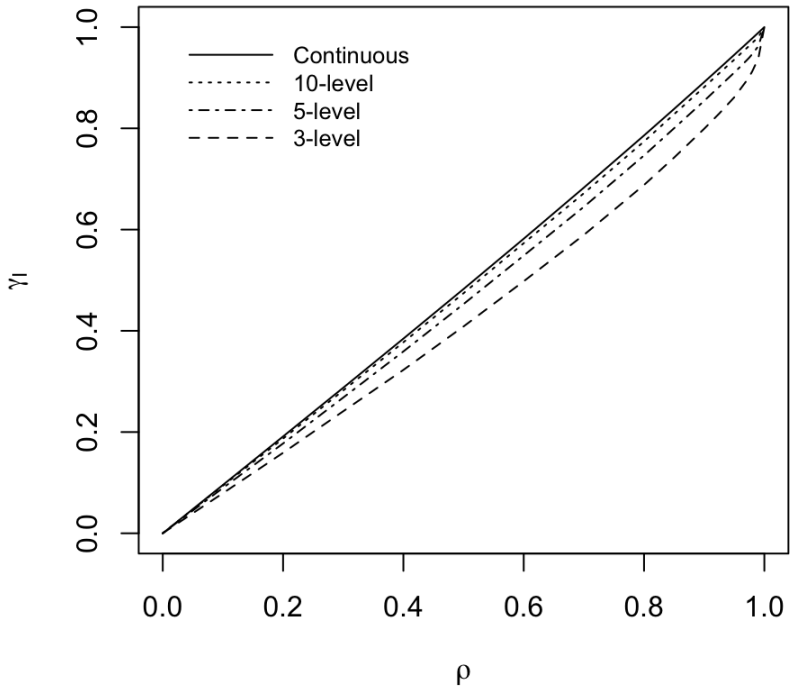


Figure 2.6: Parameters of rank ICC (γ_I) as a function of the within-cluster correlation (ρ) of X_{ij} when data are continuous or discretized into ordered categorical variables with 3, 5, or 10 levels.

2.5 Applications

2.5.1 Albumin-Creatinine Ratio

In a cross-sectional study, 598 people living with HIV in Nigeria on stable dolutegravir-based antiretroviral therapy provided first-morning void urine specimens at two visits 4 to 8 weeks apart (Wudil et al., 2021). The collected urine specimens were used to calculate the urine albumin-creatinine ratio (uACR). There is interest in estimating the intraclass correlation of uACR. Each patient is considered a cluster, and each cluster has two observations. The uACR measurements are right-skewed, and the empirical distributions of the first and second uACR measurements were comparable (Figure 2.8). The rank ICC estimate was 0.217 (95% CI: 0.140-0.295, Table 2.1). The traditional ICC estimate on the original scale obtained from a random effects model was 0.493, which was driven by a single pair of measurements with extreme values. After removing that pair, the rank ICC estimate was almost unchanged (0.213, 95% CI: 0.136-0.291) while the traditional ICC on the original scale dropped dramatically to 0.160, illustrating the robustness of the rank ICC compared to the traditional ICC. Instead of removing extreme observations, one could consider transforming the data.

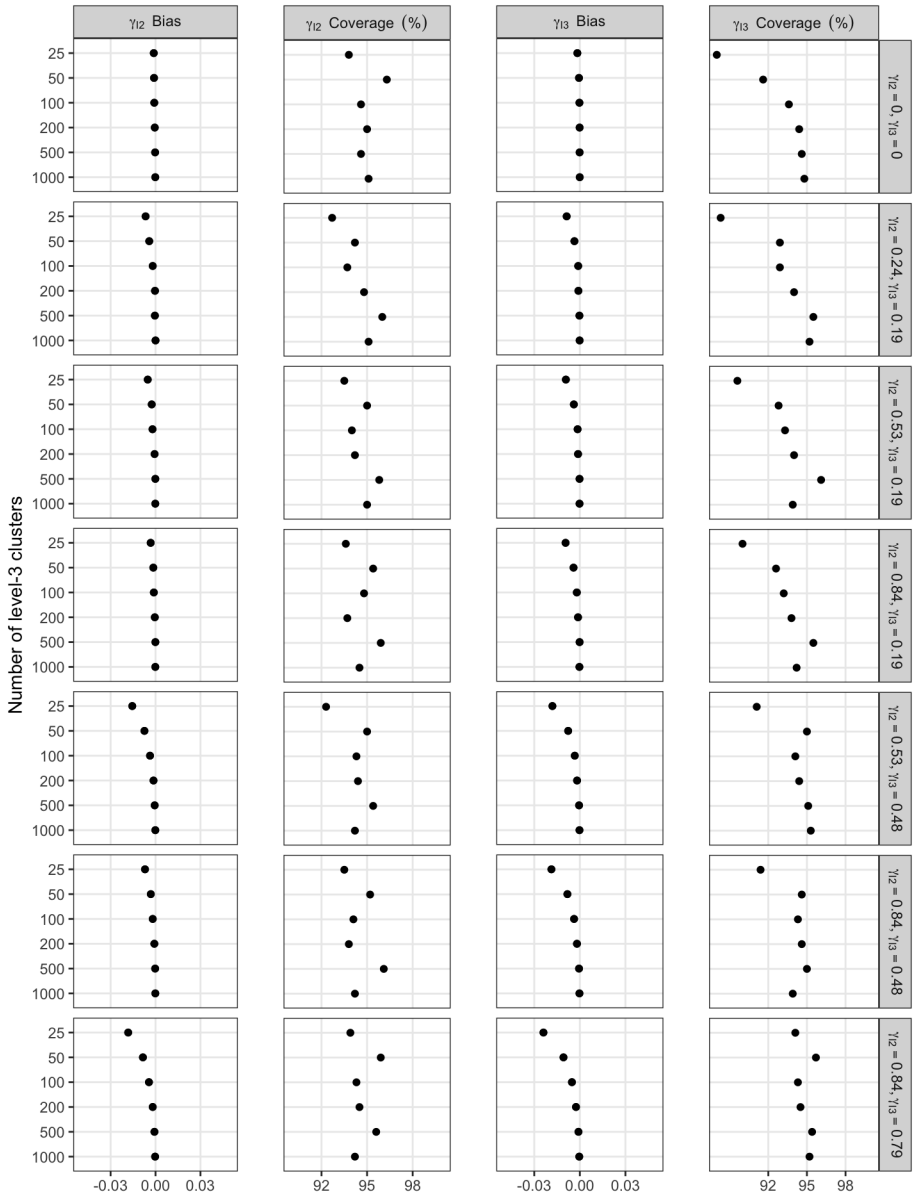


Figure 2.7: Bias and coverage of 95% CIs for $\hat{\gamma}_{I2}$ and $\hat{\gamma}_{I3}$ at different true values of γ_{I2} and γ_{I3} and different numbers of level-3 units. The number of level-2 units in a level-3 unit was set at 15. The number of level-1 units in a level-2 unit was set at 2.

The traditional ICC estimate was 0.254 after log transformation and 0.345 after square root transformation, illustrating the sensitivity of the traditional ICC to the choice of scale.

2.5.2 Status Epilepticus

The Bridging the Childhood Epilepsy Treatment Gap in Africa (BRIDGE) study is a non-inferiority randomized clinical trial of childhood epilepsy care at 60 randomly selected primary healthcare centers (PHCs) in

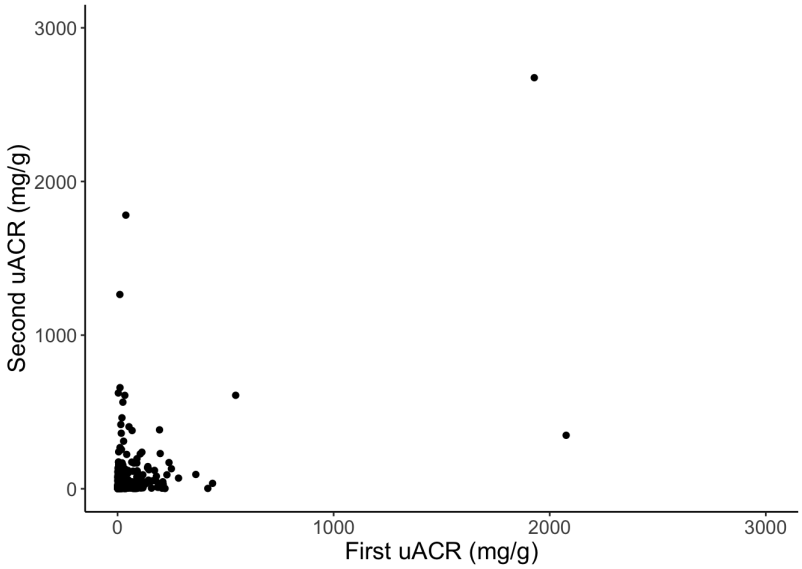


Figure 2.8: Scatter plot of the first and second uACR measurements of each person in the example of albumin-creatinine ratio.

Table 2.1: Estimates of rank ICC and traditional ICC of uACR in the example of albumin-creatinine ratio.

	Original data	Extreme values removed	Log transformation	Square root transformation
Rank ICC [95% CI]	0.217 [0.140, 0.295]	0.213 [0.136, 0.291]	0.217 [0.140, 0.295]	0.217 [0.140, 0.295]
ICC	0.493	0.160	0.254	0.345

northern Nigeria (Aliyu et al., 2019). The trial is designed to understand if task-shifting childhood epilepsy treatment by trained community health workers can be as effective at reducing seizures as treatment by trained physicians. The study recruited 1507 children with untreated epilepsy from the participating PHCs. Each child's number of seizures in the six months prior to randomization was collected (Figure 2.9), with a median of 10 (range 1-50). There is interest in estimating the ICC for the number of seizures across PHCs. Cluster size ranged from 19 to 31 children per PHC. Since the PHCs were the units of randomization in this study, it seems reasonable to treat them equally. The rank ICC based on assigning equal weights to clusters was estimated as 0.0482 (95% CI: 0.023-0.073), which suggested low association between the number of seizures in children within a PHC (Table 2.2). Other methods of weight assignment yielded similar estimates: assigning equal weights to children resulted in an estimate of 0.0514 (95% CI: 0.025-0.078), the ESS weighting approach yielded 0.0496 (95% CI: 0.024-0.075), and the combination weighting approach resulted in 0.0512 (95% CI: 0.025-0.078). For comparison, the ICC estimated using a linear random effects model was

0.0426, and the ICCs estimated using generalized linear random effects models were 0.0268 (quasi-Poisson) and 0.0168 (negative binomial) (Nakagawa et al., 2017).

Figure 2.9: Histogram of numbers of seizures of children with untreated epilepsy from the 60 primary healthcare centers in the example of status epilepticus.

Table 2.2: Estimates of rank ICC and traditional ICC for the number of seizures across the primary healthcare centers in the sample of status epilepticus.

Rank ICC [95% CI]	Equal weights for clusters	0.0482 [0.023, 0.073]
	Equal weights for observations	0.0514 [0.025, 0.078]
	Iterative weighting based on the effective sample size	0.0496 [0.024, 0.075]
	Iterative weighting based on the combination	0.0512 [0.025, 0.078]
ICC	Linear	0.0426
	Quasi-Poisson link	0.0268
	Negative binomial link	0.0168

2.5.3 Patient Health Questionnaire-9 Score

In a third example, we used baseline data from the Homens para Saúde Mais (HoPS+) study, a clustered randomized controlled trial in Zambézia Province, Mozambique (Audet et al., 2018). The trial aimed to measure the impact of incorporating male partners with HIV into prenatal care for pregnant women living with HIV on retention in care, adherence to treatment, and mother-to-child HIV transmission. The trial enrolled 813 couples living with HIV (with a pregnant female) at 24 clinical sites. Depressive symptoms at the time of study enrollment were evaluated with the Patient Health Questionnaire-9 (PHQ-9), a nine-item

scale that measures depressive symptoms over the previous two weeks. The ordinal PHQ-9 score had a median of 2 (interquartile range 0-5), ranging from 0 to 27 (Figure 2.10). The data have three levels: the innermost level is the person, the middle level is the couple, and the outer level is the clinical site. The number of couples at a clinical site ranged from 2 to 68. Our estimates assigned equal weights to couples. The estimated rank ICC at the couple level, $\hat{\gamma}_{I2}$, was 0.678 (95% CI: 0.518-0.838), suggesting substantial clustering of PHQ-9 scores within couples (Table 2.3). The estimated rank ICC at the clinical site level, $\hat{\gamma}_{I3}$, was 0.397 (95% CI: 0.242-0.552), which was higher than expected, suggesting a fairly high correlation within clinics. This was confirmed by the estimated rank ICC among females at the clinical level (0.418, 95% CI: 0.260-0.576) and among males (0.395, 95% CI: 0.243-0.548). For comparison, the ICC estimates obtained from a linear random effects model were 0.792 at the couple level and 0.474 at the clinical site level, both larger than their rank ICC counterparts.

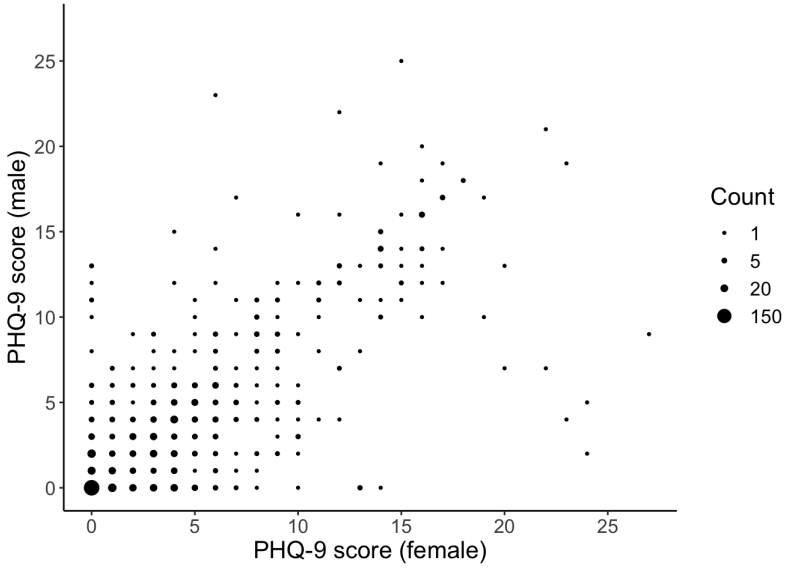


Figure 2.10: Scatter plot of PHQ-9 scores of male and female partners enrolled in the clustered randomized clinical trial in the example of Patient Health Questionnaire-9 score.

Table 2.3: Estimates of rank ICC and traditional ICC of PHQ-9 score at the couple level and the clinical site level in the example of Patient Health Questionnaire-9 score.

	The couple level	**The clinical site level**	**Females at the clinical level**	**Males at the clinical level**
Rank ICC [95% CI]	0.678 [0.518, 0.838]	0.397 [0.242, 0.552]	0.418 [0.260, 0.576]	0.395 [0.243, 0.548]
ICC	0.792	0.474	0.452	0.497

2.6 Discussion

In this Chapter, we defined the rank ICC as a natural extension of Fisher's ICC to the rank scale, and described its population parameter. Our approach maintains the spirit of Fisher's ICC while creating a nonparametric rank ICC measure analogous to Spearman's rank correlation. The rank ICC is simply interpreted as the rank correlation between a pair of observations from the same cluster. We also extended the rank ICC for distributions with more than two hierarchies (i.e., equation (2.5)). Our estimator of the rank ICC is insensitive to extreme values and skewed distributions, and does not depend on the scale of the data. It is also consistent and asymptotically normal, with low bias and good coverage in our simulations. Our framework is general, and applicable to any orderable variables with estimable distributions.

We also discussed assigning weights to clusters and observations under different cases when estimating the rank ICC for two-level data with heterogeneous cluster sizes. In general, the selection of weights should be driven by subject matter knowledge. However, in practice, there may be uncertainty in how clusters contribute to the underlying distribution, and efficiency considerations might guide the choice of weights.

There is a relationship between the rank ICC and Spearman's rank correlation when the cluster size is two. With two ordered observations per cluster following the same marginal distribution, the population parameter of the rank ICC is mathematically equal to that of Spearman's rank correlation between the first and second observations. However, their estimation procedures differ; in estimating Spearman's rank correlation, we separately estimate the variances and means of the first and second observations, but in estimating γ_I, we pool the data to estimate their overall variance and mean. For example, in the albumin-creatinine ratio study, the estimate of Spearman's rank correlation between the first and second uACR measurements was 0.236, close but not equal to the rank ICC estimate, 0.217.

Our rank ICC fills an important gap in the analysis of clustered data. Given Fisher's introduction of the ICC nearly 100 years ago, we are surprised that a rank-based ICC has not been developed until now. We suspect that some researchers may have simply ranked their data and then used the ratio of the between-cluster and total variances on the rank scale as a rank-based ICC measure, as suggested by others for ordered categorical data (Hallgren, 2012; Denham, 2016). Although not completely unreasonable, such an approach is ad hoc and does not correspond with a sensible population parameter. Alternatively, some researchers may prefer estimating the similarity within clusters via constructing models for continuous and ordered categorical clustered data, in particular random effects models (Agresti and Natarajan, 2001; Skrondal and Rabe-Hesketh, 2004; Koo and Li, 2016). With linear mixed models, the ICC is calculated using estimates of the variance of the random effects and the residuals. These model-derived ICC estimates may be sensitive to the choice of the model: e.g., the form of the linear predictor, potential response variable transformation, non-normality

of residuals, and/or non-normality of random effects. With generalized random effects models (e.g., for count or ordinal response variables), the ICC is evaluated on the continuous latent variable scale after a link transformation, which complicates interpretation and remains sensitive to model choice (Skrondal and Rabe-Hesketh, 2004). These models may also be sensitive to the method used to derive the within-cluster variance (Nakagawa et al., 2017). In contrast, our rank ICC does not require fitting a model and provides a simple and interpretable one-number summary of within-cluster similarity across many types of variables.

Our rank ICC has some limitations. It does not adjust for the effect of other variables on within-cluster similarity. For example, in the status epilepticus study (Section 2.5.2), there may be interest in measuring the intraclass correlation after adjusting for child age. Our rank ICC cannot do this, whereas a model-derived ICC estimate can. Future work could consider extensions to develop covariate-adjusted conditional and partial rank ICCs. An additional limitation is that our rank ICC appears to have slightly negative bias with small numbers of clusters when γ_I is large; this problem goes away as the number of clusters increases. Furthermore, our rank ICC can be time-consuming to calculate with very large sample sizes. In such settings, analysts may consider empirically estimating the rank ICC by randomly sampling clusters with replacement, then sampling pairs of observations from the selected clusters, and finally estimating Spearman's correlation across many sampled pairs.

In Chapter 4, we apply the rank ICC to the designs of clustered randomized controlled trials with skewed or ordered categorical outcomes.

2.7 Supplementary Material

2.7.1 Proof of asymptotic properties of $\hat{\gamma}_I$ with two hierarchies

Let $g_n(\mathbf{x})$ and $h_n(\mathbf{x})$ be two functions such that $\hat{\gamma}_I = [\frac{1}{n}\sum_{i=1}^n g_n(\mathbf{x}_i)]/[\frac{1}{n}\sum_{i=1}^n h_n(\mathbf{x}_i)]$, where $\mathbf{x}_i = (x_{i1},...,x_{ik_i})$ denotes a vector of observations in a cluster. Let $\mathbf{P}_n$ denotes an empirical measure and $\mathbf{P}$ denotes the underlying probability measure such that $\mathbf{P}_n\{g(\mathbf{x})\} = \frac{1}{n}\sum_{i=1}^n g(\mathbf{x}_i)$ and $E[\mathbf{P}_n\{g(\mathbf{x})\}] = \mathbf{P}\{g(\mathbf{x})\}$, and same for $h(\mathbf{x})$. We then have $\hat{\gamma}_I = \frac{\mathbf{P}_n\{g_n(\mathbf{x})\}}{\mathbf{P}_n\{h_n(\mathbf{x})\}}$ and $\gamma_I = \frac{\mathbf{P}\{g(\mathbf{x})\}}{\mathbf{P}\{h(\mathbf{x})\}}$. Here $\hat{F}^*$ belongs to a Donsker class which contains all distribution functions in X's support, where a Donsker class is a sets of functions with the useful property that empirical processes indexed by the class converge weakly to a certain Gaussian process.

ASSUMPTION 1 For any function $\tilde{f}$ in a Donsker class, there exits a corresponding function f in another Donsker class such that $\sum_i\sum_j w_{ij}\tilde{f}(x_{ij}) - E[\tilde{f}(x_{ij})] = (\mathbf{P}_n - \mathbf{P})\{f(\mathbf{x})\} + O_p(1/\sqrt{n})$.

REMARK 1 Assumption 1 trivially holds if $w_{ij} = \frac{1}{nk_i}$, where k_i is a bounded variable and the two Donsker classes are the same. The asymptotic properties of $\hat{\gamma}_I$ shown under $w_{ij} = \frac{1}{nk_i}$ in the followings are also valid for any w_{ij} (e.g., $\frac{1}{\sum_i kl}$) that satisfies Assumption 1.

We consider $w_{ij} = \frac{1}{nk_i}$ in the followings. Because $\hat{F}^*$ belongs to a Donsker class which contains all

distribution functions in X's support and $\hat{F}^*$ uniformly converges to F^* with probability one, g_n and h_n also belong to some Donsker classes based on Assumption 1 and converge in $L_2(P)$ distance to their limits $g(\mathbf{x})$ and $h(\mathbf{x})$, respectively, where $g(\mathbf{x}) = E\big[\sum_{j'>j} \frac{2w_{i.}}{k_i(k_i-1)}[F^*(x_{ij}) - E(F^*)][F^*(x_{ij'}) - E(F^*)]\big]$ and $h(\mathbf{x}) = E\big[\sum_{j=1}^{k_i} w_{ij}[F^*(x_{ij}) - E(F^*)]^2\big]$. Also, $|g_n - g| \to 0$, $|h_n - h| \to 0$.

Then $\sqrt{n}(\hat{\gamma}_I - \gamma_I)$ can be expressed in terms of the empirical process $\sqrt{n}(\mathbf{P}_n - \mathbf{P})$,

$$
\begin{aligned}
&\sqrt{n}(\hat{\gamma}_I - \gamma_I) = \sqrt{n}\Big(\frac{\mathbf{P}_n\{g_n(\mathbf{x})\}}{\mathbf{P}_n\{h_n(\mathbf{x})\}} - \gamma_I\Big) \\
&= \sqrt{n}\Big(\frac{\mathbf{P}_n\{g_n(\mathbf{x})\}}{\mathbf{P}_n\{h_n(\mathbf{x})\}} - \frac{\mathbf{P}\{g_n(\mathbf{x})\}}{\mathbf{P}\{h_n(\mathbf{x})\}} + \frac{\mathbf{P}\{g_n(\mathbf{x})\}}{\mathbf{P}\{h_n(\mathbf{x})\}} - \gamma_I\Big) \\
&= \sqrt{n}\Big(\frac{\mathbf{P}_n\{g_n(\mathbf{x})\}}{\mathbf{P}_n\{h_n(\mathbf{x})\}} - \frac{\mathbf{P}\{g_n(\mathbf{x})\}}{\mathbf{P}_n\{h_n(\mathbf{x})\}} + \frac{\mathbf{P}\{g_n(\mathbf{x})\}}{\mathbf{P}_n\{h_n(\mathbf{x})\}} - \frac{\mathbf{P}\{g_n(\mathbf{x})\}}{\mathbf{P}\{h_n(\mathbf{x})\}} + \frac{\mathbf{P}\{g_n(\mathbf{x})\}}{\mathbf{P}\{h_n(\mathbf{x})\}} - \gamma_I\Big) \\
&= \sqrt{n}(\mathbf{P}_n - \mathbf{P})\Big\{\frac{g_n(\mathbf{x})}{\mathbf{P}_n\{h_n(\mathbf{x})\}} - \frac{h_n(\mathbf{x})\mathbf{P}\{g_n(\mathbf{x})\}}{\mathbf{P}_n\{h_n(\mathbf{x})\}\mathbf{P}\{h_n(\mathbf{x})\}}\Big\} + \sqrt{n}\Big[\frac{\mathbf{P}\{g_n(\mathbf{x})\}}{\mathbf{P}\{h_n(\mathbf{x})\}} - \gamma_I\Big] \\
&= \sqrt{n}(\mathbf{P}_n - \mathbf{P})\Big\{\frac{g(\mathbf{x})}{E[h(\mathbf{x})]} - \frac{h(\mathbf{x})E[g(\mathbf{x})]}{(E[h(\mathbf{x})])^2}\Big\} + O_p(1) + \sqrt{n}\Big[\frac{\mathbf{P}\{g_n(\mathbf{x})\}}{\mathbf{P}\{h_n(\mathbf{x})\}} - \gamma_I\Big]
\end{aligned}
$$

We perform the linearization for $\sqrt{n}\Big[\frac{\mathbf{P}\{g_n(\mathbf{x})\}}{\mathbf{P}\{h_n(\mathbf{x})\}} - \gamma_I\Big]$ around $(F^*, \bar{F}^*)$, where $\bar{F}^*$ denotes the functional component of $E(F^*)$.

$$
\begin{aligned}
\sqrt{n}\Big[\frac{\mathbf{P}\{g_n(\mathbf{x})\}}{\mathbf{P}\{h_n(\mathbf{x})\}} - \gamma_I\Big] &= \sqrt{n}\frac{\mathbf{P}\big\{\nabla_{F^*} g_n(\mathbf{x})(\hat{F}^* - F^*)\big\}}{\mathbf{P}\{h_n(\mathbf{x})\}} \\
&\quad - \sqrt{n}\frac{\mathbf{P}\{g_n(\mathbf{x})\}}{(\mathbf{P}\{h_n(\mathbf{x})\})^2}\mathbf{P}\{\nabla_{F^*} h_n(\mathbf{x})(\hat{F}^* - F^*)\} \\
&\quad + \sqrt{n}\frac{\mathbf{P}\big\{\nabla_{\bar{F}^*} g_n(\mathbf{x})(\bar{\hat{F}}^* - E[F^*])\big\}}{\mathbf{P}\{h_n(\mathbf{x})\}} \\
&\quad - \sqrt{n}\frac{\mathbf{P}\{g_n(\mathbf{x})\}}{(\mathbf{P}\{h_n(\mathbf{x})\})^2}\mathbf{P}\{\nabla_{\bar{F}^*} h_n(\mathbf{x})(\bar{\hat{F}}^* - E[F^*])\} + O_p(1) \\
&= (1) + (2) + (3) + (4) + O_p(1)
\end{aligned}
$$

We then transform $(1) - (4)$ into expressions of $\sqrt{n}(\mathbf{P}_n - \mathbf{P})$.

$$
(1) = \sqrt{n}\frac{\mathbf{P}\big\{\nabla_{F^*} g_n(\mathbf{x})(\hat{F}^* - F^*)\big\}}{\mathbf{P}\{h_n(\mathbf{x})\}}
$$

Obtain derivative of each $F^*(x_{ij})$

$$
\begin{aligned}
(1) &= \sqrt{n}\mathbf{P}\Big\{\frac{2w_{i.}}{k_i(k_i-1)}\sum_{j'>j}(\hat{F}^*(x_{ij'}) - \bar{\hat{F}}^*)(\hat{F}^*(x_{ij}) - F^*(x_{ij}))\Big\}\Big/\mathbf{P}\{h_n(\mathbf{x})\} \\
&\quad + \sqrt{n}\mathbf{P}\Big\{\frac{2w_{i.}}{k_i(k_i-1)}\sum_{j'>j}(\hat{F}^*(x_{ij}) - \bar{\hat{F}}^*)(\hat{F}^*(x_{ij'}) - F^*(x_{ij'}))\Big\}\Big/\mathbf{P}\{h_n(\mathbf{x})\} \\
&= \sqrt{n}E\Big[\frac{2w_{i.}}{k_i(k_i-1)}\sum_{j'>j}(\hat{F}^*(x_{ij'}) - \bar{\hat{F}}^*)(\hat{F}^*(x_{ij}) - F^*(x_{ij}))\Big]\Big/E[h_n(\mathbf{x})]
\end{aligned}
$$

$$+\sqrt{n}E\Big[\frac{2w_{i.}}{k_i(k_i-1)}\sum_{j'>j}(\hat{F}^*(x_{ij})-\bar{\hat{F}}^*)(\hat{F}^*(x_{ij'})-F^*(x_{ij'}))\Big]/E[h_n(\mathbf{x})]$$

Plug in the expression of $\hat{F}^*(x_{ij})$ and $\hat{F}^*(x_{ij'})$

$$\begin{aligned}(1)=\frac{\sqrt{n}}{E[h_n(\mathbf{x})]}\Big\{E\Big[\frac{2w_{i.}}{k_i(k_i-1)}\sum_{j'>j}\Big(\hat{F}^*(x_{ij'})-\bar{\hat{F}}^*\Big)\Big(\sum_{i'}\sum_{j''}w_{i'j''}[I(x_{i'j''}<x_{ij})\\ +I(x_{i'j''}\le x_{ij})]/2-F^*(x_{ij})+\frac{2w_{i.}}{k_i(k_i-1)}\sum_{j'>j}\Big(\hat{F}^*(x_{ij})-\bar{\hat{F}}^*\Big)\\ \times\Big(\sum_{i'}\sum_{j''}w_{i'j''}[I(x_{i'j''}<x_{ij'})+I(x_{i'j''}\le x_{ij'})]/2-F^*(x_{ij'})\Big)\Big]\Big\}\end{aligned}$$

Take i' and j'' outside the expectation

$$\begin{aligned}(1)=\frac{\sqrt{n}}{E[h_n(\mathbf{x})]}\sum_{i'}\sum_{j''}w_{i'j''}E\Big\{\frac{2w_{i.}}{k_i(k_i-1)}\sum_{j'>j}\Big(\hat{F}^*(x_{ij'})-\bar{\hat{F}}^*\Big)\Big(I(x_{i'j''}<x_{ij})\\ +I(x_{i'j''}\le x_{ij})\Big)/2+\frac{2w_{i.}}{k_i(k_i-1)}\sum_{j'>j}\Big(\hat{F}^*(x_{ij})-\bar{\hat{F}}^*\Big)\Big(I(x_{i'j''}<x_{ij'})\\ +I(x_{i'j''}\le x_{ij'})\Big)/2\Big\}-\frac{\sqrt{n}}{E[h_n(\mathbf{x})]}E\Big\{\frac{2w_{i.}}{k_i(k_i-1)}\sum_{j'>j}\Big[(\hat{F}^*(x_{ij'})-\bar{\hat{F}}^*)F^*(x_{ij})\\ +(\hat{F}^*(x_{ij})-\bar{\hat{F}}^*)F^*(x_{ij'})\Big]\Big\}\end{aligned}$$

The expectation in the first component is a function of $x_{i'j''}$,

we denote it as $\tilde{a}_1(x_{i'j''})$

$$(1)=\frac{\sqrt{n}}{E[h(\mathbf{x})]}\{\sum_{i'}\sum_{j''}w_{i'j''}\tilde{a}_1(x_{i'j''})-E[\tilde{a}_1(x_{i'j''})]\}+O_p(1)$$

Because $\tilde{a}_1(x_{ij})$ belongs to a Donsker class

$$(1)=\frac{\sqrt{n}}{E[h(\mathbf{x})]}(\mathbf{P}_n-\mathbf{P})\{a_1(\mathbf{x})\}+O_p(1)\text{ (under Assumption 1)}$$

The derivation of $(2)-(4)$ is similar to that of (1) and under Assumption 1.

$$\begin{aligned}(2)&=-\sqrt{n}\frac{\mathbf{P}\{g_n(\mathbf{x})\}}{(\mathbf{P}\{h_n(\mathbf{x})\})^2}\mathbf{P}\{\nabla_{F^*}h_n(\mathbf{x})(\hat{F}^*-F^*)\}\\ &-\sqrt{n}\frac{\mathbf{P}\{g_n(\mathbf{x})\}}{(\mathbf{P}\{h_n(\mathbf{x})\})^2}\mathbf{P}\Big\{2\sum_j w_{ij}(\hat{F}^*(x_{ij})-\bar{\hat{F}}^*)(\hat{F}^*(x_{ij})-F^*(x_{ij}))\Big\}\\ &=-\sqrt{n}\frac{E[g_n(\mathbf{x})]}{(E[h_n(\mathbf{x})])^2}E\Big\{2\sum_j w_{ij}(\hat{F}^*(x_{ij})-\bar{\hat{F}}^*)(\hat{F}^*(x_{ij})-F^*(x_{ij}))\Big\}\\ &=-\sqrt{n}\frac{E[g_n(\mathbf{x})]}{(E[h_n(\mathbf{x})])^2}E\Big\{2\sum_j w_{ij}(\hat{F}^*(x_{ij})-\bar{\hat{F}}^*)\Big(\sum_{i'}\sum_{j'}w_{i'j'}[I(x_{i'j'}<x_{ij})\\ &+I(x_{i'j'}\le x_{ij})]/2-F^*(x_{ij})\Big)\Big\}\end{aligned}$$

$$= -\sqrt{n}\frac{E[g_n(\mathbf{x})]}{(E[h_n(\mathbf{x})])^2}\sum_{i'}\sum_{j'} w_{i'j'}E\Big\{2\sum_j w_{ij}(\hat{F}^*(x_{ij})-\bar{\hat{F}}^*)[I(x_{i'j'}<x_{ij})$$

$$+I(x_{i'j'}\le x_{ij})]/2\Big\}+\sqrt{n}\frac{E[g_n(\mathbf{x})]}{(E[h_n(\mathbf{x})])^2}E\Big\{2\sum_j w_{ij}(\hat{F}^*(x_{ij})-\bar{\hat{F}}^*)F^*(x_{ij})\Big\}$$

Let $\tilde{a}_2(x_{i'j'})$ denote the expectation in the first component

$$(2) = -\sqrt{n}\frac{E[g_n(\mathbf{x})]}{(E[h_n(\mathbf{x})])^2}\{\sum_{i'}\sum_{j'} w_{i'j'}\tilde{a}_2(x_{i'j'})-E[\tilde{a}_2(x_{i'j'})]\}+O_p(1)$$

$$= -\frac{\sqrt{n}E[g(\mathbf{x})]}{(E[h(\mathbf{x})])^2}(\mathbf{P}_n-\mathbf{P})\{a_2(\mathbf{x})\}+O_p(1)$$

$$(3) = \sqrt{n}\frac{\mathbf{P}\big\{\nabla_{\bar{F}^*}g_n(\mathbf{x})(\bar{\hat{F}}^*-E[F^*])\big\}}{\mathbf{P}\{h_n(\mathbf{x})\}}$$

$$= \frac{-\sqrt{n}}{\mathbf{P}\{h_n(\mathbf{x})\}}\mathbf{P}\Big\{\frac{2w_{i.}}{k_i(k_i-1)}\big[\sum_{j'>j}(\hat{F}^*(x_{ij})-\bar{\hat{F}}^*+\hat{F}^*(x_{ij'})-\bar{\hat{F}}^*)\big](\bar{\hat{F}}^*-E[F^*])\Big\}$$

$$= \frac{-\sqrt{n}}{E[h_n(\mathbf{x})]}E\Big\{\sum_{j'>j}\frac{2w_{i.}}{k_i(k_i-1)}[\hat{F}^*(x_{ij'})-\bar{\hat{F}}^*+\hat{F}^*(x_{ij})-\bar{\hat{F}}^*](\bar{\hat{F}}^*-E[F^*])\}$$

$$= -\frac{\sqrt{n}}{E[h_n(\mathbf{x})]}E\Big\{\Big(\sum_{j'>j}\frac{2w_{i.}}{k_i(k_i-1)}[\hat{F}^*(x_{ij'})-\bar{\hat{F}}^*+\hat{F}^*(x_{ij})-\bar{\hat{F}}^*]\Big)\Big(\sum_{i'}\sum_{j'} w_{i'j'}$$

$$\big[\hat{F}^*(x_{i'j'})+\frac{1}{n}\sum_i\sum_j w_{ij}[I(x_{i'j'}<x_{ij})+I(x_{i'j'}\le x_{ij})]/2\big]-E[F^*]\Big)\Big\}$$

$$= -\frac{\sqrt{n}}{E[h_n(\mathbf{x})]}\sum_{i'}\sum_{j'} w_{i'j'}E\Big\{\Big(\sum_{j'>j}\frac{2w_{i.}}{k_i(k_i-1)}[\hat{F}^*(x_{ij'})-\bar{\hat{F}}^*+\hat{F}^*(x_{ij})-\bar{\hat{F}}^*]\Big)$$

$$\times\Big(\hat{F}^*(x_{i'j'})+\frac{1}{n}\sum_i\sum_j w_{ij}[I(x_{i'j'}<x_{ij})+I(x_{i'j'}\le x_{ij})]/2\Big)\Big\}$$

$$+\frac{\sqrt{n}}{E[h_n(\mathbf{x})]}E\Big\{\sum_{j'>j}\frac{2w_{i.}}{k_i(k_i-1)}[\hat{F}^*(x_{ij'})-\bar{\hat{F}}^*+\hat{F}^*(x_{ij})-\bar{\hat{F}}^*]E[F^*]\Big\}$$

We denote the expectation in the first component as $\tilde{a}_3(x_{i'j'})$

$$(3) = -\frac{\sqrt{n}}{E[h_n(\mathbf{x})]}\{\sum_{i'}\sum_{j'} w_{i'j'}\tilde{a}_3(x_{i'j'})-E[\tilde{a}_3(x_{i'j'})]\}$$

$$= -\frac{\sqrt{n}}{E[h(\mathbf{x})]}(\mathbf{P}_n-\mathbf{P})\{a_3(\mathbf{x})\}+O_p(1)$$

$$(4) = -\sqrt{n}\frac{\mathbf{P}\{g_n(\mathbf{x})\}}{(\mathbf{P}\{h_n(\mathbf{x})\})^2}\mathbf{P}\{\nabla_{\bar{F}^*}h_n(\mathbf{x})(\bar{\hat{F}}^*-E[F^*])\}$$

$$= \sqrt{n}\frac{E[g_n(\mathbf{x})]}{(E[h_n(\mathbf{x})])^2}\mathbf{P}\Big\{2\sum_j w_{ij}(\hat{F}^*(x_{ij})-\bar{\hat{F}}^*)(\bar{\hat{F}}^*-E[F^*])\Big\}$$

$$= \sqrt{n}\frac{E[g_n(\mathbf{x})]}{(E[h_n(\mathbf{x})])^2}\sum_{i'}\sum_{j'} w_{i'j'}E\Big\{\sum_j 2w_{ij}(\hat{F}^*(x_{ij})-\bar{\hat{F}}^*)$$

$$\times \left(\hat{F}^*(x_{i'j'}) + \frac{1}{n}\sum_i\sum_j w_{ij}[I(x_{i'j'} < x_{ij}) + I(x_{i'j'} \le x_{ij})]/2)\right\}$$
$$- \sqrt{n}\frac{E[g(\mathbf{x})]}{(E[h(\mathbf{x})])^2} E\left\{2\sum_j w_{ij}(\hat{F}^*(x_{ij}) - \bar{\hat{F}}^*)E[F^*]\right\}$$

The expectation in the first component is denoted as $\tilde{a}_4(x_{i'j'})$

$$(4) = \sqrt{n}\frac{E[g_n(\mathbf{x})]}{(E[h_n(\mathbf{x})])^2}\{\sum_{i'}\sum_{j'} w_{i'j'}\tilde{a}_4(x_{i'j'}) - E[\tilde{a}_4(x_{i'j'})]\}$$
$$= \sqrt{n}\frac{E[g(\mathbf{x})]}{(E[h(\mathbf{x})])^2}(\mathbf{P}_n - \mathbf{P})\{a_4(\mathbf{x})\} + O_p(1)$$

With the derivation of $(1)-(4)$ above, we can express $\sqrt{n}\left[\frac{\mathbf{P}\{g_n(\mathbf{x})\}}{\mathbf{P}\{h_n(\mathbf{x})\}} - \gamma_I\right]$ by the empirical process. Let $a(\mathbf{x}) = \frac{1}{E[h(\mathbf{x})]}[a_1(\mathbf{x}) - a_3(\mathbf{x})] - \frac{E[g(\mathbf{x})]}{(E[h(\mathbf{x})])^2}[a_2(\mathbf{x}) - a_4(\mathbf{x})]$, note $a(\mathbf{x})$ is a deterministic function of $\mathbf{x}$. We then have $\sqrt{n}\left[\frac{\mathbf{P}\{g_n(\mathbf{x})\}}{\mathbf{P}\{h_n(\mathbf{x})\}} - \gamma_I\right] = \sqrt{n}(\mathbf{P}_n - \mathbf{P})\{a(\mathbf{x})\} + O_p(1)$.

Accordingly, we have the expression for $\sqrt{n}(\hat{\gamma}_I - \gamma_I)$ in terms of the empirical process,

$$\sqrt{n}(\hat{\gamma}_I - \gamma_I) = \sqrt{n}(\mathbf{P}_n - \mathbf{P})\{t(\mathbf{x})\} + O_p(1)$$

where $t(\mathbf{x}) = \frac{g(\mathbf{x})}{E[h(\mathbf{x})]} - \frac{h(\mathbf{x})E[g(\mathbf{x})]}{(E[h(\mathbf{x})])^2} + a(\mathbf{x})$. By the central limit theorem,

$$\sqrt{n}(\mathbf{P}_n - \mathbf{P})\{t(\mathbf{x})\} \xrightarrow{d} N(0, \sigma_t^2)$$

σ_t^2 is the variance of t. Thus, we can say that as $n \to \infty$,

$$\sqrt{n}(\hat{\gamma}_I - \gamma_I) \xrightarrow{d} N(0, \sigma_t^2)$$

The large sampling distribution of $\hat{\gamma}_I$ is $N(\gamma_I, \sigma_t^2/n)$. Since $\frac{1}{\sqrt{n}} \to 0$ as $n \to \infty$, by the Slutsky's theorem, we have

$$\hat{\gamma}_I - \gamma_I = \frac{1}{\sqrt{n}}\sqrt{n}(\hat{\gamma}_I - \gamma_I) \xrightarrow{d} 0 \Rightarrow \hat{\gamma}_I - \gamma_I \xrightarrow{p} 0$$

That is, $\hat{\gamma}_I$ converges to γ_I in probability. Therefore, $\hat{\gamma}_I$ is a consistent estimator of γ_I.

2.7.2 Variance estimation of $\hat{\gamma}_I$ with two hierarchies

Given two-level data $\{x_{ij}, i = 1,...,n, j = 1,...,k_i\}$, we can estimate the variance of $\hat{\gamma}_I$ using the sample variance of $t(\mathbf{x})$. We first obtain the estimate of $t(\mathbf{x})$ for each cluster. Let $A_n = \frac{1}{n}\sum_i^n g_n(\mathbf{x}_i)$ and $B_n = \frac{1}{n}\sum_i^n h_n(\mathbf{x}_i)$. For each cluster, we compute

$$\hat{t}(\mathbf{x}_i) = \frac{g_n(\mathbf{x}_i)}{B_n} - \frac{h_n(\mathbf{x}_i)A_n}{B_n^2} + \hat{a}(\mathbf{x}_i)$$

where $\hat{a}(\mathbf{x}_i) = I_i + II_i + III_i + IV_i$, and the four items are counterparts of components in $(1)-(4)$. We describe $I_i - IV_i$ in the followings.

$$I_i = \frac{1}{B_n}\sum_{j''=1}^{k_i} w_{ij''}\sum_{i'=1}^{n}\frac{2w_{i'.}}{k_{i'}(k_{i'}-1)}\sum_{j'>j}\Big(\{\hat{F}^*(x_{i'j'})-\bar{\hat{F}}^*\}\{[I(x_{ij''}\le x_{i'j})+$$

$$I(x_{ij''}<x_{i'j})]/2\}+\sum_{j'>j}\{\hat{F}^*(x_{i'j})-\bar{\hat{F}}^*\}\{[I(x_{ij''}\le x_{i'j'})+I(x_{ij''}<x_{i'j'})]/2\}\Big)$$

$$II_i = -\frac{A_n}{B_n^2}\sum_{j=1}^{k_i} w_{ij}\sum_{i'=1}^{n}\sum_{j'=1}^{k_{i'}} 2w_{i'j'}[\hat{F}^*(x_{i'j'})-\bar{\hat{F}}^*][I(x_{ij}\le x_{i'j'})+I(x_{ij}<x_{i'j'})]/2$$

$$III_i = -\frac{C_{ni}}{B_n}\sum_{i'=1}^{n}\frac{2w_{i'.}}{k_{i'}(k_{i'}-1)}\sum_{j'>j}[\hat{F}^*(x_{i'j})-\bar{\hat{F}}^*+\hat{F}^*(x_{i'j'})-\bar{\hat{F}}^*]$$

$$IV_i = \frac{A_nC_{ni}}{B_n^2}\sum_{i'=1}^{n}\sum_{j'=1}^{k_{i'}} 2w_{i'j'}[\hat{F}^*(x_{i'j'})-\bar{\hat{F}}^*]$$

where $C_{ni} = \sum_{j=1}^{k_i} w_{ij}\hat{F}^*(x_{ij}) + \frac{1}{n}\sum_{i'=1}^{n}\sum_{j'=1}^{k_{i'}} w_{i'j'}\big\{\sum_{j=1}^{k_i} w_{ij}[I(x_{ij}\le x_{i'j'})+I(x_{ij}<x_{i'j'})]/2\big\}$.
Then we can obtain the sample variance of $t(\mathbf{x})$ with $\{\hat{t}(\mathbf{x}_i), i=1,...,n\}$, denoted as $\hat{\sigma}_t^2$. The variance of $\hat{\gamma}_I$ is estimated by $\hat{\sigma}_t^2/n$.

2.7.3 Proof of asymptotic properties with three hierarchies

Let $g_n^{(2)}(\mathbf{x})$, $g_n^{(3)}(\mathbf{x})$ and $h_n(\mathbf{x})$ be three functions such that $\hat{\gamma}_{I2} = [\frac{1}{n}\sum_{i=1}^n g_n^{(2)}(\mathbf{x}_i)]/[\frac{1}{n}\sum_{i=1}^n h_n(\mathbf{x}_i)]$ and $\hat{\gamma}_{I3} = [\frac{1}{n}\sum_{i=1}^n g_n^{(3)}(\mathbf{x}_i)]/[\frac{1}{n}\sum_{i=1}^n h_n(\mathbf{x}_i)]$, where $\mathbf{x}_i = (x_{i11},...,x_{i1m_{i1}},...,x_{in_i1},...,x_{in_im_{ij}})$ denotes a vector of observations in a level-3 unit. We then have $\hat{\gamma}_{I2} = \frac{\mathbf{P}_n\{g_n^{(2)}(\mathbf{x})\}}{\mathbf{P}_n\{h_n(\mathbf{x})\}}$, $\gamma_{I2} = \frac{\mathbf{P}\{g^{(2)}(\mathbf{x})\}}{\mathbf{P}\{h(\mathbf{x})\}}$, $\hat{\gamma}_{I3} = \frac{\mathbf{P}_n\{g_n^{(3)}(\mathbf{x})\}}{\mathbf{P}_n\{h_n(\mathbf{x})\}}$, and $\gamma_{I3} = \frac{\mathbf{P}\{g^{(3)}(\mathbf{x})\}}{\mathbf{P}\{h(\mathbf{x})\}}$.

ASSUMPTION 2 For any function $\tilde{f}$ in a Donsker class, there exits a corresponding function f in another Donsker class such that $\sum_i\sum_j\sum_k w_{ijk}\tilde{f}(x_{ijk}) - E[\tilde{f}(x_{ijk})] = (\mathbf{P}_n - \mathbf{P})\{f(\mathbf{x})\} + O_p(1/\sqrt{n})$.

REMARK 2 Assumption 2 trivially holds if $w_{ijk} = \frac{1}{nn_im_{ij}}$, where n_i and m_{ij} are bounded variables and the two Donsker classes are the same. The asymptotic properties of $\hat{\gamma}_{I2}$ and $\hat{\gamma}_{I3}$ shown under $w_{ijk} = \frac{1}{nn_im_{ij}}$ in the followings are also valid for any w_{ijk} (e.g., $w_{ijk} = \frac{1}{\sum_i\sum_j m_{ij}}$ or $\frac{1}{m_{ij}\sum_{i=1}^n n_i}$) that satisfies Assumption 2.

We consider $w_{ijk} = \frac{1}{nn_im_{ij}}$ in the following proofs. Because $\hat{F}^*$ belongs to a Donsker class which contains all distribution functions in X's support and $\hat{F}^*$ uniformly converges to F^* with probability one, $g_n^{(2)}$, $g_n^{(3)}$, and h_n also belong to some Donsker classes based on Assumption 2 and converge in $L_2(P)$ distance to their limits $g^{(2)}(\mathbf{x})$, $g^{(3)}(\mathbf{x})$ and $h(\mathbf{x})$, respectively, where $g^{(2)}(\mathbf{x}) = E\big[\sum_j\sum_{k'>k}\frac{2w_{ij.}}{m_{ij}(m_{ij}-1)}[F^*(x_{ijk})-E(F^*)][F^*(x_{ijk'})-E(F^*)]\big]$, $g^{(3)}(\mathbf{x}) = E\big[\sum_{j'>j}\sum_k\sum_l\frac{w_{i..}}{c_i}[F^*(x_{ijk})-E(F^*)][F^*(x_{ij'l})-E(F^*)]\big]$, and $h(\mathbf{x}) = E\big[\sum_{j=1}^{n_i}\sum_{k=1}^{m_{ij}} w_{ijk}[F^*(x_{ijk})-E(F^*)]^2\big]$. Also, $|g_n^{(2)}-g^{(2)}|\to 0$, $|g_n^{(3)}-g^{(3)}|\to 0$, and $|h_n-h|\to 0$.

Then we express $\sqrt{n}(\hat{\gamma}_{I2}-\gamma_{I2})$ and $\sqrt{n}(\hat{\gamma}_{I3}-\gamma_{I3})$ in terms of the empirical process.

$$
\begin{aligned}
&\sqrt{n}(\hat{\gamma}_{I2}-\gamma_{I2}) = \sqrt{n}\Big(\frac{\mathbf{P}_n\{g_n^{(2)}(\mathbf{x})\}}{\mathbf{P}_n\{h_n(\mathbf{x})\}}-\gamma_{I2}\Big)\\
&= \sqrt{n}\Big(\frac{\mathbf{P}_n\{g_n^{(2)}(\mathbf{x})\}}{\mathbf{P}_n\{h_n(\mathbf{x})\}}-\frac{\mathbf{P}\{g_n^{(2)}(\mathbf{x})\}}{\mathbf{P}\{h_n(\mathbf{x})\}}+\frac{\mathbf{P}\{g_n^{(2)}(\mathbf{x})\}}{\mathbf{P}\{h_n(\mathbf{x})\}}-\gamma_{I2}\Big)\\
&= \sqrt{n}\Big(\frac{\mathbf{P}_n\{g_n^{(2)}(\mathbf{x})\}}{\mathbf{P}_n\{h_n(\mathbf{x})\}}-\frac{\mathbf{P}\{g_n^{(2)}(\mathbf{x})\}}{\mathbf{P}_n\{h_n(\mathbf{x})\}}+\frac{\mathbf{P}\{g_n^{(2)}(\mathbf{x})\}}{\mathbf{P}_n\{h_n(\mathbf{x})\}}-\frac{\mathbf{P}\{g_n^{(2)}(\mathbf{x})\}}{\mathbf{P}\{h_n(\mathbf{x})\}}+\frac{\mathbf{P}\{g_n^{(2)}(\mathbf{x})\}}{\mathbf{P}\{h_n(\mathbf{x})\}}-\gamma_{I2}\Big)\\
&= \sqrt{n}(\mathbf{P}_n-\mathbf{P})\Big\{\frac{g_n^{(2)}(\mathbf{x})}{\mathbf{P}_n\{h_n(\mathbf{x})\}}-\frac{h_n(\mathbf{x})\mathbf{P}\{g_n^{(2)}(\mathbf{x})\}}{\mathbf{P}_n\{h_n(\mathbf{x})\}\mathbf{P}\{h_n(\mathbf{x})\}}\Big\}+\sqrt{n}\Big[\frac{\mathbf{P}\{g_n^{(2)}(\mathbf{x})\}}{\mathbf{P}\{h_n(\mathbf{x})\}}-\gamma_{I2}\Big]\\
&= \sqrt{n}(\mathbf{P}_n-\mathbf{P})\Big\{\frac{g^{(2)}(\mathbf{x})}{E[h(\mathbf{x})]}-\frac{h(\mathbf{x})E[g^{(2)}(\mathbf{x})]}{(E[h(\mathbf{x})])^2}\Big\}+O_p(1)+\sqrt{n}\Big[\frac{\mathbf{P}\{g_n^{(2)}(\mathbf{x})\}}{\mathbf{P}\{h_n(\mathbf{x})\}}-\gamma_{I2}\Big]
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
&\sqrt{n}(\hat{\gamma}_{I3}-\gamma_{I3}) = \sqrt{n}\Big(\frac{\mathbf{P}_n\{g_n^{(3)}(\mathbf{x})\}}{\mathbf{P}_n\{h_n(\mathbf{x})\}}-\gamma_{I3}\Big)\\
&= \sqrt{n}(\mathbf{P}_n-\mathbf{P})\Big\{\frac{g_n^{(3)}(\mathbf{x})}{\mathbf{P}_n\{h_n(\mathbf{x})\}}-\frac{h_n(\mathbf{x})\mathbf{P}\{g_n^{(3)}(\mathbf{x})\}}{\mathbf{P}_n\{h_n(\mathbf{x})\}\mathbf{P}\{h_n(\mathbf{x})\}}\Big\}+\sqrt{n}\Big[\frac{\mathbf{P}\{g_n^{(3)}(\mathbf{x})\}}{\mathbf{P}\{h_n(\mathbf{x})\}}-\gamma_{I3}\Big]\\
&= \sqrt{n}(\mathbf{P}_n-\mathbf{P})\Big\{\frac{g^{(3)}(\mathbf{x})}{E[h(\mathbf{x})]}-\frac{h(\mathbf{x})E[g^{(3)}(\mathbf{x})]}{(E[h(\mathbf{x})])^2}\Big\}+O_p(1)+\sqrt{n}\Big[\frac{\mathbf{P}\{g_n^{(3)}(\mathbf{x})\}}{\mathbf{P}\{h_n(\mathbf{x})\}}-\gamma_{I3}\Big]
\end{aligned}
$$

We perform the linearization for $\sqrt{n}\Big[\frac{\mathbf{P}\{g_n^{(2)}(\mathbf{x})\}}{\mathbf{P}\{h_n(\mathbf{x})\}}-\gamma_{I2}\Big]$ and $\sqrt{n}\Big[\frac{\mathbf{P}\{g_n^{(3)}(\mathbf{x})\}}{\mathbf{P}\{h_n(\mathbf{x})\}}-\gamma_{I3}\Big]$ separately around $(F^*,\bar{F}^*)$, where $\bar{F}^*$ denotes the functional component of $E(F^*)$.

$$
\begin{aligned}
\sqrt{n}\Big[\frac{\mathbf{P}\{g_n^{(2)}(\mathbf{x})\}}{\mathbf{P}\{h_n(\mathbf{x})\}}-\gamma_{I2}\Big] &= \sqrt{n}\frac{\mathbf{P}\big\{\nabla_{F^*}g_n^{(2)}(\mathbf{x})(\hat{F}^*-F^*)\big\}}{\mathbf{P}\{h_n(\mathbf{x})\}}\\
&\quad-\sqrt{n}\frac{\mathbf{P}\{g_n^{(2)}(\mathbf{x})\}}{(\mathbf{P}\{h_n(\mathbf{x})\})^2}\mathbf{P}\{\nabla_{F^*}h_n(\mathbf{x})(\hat{F}^*-F^*)\}\\
&\quad+\sqrt{n}\frac{\mathbf{P}\big\{\nabla_{\bar{F}^*}g_n^{(2)}(\mathbf{x})(\hat{\bar{F}}^*-E[F^*])\big\}}{\mathbf{P}\{h_n(\mathbf{x})\}}\\
&\quad-\sqrt{n}\frac{\mathbf{P}\{g_n^{(2)}(\mathbf{x})\}}{(\mathbf{P}\{h_n(\mathbf{x})\})^2}\mathbf{P}\{\nabla_{\bar{F}^*}h_n(\mathbf{x})(\hat{\bar{F}}^*-E[F^*])\}+O_p(1)\\
&= (*1)+(*2)+(*3)+(*4)+O_p(1)
\end{aligned}
$$

$$
\begin{aligned}
\sqrt{n}\Big[\frac{\mathbf{P}\{g_n^{(3)}(\mathbf{x})\}}{\mathbf{P}\{h_n(\mathbf{x})\}}-\gamma_{I3}\Big] &= \sqrt{n}\frac{\mathbf{P}\big\{\nabla_{F^*}g_n^{(3)}(\mathbf{x})(\hat{F}^*-F^*)\big\}}{\mathbf{P}\{h_n(\mathbf{x})\}}\\
&\quad-\sqrt{n}\frac{\mathbf{P}\{g_n^{(3)}(\mathbf{x})\}}{(\mathbf{P}\{h_n(\mathbf{x})\})^2}\mathbf{P}\{\nabla_{F^*}h_n(\mathbf{x})(\hat{F}^*-F^*)\}
\end{aligned}
$$

$$
\begin{aligned}
&+\sqrt{n}\frac{\mathbf{P}\{\nabla_{\bar{F}^*} g_n^{(3)}(\mathbf{x})(\bar{\hat{F}}^* - E[F^*])\}}{\mathbf{P}\{h_n(\mathbf{x})\}} \\
&-\sqrt{n}\frac{\mathbf{P}\{g_n^{(3)}(\mathbf{x})\}}{(\mathbf{P}\{h_n(\mathbf{x})\})^2}\mathbf{P}\{\nabla_{\bar{F}^*} h_n(\mathbf{x})(\bar{\hat{F}}^* - E[F^*])\} + O_p(1) \\
&= (*5) + (*6) + (*7) + (*8) + O_p(1)
\end{aligned}
$$

Then transform $(*1) - (*8)$ into expressions of $\sqrt{n}(\mathbf{P}_n - \mathbf{P})$.

$$
\begin{aligned}
(*1) &= \sqrt{n}\frac{\mathbf{P}\{\nabla_{F^*} g_n^{(2)}(\mathbf{x})(\hat{F}^* - F^*)\}}{\mathbf{P}\{h_n(\mathbf{x})\}} \\
&= \frac{\sqrt{n}}{\mathbf{P}\{h_n(\mathbf{x})\}}\mathbf{P}\Big\{\sum_j \frac{2w_{ij.}}{m_{ij}(m_{ij}-1)}\sum_{k'>k}(\hat{F}^*(x_{ijk'}) - \bar{\hat{F}}^*)(\hat{F}^*(x_{ijk}) - F^*(x_{ijk}))\Big\} \\
&\quad + \frac{\sqrt{n}}{\mathbf{P}\{h_n(\mathbf{x})\}}\mathbf{P}\Big\{\sum_j \frac{2w_{ij.}}{m_{ij}(m_{ij}-1)}\sum_{k'>k}(\hat{F}^*(x_{ijk}) - \bar{\hat{F}}^*)(\hat{F}^*(x_{ijk'}) - F^*(x_{ijk'}))\Big\} \\
&= \frac{\sqrt{n}}{E[h_n(\mathbf{x})]}E\Big[\sum_j \frac{2w_{ij.}}{m_{ij}(m_{ij}-1)}\sum_{k'>k}(\hat{F}^*(x_{ijk'}) - \bar{\hat{F}}^*)(\hat{F}^*(x_{ijk}) - F^*(x_{ijk}))\Big] \\
&\quad + \frac{\sqrt{n}}{E[h_n(\mathbf{x})]}E\Big[\sum_j \frac{2w_{ij.}}{m_{ij}(m_{ij}-1)}\sum_{k'>k}(\hat{F}^*(x_{ijk}) - \bar{\hat{F}}^*)(\hat{F}^*(x_{ijk'}) - F^*(x_{ijk'}))\Big] \\
&= \frac{\sqrt{n}}{E[h_n(\mathbf{x})]}\Big\{E\Big[\sum_j \frac{2w_{ij.}}{m_{ij}(m_{ij}-1)}\sum_{k'>k}\Big(\hat{F}^*(x_{ijk'}) - \bar{\hat{F}}^*\Big) \\
&\quad \times \Big(\sum_{i'}\sum_{j'',k''} w_{i'j''k''}[I(x_{i'j''k''} < x_{ijk}) + I(x_{i'j''k''} \le x_{ijk})]/2 - F^*(x_{ijk})\Big)\Big] \\
&\quad + E\Big[\sum_j \frac{2w_{ij.}}{m_{ij}(m_{ij}-1)}\sum_{k'>k}\Big(\hat{F}^*(x_{ijk}) - \bar{\hat{F}}^*\Big) \\
&\quad \times \Big(\sum_{i'}\sum_{j'',k''} w_{i'j''k''} \times [I(x_{i'j''k''} < x_{ijk'}) + I(x_{i'j''k''} \le x_{ijk'})]/2 - F^*(x_{ijk'})\Big)\Big]\Big\} \\
&= \frac{\sqrt{n}}{E[h_n(\mathbf{x})]}\sum_{i'}\sum_{j'',k''} w_{i'j''k''}E\Big\{\sum_j \frac{2w_{ij.}}{m_{ij}(m_{ij}-1)}\sum_{k'>k}\Big[\Big(\hat{F}^*(x_{ijk'}) - \bar{\hat{F}}^*\Big) \\
&\quad \times \Big(I(x_{i'j''k''} < x_{ijk}) + I(x_{i'j''k''} \le x_{ijk})\Big)/2 + \Big(\hat{F}^*(x_{ijk}) - \bar{\hat{F}}^*\Big) \\
&\quad \times \Big(I(x_{i'j''k''} < x_{ijk'}) + I(x_{i'j''k''} \le x_{ijk'})\Big)/2\Big]\Big\} \\
&\quad - \frac{\sqrt{n}}{E[h_n(\mathbf{x})]}E\Big\{\sum_j \frac{2w_{ij.}}{m_{ij}(m_{ij}-1)}\sum_{k'>k}\Big[(\hat{F}^*(x_{ijk'}) - \bar{\hat{F}}^*)F^*(x_{ijk}) + \\
&\quad (\hat{F}^*(x_{ijk}) - \bar{\hat{F}}^*)F^*(x_{ijk'})\Big]\Big\}
\end{aligned}
$$

The expectation in the first component is a function of $x_{i'j''k''}$,

we denote it as $\tilde{a}_1^*(x_{i'j''k''})$

$$
= \frac{\sqrt{n}}{E[h(\mathbf{x})]}\{\sum_{i'}\sum_{j''}\sum_{k''} w_{i'j''k''}\tilde{a}_1^*(x_{i'j''k''}) - E[\tilde{a}_1^*(x_{i'j''k''})]\} + O_p(1)
$$

Because $\tilde{a}_1^*(x_{ijk})$ belongs to a Donsker class

$$(*1) = \frac{\sqrt{n}}{E[h(\mathbf{x})]}(\mathbf{P}_n - \mathbf{P})\{a_1^*(\mathbf{x})\} + O_p(1) \text{ (under Assumption 2)}$$

The derivation of $(*2) - (*8)$ is similar to that of $(*1)$ and under Assumption 2.

$$\begin{aligned}
(*2) &= -\sqrt{n}\frac{\mathbf{P}\{g_n^{(2)}(\mathbf{x})\}}{(\mathbf{P}\{h_n(\mathbf{x})\})^2}\mathbf{P}\{\nabla_{F^*}h_n(\mathbf{x})(\hat{F}^* - F^*)\} \\
&\quad - \sqrt{n}\frac{\mathbf{P}\{g_n^{(2)}(\mathbf{x})\}}{(\mathbf{P}\{h_n(\mathbf{x})\})^2}\mathbf{P}\Big\{2\sum_j\sum_k w_{ijk}(\hat{F}^*(x_{ijk}) - \bar{\hat{F}}^*)(\hat{F}^*(x_{ijk}) - F^*(x_{ijk}))\Big\} \\
&= -\sqrt{n}\frac{E[g_n^{(2)}(\mathbf{x})]}{(E[h_n(\mathbf{x})])^2}E\Big\{2\sum_j\sum_k w_{ijk}(\hat{F}^*(x_{ijk}) - \bar{\hat{F}}^*)(\hat{F}^*(x_{ijk}) - F^*(x_{ijk}))\Big\} \\
&= -\sqrt{n}\frac{E[g_n^{(2)}(\mathbf{x})]}{(E[h_n(\mathbf{x})])^2}E\Big\{2\sum_j\sum_k w_{ijk}(\hat{F}^*(x_{ijk}) - \bar{\hat{F}}^*)\Big(\sum_{i'}\sum_{j'k'} w_{i'j'k'} \\
&\quad \times [I(x_{i'j'k'} < x_{ijk}) + I(x_{i'j'k'} \le x_{ijk})]/2 - F^*(x_{ijk})\Big)\Big\} \\
&= -\sqrt{n}\frac{E[g_n^{(2)}(\mathbf{x})]}{(E[h_n(\mathbf{x})])^2}\sum_{i'}\sum_{j',k'} w_{i'j'k'}E\Big\{2\sum_j\sum_k w_{ijk}(\hat{F}^*(x_{ijk}) - \bar{\hat{F}}^*) \\
&\quad \times [I(x_{i'j'k'} < x_{ijk}) + I(x_{i'j'k'} \le x_{ijk})]/2\Big\} \\
&\quad + \sqrt{n}\frac{E[g_n^{(2)}(\mathbf{x})]}{(E[h_n(\mathbf{x})])^2}E\Big\{2\sum_j\sum_k w_{ijk}(\hat{F}^*(x_{ijk}) - \bar{\hat{F}}^*)F^*(x_{ijk})\Big\}
\end{aligned}$$

Let $\tilde{a}_2^*(x_{i'j'k'})$ denote the expectation in the first component

$$\begin{aligned}
(*2) &= -\sqrt{n}\frac{E[g_n^{(2)}(\mathbf{x})]}{(E[h_n(\mathbf{x})])^2}\{\sum_{i'}\sum_{j'}\sum_{k'} w_{i'j'k'}\tilde{a}_2^*(x_{i'j'k'}) - E[\tilde{a}_2^*(x_{i'j'k'})]\} + O_p(1) \\
&= -\frac{\sqrt{n}E[g^{(2)}(\mathbf{x})]}{(E[h(\mathbf{x})])^2}(\mathbf{P}_n - \mathbf{P})\{a_2^*(\mathbf{x})\} + O_p(1)
\end{aligned}$$

$$\begin{aligned}
(*3) &= \sqrt{n}\frac{\mathbf{P}\{\nabla_{\bar{F}^*}g_n^{(2)}(\mathbf{x})(\bar{\hat{F}}^* - E[F^*])\}}{\mathbf{P}\{h_n(\mathbf{x})\}} \\
&= \sqrt{n}\mathbf{P}\Big\{-\sum_j \frac{2w_{ij.}}{m_{ij}(m_{ij}-1)}[\sum_{k'>k}(\hat{F}^*(x_{ijk}) - \bar{\hat{F}}^* + \hat{F}^*(x_{ijk'}) - \bar{\hat{F}}^*)] \\
&\quad \times (\bar{\hat{F}}^* - E[F^*])\Big\}/\mathbf{P}\{h_n(\mathbf{x})\} \\
&= -\sqrt{n}E\Big\{\sum_j \frac{2w_{ij.}}{m_{ij}(m_{ij}-1)}\sum_{k'>k}[\hat{F}^*(x_{ijk'}) - \bar{\hat{F}}^* + \hat{F}^*(x_{ijk}) - \bar{\hat{F}}^*] \\
&\quad \times (\bar{\hat{F}}^* - E[F^*])\}/E[h_n(\mathbf{x})] \\
&= -\frac{\sqrt{n}}{E[h_n(\mathbf{x})]}E\Big\{\Big(\sum_j \frac{2w_{ij.}}{m_{ij}(m_{ij}-1)}\sum_{k'>k}[\hat{F}^*(x_{ijk'}) - \bar{\hat{F}}^* + \hat{F}^*(x_{ijk}) - \bar{\hat{F}}^*]\Big)
\end{aligned}$$

$$\times \Big(\sum_{i'} \sum_{j',k'} w_{i'j'k'} \big[\hat{F}^*(x_{i'j'k'}) + \frac{1}{n} \sum_{i,j,k} w_{ijk} [I(x_{i'j'k'} < x_{ijk}) + I(x_{i'j'k'} \le x_{ijk})]/2 \big] - E[F^*] \Big) \Big\}$$

$$= -\frac{\sqrt{n}}{E[h_n(\mathbf{x})]} \sum_{i'} \sum_{j',k'} w_{i'j'k'} E\Big\{ \Big(\sum_j \frac{2w_{ij.}}{m_{ij}(m_{ij}-1)} \sum_{k'>k} [\hat{F}^*(x_{ijk'}) - \bar{\hat{F}}^* + \hat{F}^*(x_{ijk}) - \bar{\hat{F}}^*] \Big) \Big(\hat{F}^*(x_{i'j'k'}) + \frac{1}{n} \sum_{i,j,k} w_{ij} [I(x_{i'j'k'} < x_{ijk}) + I(x_{i'j'k'} \le x_{ijk})]/2 \Big) \Big\}$$

$$+ \frac{\sqrt{n}}{E[h_n(\mathbf{x})]} E\Big\{ \sum_j \frac{2w_{ij.}}{m_{ij}(m_{ij}-1)} \sum_{k'>k} [\hat{F}^*(x_{ijk'}) - \bar{\hat{F}}^* + \hat{F}^*(x_{ijk}) - \bar{\hat{F}}^*] E[F^*] \Big\}$$

We denote the expectation in the first component as $\tilde{a}_3^*(x_{i'j'k'})$

$$(*3) = -\frac{\sqrt{n}}{E[h_n(\mathbf{x})]} \{ \sum_{i'} \sum_{j'} \sum_{k'} w_{i'j'k'} \tilde{a}_3^*(x_{i'j'k'}) - E[\tilde{a}_3^*(x_{i'j'k'})] \}$$

$$= -\frac{\sqrt{n}}{E[h(\mathbf{x})]} (\mathbf{P}_n - \mathbf{P}) \{ a_3^*(\mathbf{x}) \} + O_p(1)$$

$$(*4) = -\sqrt{n} \frac{\mathbf{P}\{g_n^{(2)}(\mathbf{x})\}}{(\mathbf{P}\{h_n(\mathbf{x})\})^2} \mathbf{P}\{ \nabla_{\bar{F}^*} h_n(\mathbf{x}) (\bar{\hat{F}}^* - E[F^*]) \}$$

$$= \sqrt{n} \frac{E[g_n^{(2)}(\mathbf{x})]}{(E[h_n(\mathbf{x})])^2} \mathbf{P}\Big\{ 2 \sum_j \sum_k w_{ijk} (\hat{F}^*(x_{ijk}) - \bar{\hat{F}}^*)(\bar{\hat{F}}^* - E[F^*]) \Big\}$$

$$= \sqrt{n} \frac{E[g_n(\mathbf{x})]}{(E[h_n(\mathbf{x})])^2} \sum_{i'} \sum_{j'} \sum_{k'} w_{i'j'k'} E\Big\{ \big(\sum_j \sum_k 2w_{ijk} (\hat{F}^*(x_{ijk}) - \bar{\hat{F}}^*) \big) \times \big(\hat{F}^*(x_{i'j'k'}) + \frac{1}{n} \sum_i \sum_j \sum_k w_{ijk} [I(x_{i'j'k'} < x_{ijk}) + I(x_{i'j'k'} \le x_{ijk})]/2 \big) \Big\}$$

$$- \sqrt{n} \frac{E[g(\mathbf{x})]}{(E[h(\mathbf{x})])^2} E\Big\{ 2 \sum_j \sum_k w_{ijk} (\hat{F}^*(x_{ijk}) - \bar{\hat{F}}^*) E[F^*] \Big\}$$

The expectation in the first component is denoted as $\tilde{a}_4^*(x_{i'j'k'})$

$$(*4) = \sqrt{n} \frac{E[g_n^{(2)}(\mathbf{x})]}{(E[h_n(\mathbf{x})])^2} \{ \sum_{i'} \sum_{j'} \sum_{k'} w_{i'j'k'} \tilde{a}_4^*(x_{i'j'k'}) - E[\tilde{a}_4(x_{i'j'k'})] \}$$

$$= \sqrt{n} \frac{E[g^{(2)}(\mathbf{x})]}{(E[h(\mathbf{x})])^2} (\mathbf{P}_n - \mathbf{P}) \{ a_4^*(\mathbf{x}) \} + O_p(1)$$

$$(*5) = \sqrt{n} \frac{\mathbf{P}\{ \nabla_{F^*} g_n^{(3)}(\mathbf{x}) (\hat{F}^* - F^*) \}}{\mathbf{P}\{ h_n(\mathbf{x}) \}}$$

$$= \sqrt{n} \mathbf{P}\Big\{ \frac{w_{i..}}{c_i} \sum_{j'>j} \sum_{k,l} (\hat{F}^*(x_{ijk}) - \bar{\hat{F}}^*)(\hat{F}^*(x_{ij'l}) - F^*(x_{ij'l})) \Big\} / \mathbf{P}\{ h_n(\mathbf{x}) \}$$

$$+ \sqrt{n} \mathbf{P}\Big\{ \frac{w_{i..}}{c_i} \sum_{j'>j} \sum_{k,l} (\hat{F}^*(x_{ij'l}) - \bar{\hat{F}}^*)(\hat{F}^*(x_{ijk}) - F^*(x_{ijk})) \Big\} / \mathbf{P}\{ h_n(\mathbf{x}) \}$$

$$
\begin{aligned}
&= \sqrt{n}E\Big[\frac{w_{i..}}{c_i}\sum_{j'>j}\sum_{k,\,l}(\hat{F}^*(x_{ijk})-\bar{\hat{F}}^*)(\hat{F}^*(x_{ij'l})-F^*(x_{ij'l}))\Big]/E[h_n(\mathbf{x})]\\
&\quad+\sqrt{n}E\Big[\frac{w_{i..}}{c_i}\sum_{j'>j}\sum_{k,\,l}(\hat{F}^*(x_{ij'l})-\bar{\hat{F}}^*)(\hat{F}^*(x_{ijk})-F^*(x_{ijk}))\Big]/E[h_n(\mathbf{x})]\\
&= \frac{\sqrt{n}}{E[h_n(\mathbf{x})]}\Big\{E\Big[\frac{w_{i..}}{c_i}\sum_{j'>j}\sum_{k,\,l}\Big(\hat{F}^*(x_{ijk})-\bar{\hat{F}}^*\Big)\\
&\quad\times\Big(\sum_{i'}\sum_{j'',\,k''}w_{i'j''k''}[I(x_{i'j''k''}<x_{ij'l})+I(x_{i'j''k''}\le x_{ij'l})]/2-F^*(x_{ij'l})\Big)\Big]\\
&\quad+E\Big[\frac{w_{i..}}{c_i}\sum_{j'>j}\sum_{k,\,l}\Big(\hat{F}^*(x_{ij'l})-\bar{\hat{F}}^*\Big)\\
&\quad\times\Big(\sum_{i'}\sum_{j'',\,k''}w_{i'j''k''}[I(x_{i'j''k''}<x_{ijk})+I(x_{i'j''k''}\le x_{ijk})]/2-F^*(x_{ijk})\Big)\Big]\Big\}\\
&= \frac{\sqrt{n}}{E[h_n(\mathbf{x})]}\sum_{i'}\sum_{j'',k''}w_{i'j''k''}E\Big\{\frac{w_{i..}}{c_i}\sum_{j'>j}\sum_{k,\,l}\Big[\Big(\hat{F}^*(x_{ijk})-\bar{\hat{F}}^*\Big)\\
&\quad\times\Big(I(x_{i'j''k''}<x_{ij'l})+I(x_{i'j''k''}\le x_{ij'l})\Big)/2\\
&\quad+\Big(\hat{F}^*(x_{ij'l})-\bar{\hat{F}}^*\Big)\Big(I(x_{i'j''k''}<x_{ijk})+I(x_{i'j''k''}\le x_{ijk})\Big)/2\Big]\Big\}\\
&\quad-\frac{\sqrt{n}}{E[h_n(\mathbf{x})]}E\Big\{\frac{w_{i..}}{c_i}\sum_{j'>j}\sum_{k,\,l}\Big[(\hat{F}^*(x_{ijk})-\bar{\hat{F}}^*)F^*(x_{ij'l})\\
&\quad+(\hat{F}^*(x_{ij'l})-\bar{\hat{F}}^*)F^*(x_{ijk})\Big]\Big\}
\end{aligned}
$$

We denote the expectation in the first component as $\tilde{a}_5^*(x_{i'j''k''})$

$$
\begin{aligned}
&= \frac{\sqrt{n}}{E[h(\mathbf{x})]}\{\sum_{i'}\sum_{j''}\sum_{k''}w_{i'j''k''}\tilde{a}_5^*(x_{i'j''k''})-E[\tilde{a}_5^*(x_{i'j''k''})]\}+O_p(1)\\
&= \frac{\sqrt{n}}{E[h(\mathbf{x})]}(\mathbf{P}_n-\mathbf{P})\{a_5^*(\mathbf{x})\}+O_p(1)\text{ (under Assumption 2)}
\end{aligned}
$$

$$
\begin{aligned}
(*6) &= -\sqrt{n}\frac{\mathbf{P}\{g_n^{(3)}(\mathbf{x})\}}{(\mathbf{P}\{h_n(\mathbf{x})\})^2}\mathbf{P}\{\nabla_{F^*}h_n(\mathbf{x})(\hat{F}^*-F^*)\}\\
&\quad-\sqrt{n}\frac{\mathbf{P}\{g_n^{(3)}(\mathbf{x})\}}{(\mathbf{P}\{h_n(\mathbf{x})\})^2}\mathbf{P}\Big\{2\sum_j\sum_k w_{ijk}(\hat{F}^*(x_{ijk})-\bar{\hat{F}}^*)(\hat{F}^*(x_{ijk})-F^*(x_{ijk}))\Big\}\\
&= -\sqrt{n}\frac{E[g_n^{(3)}(\mathbf{x})]}{(E[h_n(\mathbf{x})])^2}E\Big\{2\sum_j\sum_k w_{ijk}(\hat{F}^*(x_{ijk})-\bar{\hat{F}}^*)(\hat{F}^*(x_{ijk})-F^*(x_{ijk}))\Big\}\\
&= -\sqrt{n}\frac{E[g_n^{(3)}(\mathbf{x})]}{(E[h_n(\mathbf{x})])^2}E\Big\{2\sum_j\sum_k w_{ijk}(\hat{F}^*(x_{ijk})-\bar{\hat{F}}^*)\\
&\quad\times\Big(\sum_{i'}\sum_{j'k'}w_{i'j'k'}[I(x_{i'j'k'}<x_{ijk})+I(x_{i'j'k'}\le x_{ijk})]/2-F^*(x_{ijk})\Big)\Big\}\\
&= -\sqrt{n}\frac{E[g_n^{(3)}(\mathbf{x})]}{(E[h_n(\mathbf{x})])^2}\sum_{i'}\sum_{j',k'}w_{i'j'k'}E\Big\{2\sum_j\sum_k w_{ijk}(\hat{F}^*(x_{ijk})-\bar{\hat{F}}^*)
\end{aligned}
$$

$$\times [I(x_{i'j'k'} < x_{ijk}) + I(x_{i'j'k'} \le x_{ijk})]/2\Big\}$$
$$+ \sqrt{n}\frac{E[g_n^{(3)}(\mathbf{x})]}{(E[h_n(\mathbf{x})])^2} E\Big\{2\sum_j\sum_k w_{ijk}(\hat{F}^*(x_{ijk}) - \bar{\hat{F}}^*)F^*(x_{ijk})\Big\}$$

Let $\tilde{a}_6^*(x_{i'j'k'})$ denote the expectation in the first component

$$(*6) = -\sqrt{n}\frac{E[g_n^{(3)}(\mathbf{x})]}{(E[h_n(\mathbf{x})])^2}\{\sum_{i'}\sum_{j'}\sum_{k'} w_{i'j'k'}\tilde{a}_6^*(x_{i'j'k'}) - E[\tilde{a}_6^*(x_{i'j'k'})]\} + O_p(1)$$
$$= -\frac{\sqrt{n}E[g^{(3)}(\mathbf{x})]}{(E[h(\mathbf{x})])^2}(\mathbf{P}_n - \mathbf{P})\{a_6^*(\mathbf{x})\} + O_p(1)$$

$$(*7) = \sqrt{n}\frac{\mathbf{P}\{\nabla_{\bar{F}^*} g_n^{(3)}(\mathbf{x})(\bar{\hat{F}}^* - E[F^*])\}}{\mathbf{P}\{h_n(\mathbf{x})\}}$$
$$= \sqrt{n}\mathbf{P}\Big\{-\frac{w_{i..}}{c_i}\sum_{j'>j}[\sum_{k,l}(\hat{F}^*(x_{ijk}) - \bar{\hat{F}}^* + \hat{F}^*(x_{ij'l}) - \bar{\hat{F}}^*)]$$
$$\times (\bar{\hat{F}}^* - E[F^*])\Big\}/\mathbf{P}\{h_n(\mathbf{x})\}$$
$$= -\frac{\sqrt{n}}{E[h_n(\mathbf{x})]}E\Big\{\frac{w_{i..}}{c_i}\sum_{j'>j}[\sum_{k,l}[\hat{F}^*(x_{ijk}) - \bar{\hat{F}}^* + \hat{F}^*(x_{ij'l}) - \bar{\hat{F}}^*](\bar{\hat{F}}^* - E[F^*])\}$$
$$= -\frac{\sqrt{n}}{E[h_n(\mathbf{x})]}E\Big\{\frac{w_{i..}}{c_i}\sum_{j'>j}[\sum_{k,l}[\hat{F}^*(x_{ijk}) - \bar{\hat{F}}^* + \hat{F}^*(x_{ij'l}) - \bar{\hat{F}}^*]$$
$$\times \Big(\sum_{i'}\sum_{j'',k''} w_{i'j''k''}[\hat{F}^*(x_{i'j''k''}) + \frac{1}{n}\sum_{i,j,k} w_{ijk}[I(x_{i'j''k''} < x_{ijk})$$
$$+ I(x_{i'j''k''} \le x_{ijk})]/2] - E[F^*]\Big)\Big\}$$
$$= -\frac{\sqrt{n}}{E[h_n(\mathbf{x})]}\sum_{i'}\sum_{j'',k''} w_{i'j''k''}E\Big\{\frac{w_{i..}}{c_i}\sum_{j'>j}\sum_{k,l}[\hat{F}^*(x_{ijk}) - \bar{\hat{F}}^* + \hat{F}^*(x_{ij'l}) - \bar{\hat{F}}^*]$$
$$\times \Big(\hat{F}^*(x_{i'j''k''}) + \frac{1}{n}\sum_{i,j,k} w_{ij}[I(x_{i'j''k''} < x_{ijk}) + I(x_{i'j''k''} \le x_{ijk})]/2\Big)\Big\}$$
$$+ \frac{\sqrt{n}}{E[h_n(\mathbf{x})]}E\Big\{\frac{w_{i..}}{c_i}\sum_{j'>j}[\sum_{k,l}[\hat{F}^*(x_{ijk}) - \bar{\hat{F}}^* + \hat{F}^*(x_{ij'l}) - \bar{\hat{F}}^*]E[F^*]\Big\}$$

We denote the expectation in the first component as $\tilde{a}_7^*(x_{i'j''k''})$

$$(*7) = -\frac{\sqrt{n}}{E[h_n(\mathbf{x})]}\{\sum_{i'}\sum_{j''}\sum_{k''} w_{i'j'k''}\tilde{a}_7^*(x_{i'j''k''}) - E[\tilde{a}_7^*(x_{i'j''k''})]\}$$
$$= -\frac{\sqrt{n}}{E[h(\mathbf{x})]}(\mathbf{P}_n - \mathbf{P})\{a_7^*(\mathbf{x})\} + O_p(1)$$

$$(*8) = -\sqrt{n}\frac{\mathbf{P}\{g_n^{(3)}(\mathbf{x})\}}{(\mathbf{P}\{h_n(\mathbf{x})\})^2}\mathbf{P}\{\nabla_{\bar{F}^*} h_n(\mathbf{x})(\bar{\hat{F}}^* - E[F^*])\}$$
$$= \sqrt{n}\frac{E[g_n^{(3)}(\mathbf{x})]}{(E[h_n(\mathbf{x})])^2}\mathbf{P}\Big\{2\sum_j\sum_k w_{ijk}(\hat{F}^*(x_{ijk}) - \bar{\hat{F}}^*)(\bar{\hat{F}}^* - E[F^*])\Big\}$$

$$= \sqrt{n}\frac{E[g_n^{(3)}(\mathbf{x})]}{(E[h_n(\mathbf{x})])^2}\sum_{i'}\sum_{j'}\sum_{k'} w_{i'j'k'} E\Big\{\sum_j\sum_k 2w_{ijk}(\hat{F}^*(x_{ijk}) - \bar{\hat{F}}^*)$$
$$\times\big(\hat{F}^*(x_{i'j'k'}) + \frac{1}{n}\sum_i\sum_j\sum_k w_{ijk}[I(x_{i'j'k'} < x_{ijk}) + I(x_{i'j'k'} \le x_{ijk})]/2\big)\Big\}$$
$$-\sqrt{n}\frac{E[g^{(3)}(\mathbf{x})]}{(E[h(\mathbf{x})])^2}E\Big\{2\sum_j\sum_k w_{ijk}(\hat{F}^*(x_{ijk}) - \bar{\hat{F}}^*)E[F^*]\Big\}$$

The expectation in the first component is denoted as $\tilde{a}_4^*(x_{i'j'k'})$

$$(*8) = \sqrt{n}\frac{E[g_n^{(3)}(\mathbf{x})]}{(E[h_n(\mathbf{x})])^2}\{\sum_{i'}\sum_{j'}\sum_{k'} w_{i'j'k'}\tilde{a}_8^*(x_{i'j'k'}) - E[\tilde{a}_8(x_{i'j'k'})]\}$$
$$= \sqrt{n}\frac{E[g^{(3)}(\mathbf{x})]}{(E[h(\mathbf{x})])^2}(\mathbf{P}_n - \mathbf{P})\{a_8^*(\mathbf{x})\} + O_p(1)$$

Let $a^{(2)}(\mathbf{x}) = \frac{1}{E[h(\mathbf{x})]}[a_1^*(\mathbf{x}) - a_3^*(\mathbf{x})] - \frac{E[g^{(2)}(\mathbf{x})]}{(E[h(\mathbf{x})])^2}[a_2^*(\mathbf{x}) - a_4^*(\mathbf{x})]$ and $a^{(3)}(\mathbf{x}) = \frac{1}{E[h(\mathbf{x})]}[a_5^*(\mathbf{x}) - a_7^*(\mathbf{x})] - \frac{E[g^{(3)}(\mathbf{x})]}{(E[h(\mathbf{x})])^2}[a_6^*(\mathbf{x}) - a_8^*(\mathbf{x})]$. With the derivation of $(*1) - (*8)$ above, we have $\sqrt{n}\Big[\frac{\mathbf{P}\{g_n^{(2)}(\mathbf{x})\}}{\mathbf{P}\{h_n(\mathbf{x})\}} - \gamma_{I2}\Big] = \sqrt{n}(\mathbf{P}_n - \mathbf{P})\{a^{(2)}(\mathbf{x})\} + O_p(1)$ and $\sqrt{n}\Big[\frac{\mathbf{P}\{g_n^{(3)}(\mathbf{x})\}}{\mathbf{P}\{h_n(\mathbf{x})\}} - \gamma_{I3}\Big] = \sqrt{n}(\mathbf{P}_n - \mathbf{P})\{a^{(3)}(\mathbf{x})\} + O_p(1)$.

Accordingly, we have the expressions for $\sqrt{n}(\hat{\gamma}_{I2} - \gamma_{I2})$ and $\sqrt{n}(\hat{\gamma}_{I3} - \gamma_{I3})$ in terms of the empirical process,

$$\sqrt{n}(\hat{\gamma}_{I2} - \gamma_{I2}) = \sqrt{n}(\mathbf{P}_n - \mathbf{P})\{t_2(\mathbf{x})\} + O_p(1)$$

$$\sqrt{n}(\hat{\gamma}_{I3} - \gamma_{I3}) = \sqrt{n}(\mathbf{P}_n - \mathbf{P})\{t_3(\mathbf{x})\} + O_p(1)$$

where $t_2(\mathbf{x}) = \frac{g^{(2)}(\mathbf{x})}{E[h(\mathbf{x})]} - \frac{h(\mathbf{x})E[g^{(2)}(\mathbf{x})]}{(E[h(\mathbf{x})])^2} + a^{(2)}(\mathbf{x})$ and $t_3(\mathbf{x}) = \frac{g^{(3)}(\mathbf{x})}{E[h(\mathbf{x})]} - \frac{h(\mathbf{x})E[g^{(3)}(\mathbf{x})]}{(E[h(\mathbf{x})])^2} + a^{(3)}(\mathbf{x})$. By the central limit theorem,

$$\sqrt{n}(\mathbf{P}_n - \mathbf{P})\{t_2(\mathbf{x})\} \xrightarrow{d} N(0, \sigma_{t_2}^2)$$

$$\sqrt{n}(\mathbf{P}_n - \mathbf{P})\{t_3(\mathbf{x})\} \xrightarrow{d} N(0, \sigma_{t_3}^2)$$

$\sigma_{t_2}^2$ is the variance of t_2 and $\sigma_{t_3}^2$ is the variance of t_3. Thus, we can say that as $n \to \infty$,

$$\sqrt{n}(\hat{\gamma}_{I2} - \gamma_{I2}) \xrightarrow{d} N(0, \sigma_{t_2}^2)$$

$$\sqrt{n}(\hat{\gamma}_{I3} - \gamma_{I3}) \xrightarrow{d} N(0, \sigma_{t_3}^2)$$

The large sampling distribution of $\hat{\gamma}_{I2}$ is $N(\gamma_I, \sigma_{t_2}^2/n)$ and of $\hat{\gamma}_{I3}$ is $N(\gamma_I, \sigma_{t_3}^2/n)$. Since $\frac{1}{\sqrt{n}} \to 0$ as $n \to \infty$, by the Slutsky's theorem, we have

$$\hat{\gamma}_{I2} - \gamma_{I2} = \frac{1}{\sqrt{n}}\sqrt{n}(\hat{\gamma}_{I2} - \gamma_{I2}) \xrightarrow{d} 0 \Rightarrow \hat{\gamma}_{I2} - \gamma_{I2} \xrightarrow{p} 0$$

Similarly, $\hat{\gamma}_{I3} - \gamma_{I3} \xrightarrow{p} 0$. Therefore, $\hat{\gamma}_{I2}$ is a consistent estimator of γ_{I2}, and $\hat{\gamma}_{I3}$ is a consistent estimator of γ_{I3}.

2.7.4 Variance estimation with three hierarchies

With a three-level dataset $\{x_{ijk}, i = 1,...,n, j = 1,...,n_i, k = 1,...,m_{ij}\}$, we can use the sample variances of $t_2(\mathbf{x})$ and $t_3(\mathbf{x})$ to estimate the variances of $\hat{\gamma}_{I2}$ and $\hat{\gamma}_{I3}$. We first obtain the estimates of $t_2(\mathbf{x})$ and $t_3(\mathbf{x})$ for each cluster. Let $A_n^{(2)} = \frac{1}{n}\sum_i^n g_n^{(2)}(\mathbf{x}_i)$, $A_n^{(3)} = \frac{1}{n}\sum_i^n g_n^{(3)}(\mathbf{x}_i)$ and $B_n = \frac{1}{n}\sum_i^n h_n(\mathbf{x}_i)$. For each cluster, we compute

$$\hat{t}_2(\mathbf{x}_i) = \frac{g_n^{(2)}(\mathbf{x}_i)}{B_n} - \frac{h_n(\mathbf{x}_i)A_n^{(2)}}{B_n^2} + \hat{a}^{(2)}(\mathbf{x}_i)$$

$$\hat{t}_3(\mathbf{x}_i) = \frac{g_n^{(3)}(\mathbf{x}_i)}{B_n} - \frac{h_n(\mathbf{x}_i)A_n^{(3)}}{B_n^2} + \hat{a}^{(3)}(\mathbf{x}_i)$$

where $\hat{a}^{(2)}(\mathbf{x}_i) = I_i + II_i + III_i + IV_i$, $\hat{a}^{(3)}(\mathbf{x}_i) = V_i + VI_i + VII_i + VIII_i$, and the eight items are counterparts of components in $(*1) - (*8)$. We describe $I_i - VIII_i$ in the followings.

$$\begin{aligned}
I_i &= \frac{1}{B_n}\sum_{j=1}^{n_i}\sum_{k=1}^{m_{ij}} w_{ijk}\sum_{i'=1}^{n}\sum_{j'=1}^{n_{i'}}\frac{2w_{i'j'.}}{m_{i'j}(k_{i'}-1)}\sum_{k'>k''}\Big(\{\hat{F}^*(x_{i'j'k'}) - \bar{\hat{F}}^*\} \\
&\quad \times \{[I(x_{ijk} \le x_{i'j'k''}) + I(x_{ijk} < x_{i'j'k''})]/2\} \\
&\quad + \{\hat{F}^*(x_{i'j'k''}) - \bar{\hat{F}}^*\}\{[I(x_{ijk} \le x_{i'j'k'}) + I(x_{ijk} < x_{i'j'k'})]/2\}\Big) \\
II_i &= -\frac{A_n^{(2)}}{B_n^2}\sum_{j=1}^{n_i}\sum_{k=1}^{m_{ij}} w_{ijk}\sum_{i'=1}^{n}\sum_{j'=1}^{n_{i'}}\sum_{k'=1}^{m_{i'j'}} 2w_{i'j'k'}[\hat{F}^*(x_{i'j'k'}) - \bar{\hat{F}}^*] \\
&\quad \times [I(x_{ijk} \le x_{i'j'k'}) + I(x_{ijk} < x_{i'j'k'})]/2 \\
III_i &= -\frac{C_{ni}}{B_n}\sum_{i'=1}^{n}\sum_{j'=1}^{n_{i'}}\frac{2w_{i'j'.}}{m_{i'j'}(m_{i'j'}-1)}\sum_{k'>k}[\hat{F}^*(x_{i'j'k}) - \bar{\hat{F}}^* + \hat{F}^*(x_{i'j'k'}) - \bar{\hat{F}}^*] \\
IV_i &= \frac{A_n^{(2)}C_{ni}}{B_n^2}\sum_{i'=1}^{n}\sum_{j'=1}^{n_{i'}}\sum_{k'=1}^{m_{i'j'}} 2w_{i'j'k'}[\hat{F}^*(x_{i'j'k'}) - \bar{\hat{F}}^*] \\
V_i &= \frac{1}{B_n}\sum_{j=1}^{n_i}\sum_{k=1}^{m_{ij}} w_{ijk}\sum_{i'=1}^{n}\frac{w_{i'..}}{w_{i'}}\sum_{j'>j''}\sum_{k',l}\Big(\{\hat{F}^*(x_{i'j'k'}) - \bar{\hat{F}}^*\} \\
&\quad \times \{[I(x_{ijk} \le x_{i'j''l}) + I(x_{ijk} < x_{i'j''l})]/2\} \\
&\quad + \{\hat{F}^*(x_{i'j''l}) - \bar{\hat{F}}^*\}\{[I(x_{ijk} \le x_{i'j'k'}) + I(x_{ijk} < x_{i'j'k'})]/2\}\Big) \\
VI_i &= -\frac{A_n^{(3)}}{B_n^2}\sum_{j=1}^{n_i}\sum_{k=1}^{m_{ij}} w_{ijk}\sum_{i'=1}^{n}\sum_{j'=1}^{n_{i'}}\sum_{k'=1}^{m_{i'j'}} 2w_{i'j'k'}[\hat{F}^*(x_{i'j'k'}) - \bar{\hat{F}}^*] \\
&\quad \times [I(x_{ijk} \le x_{i'j'k'}) + I(x_{ijk} < x_{i'j'k'})]/2 \\
VII_i &= -\frac{C_{ni}}{B_n}\sum_{i'=1}^{n}\frac{w_{i'..}}{w_{i'}}\sum_{j'>j}\sum_{k,l}[\hat{F}^*(x_{i'j'k}) - \bar{\hat{F}}^* + \hat{F}^*(x_{i'j''l}) - \bar{\hat{F}}^*]
\end{aligned}$$

$$VIII_i = \frac{A_n^{(3)} C_{ni}}{B_n^2} \sum_{i'=1}^{n} \sum_{j'=1}^{n_{i'}} \sum_{k'=1}^{m_{i'j'}} 2 w_{i'j'k'} [\hat{F}^*(x_{i'j'k'}) - \bar{\hat{F}}^*]$$

where $C_{ni} = \sum_{j=1}^{n_i} \sum_{k=1}^{m_{ij}} w_{ijk} \hat{F}^*(x_{ijk}) + \frac{1}{n} \sum_{i'=1}^{n} \sum_{j'=1}^{n_{i'}} \sum_{k'=1}^{m_{i'j'}} w_{i'j'k'} \Big\{ \sum_{j=1}^{n_i} \sum_{k=1}^{m_{ij}} w_{ijk} [I(x_{ijk} \le x_{i'j'k'}) + I(x_{ijk} < x_{i'j'k'})]/2 \Big\}$. Then we can obtain the sample variances of $t_2(\mathbf{x})$ and $t_3(\mathbf{x})$ with $\{\hat{t}_2(\mathbf{x}_i), i = 1, ..., n\}$ and $\{\hat{t}_3(\mathbf{x}_i), i = 1, ..., n\}$, denoted as $\hat{\sigma}_{t_2}^2$ and $\hat{\sigma}_{t_3}^2$. The asymptotic variances of $\hat{\gamma}_{I2}$ and $\hat{\gamma}_{I3}$ are estimated by $\hat{\sigma}_{t_2}^2/n$ and $\hat{\sigma}_{t_3}^2/n$, respectively.

Table 2.4: Bias, standard error (SE) and coverage of 95% CIs based on the six possible inference approaches for $\hat{\gamma}_I$ at $\gamma_I = 0.48$ and different numbers of clusters under Scenario I (normality). We set the cluster size to be 30. There were 200 replicates per bootstrap.

(a) Coverage probabilities of 95% CIs

Size	Asymptotic SE	Fisher transformation	Cluster bootstrap percentile	Two-stage bootstrap percentile	Cluster bootstrap SE	Two-stage bootstrap SE
25	0.913	0.911	0.860	0.911	0.916	0.923
50	0.945	0.935	0.915	0.965	0.949	0.950
100	0.939	0.938	0.922	0.953	0.935	0.941
200	0.944	0.945	0.921	0.923	0.943	0.948
500	0.950	0.947	0.944	0.841	0.946	0.947
1000	0.949	0.950	0.934	0.679	0.949	0.954

(b) Bias and SE of $\hat{\gamma}_I$

Size	Percent bias (%)	Empirical SE	Averaged asymptotic SE	Averaged cluster bootstrap SE	Averaged two-stage bootstrap SE
25	-3.859	0.077	0.075	0.074	0.076
50	-1.684	0.053	0.054	0.054	0.054
100	-0.726	0.039	0.038	0.038	0.039
200	-0.274	0.027	0.027	0.027	0.027
500	-0.069	0.017	0.017	0.017	0.017
1000	-0.015	0.012	0.012	0.012	0.012

(c) Non-Coverage at the tails of 95% CIs of γ_I

Size	Asymptotic SE	Fisher transformation	Cluster bootstrap percentile	Two-stage bootstrap percentile	Cluster bootstrap SE	Two-stage bootstrap SE
25	R=0.07, L=0.017	R=0.08, L=0.009	R=0.073, L=0.011	R=0.067, L=0.01	R=0.133, L=0.007	R=0.081, L=0.008
50	R=0.04, L=0.015	R=0.055, L=0.01	R=0.039, L=0.012	R=0.038, L=0.012	R=0.081, L=0.004	R=0.027, L=0.008
100	R=0.035, L=0.026	R=0.042, L=0.02	R=0.042, L=0.023	R=0.037, L=0.022	R=0.061, L=0.017	R=0.012, L=0.035
200	R=0.034, L=0.022	R=0.037, L=0.018	R=0.038, L=0.019	R=0.033, L=0.019	R=0.059, L=0.02	R=0.008, L=0.069
500	R=0.02, L=0.03	R=0.023, L=0.03	R=0.024, L=0.03	R=0.022, L=0.031	R=0.029, L=0.027	R=0.005, L=0.154
1000	R=0.029, L=0.022	R=0.029, L=0.021	R=0.029, L=0.022	R=0.026, L=0.02	R=0.039, L=0.027	R=0.003, L=0.318

Table 2.5: Coverage and non-Coverage at the tails of 95% CIs based on the four bootstrapping inference approaches for $\hat{\gamma}_I$ under Scenario I at $\gamma_I = 0.48$. We set the cluster size to be 30. There were 1000 replicates per bootstrap.

(a) Coverage of bootstrap methods with 1000 replications each bootstrap

Size	Cluster bootstrap percentile	Two-stage bootstrap percentile	Cluster bootstrap SE	Two-stage bootstrap SE
25	0.889	0.932	0.913	0.918
50	0.930	0.960	0.939	0.950
100	0.934	0.952	0.937	0.942
200	0.940	0.935	0.944	0.946
500	0.952	0.855	0.950	0.954

(b) Non-Coverage at the tails of 95% CIs of γ_I

Size	Cluster bootstrap percentile	Two-stage bootstrap percentile	Cluster bootstrap SE	Two-stage bootstrap SE
25	R=0.072, L=0.015	R=0.067, L=0.015	R=0.106, L=0.005	R=0.063, L=0.005
50	R=0.046, L=0.015	R=0.035, L=0.015	R=0.064, L=0.006	R=0.026, L=0.014
100	R=0.037, L=0.026	R=0.032, L=0.026	R=0.047, L=0.019	R=0.012, L=0.036
200	R=0.034, L=0.022	R=0.032, L=0.022	R=0.044, L=0.016	R=0.006, L=0.059
500	R=0.019, L=0.031	R=0.018, L=0.028	R=0.023, L=0.025	R=0.003, L=0.142

Table 2.6: Bias, SE, and coverage of 95% CIs based on the six possible inference approaches for $\hat{\gamma}_I$ at $\gamma_I = 0.09$ and different numbers of clusters under Scenario I. We set the cluster size to be 30. There were 200 replicates per bootstrap.

(a) Coverage of 95% CIs

Size	Asymptotic SE	Fisher transformation	Cluster bootstrap percentile	Two-stage bootstrap percentile	Cluster bootstrap SE	Two-stage bootstrap SE
25	0.872	0.873	0.852	0.973	0.868	0.937
50	0.912	0.912	0.900	0.898	0.907	0.965
100	0.910	0.911	0.910	0.724	0.906	0.970
200	0.942	0.941	0.930	0.372	0.942	0.974
500	0.938	0.938	0.927	0.018	0.934	0.977
1000	0.952	0.952	0.945	0.001	0.952	0.975

(b) Bias and SE of $\hat{\gamma}_I$

Size	Percent bias (%)	Empirical SE	Averaged asymptotic SE	Averaged cluster bootstrap SE	Averaged two-stage bootstrap SE
25	-6.199	0.032	0.030	0.029	0.036
50	-2.340	0.023	0.022	0.022	0.027
100	-1.160	0.017	0.016	0.016	0.019
200	-0.311	0.012	0.011	0.011	0.014
500	-0.151	0.007	0.007	0.007	0.009
1000	-0.002	0.005	0.005	0.005	0.006

Table 2.7: Bias, SE and coverage of 95% CIs based on the six possible inference approaches for $\hat{\gamma}_I$ at $\gamma_I = 0.89$ and different numbers of clusters under Scenario I. We set the cluster size to be 30. There were 200 replicates per bootstrap.

(a) Coverage of 95% CIs

Size	Asymptotic SE	Fisher transformation	Cluster bootstrap percentile	Two-stage bootstrap percentile	Cluster bootstrap SE	Two-stage bootstrap SE
25	0.960	0.924	0.816	0.852	0.975	0.974
50	0.969	0.949	0.893	0.921	0.978	0.977
100	0.956	0.939	0.902	0.941	0.963	0.958
200	0.953	0.952	0.932	0.949	0.954	0.953
500	0.951	0.948	0.931	0.932	0.944	0.944
1000	0.950	0.945	0.937	0.885	0.949	0.947

(b) Bias and SE of $\hat{\gamma}_I$

Size	Percent bias (%)	Empirical SE	Averaged asymptotic SE	Averaged cluster bootstrap SE	Averaged two-stage bootstrap SE
25	-2.507	0.039	0.039	0.044	0.043
50	-1.161	0.024	0.025	0.027	0.027
100	-0.566	0.017	0.017	0.018	0.017
200	-0.268	0.012	0.012	0.012	0.012
500	-0.099	0.007	0.007	0.007	0.007
1000	-0.048	0.005	0.005	0.005	0.005

Table 2.8: Bias, SE, and coverage of 95% CIs based on the six possible inference approaches and for $\hat{\gamma}_I$ at $\gamma_I = 0.48$ and different numbers of clusters under Scenario I. The cluster size followed a uniform distribution from 2 to 50. There were 200 replicates per bootstrap. "Equal clusters" refers to assigning equal weights to clusters. "Equal obs" refers to assigning equal weights to observations. "PR." refers to the approaches using bootstrap percentiles. "Asym. SE" refers to the approach using asymptotic standard error. "Emp. SE" refers to the empirical standard error. "Avg. SE" refers to the average of standard errors. "boot." means bootstrap.

(a) Coverage of 95% CIs of γ_I

Size	Equal clusters					Equal obs				
	Asym. SE	Cluster bootstrap		Two-stage bootstrap		Asym. SE	Cluster bootstrap		Two-stage bootstrap	
		PR.	SE	PR.	SE		PR.	SE	PR.	SE
25	0.910	0.893	0.909	0.952	0.919	0.900	0.872	0.908	0.917	0.908
50	0.934	0.919	0.926	0.954	0.939	0.910	0.899	0.913	0.931	0.912
100	0.953	0.941	0.948	0.923	0.950	0.950	0.941	0.951	0.955	0.952
200	0.953	0.948	0.950	0.831	0.952	0.948	0.934	0.952	0.928	0.952
500	0.942	0.932	0.944	0.525	0.944	0.945	0.938	0.945	0.861	0.945
1000	0.947	0.935	0.945	0.218	0.946	0.938	0.932	0.938	0.717	0.942

(b) Bias and SE of $\hat{\gamma}_I$

Size	Equal clusters					Equal obs				
	Bias (%)	Emp. SE	Avg. asym. SE	Avg. cluster boot. SE	Avg. two-stage boot. SE	Bias (%)	Emp. SE	Avg. asym. SE	Avg. cluster boot. SE	Avg. two-stage boot. SE
25	-4.033	0.084	0.082	0.081	0.082	-4.605	0.086	0.083	0.082	0.083
50	-2.029	0.059	0.058	0.058	0.058	-2.375	0.063	0.060	0.059	0.060
100	-0.783	0.040	0.041	0.041	0.041	-1.059	0.041	0.043	0.043	0.043
200	-0.713	0.029	0.029	0.029	0.029	-0.708	0.031	0.031	0.030	0.031
500	-0.258	0.019	0.019	0.019	0.019	-0.258	0.019	0.019	0.019	0.019
1000	-0.174	0.013	0.013	0.013	0.013	-0.181	0.014	0.014	0.014	0.014

Table 2.9: Bias, SE, and coverage of 95% CIs based on the eight possible inference approaches for $\hat{\gamma}_{I2}$ at $(\gamma_{I2}, \gamma_{I3}) = (0.53, 0.19)$ and different number of level-3 units. We set the number of level-2 units in a level-3 unit to be 15, and the number of level-1 units in a level-2 unit to be 2. There were 200 replicates per bootstrap. "Avg. SE" refers to the average of standard errors.

(a) Coverage of 95% CIs of γ_{I2}

Size	Asymptotic SE	Fisher transformation	Bootstrap SE			Bootstrap percentiles		
			One-stage	Two-stage	Three-stage	One-stage	Two-stage	Three-stage
25	0.935	0.931	0.925	0.982	0.875	0.908	0.975	0
50	0.950	0.946	0.946	0.991	0.887	0.934	0.985	0
100	0.940	0.937	0.939	0.992	0.869	0.938	0.983	0
200	0.942	0.938	0.940	0.980	0.867	0.929	0.977	0
500	0.958	0.961	0.958	0.995	0.899	0.955	0.989	0
1000	0.950	0.952	0.948	0.987	0.871	0.945	0.984	0

(b) Bias and SE of $\hat{\gamma}_{I2}$

Size	Percent bias (%)	Emp.SE	Avg. asymptotic SE	Avg. cluster bootstrap SE	Avg. two-stage bootstrap SE	Avg. three-stage bootstrap SE
25	-0.981	0.047	0.045	0.044	0.058	0.037
50	-0.480	0.031	0.032	0.032	0.041	0.026
100	-0.381	0.024	0.023	0.023	0.029	0.018
200	-0.120	0.017	0.016	0.016	0.021	0.013
500	-0.025	0.010	0.010	0.010	0.013	0.008
1000	-0.031	0.007	0.007	0.007	0.009	0.006

Table 2.10: Bias, SE, and coverage of 95% CIs based on the eight possible inference approaches for $\hat{\gamma}_{I3}$ at $(\gamma_{I2}, \gamma_{I3}) = (0.53, 0.19)$ and different number of level-3 units. We set the number of level-2 units in a level-3 unit to be 15, and the number of level-1 units in a level-2 unit to be 2. There were 200 replicates per bootstrap. "Avg. SE" refers to the average of standard errors.

(a) Coverage of 95% CIs of γ_{I3}

Size	Asymptotic SE	Fisher transformation	Bootstrap SE			Bootstrap percentiles		
			One-stage	Two-stage	Three-stage	One-stage	Two-stage	Three-stage
25	0.896	0.895	0.885	0.934	0.941	0.861	0.968	0.970
50	0.928	0.928	0.924	0.958	0.963	0.915	0.939	0.951
100	0.933	0.934	0.933	0.955	0.962	0.922	0.809	0.831
200	0.940	0.939	0.932	0.967	0.970	0.924	0.584	0.628
500	0.961	0.960	0.953	0.975	0.978	0.946	0.136	0.162
1000	0.939	0.939	0.939	0.958	0.965	0.935	0.007	0.008

(b) Bias and SE of $\hat{\gamma}_{I3}$

Size	Percent bias (%)	Emp.SE	Avg. asymptotic SE	Avg. cluster bootstrap SE	Avg. two-stage bootstrap SE	Avg. three-stage bootstrap SE
25	-4.809	0.053	0.050	0.049	0.056	0.058
50	-2.064	0.037	0.037	0.036	0.041	0.042
100	-0.774	0.027	0.027	0.026	0.029	0.030
200	-0.617	0.019	0.019	0.019	0.021	0.022
500	-0.112	0.012	0.012	0.012	0.013	0.014
1000	-0.114	0.009	0.008	0.008	0.009	0.010

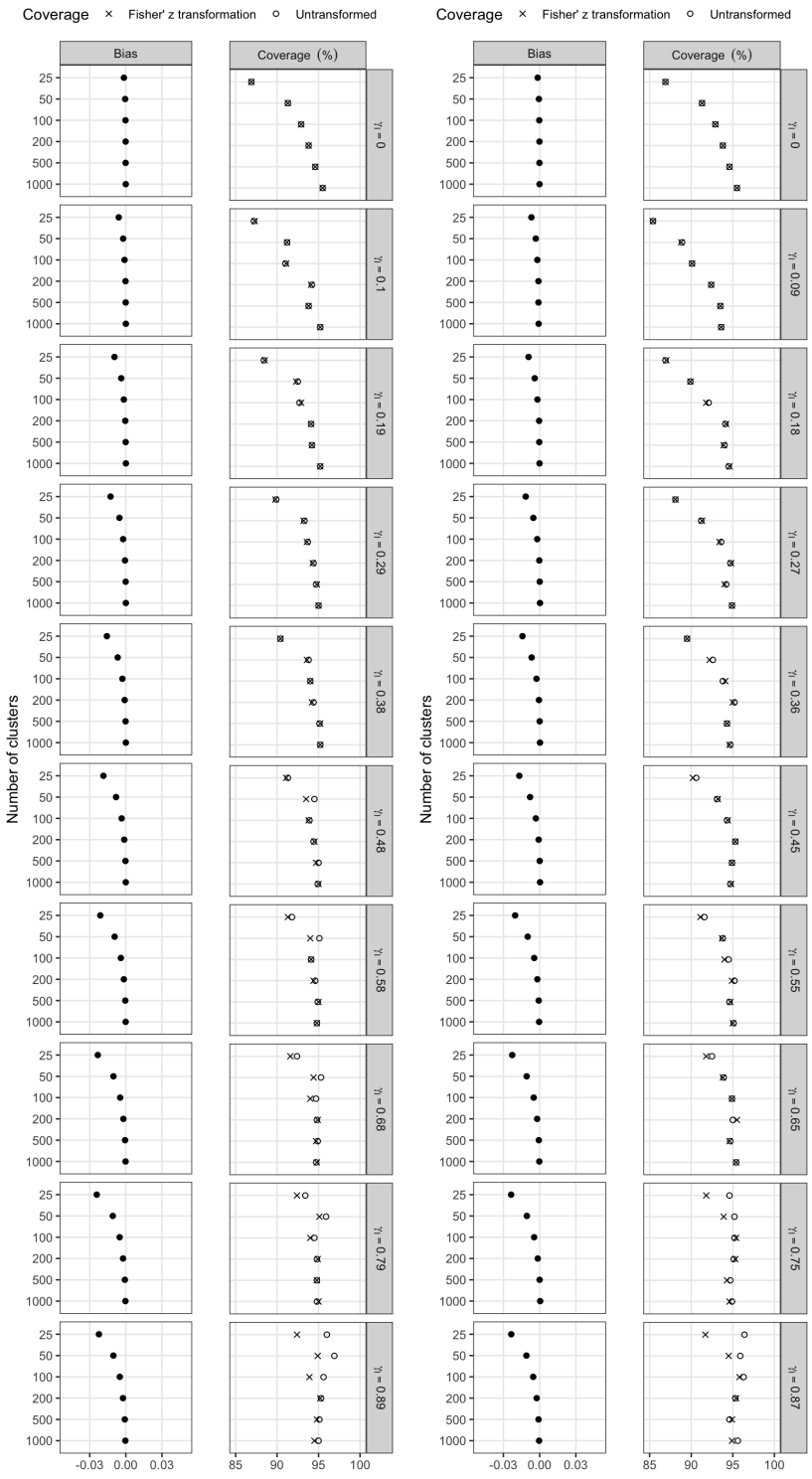
Table 2.11: Bias, SE, and coverage of 95% CIs based on the eight possible inference approaches for $\hat{\gamma}_{I2}$ at $(\gamma_{I2}, \gamma_{I3}) = (0.53, 0.19)$ and unequal numbers of level-2 units per level-3 unit. We set the number of level-2 units in a level-3 unit to follow a uniform distribution between 2 and 15, and the number of level-1 units in a level-2 unit to be 2. There were 200 replicates per bootstrap. "Equal level-3 units" refers to assigning equal weights to level-3 units. "Equal level-2/1 units" refers to assigning equal weights to level-2/1 units. "Asym. SE" refers to the approach using asymptotic standard error.

(a) Coverage of 95% CIs of γ_{I2}

Size	Equal level-3 units				Equal level-2/1 units			
	Bias (%)	Asym. SE	Bootstrap SE	Bootstrap percentiles	Bias (%)	Asym. SE	Bootstrap SE	Bootstrap percentiles
25	-1.772	0.932	0.925	0.921	-1.576	0.915	0.915	0.899
50	-0.655	0.948	0.946	0.939	-0.689	0.952	0.946	0.934
100	-0.529	0.946	0.946	0.933	-0.345	0.944	0.945	0.938
200	-0.059	0.950	0.949	0.941	-0.066	0.953	0.947	0.944
500	-0.108	0.951	0.953	0.942	-0.077	0.950	0.949	0.941
1000	-0.047	0.946	0.949	0.940	-0.064	0.954	0.952	0.942

(b) Coverage of 95% CIs of γ_{I3}

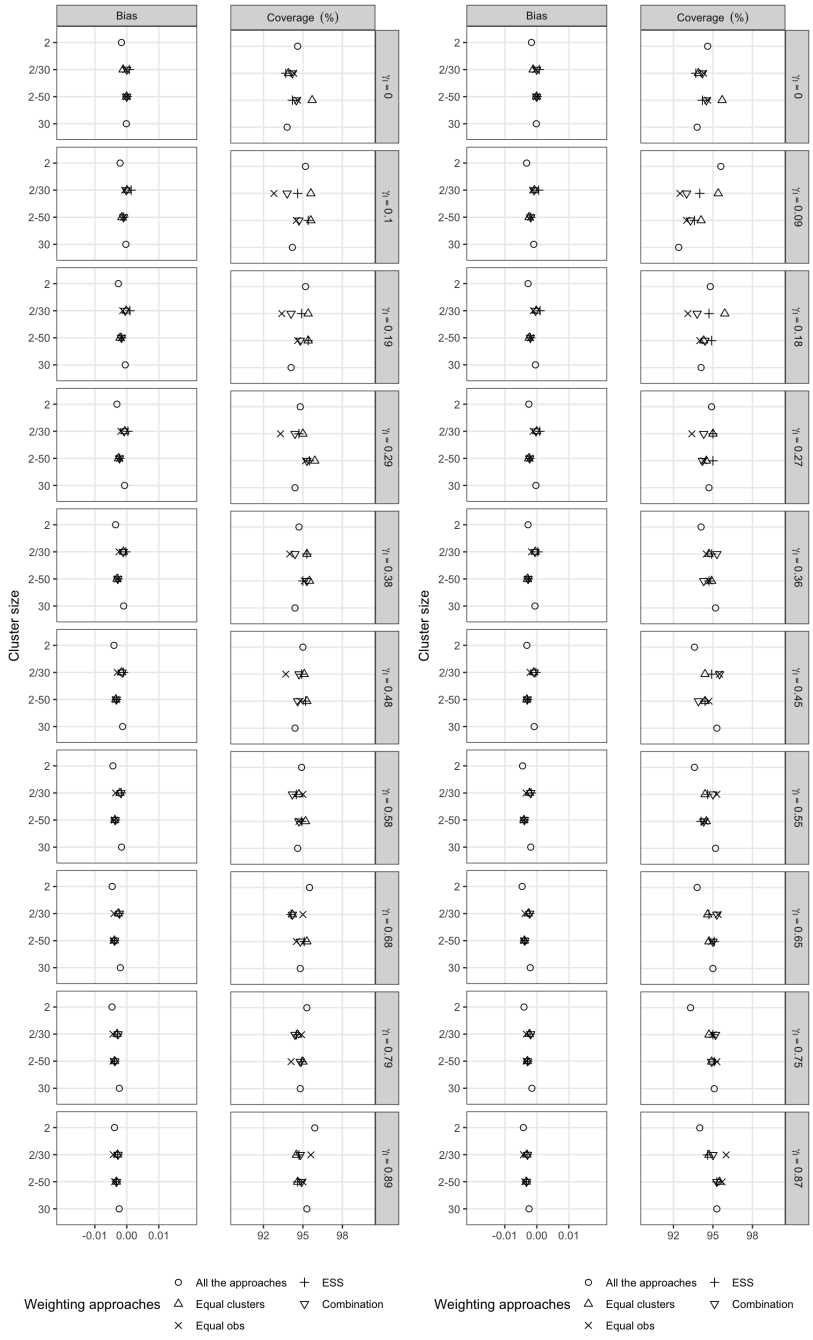
Size	Equal level-3 units				Equal level-2/1 units			
	Bias (%)	Asym. SE	Bootstrap SE	Bootstrap percentiles	Bias (%)	Asym. SE	Bootstrap SE	Bootstrap percentiles
25	-7.010	0.898	0.891	0.864	-6.003	0.875	0.867	0.844
50	-3.672	0.928	0.923	0.912	-3.204	0.911	0.901	0.896
100	-2.087	0.932	0.932	0.919	-1.532	0.927	0.925	0.918
200	-0.273	0.944	0.939	0.938	0.113	0.935	0.929	0.922
500	-0.628	0.945	0.944	0.923	-0.536	0.938	0.939	0.929
1000	-0.439	0.954	0.953	0.942	-0.239	0.954	0.950	0.947



(a) Scenario I and II

(b) Scenario III

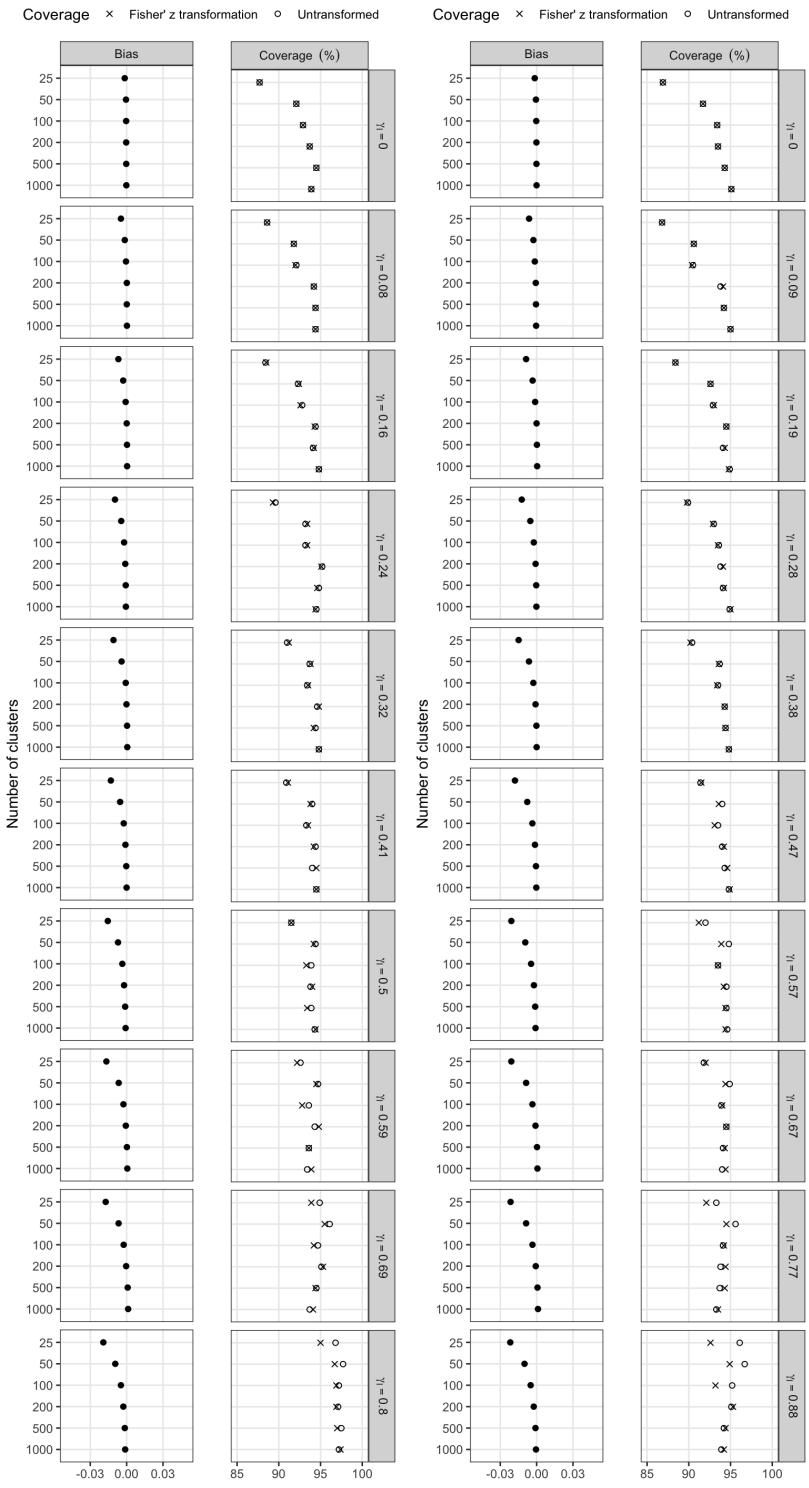
Figure 2.11: Bias and coverage of 95% CIs (i.e., based on the asymptotic SE only, based on the asymptotic SE and Fisher' z transformation) for $\hat{\gamma}_1$ at different true values of γ_1 and different numbers of clusters under Scenarios I, II, and III. The number of observations per cluster was set at 30.



(a) Scenarios I and II

(b) Scenario III

Figure 2.12: Bias and coverage of 95% CIs for $\hat{\gamma}_I$ at different true values of γ_I and different cluster sizes under Scenarios I, II, and III. The number of clusters was set at 200. "2-50" means the cluster size follows a uniform distribution from 2 to 50, "2/30" means half of the clusters have size 2 and half have 30. "Equal clusters" refers to assigning equal weights to clusters, "Equal obs" refers to assigning equal weights to observations, "ESS" refers to the iterative weighting approach based on the effective sample size, and "Combination" refers to the iterative weighting approach based on the linear combination of equal weights for clusters and equal weights for observations. We set the tolerance of the two iterative approaches to be 0.00001. The estimates of the four approaches were identical when cluster sizes were equal.



(a) 3-level (b) 10-level

Figure 2.13: Bias and coverage of 95% CIs (i.e., based on the asymptotic SE only, based on the asymptotic SE and Fisher' z transformation) for $\hat{\gamma}_I$ at different true values of γ_I of 3-level and 10-level ordered categorical variables. The number of observations per cluster was set at 30.

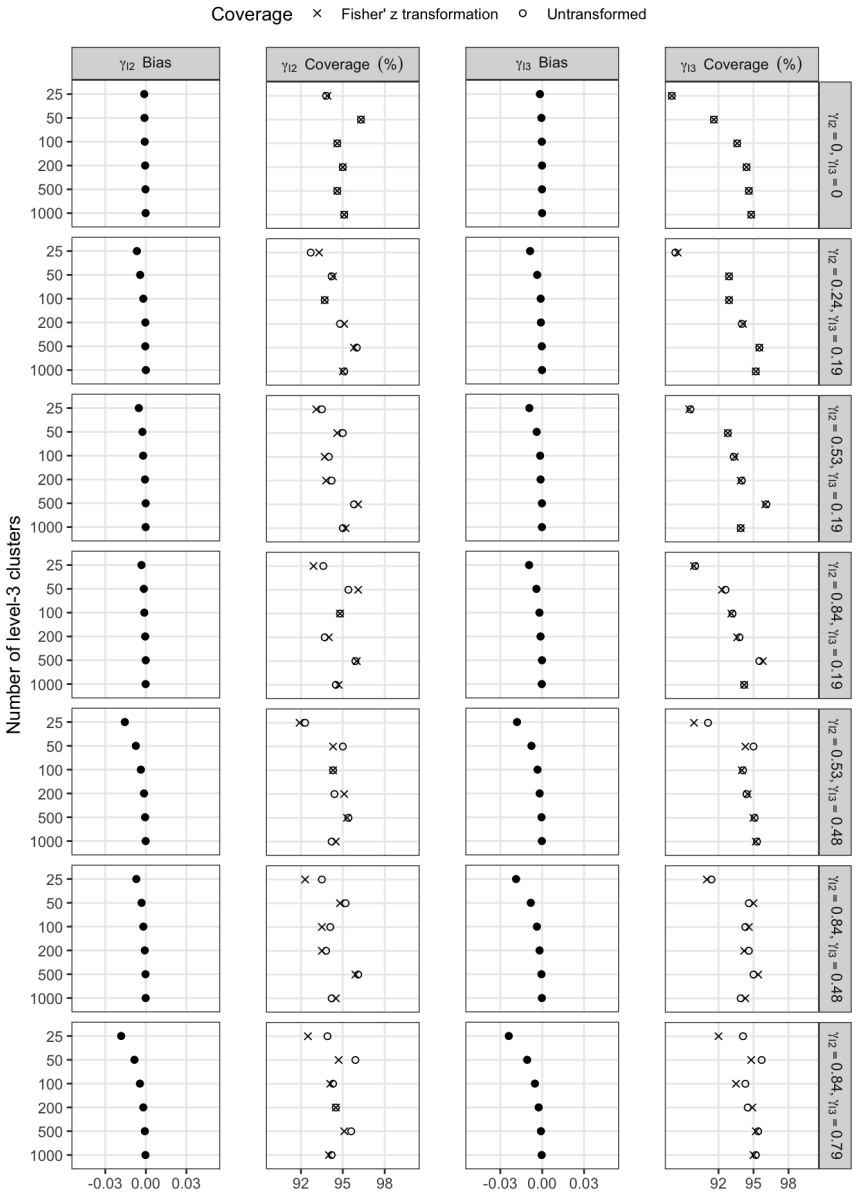


Figure 2.14: Bias and coverage of 95% CIs (i.e., based on the asymptotic SE only, based on the asymptotic SE and Fisher' z transformation) for $\hat{\gamma}_{I2}$ and $\hat{\gamma}_{I3}$ at different true values of γ_{I2} and γ_{I3} and different numbers of level-3 units. The number of level-2 units in a level-3 unit was set at 15. The number of level-1 units in a level-2 unit was set at 2.

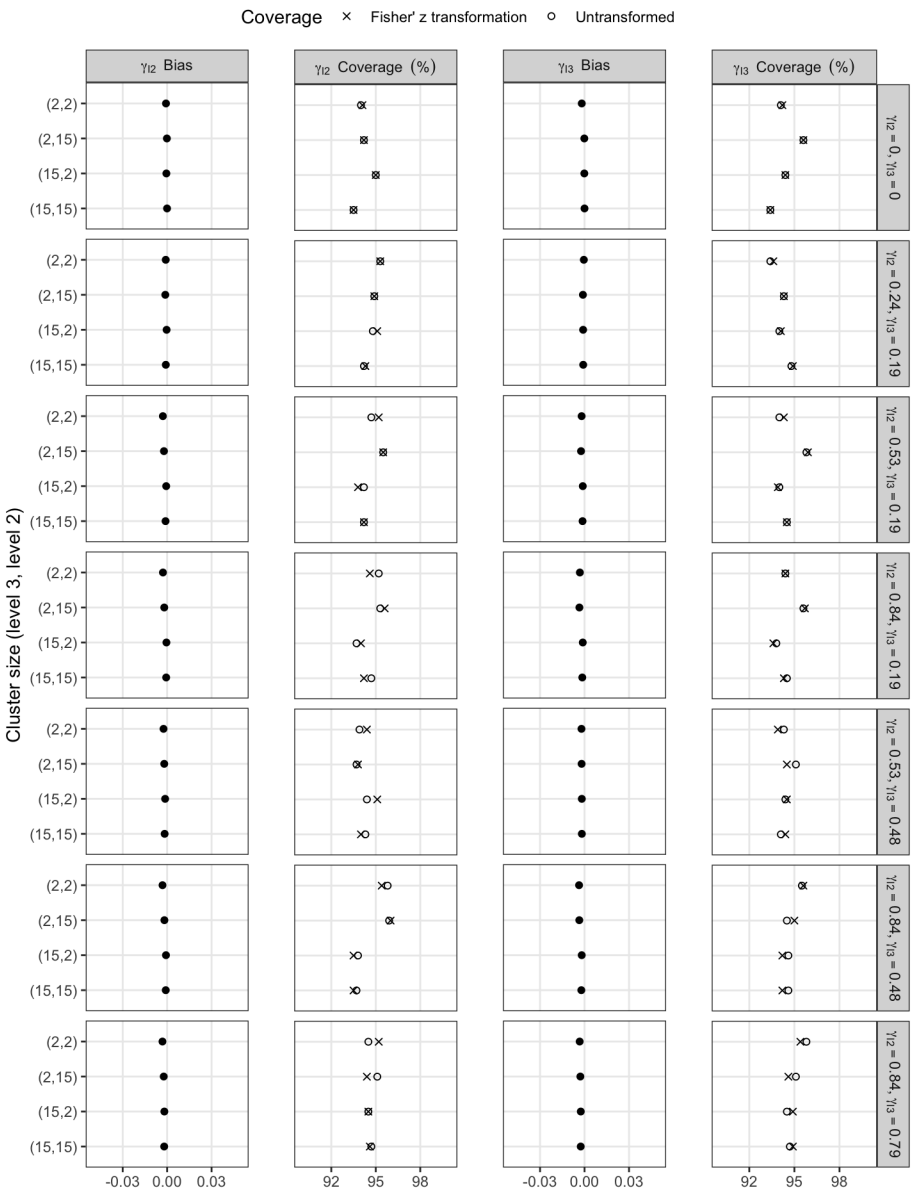


Figure 2.15: Bias and coverage of 95% CIs (i.e., based on the asymptotic SE only, based on the asymptotic SE and Fisher' z transformation) for $\hat{\gamma}_{l2}$ and $\hat{\gamma}_{l3}$ at different true values of γ_{l2} and γ_{l3} and different sizes of level-3 units and level-2 units. The number of level-3 units was set at 200.

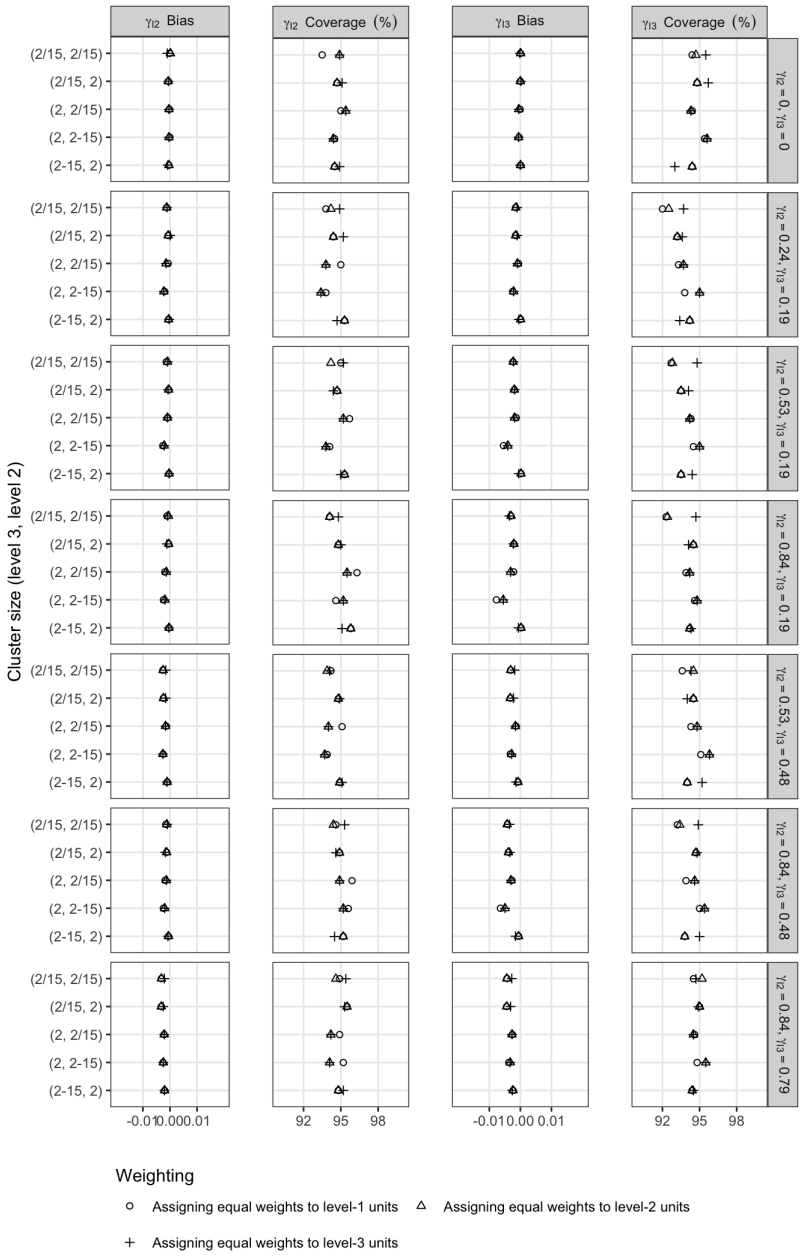


Figure 2.16: Bias and coverage of 95% CIs for $\hat{\gamma}_{I2}$ and $\hat{\gamma}_{I3}$ at different true values of γ_{I2} and γ_{I3} and different unequal sizes of level-3 units and level-2 units. The number of level-3 units was set at 200. "2-15" means the cluster size follows a uniform distribution from 2 to 15, "2/15" means half of the clusters have size 2 and a half have 15.

CHAPTER 3

Between- and Within-Cluster Spearman Rank Correlations

3.1 Introduction

Clustered data are common in practice. Clustering arises when subjects (e.g., people) are measured repeatedly, or subjects are nested in clusters (e.g., households, schools) and measured only once. The total, between-, and within-cluster Pearson correlations are frequently used in the analysis of clustered data (Snijders and Bosker, 1999; Ferrari et al., 2005). The total correlation measures the overall correlation but fails to acknowledge the clustered nature of the data. The between-cluster correlation measures the association between underlying variables at the cluster level, while the within-cluster correlation is the correlation after controlling for clustering.

For example, in an observational study, people living with HIV on antiretroviral therapy (ART) had repeated measurements of CD4 and CD8 counts (Castilho et al., 2016). There is interest in measuring the correlation between CD4 and CD8 counts subject to clustering. The between-cluster correlation measures the association between the underlying CD4 and CD8 counts in individuals. The within-cluster correlation describes the correlation between variations in CD4 and CD8 measures due to changes over time or measurement errors. Together, the total, between-, and within-cluster correlations provide a more complete picture of the relationship between CD4 and CD8 counts.

However, Pearson correlations are sensitive to extreme values and skewed distributions, and they depend on the scale of the data. For example, CD4 and CD8 counts are both right-skewed and sometimes transformed prior to analyses; estimates of the total, between-, and within-cluster Pearson correlations will vary with the transformation. Some recent studies have proposed nonparametric measures of correlation for clustered data. Rosner and Glynn (2017) proposed a regression-based approach to obtain the maximum likelihood estimate of Pearson correlation for clustered data and then compute Spearman rank correlation by using its relationship with Pearson correlation under bivariate normality. Shih and Fay (2017) defined Spearman rank correlation for clustered data as the Pearson correlation between the population versions of ridits (Bross, 1958), and applied within-cluster resampling and U-statistics for estimation and inference. Hunsberger, et al (2022) extended the work of Shih and Fay by improving the nominal level of the tests for clustered data with small sample sizes. However, these nonparametric measures are only for the total correlation. There is a need to develop rank-based between- and within-cluster correlations.

In this Chapter, we define population parameters of between- and within-cluster Spearman rank correla-

tions, which are natural extensions of the traditional between- and within-cluster Pearson correlations to the rank scale. We show that the total Spearman rank correlation approximates a weighted sum of the between- and within-cluster Spearman rank correlations, where the weights are functions of rank intraclass correlation coefficients (Tu et al., 2023). We also show the equivalence between the within-cluster Spearman's rank correlation and the covariate-adjusted partial Spearman's rank correlation with cluster indicators as covariates (Liu et al., 2018).

3.2 Review of Pearson correlations for clustered data

Let X_{ij} and Y_{ij} denote two random variables from a two-level hierarchical joint distribution, where i represents cluster and j is the index within cluster i. For the review in this section, we assume an additive model and equal within-cluster covariance matrices across clusters. Specifically, consider a bivariate population model of $(X_{ij}, Y_{ij})^T$ in an infinite population,

$$\begin{pmatrix} X_{ij} \\ Y_{ij} \end{pmatrix} = \begin{pmatrix} U_{Xi} \\ U_{Yi} \end{pmatrix} + \begin{pmatrix} R_{Xij} \\ R_{Yij} \end{pmatrix}, \tag{3.1}$$

where $(U_{Xi}, U_{Yi})^T$ is the cluster mean of the ith cluster, $(R_{Xij}, R_{Yij})^T$ is the within-cluster deviation of the jth observation in the ith cluster with zero means, and $(U_{Xi}, U_{Yi})^T \perp (R_{Xij}, R_{Yij})^T$. The covariance matrix of $(U_{Xi}, U_{Yi})^T$ is denoted as $\begin{pmatrix} \sigma_u^2 & \rho_b \sigma_u \eta_u \\ \rho_b \sigma_u \eta_u & \eta_u^2 \end{pmatrix}$, and the covariance matrix of $(R_{Xij}, R_{Yij})^T$ is denoted as $\begin{pmatrix} \sigma_r^2 & \rho_w \sigma_r \eta_r \\ \rho_w \sigma_r \eta_r & \eta_r^2 \end{pmatrix}$. The intraclass correlation coefficient (ICC) of X is $\rho_{I_X} = \sigma_u^2/(\sigma_u^2 + \sigma_r^2)$ and of Y is $\rho_{I_Y} = \eta_u^2/(\eta_u^2 + \eta_r^2)$, evaluating the correlation between two random observations in a random cluster.

With the above bivariate population model, the between-cluster correlation is the correlation between the cluster means, $\rho_b = corr(U_{Xi}, U_{Yi})$. The within-cluster correlation is the correlation between the within-cluster deviations, $\rho_w = corr(R_{Xij}, R_{Yij})$. Since

$$\begin{aligned} \rho_t &= \rho(X_{ij}, Y_{ij}) \\ &= \frac{\text{cov}(X_{ij}, Y_{ij})}{\sqrt{\text{var}(X_{ij})\text{var}(Y_{ij})}} \\ &= \frac{\text{cov}(U_{Xi}, U_{Yi}) + \text{cov}(R_{Xij}, R_{Yij})}{\sqrt{\text{var}(X_{ij})\text{var}(Y_{ij})}} \\ &= \frac{\rho_b \sigma_u \eta_u + \rho_w \sigma_r \eta_r}{\sqrt{(\sigma_u^2 + \sigma_r^2)(\eta_u^2 + \eta_r^2)}} \\ &= \rho_b \sqrt{\rho_{I_X} \rho_{I_Y}} + \rho_w \sqrt{(1 - \rho_{I_X})(1 - \rho_{I_Y})}, \end{aligned} \tag{3.2}$$

the total correlation is a weighted sum of the between- and within-cluster correlations, where the weights depend on the ICCs of X and Y.

3.3 Population parameters of Spearman rank correlations for clustered data

Spearman rank correlation is essentially the correlation between cumulative distribution functions (CDFs) for continuous variables, also known as the grade correlation (Kruskal, 1958), or more broadly, the correlation between population versions of midranks or ridits (Kendall, 1970; Bross, 1958). Generically, let F be a CDF, $F(x-) = \lim_{t\uparrow x} F(t)$, and $F^*(x) = \{F(x) + F(x-)\}/2$. If the distribution is continuous, then $F^*(x) = F(x)$. If the distribution is discrete or mixed, $F^*(x)$ corresponds to the population versions of ridits (Bross, 1958). The population parameter of Spearman rank correlation between two random variables X and Y with CDFs F_X and F_Y is denoted as $\gamma(X,Y) = corr\{F_X^*(X), F_Y^*(Y)\}$ (Kendall, 1970; Liu et al., 2018).

Let X_{ij} and Y_{ij} denote two random variables from a two-level hierarchical joint distribution, where i represents cluster and j is the index within cluster i. The total Spearman rank correlation is the overall rank correlation between X_{ij} and Y_{ij}. We define its population parameter as

$$\gamma_t = \gamma(X_{ij}, Y_{ij}) = corr\{F_X^*(X_{ij}), F_Y^*(Y_{ij})\}. \tag{3.3}$$

With continuous X_{ij} and Y_{ij}, $\gamma_t = 12cov\{F_X(X_{ij}), F_Y(Y_{ij})\}$, because $F_X^*(X_{ij}) = F_X(X_{ij}) \sim Unif(0,1)$, $F_Y^*(Y_{ij}) = F_Y(Y_{ij}) \sim Unif(0,1)$, and their variances equal $1/12$.

Let $F_{X|i}$ and $F_{Y|i}$ be the CDFs of X and Y conditional on being in cluster i, respectively. The population parameter of the within-cluster Spearman rank correlation is defined as

$$\gamma_w = corr\{F_{X|i}^*(X_{ij}), F_{Y|i}^*(Y_{ij})\}. \tag{3.4}$$

Note that γ_w is not a function of cluster index and that it does not assume an equal variance structure across clusters. In fact, γ_w is identical to the covariate-adjusted partial Spearman rank correlation (Liu et al., 2018), where the covariates are cluster indicators. Since the partial Spearman rank correlation can be expressed using probability-scale residuals (PSRs) (Li and Shepherd, 2012; Shepherd et al., 2016), we can express γ_w similarly. The PSRs of $X_{ij} = x$ and $Y_{ij} = y$ are defined as $r(x, F_{X|i}) = 2F_{X|i}^*(x) - 1$ and $r(y, F_{Y|i}) = 2F_{Y|i}^*(y) - 1$, respectively. Then we have

$$\gamma_w = corr(r(X_{ij}, F_{X|i}), r(Y_{ij}, F_{Y|i})).$$

This connection allows us to derive an estimator for γ_w, which will be described in Section 3.5.

The usage of cluster means is not desirable for the between-cluster Spearman rank correlation because

means are scale-dependent and sensitive to outliers and skewness. We use the general concept of cluster centroids to define the between-cluster Spearman rank correlation. A cluster centroid defines the central tendency of random variables in the same cluster. It is usually the median. Let $\tilde{X}_i$ and $\tilde{Y}_i$ be the cluster centroid parameters of the ith cluster, with marginal CDFs denoted as $F_{\tilde{X}}$ and $F_{\tilde{Y}}$, respectively. Assuming that clusters are independent, the between-cluster Spearman rank correlation treats clusters as units of interest and measures the association between cluster centroids. We define its population parameter as

$$\gamma_b = \gamma(\tilde{X}_i, \tilde{Y}_i) = corr\{F^*_{\tilde{X}}(\tilde{X}_i), F^*_{\tilde{Y}}(\tilde{Y}_i)\}. \tag{3.5}$$

Our definitions of γ_t, γ_b, and γ_w are easily interpreted as rank correlations. In the special case where $(X_{ij}, Y_{ij})^T$ has a similar hierarchical population model as (3.1) in Section 3.2 except that $(U_{Xi}, U_{Yi})^T$ is the cluster median and $(R_{Xij}, R_{Yij})^T$ has a median of zero, then $\gamma_t = \gamma(X_{ij}, Y_{ij})$, $\gamma_b = \gamma(U_{Xi}, U_{Yi})$, and $\gamma_w = \gamma(R_{Xij}, R_{Yij})$.

Furthermore, our definitions of γ_t, γ_b, and γ_w are also applicable to ordered categorical data. While the definitions of γ_t and γ_w in (3.3) and (3.4) can be directly applied, the definition of γ_b in (3.5) needs an extension. For an ordered categorical variable X, the median is defined as any category c for which $P(X \leq c) \geq 0.5$ and $P(X \geq c) \geq 0.5$. The median is often a unique value. In the rare situation where $P(X \leq c) = 0.5$, both the category c and the next higher category (denoted as $c+$) are the medians, and we define the cluster centroid $\tilde{X}$ as c with a probability of 0.5 and $c+$ with a probability of 0.5. If there are clusters like this with two cluster medians for a variable, we define $\gamma_b = E[\gamma(\tilde{X}_i, \tilde{Y}_i)]$, the expectation of the Spearman rank correlation over all possible combinations of cluster medians in the population. If no clusters have two cluster medians, the definition in (3.5) can be directly applied.

3.4 Relationship between the total, between-, and within-cluster Spearman rank correlations

The total Spearman rank correlation can be decomposed into two weighted components. The weights are functions of the rank ICC, which is a natural extension of Fisher's ICC (Fisher, 1925) to the rank scale (Tu et al., 2023). The rank ICC of X is

$$\begin{aligned}
\gamma_{I_X} &= corr[F^*_X(X_{ij}), F^*_X(X_{ij'})] \\
&= cov[F^*_X(X_{ij}), F^*_X(X_{ij'})]/var[F^*_X(X_{ij})] \\
&= cov\{E[F^*_X(X_{ij})|i], E[F^*_X(X_{ij'})|i]\}/var[F^*_X(X_{ij})] \\
&\quad + E\{cov[F^*_X(X_{ij}), F^*_X(X_{ij'})|i]\}/var[F^*_X(X_{ij})] \\
&= var\{E[F^*_X(X_{ij})|i]\}/var[F^*_X(X_{ij})] + D_X,
\end{aligned}$$

where $(X_{ij}, X_{ij'})$ is a random pair drawn from a random cluster and $j \neq j'$, and $D_X = E\{cov[F_X^*(X_{ij}), F_X^*(X_{ij'})|i]\}/var[F_X^*(X_{ij})]$. When cluster sizes in the population are finite, D_X is negative. When cluster sizes in the population are infinite, D_X is equal to 0. The rank ICC of Y, γ_{I_Y}, is similarly defined.

The decomposition of the total Spearman rank correlation is

$$\begin{aligned}
\gamma_t &= corr\{F_X^*(X_{ij}), F_Y^*(Y_{ij})\} \\
&= \frac{cov\{F_X^*(X_{ij}), F_Y^*(Y_{ij})\}}{\sqrt{var[F_X^*(X_{ij})]var[F_Y^*(Y_{ij})]}} \\
&= \frac{cov\{E[F_X^*(X_{ij})|i], E[F_Y^*(Y_{ij})|i]\} + E\{cov[F_X^*(X_{ij}), F_Y^*(Y_{ij})|i]\}}{\sqrt{var[F_X^*(X_{ij})]var[F_Y^*(Y_{ij})]}} \\
&= \frac{cov\{E[F_X^*(X_{ij})|i], E[F_Y^*(Y_{ij})|i]\}}{var\{E[F_X^*(X_{ij})|i]\}var\{E[F_Y^*(Y_{ij})|i]\}} \sqrt{\frac{var\{E[F_X^*(X_{ij})|i]\}}{var[F_X^*(X_{ij})]}} \sqrt{\frac{var\{E[F_Y^*(Y_{ij})|i]\}}{var[F_Y^*(Y_{ij})]}} \\
&+ \frac{E\{cov[F_X^*(X_{ij}), F_Y^*(Y_{ij})|i]\}}{\sqrt{E\{var[F_X^*(X_{ij})|i]\}E\{var[F_Y^*(Y_{ij})|i]\}}} \sqrt{\frac{E\{var[F_X^*(X_{ij})|i]\}}{var[F_X^*(X_{ij})]}} \sqrt{\frac{E\{var[F_Y^*(Y_{ij})|i]\}}{var[F_Y^*(Y_{ij})]}} \\
&= corr\{E[F_X^*(X_{ij})|i], E[F_Y^*(Y_{ij})|i]\} \sqrt{(\gamma_{I_X} - D_X)(\gamma_{I_Y} - D_Y)} \\
&+ \frac{E\{cov[F_X^*(X_{ij}), F_Y^*(Y_{ij})|i]\}}{\sqrt{E\{var[F_X^*(X_{ij})|i]\}E\{var[F_Y^*(Y_{ij})|i]\}}} \sqrt{(1 - \gamma_{I_X} + D_X)(1 - \gamma_{I_Y} + D_Y)} \\
&= S_1 \sqrt{(\gamma_{I_X} - D_X)(\gamma_{I_Y} - D_Y)} + S_2 \sqrt{(1 - \gamma_{I_X} + D_X)(1 - \gamma_{I_Y} + D_Y)},
\end{aligned}$$

where $S_1 = corr\{E[F_X^*(X_{ij})|i], E[F_Y^*(Y_{ij})|i]\}$ and $S_2 = \frac{E\{cov[F_X^*(X_{ij}), F_Y^*(Y_{ij})|i]\}}{\sqrt{E[var[F_X^*(X_{ij})|i]E[var[F_Y^*(Y_{ij})|i]}}$. When the cluster size in the population is infinite, then $D_X = D_Y = 0$ and $\gamma_t = S_1\sqrt{\gamma_{I_X}\gamma_{I_Y}} + S_2\sqrt{(1-\gamma_{I_X})(1-\gamma_{I_Y})}$. Simulations suggest that S_1 and S_2 can be approximated by γ_b and γ_w, respectively. That is,

$$\gamma_t \approx \gamma_b \sqrt{(\gamma_{I_X} - D_X)(\gamma_{I_Y} - D_Y)} + \gamma_w \sqrt{(1 - \gamma_{I_X} + D_X)(1 - \gamma_{I_Y} + D_Y)}. \tag{3.6}$$

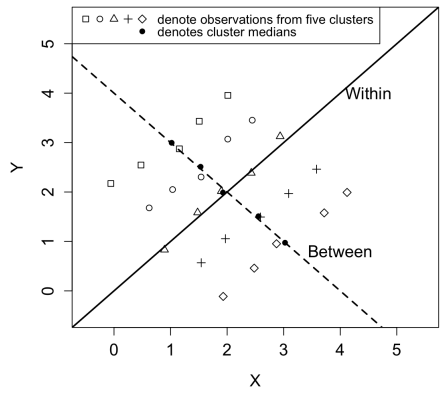
If the cluster size in the population is large, then $D_X \approx D_Y \approx 0$ and we have

$$\gamma_t \approx \gamma_b \sqrt{\gamma_{I_X}\gamma_{I_Y}} + \gamma_w \sqrt{(1-\gamma_{I_X})(1-\gamma_{I_Y})}. \tag{3.7}$$

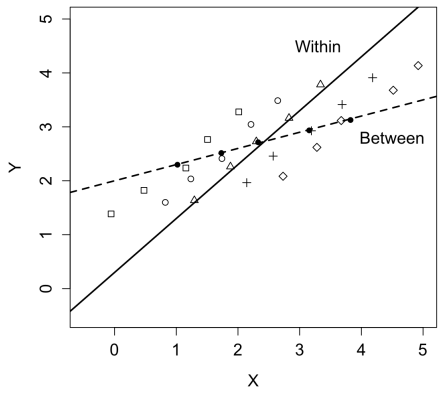
This relationship is similar to that for Pearson correlations in (3.2), which was derived for the additive model (3.1) with infinite cluster sizes.

We provide some toy examples to illustrate the relationship between γ_t, γ_b, and γ_w under different rank ICCs of X and Y (Figure 3.1). Figures 3.1a and 3.1b show examples where γ_b and γ_w are in the opposite or same directions, respectively. If X and Y have moderate rank ICCs of 0.5, then γ_b and γ_w contribute equally

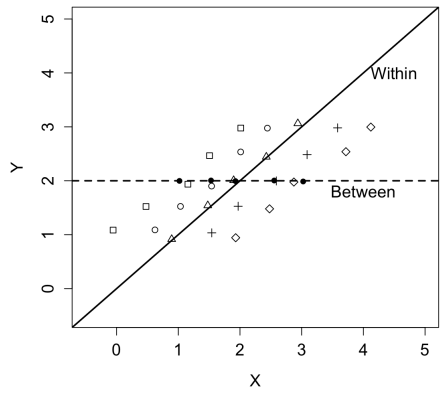
to γ_t, and γ_t is the average of γ_b and γ_w (Figure 3.1a and 3.1e). If the rank ICCs are large, γ_t is dominated by γ_b, while if the rank ICCs are low, γ_t is dominated by γ_w. More extremely, if one of the rank ICCs is close to 1, γ_t is close to $\gamma_b\sqrt{\gamma_{I_X}\gamma_{I_Y}}$ (Figure 3.1d). On the contrary, if one of the rank ICCs is near 0, which means that the observations in a cluster are nearly independent, then γ_t is close to $\gamma_w\sqrt{(1-\gamma_{I_X})(1-\gamma_{I_Y})}$ (Figure 3.1c). When cluster sizes in the population are finite, the rank ICCs can be negative. If any of the rank ICCs is negative, the relationship between γ_t, γ_b, and γ_w is (3.6) rather than the simpler (3.7). Figure 3.1f illustrates an extreme example where the rank ICCs are both -1. (This happens when cluster sizes are two.) In this example, γ_b and γ_w are strong and opposite whereas the total correlation is zero.



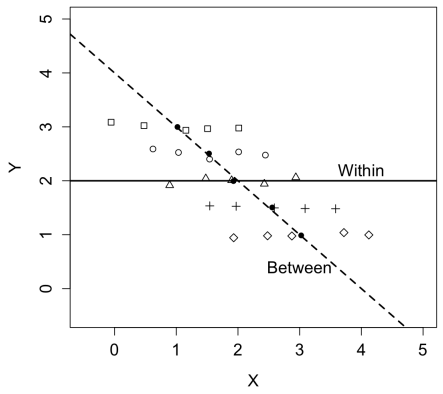
(a) $(\gamma_t, \gamma_b, \gamma_w, \gamma_{I_X}, \gamma_{I_Y}) = (0, -1, 1, 0.5, 0.5)$



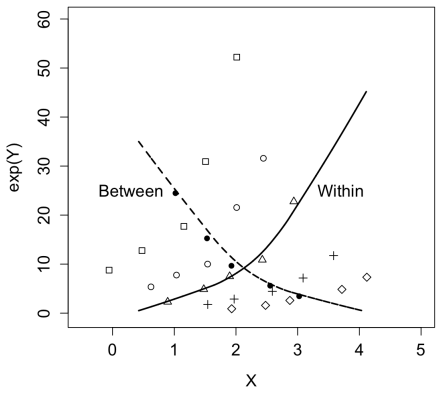
(b) $(\gamma_t, \gamma_b, \gamma_w, \gamma_{I_X}, \gamma_{I_Y}) = (0.84, 1, 1, 0.6, 0.1)$



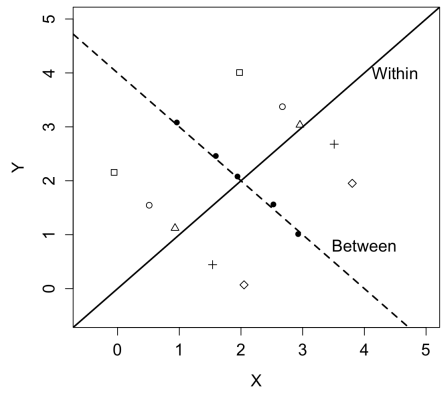
(c) $(\gamma_t, \gamma_b, \gamma_w, \gamma_{I_X}, \gamma_{I_Y}) = (0.71, 0, 1, 0.5, 0)$



(d) $(\gamma_t, \gamma_b, \gamma_w, \gamma_{I_X}, \gamma_{I_Y}) = (-0.71, -1, 0, 0.5, 1)$



(e) $(\gamma_t, \gamma_b, \gamma_w, \gamma_{I_X}, \gamma_{I_Y}) = (0, -1, 1, 0.5, 0.5)$



(f) $(\gamma_t, \gamma_b, \gamma_w, \gamma_{I_X}, \gamma_{I_Y}) = (0, -1, 1, -1, -1)$

Figure 3.1: Toy examples for the relationship between total (γ_t), between-cluster (γ_b), within-cluster (γ_w) Spearman rank correlations and the rank intraclass correlations (γ_{I_X}, γ_{I_Y}). For illustration purposes, we show five clusters and five (two for (f)) observations to represent the distribution of each cluster. Black dots represent cluster medians, and the other symbols represent the five clusters. The solid lines show the direction of the within-cluster correlation and the dashed lines show the direction of the between-cluster correlation.

3.5 Estimation

Since the total Spearman rank correlation is the Pearson correlation between F_X^* and F_Y^*, our estimator of γ_t is $\hat{\gamma}_t = corr(\hat{F}_X^*, \hat{F}_Y^*)$, which is a plug-in estimator. Given two-level data $\{(x_{ij}, y_{ij}) : i = 1,2,...,n, j = 1,2,...,k_i\}$ with a total number of observations of $N = \sum_{i=1}^{n} k_i$, a nonparametric estimator of the CDF of X is $\hat{F}_X(x) = \sum_{i=1}^{n}\sum_{j=1}^{k_i} w_{ij} I(x_{ij} \le x)$, where w_{ij} is the weight of observation x_{ij} and $\sum_{i=1}^{n}\sum_{j=1}^{k_i} w_{ij} = 1$. The weight w_{ij} depends on how we believe the data reflect the composition of the underlying hierarchical distribution; for example, $w_{ij} = 1/(nk_i)$ corresponds to equal weights for clusters and $w_{ij} = 1/N$ corresponds to equal weights for observations (Tu et al., 2023). Similarly, we estimate $\hat{F}_X(x-) = \sum_{i=1}^{n}\sum_{j=1}^{k_i} w_{ij} I(x_{ij} < x)$, and define $\hat{F}_X^*(x) = \{\hat{F}_X(x) + \hat{F}_X(x-)\}/2$. Estimation is the same for F_Y^*. The general form of our estimator of γ_t is

$$\hat{\gamma}_t = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k_i} w_{ij}(\hat{F}_X^*(x_{ij}) - \bar{\hat{F}}_X^*)(\hat{F}_Y^*(y_{ij}) - \bar{\hat{F}}_Y^*)}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{k_i} w_{ij}(\hat{F}_X^*(x_{ij}) - \bar{\hat{F}}_X^*)^2}\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{k_i} w_{ij}(\hat{F}_Y^*(y_{ij}) - \bar{\hat{F}}_Y^*)^2}},$$

where $\bar{\hat{F}}_X^* = \sum_{i=1}^{n}\sum_{j=1}^{k_i} w_{ij}\hat{F}_X^*(x_{ij})$ and $\bar{\hat{F}}_Y^* = \sum_{i=1}^{n}\sum_{j=1}^{k_i} w_{ij}\hat{F}_Y^*(y_{ij})$. If we assign equal weights to clusters (i.e., $w_{ij} = 1/(nk_i)$), our estimator of the total Spearman rank correlation is equal to the estimator of Shih and Fay (2017).

Section 3.3 shows that γ_w is identical to the covariate-adjusted partial Spearman rank correlation and can be expressed in terms of PSRs, suggesting that γ_w can be estimated by sample PSRs (Liu et al., 2017). Hence, our estimator of γ_w is

$$\hat{\gamma}_w = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k_i} w_{ij}(x_{ij,res} - \bar{x}_{res})(y_{ij,res} - \bar{y}_{res})}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{k_i} w_{ij}(x_{ij,res} - \bar{x}_{res})^2}\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{k_i} w_{ij}(y_{ij,res} - \bar{y}_{res})^2}}, \tag{3.8}$$

where $x_{ij,res} = r(x_{ij}, \hat{F}_{X|i})$, $y_{ij,res} = r(y_{ij}, \hat{F}_{Y|i})$, $\bar{x}_{res} = \sum_{i=1}^{n}\sum_{j=1}^{k_i} w_{ij}x_{ij,res}$, $\bar{y}_{res} = \sum_{i=1}^{n}\sum_{j=1}^{k_i} w_{ij}y_{ij,res}$. We can obtain PSRs using nonparametric, parametric, or semiparametric models. A nonparametric estimator of $F_{X|i}$ is simply the empirical CDF of X in cluster i. Estimators from nonparametric models are the most robust but can be inefficient and unstable if cluster sizes are small. Parametric models are the most efficient under correct assumptions but less robust to extreme values, sensitive to model misspecification or outcome transformation, and not congruent with the spirit of Spearman rank correlation. To achieve a compromise between robustness and efficiency, we employ semiparametric models in which only the order information of outcomes is used and the clusters share a common latent variable distribution except for cluster-specific shifts. This way we can borrow information across clusters and still maintain the rank-based nature of Spearman rank correlation.

Specifically, we designate cluster 1 as the reference cluster, and define Z_i ($i = 2,\ldots,n$) as an indicator vari-

able such that $Z_i = 1$ when the observation is in cluster i and $Z_i = 0$ otherwise. We then model X and Y on $Z = (Z_2, \ldots, Z_n)^T$ to obtain PSRs for X and Y, respectively. Here we incorporate the semiparametric linear transformation model where the monotonic transformation, $H_X(\cdot)$, is unspecified, $F_{X|Z}(x) = P\{H_X(\beta_X^T Z + \varepsilon) \leq x|Z\} = F_\varepsilon\{H_X^{-1}(x) - \beta_X^T Z\}$, where ε follows a known distribution and $\beta_X = (\beta_{X2}, \ldots, \beta_{Xn})^T$. The semiparametric transformation model can be written in the form of the ordinal cumulative probability model (CPM), $g_X\{F_{X|Z}(x)\} = \alpha_X(x) - \beta_X^T Z$, where $\alpha_X(x) = H_X^{-1}(x)$ is estimated with a step function, and $g_X(\cdot) = F_\varepsilon^{-1}$ is a link function (Liu et al., 2017). A similar model is fit for Y on Z. Model estimation can be implemented using software for fitting ordinal cumulative probability ("link") models with each unique outcome representing a separate ordinal category. For example, the `orm()` function in the `rms` package of R can be used (Harrell, 2015). After obtaining PSRs from the CPMs of X on Z and of Y on Z, we then simply estimate γ_w as in (3.8).

As mentioned in Section 3.3, we often use the cluster median as the cluster centroid, so γ_b is Spearman rank correlation between cluster medians. One simple estimation approach is to estimate γ_b as Spearman rank correlation between the sample cluster medians (i.e., $\{(\hat{\bar{x}}_i, \hat{\bar{y}}_i) : i = 1, 2, ..., n\}$, where $\hat{\bar{x}}_i$ and $\hat{\bar{y}}_i$ are the medians of $\{x_{i1}, ..., x_{ik_i}\}$ and $\{y_{i1}, ..., y_{ik_i}\}$, respectively). However, this approach only uses information within clusters, which can have high variations with small cluster sizes. Thus, we consider estimating the cluster medians using CPMs of X on Z and of Y on Z. The CPMs borrow information across clusters and their estimates of cluster medians are less variable than the simple estimates. Moreover, for ordered categorical data, the CPMs allow us to obtain cluster medians on the latent variable scale, thus simplifying the estimation of γ_b by eliminating the need to consider all possible combinations of cluster medians on the original scale in the presence of clusters with two medians.

Let us consider a CPM of X on Z, $g_X\{F_{X|Z}(x)\} = \alpha_X(x) - \beta_X^T Z$, where g_X is a symmetric link function such as logit or probit. For any $Z = z$, let x_z be the true median of X given $Z = z$. Since $F_{X|Z=z}(x_z) = 0.5$ and $g_X(0.5) = 0$, we have $0 = \alpha_X(x_z) - \beta_X^T z$ and $\alpha_X(x_z) = \beta_X^T z$. That is, the monotone function α_X transforms the median x_z to $\beta_X^T z$. In the setting of clustered data, $Z = (Z_2, \ldots, Z_n)^T$ is a vector of indicator variables for the clusters, and thus the cluster medians are 0 for cluster 1 and β_{Xi} for cluster i ($i = 2, \ldots, n$). Since α_X is a monotonic increasing transformation, a Spearman rank correlation that involves the cluster medians of X can be computed with $(0, \beta_{X2}, \ldots, \beta_{Xn})^T$. Similarly, a Spearman rank correlation that involves the cluster medians of Y can be computed with $(0, \beta_{Y2}, \ldots, \beta_{Yn})^T$. All these values can be estimated from the CPMs. Thus, our estimator of γ_b is the rank correlation over the n pairs of estimated cluster medians, $\{(0,0), (\hat{\beta}_{X2}, \hat{\beta}_{Y2}), \ldots, (\hat{\beta}_{Xn}, \hat{\beta}_{Yn})\}$. Furthermore, we also consider weighting clusters in the estimation procedures for γ_b. Let $w_{i\cdot}$ denote the weight of cluster i and $w_{i\cdot} = \sum_{j=1}^{k_i} w_{ij}$. A nonparametric estimator of the CDF of β_X is $\hat{F}_{\beta_X}(t) = \sum_{i=1}^n w_{i\cdot} I(\hat{\beta}_{Xi} \leq t)$, similarly $\hat{F}_{\beta_X}(t-) = \sum_{i=1}^n w_{i\cdot} I(\hat{\beta}_{Xi} < t)$, and we define

$\hat{F}^*_{\beta_X}(t) = \{\hat{F}_{\beta_X}(t) + \hat{F}_{\beta_X}(t-)\}/2$. Estimation for $F^*_{\beta_Y}$ is similar. Therefore, one estimator of γ_b is

$$\hat{\gamma}_{b_M} = \frac{\sum_{i=1}^{n} w_{i\cdot}(\hat{F}^*_{\beta_X}(\hat{\beta}_{Xi}) - \bar{\hat{F}}^*_{\beta_X})(\hat{F}^*_{\beta_Y}(\hat{\beta}_{Yi}) - \bar{\hat{F}}^*_{\beta_Y})}{\sqrt{\sum_{i=1}^{n} w_{i\cdot}(\hat{F}^*_{\beta_X}(\hat{\beta}_{Xi}) - \bar{\hat{F}}^*_{\beta_X})^2}\sqrt{\sum_{i=1}^{n} w_{i\cdot}(\hat{F}^*_{\beta_Y}(\hat{\beta}_{Yi}) - \bar{\hat{F}}^*_{\beta_Y})^2}},$$

where $\hat{\beta}_{X1} = \hat{\beta}_{Y1} = 0$, $\bar{\hat{F}}^*_{\beta_X} = \sum_{i=1}^{n} w_{i\cdot}\hat{F}^*_{\beta_X}(\hat{\beta}_{Xi})$, and $\bar{\hat{F}}^*_{\beta_Y} = \sum_{i=1}^{n} w_{i\cdot}\hat{F}^*_{\beta_Y}(\hat{\beta}_{Yi})$. When $w_{i\cdot} = 1/n$, $\bar{\hat{F}}^*_{\beta_X} = \bar{\hat{F}}^*_{\beta_Y} = 1/2$.

If cluster sizes are very small, the estimates of β_X and β_Y, and thus $\hat{\gamma}_{b_M}$, may be poor. We consider another estimation approach. As shown in Section 3.4 equation (3.6), γ_t is approximated by a weighted sum of γ_w and γ_b, where the weights are functions of γ_{I_X} and γ_{I_Y}. We can use this relationship to obtain an estimate of γ_b,

$$\hat{\gamma}_{b_A} = \frac{\hat{\gamma}_t - \sqrt{(1-\hat{\gamma}_{I_X} + \hat{D}_X)(1-\hat{\gamma}_{I_Y} + \hat{D}_Y)}\hat{\gamma}_w}{\sqrt{(\hat{\gamma}_{I_X} - \hat{D}_X)(\hat{\gamma}_{I_Y} - \hat{D}_Y)}},$$

where $\hat{\gamma}_{I_X}$ and $\hat{\gamma}_{I_Y}$ are nonparametric estimators of γ_{I_X} and γ_{I_X} (Tu et al., 2023), $\hat{D}_X = \frac{\sum_{i=1}^{n} w_{i\cdot}\sum_{j<j'}\frac{2}{k_i(k_i-1)}[\hat{F}^*(x_{ij}) - \bar{F}^*_i][\hat{F}^*(x_{ij'}) - \bar{F}^*_i]}{\sum_{i=1}^{n}\sum_{j=1}^{k_i} w_{ij}[\hat{F}^*(x_{ij}) - \bar{F}^*]^2}$, $\bar{F}^*_i = \sum_j F^*(x_{ij})/k_i$, and similar for $\hat{D}_Y$. If the cluster size in the population is infinite, $D_X = D_Y = 0$, then $\hat{\gamma}_{b_A} = \frac{\hat{\gamma}_t - \sqrt{(1-\hat{\gamma}_{I_X})(1-\hat{\gamma}_{I_Y})}\hat{\gamma}_w}{\sqrt{\hat{\gamma}_{I_X}\hat{\gamma}_{I_Y}}}$. Note that $\hat{\gamma}_{b_A}$ can be greater than 1 or less than -1; in those cases, we define $\hat{\gamma}_{b_A}$ to be 1 or -1, respectively. When cluster sizes are very small, $\hat{\gamma}_{b_A}$ may be preferable over $\hat{\gamma}_{b_M}$. If either of the rank ICCs is very small, $\sqrt{\hat{\gamma}_{I_X}\hat{\gamma}_{I_Y}} \approx 0$ and $\hat{\gamma}_{b_A}$ can be unstable.

3.6 Inference

The large sample distribution of $\hat{\gamma}_w$ can be obtained by bootstrapping or large sample approximation. Here we focus on the large sample approach using M-estimation (Stefanski and Boos, 2002). The CPM is fit by minimizing the multinomial/nonparametric likelihood, and then the variance of parameter estimates can be estimated using a sandwich variance estimator that accounts for clustering. This is equivalent to fitting generalized estimating equation (GEE) methods for ordinal response variables with independence working correlation (Tian et al., 2023). Let $\psi_X(\cdot) = \mathbf{U}_X(\theta)$ denote the estimating function for the CPM of X on Z with a vector of parameters θ_X, and $\psi_Y(\cdot) = \mathbf{U}_Y(\theta)$ denote the estimating function for the CPM of Y on Z with a vector of parameters θ_Y. See the Supplementary Materials for details about these estimating functions. The components necessary for computing γ_w are denoted by θ_{w1}, θ_{w2}, θ_{w3}, θ_{w4}, and θ_{w5} such that $\gamma_w = (\theta_{w3} - \theta_{w1}\theta_{w2})/\sqrt{(\theta_{w4} - \theta^2_{w1})(\theta_{w5} - \theta^2_{w2})}$, where $\theta_{w1} = E(X_{ij,res})$, $\theta_{w2} = E(Y_{ij,res})$, $\theta_{w3} = E(X_{ij,res}Y_{ij,res})$, $\theta_{w4} = E(X^2_{ij,res})$, $\theta_{w5} = E(Y^2_{ij,res})$. We can stack $\psi_X(\cdot)$ and $\psi_Y(\cdot)$ together with these components and then

have the following estimating function,

$$\begin{aligned}
&\psi_w(\mathbf{X}_i, \mathbf{Y}_i, \mathbf{Z}_i, \theta_w) \\
&= \{\psi_X(\mathbf{X}_i, \mathbf{Z}_i, \theta_X), \psi_Y(\mathbf{Y}_i, \mathbf{Z}_i, \theta_Y), I_i^T \mathbf{X}_{i,res}/k_i - \theta_{w1}, I_i^T \mathbf{Y}_{i,res}/k_i - \theta_{w2}, \\
&\mathbf{X}_{i,res}^T \mathbf{Y}_{i,res}/k_i - \theta_{w3}, \mathbf{X}_{i,res}^T \mathbf{X}_{i,res}/k_i - \theta_{w4}, \mathbf{Y}_{i,res}^T \mathbf{Y}_{i,res}/k_i - \theta_{w5}\}^T,
\end{aligned}$$

where $\theta_w = (\theta_X, \theta_Y, \theta_{w1}, \theta_{w2}, \theta_{w3}, \theta_{w4}, \theta_{w5})$, I_i is a vector of ones with a length of k_i. The estimating equations are $\sum_{i=1}^n \psi_w(\mathbf{X}_i, \mathbf{Y}_i, \mathbf{Z}_i; \hat{\theta}_w) = 0$. Under standard regularity conditions (Stefanski and Boos, 2002), then we have $\sqrt{n}(\hat{\theta}_w - \theta_w) \xrightarrow{d} MVN(\mathbf{0}, \mathbf{V}(\theta_w))$, where $\mathbf{V}(\theta_w) = \mathbf{A}(\theta_w)^{-1}\mathbf{B}(\theta_w)\{\mathbf{A}(\theta_w)^{-1}\}^T$, $\mathbf{A}(\theta_w) = E[-\partial\psi_w(\mathbf{X}_i, \mathbf{Y}_i, \mathbf{Z}_i, \theta_w)/\partial\theta_w^T]$, and $\mathbf{B}(\theta_w) = E[\psi_w(\mathbf{X}_i, \mathbf{Y}_i, \mathbf{Z}_i, \theta_w)\psi_w(\mathbf{X}_i, \mathbf{Y}_i, \mathbf{Z}_i, \theta_w)^T]$. Since our estimator of γ_w is a function of $\hat{\theta}_{w1}$, $\hat{\theta}_{w2}$, $\hat{\theta}_{w3}$, $\hat{\theta}_{w4}$, and $\hat{\theta}_{w5}$, the delta method can be used to obtain its large sample distribution. Then we can compute the asymptotic standard error (SE) of γ_w and construct confidence intervals (CIs) for γ_w.

We use a similar approach to obtain the large sample distribution of $\hat{\gamma}_{b_M}$. Let $\beta_X = (0, \beta_{X2}, ..., \beta_{Xn})^T$ and $\beta_Y = (0, \beta_{Y2}, ..., \beta_{Yn})^T$ denote the coefficients of cluster index in the CPMs of X and Y, respectively. Note that the coefficient of the reference cluster is zero. To obtain the asymptotic variance of γ_b, we treat β_X and β_Y as random effects, for simplicity assuming that $\beta_{Xi} \overset{i.i.d}{\sim} N(\mu_{\beta_X}, \sigma^2_{\beta_X})$ and $\beta_{Yi} \overset{i.i.d}{\sim} N(\mu_{\beta_Y}, \sigma^2_{\beta_Y})$. The components necessary for computing γ_b are denoted by θ_{b1}, θ_{b2}, θ_{b3}, θ_{b4}, and θ_{b5} such that $\gamma_b = (\theta_{b3} - \theta_{b1}\theta_{b2})/\sqrt{(\theta_{b4} - \theta_{b1}^2)(\theta_{b5} - \theta_{b2}^2)}$, where $\theta_{b1} = E[F_{\beta_X}(\beta_{Xi})]$, $\theta_{b2} = E[F_{\beta_Y}(\beta_{Yi})]$, $\theta_{b3} = E[F_{\beta_X}(\beta_{Xi})F_{\beta_Y}(\beta_{Yi})]$, $\theta_{b4} = E\{[F_{\beta_X}(\beta_{Xi})]^2\}$, $\theta_{b5} = E\{[F^2_{\beta_Y}(\beta_{Yi})]^2\}$, and F_{β_X} and F_{β_Y} are the CDFs of normal distributions. Note that $E[F_{\beta_X}(\beta_{Xi})] = E[F_{\beta_X}(\beta_{Yi})] = 1/2$ in theory but they may not be $1/2$ in estimation if $w_{i\cdot} \neq 1/n$. Similar to the inference procedure of γ_w above, we stack $\psi_X(\cdot)$ and $\psi_Y(\cdot)$ with the components needed to compute γ_b stacked together, yielding the following estimating function,

$$\begin{aligned}
&\psi_b(\mathbf{X}_i, \mathbf{Y}_i, \mathbf{Z}_i, \theta_b) \\
&= \{\psi_X(\mathbf{X}_i, \mathbf{Z}_i, \theta_X), \psi_Y(\mathbf{Y}_i, \mathbf{Z}_i, \theta_Y), \beta_{Xi} - \mu_{\beta_X}, \beta_{Yi} - \mu_{\beta_Y}, \beta_{Xi}^2 - M_{\beta_X}, \beta_{Yi}^2 - M_{\beta_Y}, \\
&F_{\beta_X}(\beta_{Xi}) - \theta_{b1}, F_{\beta_Y}(\beta_{Yi}) - \theta_{b2}, F_{\beta_X}(\beta_{Xi})F_{\beta_Y}(\beta_{Yi}) - \theta_{b3}, [F_{\beta_X}(\beta_{Xi})]^2 - \theta_{b4}, [F_{\beta_X}(\beta_{Yi})]^2 - \theta_{b5}\}^T,
\end{aligned}$$

where $\theta_b = (\theta_X, \theta_Y, \mu_{\beta_X}, \mu_{\beta_Y}, M_{\beta_X}, M_{\beta_Y}, \theta_{b1}, \theta_{b2}, \theta_{b3}, \theta_{b4}, \theta_{b5})$, $M_{\beta_X} = E(\beta_X^2) = \mu^2_{\beta_X} + \sigma^2_{\beta_X}$, and $M_{\beta_Y} = E(\beta_Y^2) = \mu^2_{\beta_Y} + \sigma^2_{\beta_Y}$. The estimating equations are $\sum_{i=1}^n \psi_b(\mathbf{X}_i, \mathbf{Y}_i, \mathbf{Z}_i; \hat{\theta}_b) = 0$. We have $\sqrt{n}(\hat{\theta}_b - \theta_b) \xrightarrow{d} MVN(\mathbf{0}, \mathbf{V}(\theta_b))$ under standard regularity conditions (Stefanski and Boos, 2002), where $\mathbf{V}(\theta_b) = \mathbf{A}(\theta_b)^{-1}\mathbf{B}(\theta_b)\{\mathbf{A}(\theta_b)^{-1}\}^T$, $\mathbf{A}(\theta_b) = E[-\partial\psi_b(\mathbf{X}_i, \mathbf{Y}_i, \mathbf{Z}_i, \theta)/\partial\theta_b^T]$, and $\mathbf{B}(\theta) = E[\psi_b(\mathbf{X}_i, \mathbf{Y}_i, \mathbf{Z}_i, \theta_b)\psi_b(\mathbf{X}_i, \mathbf{Y}_i, \mathbf{Z}_i, \theta_b)^T]$. The large sample distribution of $\hat{\gamma}_{b_M}$ can be derived from the large sample distribution of $\hat{\theta}_b$ using the delta method. We

also use this expression for inference for $\hat{\gamma}_{b_A}$.

As mentioned in Section 3.5, if the same weight is assigned to all the observations, our estimator of the total Spearman rank correlation $\hat{\gamma}_t$ equals the estimator of Shih and Fay (2017). Hence, we adapt the inference method of Shih and Fay (2017) via incorporating weighting into the estimation procedures to obtain the asymptotic variance of $\hat{\gamma}_t$. Shih and Fay (2017) have provided an analytical form for the asymptotic distribution of the estimator of γ_t, which is a function of X_{ij}, Y_{ij}, F_X, F_Y, and F_{XY}. The asymptotic variance can be estimated with $\hat{F}_X$, $\hat{F}_Y$, and $\hat{F}_{XY}$ plugged in for F_X, F_Y, and F_{XY}. Here we allow $\hat{F}_X$, $\hat{F}_Y$, and $\hat{F}_{XY}$ to be obtained based on either assigning equal weights to observations or assigning equal weights to clusters, and then plug them in to estimate the asymptotic variance of $\hat{\gamma}_t$.

3.7 Simulations

We used a bivariate additive model for data generation: $\begin{pmatrix} X_{0ij} \\ Y_{0ij} \end{pmatrix} = \begin{pmatrix} U_{Xi} \\ U_{Yi} \end{pmatrix} + \begin{pmatrix} R_{Xij} \\ R_{Yij} \end{pmatrix}$, where $\begin{pmatrix} U_{Xi} \\ U_{Yi} \end{pmatrix} \overset{i.i.d}{\sim} N\left(\begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{0b} \\ \rho_{0b} & 1 \end{pmatrix}\right)$, $\begin{pmatrix} R_{Xij} \\ R_{Yij} \end{pmatrix} \overset{i.i.d}{\sim} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{0w} \\ \rho_{0w} & 1 \end{pmatrix}\right)$. Here, $\rho_{0t} = (\rho_{0b} + \rho_{0w})/2$. Let (X_{ij}, Y_{ij}) be the observation of the jth individual in the ith cluster, where $i = 1, 2, ..., n$; $j = 1, 2, ..., k_i$; and k_i is the size of the ith cluster. We considered three scenarios: (I) $X_{ij} = X_{0ij}$ and $Y_{ij} = Y_{0ij}$; (II) $X_{ij} = X_{0ij}$ and $Y_{ij} = \exp(Y_{0ij})$; (III) $X_{ij} = \exp(U_{Xi}) + R_{Xij}$ and $Y_{ij} = \exp(\exp(U_{Yi}) + R_{Yij})$. Under Scenarios I and II, since $(X_{0ij}, Y_{0ij})^T$ is bivariate normal, the true total, between-, and within-cluster Spearman rank correlations are $\gamma_t = 6\arcsin(\rho_{0t}/2)/\pi$, $\gamma_b = 6\arcsin(\rho_{0b}/2)/\pi$, and $\gamma_w = 6\arcsin(\rho_{0w}/2)/\pi$ (Pearson, 1907). Under Scenario III, γ_b and γ_w are the same as those in Scenarios I and II, but γ_t is different because $(X_{ij}, Y_{ij})^T$ is not normally distributed. We empirically computed γ_t under Scenario III by generating one million clusters each with 100 observations, and then computing γ_t. We also empirically computed the total, between- and within-cluster Pearson correlations (i.e., ρ_t, ρ_b, and ρ_w) under Scenarios II and III. While γ_b and γ_w are identical under the three scenarios, ρ_b and ρ_w are sensitive to skewness and depend on the scale of interest (Table 3.1).

Table 3.1: The total, between-cluster, and within-cluster Spearman rank correlations $(\gamma_t, \gamma_b, \gamma_w)$ and Pearson correlations (ρ_t, ρ_b, ρ_w) under Scenarios I (normality), II (exponentiated Y), and Scenario III (exponentiated cluster means and exponentiated Y) with 5 simulation settings

$(\rho_{0t}, \rho_{0b}, \rho_{0w})$	$(\gamma_t, \gamma_b, \gamma_w)$		(ρ_t, ρ_b, ρ_w)		
I, II, III	I, II	III	I	II	III
(0.75, 0.80, 0.70)	(0.73, 0.79, 0.68)	(0.53, 0.79, 0.68)	(0.75, 0.80, 0.70)	(0.42, 0.61, 0.33)	(0.02, 0.03, 0)
(0.40, 0.80, 0)	(0.38, 0.79, 0)	(0.31, 0.79, 0)	(0.40, 0.80, 0)	(0.22, 0.06, 0)	(0.01, 0.02, 0)
(0.40, 0, 0.80)	(0.38, 0, 0.79)	(0.25, 0, 0.79)	(0.40, 0, 0.80)	(0.22, 0.01, 0.37)	(0, 0, 0)
(0, 0, 0)	(0, 0, 0)	(0, 0, 0)	(0, 0, 0)	(0, 0, 0)	(0, 0, 0)
(0.05, 0.80, -0.7)	(0.05, 0.79, -0.68)	(0.09, 0.79, -0.68)	(0.05, 0.80, -0.70)	(0.03, 0.59, -0.32)	(0.01, 0.01, 0)

We first evaluated the performance of our estimators of γ_b, γ_w, and γ_t for continuous data. The simulations were conducted under Scenarios I, II, and III at $n = 100$, and $(\rho_{0b}, \rho_{0w}) \in \{(0.8, 0.7), (0.8, 0), (0, 0.8), (0, 0), (0.8, -0.7)\}$. In Scenarios I and II, the rank ICCs of X and Y are both 0.48. In Scenario III, the rank ICC of X is 0.97 while that of Y is 0.37. We considered various configurations of cluster size under Scenarios I and II: $k_i = 10$, $k_i = 20$, $k_i = 30$, and k_i uniformly ranging from 1 to 50. We compared our estimators with naive nonparametric estimators: $\hat{\gamma}_{b_n}$ estimated by Spearman rank correlation between sample cluster medians and $\hat{\gamma}_{w_n}$ estimated by the rank correlation of within-cluster deviations (differences) from sample cluster medians. Furthermore, we compared the Spearman rank correlations with the Pearson correlations in Scenario I. The estimators of ρ_b and ρ_w are based on one-way random effects models: ρ_b is estimated by Pearson correlation between the estimated cluster means from the random effects models and ρ_w is estimated by Pearson correlation between the individual deviations from the estimated cluster means (Snijders and Bosker, 1999).

In general, our estimators of γ_b, γ_w, and γ_t had low bias and good coverage with modest numbers of clusters in Scenarios I and II (Figure 3.2). They were also robust to the skewed data in Scenario II and they had lower bias than $\hat{\gamma}_{b_n}$ and $\hat{\gamma}_{w_n}$ (Table 3.3). In the extreme case where γ_b and γ_w are both strong but opposite (i.e., last row of Figure 2), our estimators of γ_b were biased. In the other settings where γ_b and γ_w greatly differed (i.e., rows 2-3 of Figure 2), the estimators of γ_b were also biased, although to a lesser extent. In these settings, the bias of $\hat{\gamma}_{b_A}$ was relatively smaller than that of $\hat{\gamma}_{b_M}$, particularly with small cluster sizes. As the cluster size increased, the bias of $\hat{\gamma}_{b_M}$ decreased, whereas the bias of $\hat{\gamma}_{b_A}$ remained relatively stable (also seen in Table 3.4). It is worth noting that in the extreme case (last row of Figure 2), the estimator of the between-cluster Pearson correlation based on random effects models also had similar bias, even when the data were normally distributed (Table 3.5).

In Scenario III, our estimator of γ_w still had low bias and good coverage (Table 3.6). In our setup, $E[U_{Yi}] = -1$, our estimators of γ_b and γ_t had more bias under Scenario III than under Scenarios I and II. This is because U_{Yi} was exponentiated in Scenario III, which led to cluster means that had a much smaller variance than that of the within-cluster deviations. In this setting, the within-cluster deviation often dominated the value of Y_{ij} creating data where it is difficult to see the effect of clustering over the within-cluster variance. Our estimator of γ_t struggled in this setting, producing biased estimates of γ_t and thus biased estimates of γ_b based on the approximation (3.7). In addition, estimation of γ_b using $\hat{\gamma}_{b_M}$ also was biased, as estimated cluster medians, even with fairly larger cluster sizes, often were far from their true rankings due to the large residual noise. When $E[U_{Yi}]$ was changed from -1 to 1, the cluster means had a larger variance than that of the within-cluster deviations, leading to much smaller bias in the estimates of γ_t and γ_b.

We then evaluated the performance of our estimators when the rank ICC was negative, which occurs when

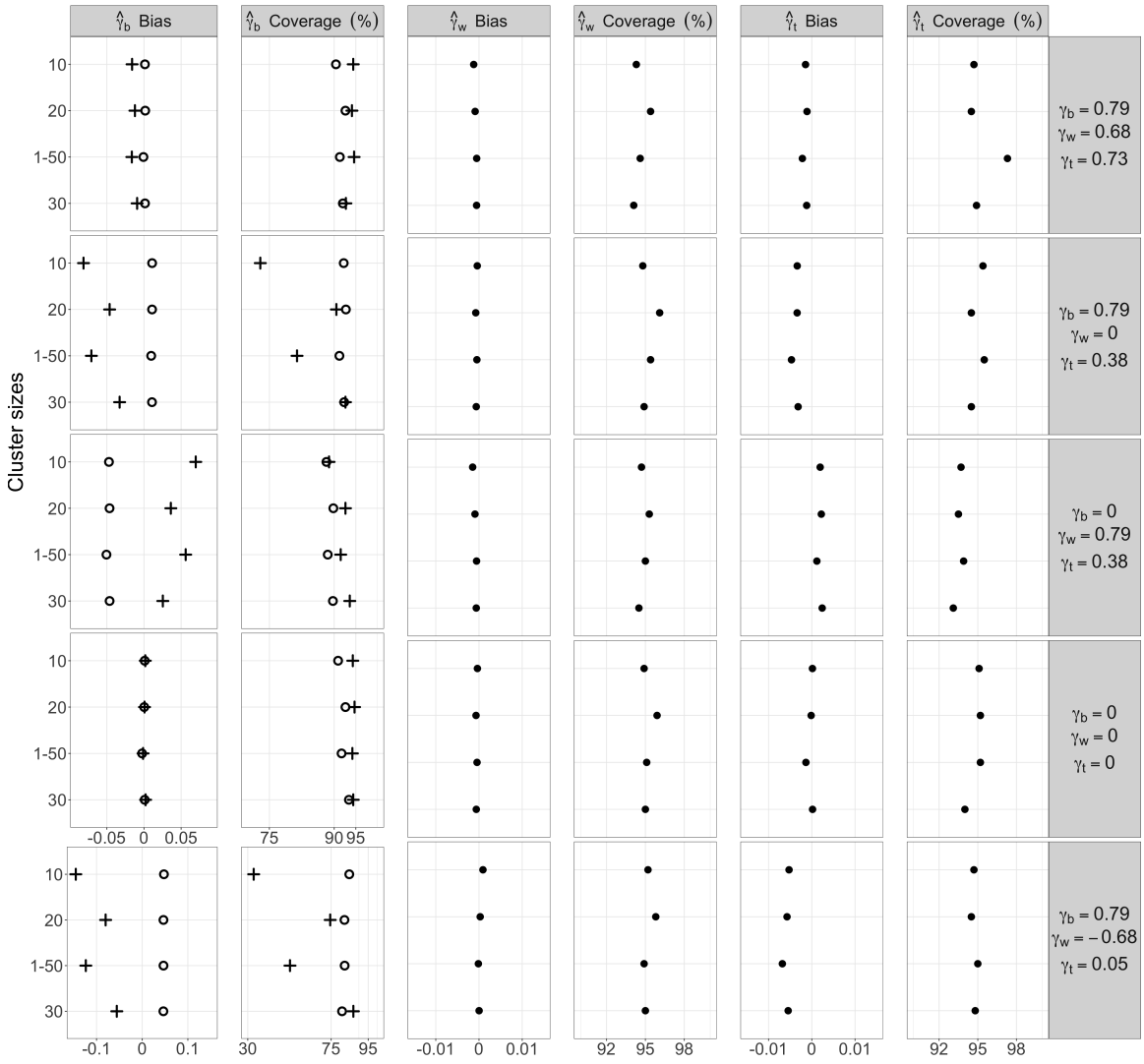


Figure 3.2: Bias and coverage of 95% CIs for our estimators of γ_b, γ_w, and γ_t at different true values and different cluster sizes under Scenarios I (normality) and II (exponentiated Y). The circle sign stands for the approximation-based estimator ($\hat{\gamma}_{b_A}$) of γ_b, and the plus sign stands for the cluster-median-based estimator ($\hat{\gamma}_{b_M}$) of γ_b. The number of clusters was set at 100. "1-50" means the cluster size follows a uniform distribution from 1 to 50.

$k_i = 2$. Our estimators of γ_t, γ_b, and γ_w had very low bias and good coverage. Details are in the Supplementary Materials Table 3.2.

Furthermore, we investigated the performance of our estimators when the link function of the CPM was misspecified as logit, loglog, and cloglog under Scenario II. We conducted 1000 simulations at $n = 100$ and $k_i = 20$. Our estimators of γ_b, γ_w, and γ_t performed similarly under the logit link as they did under the correct probit link function (Table 3.7). When the link function was misspecified as loglog or cloglog, if γ_w was large and had the opposite direction of γ_b, our estimator of γ_w had bias toward the direction of γ_b.

We also evaluated the performance of our estimators for ordered categorical data. We simulated 5-level and 10-level ordered categorical data by discretizing X_{ij} and Y_{ij} in Scenario I with cutoffs at quantiles (i.e., using the 0.2, 0.4, 0.6, 0.8 quantiles for 5 levels; and the 0.1, 0.2, ..., 0.8, 0.9 quantiles for 10 levels). We empirically computed γ_t, γ_b, and γ_w by generating one million clusters and 100 observations per cluster, with cluster medians analytically derived. The values of γ_t, γ_b, and γ_w of the 10-level ordered categorical variables are close to those of the continuous variables, while those of the 5-level ordered categorical variables are slightly smaller (Table 3.8). We conducted 1000 simulations at $n = 100$ and $k_i = 20$. Our estimators of γ_b, γ_w, and γ_t had very low bias and good coverage (Table 3.8). When γ_b was large, $\hat{\gamma}_{b_A}$ had bias, which might be due to equation (3.7) being a poor approximation of γ_t with ordered categorical data. This bias decreased as the number of ordered categories increased.

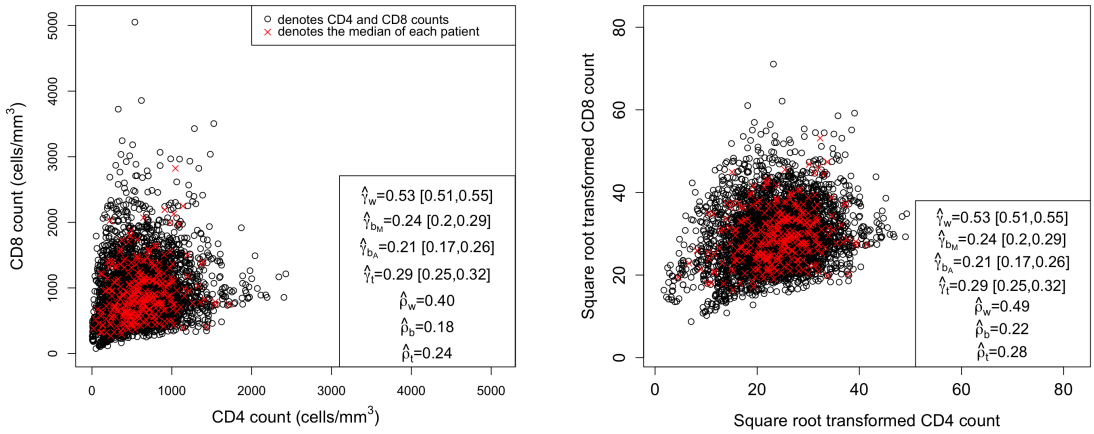
3.8 Applications

3.8.1 Longitudinal biomarker data

Repeated measures of CD4 and CD8 lymphocyte counts (cells/mm^3) were taken on 325 women living with HIV who started antiretroviral therapy (ART) at the Vanderbilt Comprehensive Care Clinic between 1998 and 2012 (Castilho et al., 2016). There is interest in evaluating the correlation between same-day CD4 and CD8 counts while considering the potential clustering in the data. All same-day CD4 and CD8 measurements taken within $\pm$4 months of ART initiation were included in analyses; the number of observations per woman ranged from 1 to 54. In this case, the cluster is the person, so it makes sense to assign equal weights to people rather than measurements. The data were very skewed, especially the CD8 count (Figure 3.3).

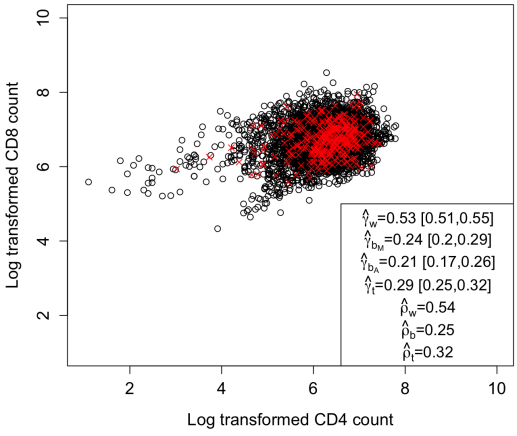
The rank ICC estimates of CD4 and CD8 counts were 0.77 and 0.76, respectively, suggesting strong similarity between measurements from the same woman. The between-cluster Spearman rank correlation was estimated to be $\hat{\gamma}_{b_M} = 0.24$ (95% CI: [0.20,0.29]) via cluster medians obtained from CPMs and was estimated to be $\hat{\gamma}_{b_A} = 0.21$ (95% CI: [0.17,0.26]) via the approximation approach, indicating a weak but positive correlation between median CD4 and CD8 counts (Figure 3.3). The Spearman rank correlation between the sample cluster medians was 0.24, close to our between-cluster Spearman rank correlation estimates. The within-cluster Spearman rank correlation estimate was 0.53 (95%: [0.51,0.55]), suggesting moderate correlation between the fluctuations in the repeated CD4 and CD8 measurements. The total Spearman rank correlation estimate, 0.29 (95% CI: [0.25,0.32]), suggests a weak to moderate overall correlation after combining between-cluster and within-cluster correlations.

The between-cluster, within-cluster, and total Pearson correlation estimates on the original scale obtained from a random effects model were estimated to be 0.18, 0.40, and 0.24, respectively, which were impacted by some extreme measurements. The three Pearson correlations were estimated to be 0.22, 0.49, and 0.28,



(a) Original scale

(b) Square-root transformation



(c) Log transformatio

Figure 3.3: Scatter plot of CD4 and CD8 counts (cells/mm^3) and estimates of total, between-, and within-cluster Spearman rank (γ_t, γ_b, γ_w, [95% CIs]) and Pearson (ρ_t, ρ_b, ρ_w) correlations. The red cross sign represents the sample cluster median and the circle sign represents the observation. $\hat{\gamma}_{b_M}$ is the estimator of γ_b based on cluster medians obtained from CPM and $\hat{\gamma}_{b_A}$ is the estimator of γ_b based on the linear approximation.

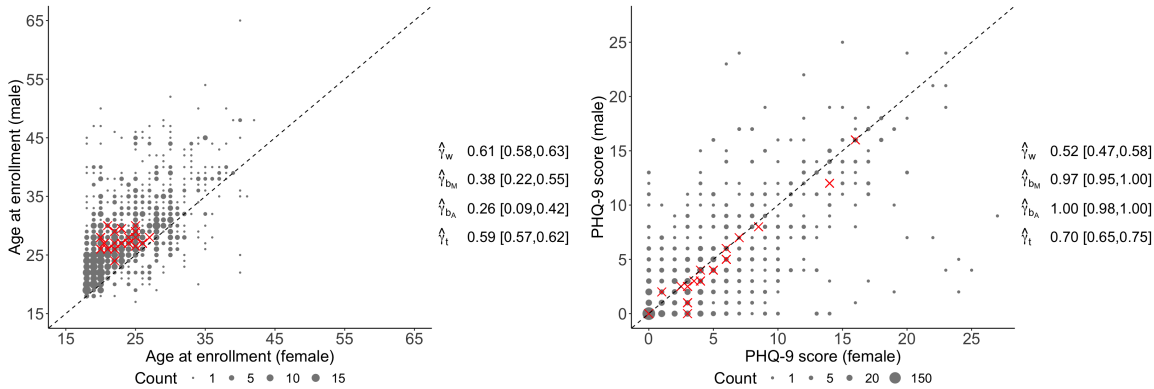
respectively, after square root transformation, and 0.25, 0.54, and 0.32, respectively, after log transformation. The notable differences in the three Pearson correlation estimates after data transformation demonstrate the sensitivity of Pearson correlation to the choice of scale. In contrast, our estimates of between-cluster, within-cluster, and total Spearman rank correlations are invariant to any monotonic transformation.

3.8.2 Cluster randomized controlled trial data

The Homens para Saúde Mais (HoPS+) study is a cluster randomized controlled trial in Zambézia Province, Mozambique (Audet et al., 2018). The trial was designed to measure the impact of incorporating male

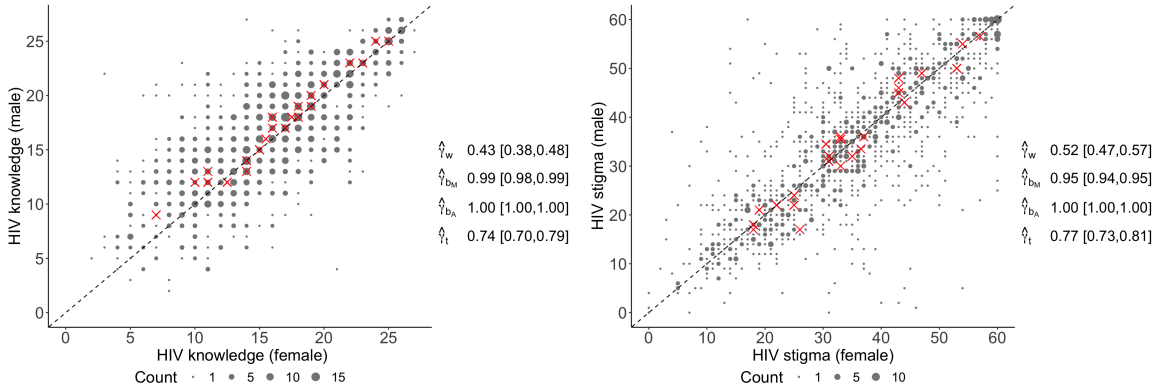
partners with HIV into prenatal care for pregnant women living with HIV on adherence to treatment. The trial enrolled 1073 participating couples living with HIV at 24 clinical sites. The number of couples at clinical sites ranged from 15 to 71. At the time of randomization (baseline), age, depressive symptoms measured by Patient Health Questionnaire-9 (PHQ-9) score, HIV knowledge, and HIV stigma were captured. We are interested in the correlation of these baseline measures within couples. We are also interested in the correlation of 12-month adherence to ART within couples. Figure 3.4 shows scatter plots of these measures.

In this example, the cluster is the clinical site and the observations are made on couples (e.g., X = age of female partners, Y = age of male partners). Hence, it is reasonable to assign equal weights to couples. The estimates of the total, between-cluster, and within-cluster Spearman rank correlations are shown in Figure 3.4. The total Spearman rank correlation for age, 0.59, was moderate to strong. The between-cluster Spearman rank correlation for age was $\hat{\gamma}_{b_M} = 0.38$, suggesting weak to moderate correlation between median male and female ages within clinical sites. The within-cluster Spearman rank correlation was 0.61, implying that after controlling for clinical site, the correlation of age between couples remained high. For PHQ-9 scores, HIV knowledge, and HIV stigma, the total Spearman rank correlation between couples was strong, ranging from 0.70 to 0.77. The correlation became moderate after controlling for clinical sites, with $\hat{\gamma}_w$ varying from 0.43 to 0.52. The between-cluster Spearman rank correlations of the three measures were extremely strong, which can be seen in Figure 3.4. The approximation-based estimates ($\hat{\gamma}_{b_A}$) of the between-cluster correlation hit the boundary and were thus set to be 1, and the cluster-median-based estimates ($\hat{\gamma}_{b_M}$) were close to 1. Taken as a whole, these estimates suggest that the scores between male and female partners for these measures are highly correlated but that some of the correlation is due to similarities within sites. This may reflect differences between participants across sites or perhaps differences in the ways the questionnaires were administrated across study sites. Finally, the total correlation for 12-month adherence was moderate, 0.47. After controlling for clinical sites, the correlation remained moderate, $\hat{\gamma}_w = 0.46$. The between-cluster correlation was moderate to high, $\hat{\gamma}_{b_M} = 0.40$ and $\hat{\gamma}_{b_A} = 0.61$; this difference might be due to the small intraclass correlation for this variable (the rank ICC of 12-month adherence for males was 0.06 and for females was 0.07).



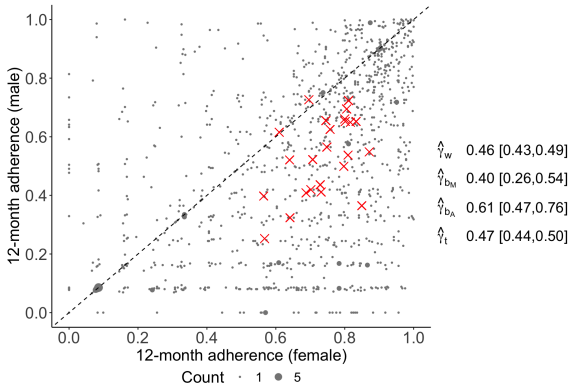
(a) Age at enrollment

(b) PHQ-9 scores



(c) HIV knowledge

(d) HIV stigma



(e) 12-month adherence

Figure 3.4: Scatter plots of PHQ-9 scores, age at enrollment (years), HIV knowledge, HIV stigma, and 12-month adherence (%) of female and male partners enrolled in the clustered randomized clinical trial. The red cross sign represents the sample cluster median and the dot sign represents the observation. The right side of each subfigure shows the estimates of total, between-, and within-cluster Spearman rank correlations (γ_t, γ_b, γ_w, [95% CIs]). $\hat{\gamma}_{b_M}$ is the estimator of γ_b based on cluster medians obtained from CPM and $\hat{\gamma}_{b_A}$ is the estimator of γ_b based on the linear approximation.

3.9 Discussion

In this Chapter, we defined the population parameters of the between- and within-cluster Spearman rank correlations for clustered data, which are natural extensions of the between- and within-cluster Pearson correlations to the rank scale. We also approximated their relationship with the total Spearman rank correlation and the rank intraclass correlation coefficient. Compared with the traditional Pearson correlation, our method is insensitive to extreme values and skewed distributions, and does not depend on the scale of the data. Our framework is general, and is applicable to any orderable variables. Our estimators are asymptotically normal, with generally low bias and good coverage in our simulations. We have developed an R package, `rankCorr`, available on CRAN, which implements our new method.

Our method has some limitations. Our method requires fitting models for the conditional distributions of X and Y given cluster index, which need to be approximately correct to get unbiased estimates. We suggest using semiparametric cumulative probability models which maintain the rank-based nature of Spearman's rank correlation. In addition, our estimator of the between-cluster Spearman rank correlation has bias when the between- and within-cluster Spearman rank correlations are in opposite directions. This problem also exists when estimating the between-cluster Pearson correlation. As the cluster size increases, the problem goes away.

In practice, one may be interested in estimating covariate-adjusted rank correlations. For example, in the application of CD4 and CD8 data, there may be interest in measuring the rank correlations after adjusting for age. The methods in this manuscript could be extended to allow for covariate adjustment by fitting CPMs that include the covariate, in addition to the cluster indicators. We suspect that the correlation between probability-scale residuals from these fitted models could be used to estimate covariate-adjusted within-cluster rank correlations and that the correlation between cluster indicator coefficients from these fitted models could be used to estimate covariate-adjusted between-cluster rank correlations. This approach is somewhat similar to random effects approaches used for estimating covariate-adjusted within- and between-cluster Pearson correlations (Ferrari et al., 2005). Such an approach, as well as Spearman rank correlation as a function of time with longitudinal data, warrants further investigation.

3.10 Supplementary Materials

3.10.1 Estimating functions of the CPMs

Let C denote the number of distinct values of X, $O_{ij,c} = I(X_{ij} \leq x_{(c)})$, and $\mu_{ij,c} = E[O_{ij,c}|\mathbf{Z}_{ij}] = P(X_{ij} \leq x_{(c)}|\mathbf{Z}_{ij})$. Then $\mathbf{O}_{ij} = (O_{ij,1},...,O_{ij,C-1})^T$, and $\mu_{ij} = (\mu_{ij,1},...,\mu_{ij,C-1})^T$. We define $\mathbf{O}_i = (\mathbf{O}_{ij}^T,...,\mathbf{O}_{ik_i}^T)^T$ and $\mu_i = (\mu_{i1}^T,...,\mu_{ik_i}^T)^T$, which are both vectors with a length of $(C-1)k_i$. Let $\pi_{ij,c} = \mu_{ij,c} - \mu_{ij,c-1} = E(I(X_{ij} = x_{(c)})|\mathbf{Z}_{ij})$ and $\pi_{ij} = (\pi_{ij,1},...,\pi_{ij,C-1})^T$. The CPM of X on Z has parameters $\theta_X = (\alpha_X^T, \beta_X^T)^T$,

where $\alpha_X = (\alpha_{X,1}, ..., \alpha_{X,C-1})$. The estimating function of the CPM of X on Z is

$$\begin{aligned}\sum_{i=1}^{n} \mathbf{U}_i(\theta_X) &= \sum_{i=1}^{n} \Big(\frac{\partial \mu_i}{\partial \theta_X}\Big)^T \mathbf{V}_i^{-1}(\mathbf{O}_i - \mu_i(\theta_X)) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{k_i} \Big(\frac{\partial \mu_{ij}}{\partial \theta_X}\Big)^T \mathbf{V}_{ij}^{-1}(\mathbf{O}_{ij} - \mu_{ij}(\theta_X)) \\ &= \mathbf{0},\end{aligned}$$

where $\mathbf{V}_{ij} = \text{Cov}(\mathbf{O}_{ij}\}$ with diagonal elements as $v_{ll} = \mu_{ij,l}(1-\mu_{ij,l})$ and off-diagonal elements as $v_{lk} = \mu_{ij,min(l,k)}(1-\mu_{ij,max(l,k)})$, $l,k = 1,...,C-1$. This estimating function is equivalent to the estimating function of GEE methods for ordinal response variables with independence working correlation. The information matrix is

$$I(\theta_X) = -E[\partial \mathbf{U}_i(\theta_X)/\partial \theta_X] = E[\Big(\frac{\partial \mu_i}{\partial \theta_X}\Big)^T \mathbf{V}_i^{-1}\Big(\frac{\partial \mu_i}{\partial \theta_X}\Big)].$$

The estimating function of the CPM of Y on Z is similar.

3.10.2 Additional simulations under negative rank ICCs

We here consider the performance of our estimators when the cluster size is 2 in the population and the rank ICC is negative. Let (X_{i1}, Y_{i1}) and (X_{i2}, Y_{i2}) be the observations of the two individual in the ith cluster. The observations in cluster i were generated as follows, $\begin{pmatrix} X_{i1} \\ Y_{i1} \end{pmatrix} = \begin{pmatrix} U_{Xi} \\ U_{Yi} \end{pmatrix} + \begin{pmatrix} R_{Xi} \\ R_{Yi} \end{pmatrix}$ and $\begin{pmatrix} X_{i2} \\ Y_{i2} \end{pmatrix} = \begin{pmatrix} U_{Xi} \\ U_{Yi} \end{pmatrix} - \begin{pmatrix} R_{Xi} \\ R_{Yi} \end{pmatrix}$, where $\begin{pmatrix} U_{Xi} \\ U_{Yi} \end{pmatrix} \overset{i.i.d}{\sim} N\Big(\begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 & \rho_b \\ \rho_b & 1 \end{pmatrix}\Big)$, $\begin{pmatrix} R_{xi} \\ R_{yi} \end{pmatrix} \overset{i.i.d}{\sim} N\Big(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 3 & 9\rho_w \\ 9\rho_w & 1 \end{pmatrix}\Big)$, $\rho_t = \rho_b/4 + 3\rho_w/4$, and $\rho_{I_X} = \rho_{I_Y} = -0.5$. We conducted 1000 simulations at $n = 100$ and $(\rho_{0b}, \rho_{0w}) \in \{(0.8, 0.7), (0, 0.8), (0,0), (0.8, -0.7)\}$. Since the sample cluster size was the same as the cluster size of the population, the data provided all the information of each sample cluster and we can use CPM to accurately estimate the cluster medians. Therefore, we used $\hat{\gamma}_{b_M}$ to estimate γ_b. Our estimators of γ_t, γ_b, γ_w had very low bias and good coverage.

Table 3.2: Bias, coverage of 95% CIs, empirical SE, and median SE for our estimators of γ_b, γ_w, and γ_t at different true values when the rank ICC was negative. The number of clusters was set at 100 and the cluster size was set to 2.

$\gamma_b, \gamma_w, \gamma_t$	$\hat{\gamma}_{b_A}$			$\hat{\gamma}_{b_M}$				$\hat{\gamma}_w$				$\hat{\gamma}_t$			
	Bias	Coverage	emp.SE	Bias	Coverage	emp.SE	mdn.SE	Bias	Coverage	emp.SE	mdn.SE	Bias	Coverage	emp.SE	mdn.SE
(0.786, 0.683, 0.708)	-0.011	0.824	0.062	-0.014	0.939	0.046	0.042	-0.005	0.935	0.058	0.057	-0.005	0.942	0.043	0.043
(0, 0.786, 0.582)	-0.001	0.929	0.109	-0.002	0.945	0.1	0.101	-0.004	0.933	0.043	0.042	-0.002	0.947	0.055	0.055
(0, 0, 0)	0.002	0.913	0.113	-0.001	0.946	0.1	0.101	0	0.944	0.099	0.099	0	0.948	0.077	0.078
(0.786, -0.683, -0.312)	-0.039	0.823	0.052	-0.013	0.94	0.045	0.042	0.005	0.93	0.058	0.057	0.001	0.937	0.078	0.078

Table 3.3: Bias and empirical SE (emp.SE) of $\hat{\gamma}_{b_n}$ and $\hat{\gamma}_{w_n}$, the naive estimators of γ_b and γ_w, at different true values under Scenarios I and II. The number of clusters was set at 100 and the cluster size was set to 20.

$(\gamma_b, \gamma_w, \gamma_t)$	$\hat{\gamma}_{b_n}$		$\hat{\gamma}_{w_n}$		$\hat{\gamma}_t$	
	Bias	emp.SE	Bias	emp.SE	Bias	emp.SE
(0.786,0.683, 0.734)	-0.026	0.048	-0.027	0.014	-0.001	0.024
(0.786,0, 0.385)	-0.063	0.052	-0.001	0.023	-0.003	0.043
(0,0.786, 0.385)	0.042	0.101	-0.028	0.011	0.002	0.055
(0,0, 0)	0	0.099	-0.001	0.023	0	0.052
(0.786,-0.683, 0.048)	-0.099	0.057	0.027	0.014	-0.006	0.063

Table 3.4: Estimates (EST), bias, empirical SE (emp.SE) of our estimators of γ_b, γ_w and γ_t under Scenarios I and II at the extreme case when $(\gamma_b, \gamma_w, \gamma_t)$=(0.786, -0.683, 0.048). The cluster size was 60 and the number of clusters was 200.

	$\hat{\gamma}_b$		$\hat{\gamma}_w$	$\hat{\gamma}_t$
	$\hat{\gamma}_{b_A}$	$\hat{\gamma}_{b_M}$		
EST	0.831	0.758	-0.683	0.046
Bias	0.045	-0.028	0.000	-0.002
emp.SE	0.027	0.034	0.005	0.044

Table 3.5: Estimate (EST) and bias of the between-cluster Pearson correlation estimator based on random-effect models under Scenario I at the extreme case when (ρ_b, ρ_w, ρ_t)=(0.8, -0.7, 0.05). The cluster size was denoted as k_i and the number of clusters was denoted as n.

(n, k_i)	EST	Bias
(100, 30)	0.751	-0.049
(200, 30)	0.751	-0.049
(200, 60)	0.775	-0.025

Table 3.6: Bias, coverage of 95% CIs, empirical SE, and median SE for our estimators of γ_b, γ_w, and γ_t at different true values under Scenario III. The number of clusters was set at 100 and the cluster size was set to 20. "emp.SE" refers to empirical standard error. "mdn.SE" refers to the median SE of the 1000 replicates.

(a) $E[U_{Yi}] = -1$

$\gamma_b, \gamma_w, \gamma_t$	$\hat{\gamma}_{b_A}$			$\hat{\gamma}_{b_M}$				$\hat{\gamma}_w$				$\hat{\gamma}_t$			
	Bias	Coverage	emp.SE	Bias	Coverage	emp.SE	mdn.SE	Bias	Coverage	emp.SE	mdn.SE	Bias	Coverage	emp.SE	mdn.SE
(0.786, 0.683, 0.533)	-0.113	0.569	0.056	-0.11	0.596	0.058	0.059	-0.001	0.93	0.014	0.013	-0.008	0.933		
(0.786, 0, 0.308)	-0.068	0.911	0.055	-0.152	0.304	0.064	0.063	0.001	0.937	0.023	0.023	-0.015	0.91	0.046	0.045
(0, 0.786, 0.246)	-0.079	0.82	0.117	0.04	0.928	0.1	0.101	-0.001	0.953	0.01	0.01	0.006	0.933	0.053	0.053
(0, 0, 0)	0.003	0.914	0.109	0.004	0.937	0.099	0.1	0.001	0.956	0.023	0.023	0.001	0.949	0.046	0.046
(0.786, -0.683, 0.086)	-0.012	0.965	0.057	-0.188	0.124	0.068	0.066	0.001	0.959	0.013	0.013	-0.015	0.916	0.066	0.063

(b) $E[U_{Yi}] = 1$

$\gamma_b, \gamma_w, \gamma_t$	$\hat{\gamma}_{b_A}$			$\hat{\gamma}_{b_M}$				$\hat{\gamma}_w$				$\hat{\gamma}_t$			
	Bias	Coverage	emp.SE	Bias	Coverage	emp.SE	mdn.SE	Bias	Coverage	emp.SE	mdn.SE	Bias	Coverage	emp.SE	mdn.SE
(0.786,0.683, 0.758)	-0.016	0.963	0.049	-0.01	0.971	0.045	0.051	-0.003	0.958	0.013	0.013	-0.003	0.941	0.041	0.039
(0.786,0, 0.653)	-0.009	0.968	0.047	-0.018	0.974	0.046	0.052	-0.001	0.96	0.022	0.023	-0.006	0.951	0.049	0.048
(0,0.786, 0.103)	-0.028	0.921	0.103	0.009	0.936	0.1	0.101	-0.003	0.949	0.009	0.01	0.004	0.94	0.087	0.087
(0,0, 0)	0.001	0.936	0.101	0.001	0.939	0.099	0.101	-0.001	0.957	0.022	0.023	0.001	0.945	0.084	0.084
(0.786,-0.683, 0.551)	0.001	0.958	0.046	-0.027	0.964	0.047	0.052	0.002	0.953	0.013	0.013	-0.008	0.946	0.061	0.059

Table 3.7: Bias, coverage of 95% CIs, empirical SE, and median SE for our estimators of γ_b, γ_w, and γ_t at different true values when CPM link function was misspecified (the correct link function is probit). The number of clusters was set at 100 and the cluster size was set to 20. "emp.SE" refers to empirical standard error. "mdn.SE" refers to the median SE of the 1000 simulations.

(a) logit link

$\gamma_b, \gamma_w, \gamma_t$	$\hat{\gamma}_{b_A}$			$\hat{\gamma}_{b_M}$				$\hat{\gamma}_w$				$\hat{\gamma}_t$			
	Bias	Coverage	emp.SE	Bias	Coverage	emp.SE	mdn.SE	Bias	Coverage	emp.SE	mdn.SE	Bias	Coverage	emp.SE	mdn.SE
(0.786, 0.683, 0.734)	-0.003	0.939	0.045	-0.017	0.942	0.046	0.043	-0.005	0.924	0.014	0.013	-0.007	0.944	0.024	0.023
(0.786, 0, 0.385)	0.001	0.948	0.045	-0.055	0.889	0.049	0.049	0	0.956	0.023	0.023	-0.017	0.927	0.042	0.042
(0, 0.786, 0.385)	-0.045	0.9	0.109	0.035	0.941	0.099	0.101	-0.004	0.94	0.01	0.01	0.007	0.935	0.055	0.054
(0, 0, 0)	-0.002	0.927	0.108	-0.002	0.949	0.102	0.101	0	0.957	0.023	0.023	-0.001	0.938	0.052	0.051
(0.786, -0.683, 0.048)	0.038	0.865	0.048	-0.087	0.732	0.056	0.054	0.004	0.949	0.013	0.013	-0.022	0.924	0.062	0.06

(b) loglog

$\gamma_b, \gamma_w, \gamma_t$	$\hat{\gamma}_{b_A}$			$\hat{\gamma}_{b_M}$				$\hat{\gamma}_w$				$\hat{\gamma}_t$			
	Bias	Coverage	emp.SE	Bias	Coverage	emp.SE	mdn.SE	Bias	Coverage	emp.SE	mdn.SE	Bias	Coverage	emp.SE	mdn.SE
(0.786, 0.683, 0.734)	0.006	0.93	0.043	-0.018	0.948	0.046	0.043	-0.005	0.938	0.013	0.013	-0.002	0.953	0.023	0.023
(0.786, 0, 0.385)	0.004	0.953	0.045	-0.06	0.852	0.053	0.05	0.005	0.944	0.023	0.022	-0.006	0.945	0.043	0.042
(0, 0.786, 0.385)	-0.028	0.913	0.109	0.05	0.91	0.101	0.101	-0.014	0.717	0.01	0.01	0.006	0.93	0.056	0.054
(0, 0, 0)	0.004	0.937	0.103	0.005	0.952	0.099	0.101	-0.001	0.956	0.022	0.022	0.002	0.948	0.051	0.051
(0.786, -0.683, 0.048)	0.011	0.936	0.048	-0.087	0.717	0.057	0.054	0.031	0.359	0.014	0.013	-0.012	0.944	0.061	0.062

(c) cloglog link

$\gamma_b, \gamma_w, \gamma_t$	$\hat{\gamma}_{b_A}$			$\hat{\gamma}_{b_M}$				$\hat{\gamma}_w$				$\hat{\gamma}_t$			
	Bias	Coverage	emp.SE	Bias	Coverage	emp.SE	mdn.SE	Bias	Coverage	emp.SE	mdn.SE	Bias	Coverage	emp.SE	mdn.SE
(0.786, 0.683, 0.734)	0.004	0.929	0.054	-0.022	0.939	0.048	0.044	-0.007	0.946	0.028	0.013	-0.005	0.948	0.024	0.023
(0.786, 0, 0.385)	-0.001	0.943	0.046	-0.063	0.847	0.051	0.051	0.006	0.965	0.021	0.023	-0.016	0.92	0.042	0.042
(0, 0.786, 0.385)	-0.033	0.912	0.111	0.046	0.929	0.1	0.1	-0.016	0.747	0.028	0.01	0.009	0.959	0.053	0.054
(0, 0, 0)	0.002	0.951	0.105	0.002	0.952	0.1	0.1	0.001	0.957	0.022	0.022	0.002	0.949	0.05	0.051
(0.786, -0.683, 0.048)	-0.002	0.972	0.056	-0.099	0.66	0.055	0.055	0.034	0.334	0.027	0.013	-0.032	0.909	0.061	0.06

Table 3.8: Bias, coverage of 95% CIs, empirical SE, and median SE for our estimators of γ_b, γ_w, and γ_t for ordered categorical data. The number of clusters was set at 100 and the cluster size was set to 20.

(a) 5 levels

$\gamma_b, \gamma_w, \gamma_t$	$\hat{\gamma}_{b_A}$			$\hat{\gamma}_{b_M}$				$\hat{\gamma}_w$				$\hat{\gamma}_t$			
	Bias	Coverage	emp.SE	Bias	Coverage	emp.SE	mdn.SE	Bias	Coverage	emp.SE	mdn.SE	Bias	Coverage	emp.SE	mdn.SE
(0.744,0.602, 0.697)	0.063	0.615	0.045	0.025	0.811	0.046	0.042	0.001	0.944	0.016	0.016	-0.001	0.941	0.024	0.023
(0.74,0.001, 0.358)	0.054	0.731	0.046	-0.006	0.939	0.051	0.047	-0.001	0.95	0.023	0.023	-0.002	0.946	0.042	0.041
(-0.003,0.671, 0.357)	-0.016	0.919	0.109	0.04	0.926	0.1	0.101	0	0.946	0.015	0.015	0.004	0.937	0.053	0.053
(0.007,0, 0.003)	-0.007	0.93	0.106	-0.006	0.946	0.1	0.101	0	0.949	0.022	0.023	-0.003	0.95	0.049	0.049
(0.743,-0.583, 0.045)	0.063	0.735	0.047	-0.043	0.922	0.055	0.053	-0.001	0.944	0.017	0.017	-0.005	0.942	0.061	0.06

(b) 10 levels

$\gamma_b, \gamma_w, \gamma_t$	$\hat{\gamma}_{b_A}$			$\hat{\gamma}_{b_M}$				$\hat{\gamma}_w$				$\hat{\gamma}_t$			
	Bias	Coverage	emp.SE	Bias	Coverage	emp.SE	mdn.SE	Bias	Coverage	emp.SE	mdn.SE	Bias	Coverage	emp.SE	mdn.SE
(0.775,0.65, 0.724)	0.028	0.822	0.044	-0.004	0.929	0.046	0.042	0	0.958	0.014	0.014	-0.002	0.946	0.024	0.023
(0.773,0, 0.376)	0.023	0.887	0.045	-0.036	0.927	0.051	0.048	-0.001	0.959	0.022	0.023	-0.002	0.947	0.043	0.042
(-0.003,0.742, 0.375)	-0.027	0.911	0.109	0.04	0.92	0.1	0.101	0	0.948	0.012	0.012	0.004	0.929	0.055	0.055
(0.006,0, 0.003)	-0.007	0.928	0.106	-0.006	0.941	0.1	0.101	0	0.966	0.022	0.023	-0.003	0.95	0.051	0.051
(0.774,-0.643, 0.048)	0.04	0.847	0.046	-0.072	0.8	0.055	0.053	-0.001	0.95	0.014	0.015	-0.007	0.942	0.063	0.062

Table 3.9: Bias and coverage (Cvrg.) of 95% CIs of our estimators of γ_b, γ_w and γ_t at different true values under Scenarios I and II with 200 clusters and 20 per cluster.

$(\gamma_b, \gamma_w, \gamma_t)$	$\hat{\gamma}_{b_A}$		$\hat{\gamma}_{b_M}$		$\hat{\gamma}_w$		$\hat{\gamma}_t$	
	Bias	Cvrg.	Bias	Cvrg.	Bias	Cvrg.	Bias	Cvrg.
(0.786,0.683, 0.734)	0.002	0.943	-0.006	0.95	-0.001	0.945	-0.001	0.945
(0.786,0, 0.385)	0.01	0.938	-0.03	0.882	0	0.941	-0.002	0.944
(0,0.786, 0.385)	-0.045	0.881	0.026	0.937	0	0.947	0.001	0.936
(0,0, 0)	0	0.936	0.001	0.941	0	0.942	0	0.944
(0.786,-0.683, 0.048)	0.045	0.742	-0.053	0.735	0	0.938	-0.002	0.942

Table 3.10: Bias and coverage (Cvrg.) of 95% CIs of our estimators of γ_b, γ_w and γ_t at different true values under Scenarios I and II with 200 clusters and 30 per cluster.

$(\gamma_b, \gamma_w, \gamma_t)$	$\hat{\gamma}_{b_A}$		$\hat{\gamma}_{b_M}$		$\hat{\gamma}_w$		$\hat{\gamma}_t$	
	Bias	Cvrg.	Bias	Cvrg.	Bias	Cvrg.	Bias	Cvrg.
(0.786,0.683, 0.734)	0.002	0.926	-0.012	0.941	-0.001	0.954	-0.001	0.945
(0.786,0, 0.385)	0.011	0.927	-0.046	0.905	-0.001	0.961	-0.003	0.945
(0,0.786, 0.385)	-0.046	0.898	0.036	0.926	-0.001	0.953	0.002	0.935
(0,0,0)	0	0.926	0.001	0.947	-0.001	0.959	0	0.952
(0.786,-0.683, 0.048)	0.047	0.822	-0.08	0.745	0	0.958	-0.006	0.945

Table 3.11: Bias and empirical SE (emp.SE) of the naive estimators of γ_b and γ_w at different true values under Scenario III with 100 clusters and 20 per cluster.

$(\gamma_b, \gamma_w, \gamma_t)$	$\hat{\gamma}_{b_n}$		$\hat{\gamma}_{w_n}$		$\hat{\gamma}_t$	
	Bias	emp.SE	Bias	emp.SE	Bias	emp.SE
(0.786,0.683, 0.533)	-0.146	0.064	-0.027	0.014	-0.008	0.035
(0.786,0, 0.308)	-0.188	0.069	0	0.023	-0.015	0.046
(0,0.786, 0.246)	0.042	0.1	-0.028	0.011	0.006	0.053
(0,0, 0)	0.001	0.099	0.001	0.022	0.001	0.046
(0.786,-0.683, 0.086)	-0.221	0.072	0.027	0.014	-0.015	0.066

CHAPTER 4

Unified and Simple Sample Size Calculations for Cluster Randomized Controlled Trials with Skewed or Ordinal Outcomes

4.1 Introduction

Cluster randomized controlled trials (RCTs) are widely used in biomedical research (Donner and Klar, 2000; Hayes and Moulton, 2009; Eldridge and Kerry, 2012). Observations from individuals within the same cluster tend to be more similar than observations from different clusters, which introduces complexity to the design of cluster RCTs. A conventional and simple approach to calculating sample sizes in cluster RCTs is to inflate the sample size of an adequately powered individual RCT by the design effect (DE) based on the intraclass correlation coefficient (ICC) (Kish, 1965; Donner et al., 1981). Although this DE is commonly used in sample size calculations for cluster RCTs (Campbell and Walters, 2014; Rutterford et al., 2015), it was originally derived for comparisons of means and may not be applicable to skewed or ordinal outcomes.

In practice, the outcome of interest is often a skewed continuous variable, an ordinal variable, or a mixture of the two (e.g. outcomes subject to a detection limit). For example, a cluster RCT was conducted to measure the impact of a multi-component intervention on adherence to antiretroviral therapy (ART) of pregnant women living with HIV (Audet et al., 2018, 2024). The primary outcome was adherence to treatment (i.e., the proportion of medications taken within 1 year), which is pseudo-continuous ranging from 0 to 1 with a left-skewed distribution. As another example, in a cluster RCT on childhood epilepsy care (Aliyu et al., 2019), the outcome of interest may be the number of seizures from 18 to 24 months, which is an irregularly distributed count variable.

Preliminary data or prior knowledge are often used to inform the design of cluster RCTs. If these sources indicate skewness in the forthcoming data, it is common to apply transformations such as logarithmic or square root transformations before calculating sample sizes for cluster RCTs with the conventional approach described above. If preliminary data are not available, it is common to compute sample sizes under implicit assumptions that the data will be appropriately transformed prior to analyses. However, sample size calculations and analysis results are often sensitive to the data transformation, which may be difficult to select in practice, and this approach is not applicable to ordinal data. Two rank-based sample size calculation approaches have been proposed as alternatives. For skewed continuous outcomes, one approach has been developed based on clustered Wilcoxon rank-sum tests (Rosner and Glynn, 2011). For ordinal outcomes, another approach, based on generalized estimating equations (GEE), has been proposed (Kim et al., 2005).

However, these rank-based approaches lack closed forms and must be implemented numerically, involving numerous calculations. In addition, the approach based on clustered Wilcoxon rank-sum tests assumes the outcome to be continuous with no ties and may not be applicable to ordinal data. The GEE-based approach requires a detailed description of the distribution and correlation structure.

In this Chapter, we propose unified and simple sample size calculations for cluster RCTs with skewed or ordinal outcomes. Our calculations involve inflating the sample size for an adequately powered individual RCT for an ordinal outcome with a DE that incorporates the rank ICC. The rank ICC is a rank-based correlation measuring the degree of similarity within clusters (Tu et al., 2023). We show that in most scenarios, this DE is close to the ratio of the variance of the ordinary Wilcoxon rank-sum statistic to the variance of the clustered Wilcoxon rank-sum statistic. With continuous data, we show that our calculations closely approximate those more complicated calculations based on clustered Wilcoxon rank-sum tests. Furthermore, our calculations can be applied to compute either the number of clusters with predetermined cluster sizes or compute the cluster sizes with a predetermined number of clusters.

4.2 Review of rank-based tests for individual and cluster RCTs

Rank-based tests are usually used to analyze skewed or ordinal data, given their nonparametric nature and robustness to the shape of the distribution. The Wilcoxon rank-sum test, also known as the Mann–Whitney U test, is a widely used rank-based approach for evaluating treatment effects with skewed or ordinal data in the absence of clustering (Mann and Whitney, 1947; Lehmann, 1975). The parameter of the Wilcoxon rank-sum test can be formulated in terms of the probabilistic index. Let θ denote the probabilistic index and $\theta = P(X_i < Y_j) + P(X_i = Y_j)/2$, where X_i is a random variable from the control population and Y_j is a random variable from the experiment population (Hollander and Wolfe, 1999). With continuous X_i and Y_j, $\theta = P(X_i < Y_j)$. The hypothesis formulated in terms of θ is $H_0 : \theta = 1/2$ vs. $H_1 : \theta \neq 1/2$. The estimator of θ, denoted as $\hat{\theta}$, is equal to the Mann-Whitney U statistic divided by the product of the sample sizes of the two arms.

Proportional odds (PO) models are also commonly applied in the analysis of ordinal outcomes (McCullagh, 1980). PO models can also be fit for a robust and rank-based analysis of continuous outcomes, where each unique continuous outcome is assigned to be a separate ordinal category (Liu et al., 2017). Whitehead (1993) has shown that the test statistic for treatment effect, derived from the likelihood of an unadjusted PO model, is exactly equal to a version of the Mann-Whitney U test statistic presented by Siegel (1956). That is, unadjusted PO models with a single binary covariate are essentially Wilcoxon rank-sum/Mann–Whitney U tests. Additionally, there is a numerical relationship between θ and the log odds ratio (OR) regarding the treatment effect in unadjusted PO models (De Neve et al., 2019): $\theta = \exp(\delta)[\exp(\delta) - \delta - 1]/(\exp(\delta) - 1)^2$,

where δ denotes the log OR.

To account for clustering, Rosner et al. (2003) developed the clustered Wilcoxon rank-sum test, which incorporates a correction to the variance of the Wilcoxon rank-sum test statistic. The clustered Wilcoxon rank-sum test can also be expressed in terms of θ. The definition of θ for clustered data is $\theta = P(X_{ij} < Y_{kl}) + P(X_{ij} = Y_{kl})/2$, where X_{ij} denotes a random variable of the jth individual in the ith cluster from the control population and Y_{kl} denotes a random variable of the lth individual in the kth cluster from the experiment population. This θ has been used in power and sample size estimation for the clustered Wilcoxon rank-sum test with continuous data (Rosner and Glynn, 2011).

PO models have also been extended to handle clustered ordinal or continuous outcomes (Heagerty and Zeger, 1996; Parsons et al., 2006; Tian et al., 2023), employing GEE-based estimation. Commonly used working correlation structures include independent, exchangeable, and first-order autoregressive (AR1) correlations. Tian et al. (2023) demonstrated that clustered continuous outcomes could also be analyzed using PO GEE-based methods. Fitting a PO model to the continuous outcome and then fixing standard error using a Huber-White sandwich estimator for covariance to correct for within-cluster correlation is equivalent to GEE with independent working correlation and is straightforward to implement. Exchangeable/AR1 working correlation structures can be statistically more efficient than independent working correlation in some settings with continuous outcomes but are more computationally burdensome. In this Chapter, we focus on PO models with independent working correlation, which we henceforth refer to as cluster PO models for simplicity. Furthermore, we show via simulations (Section 2.4) that, under the null hypothesis, the p-values of clustered Wilcoxon rank-sum tests and unadjusted cluster PO models approximate each other with a large number of clusters.

4.3 Sample size calculations

4.3.1 The design effect of cluster RCTs

The DE was initially introduced by Kish (1965) as a measure of the expected impact of a sampling design on the variance of an estimator. Subsequently, it was applied by Donner et al. (1981) to inflate sample sizes calculated under individual randomization to achieve the required statistical power under cluster randomization. The DE of cluster RCTs with respect to an estimator T is defined as

$$D_{\text{eff}}(T) = \frac{\text{var}(T)}{\text{var}(T_{\text{srs}})},$$

where $\text{var}(T)$ is the variance of T under cluster randomization and $\text{var}(T_{\text{srs}})$ is the variance of a comparable estimator under simple random sampling with replacement (SRS, or individual randomization).

Let ρ_I denote the ICC and $\rho_I = corr(X_{ij}, X_{ij'})$, where $(X_{ij}, X_{ij'})$ is a random pair from a random cluster (Fisher, 1925). The DE of cluster RCTs concerning the mean estimator $\bar{X}$ is

$$D_{\text{eff}}(\bar{X}) = 1 + \rho_I(k-1), \tag{4.1}$$

where k is the cluster size or the average of cluster sizes (Kish, 1965, 1987). The DE in (4.1) is often used as the inflation factor for sample size calculations in cluster RCTs (Campbell and Walters, 2014; Rutterford et al., 2015). Let n_{srs} denote the total sample size for an adequately powered individual RCT. Then the total sample size for a cluster RCT is calculated as

$$n = n_{\text{srs}} D_{\text{eff}}(\bar{X}) = n_{\text{srs}}\{1 + \rho_I(k-1)\}.$$

However, ρ_I is sensitive to extreme values and skewed distributions, and it depends on the scale of the data. ρ_I also lacks a clear definition when applied to ordinal data. While ordinal regression models with random effects may be used to estimate variance components, the total variance remains undefined unless numbers are assigned to levels of the ordinal response (Denham, 2016). Hence, the DE in (4.1) may not be applicable to skewed or ordinal data.

Let $\hat{\theta}$ denote the estimator of θ under clustered randomization and $\hat{\theta}_{\text{srs}}$ denote the comparable estimator under individual randomization. Then the DE associated with the clustered Wilcoxon rank-sum statistic can be expressed as $D_{\text{eff}}(\hat{\theta}) = \text{var}(\hat{\theta})/\text{var}(\hat{\theta}_{\text{srs}})$. Under the null hypothesis ($\theta = 1/2$), the analytical formulas of $\text{var}(\hat{\theta})$ and $\text{var}(\hat{\theta}_{\text{srs}})$ can be derived, enabling the analytical computation of $D_{\text{eff}}(\hat{\theta})$. However, under the alternative hypothesis ($\theta \neq 1/2$), the derivation of both $\text{var}(\hat{\theta})$ and $\text{var}(\hat{\theta}_{\text{srs}})$ requires continuous distributions (Lehmann, 1975; Rosner and Glynn, 2011), and so does the analytical computation of $D_{\text{eff}}(\hat{\theta})$. The analytical formula of $D_{\text{eff}}(\hat{\theta})$ is also complex, involving the ICC on the probit scale of the cumulative distribution function (CDF). Hence, because of its complexity, $D_{\text{eff}}(\hat{\theta})$ cannot be used as a simple inflation factor in sample size calculations with skewed or ordinal outcomes.

We consider an alternative inflation factor that closely approximates $D_{\text{eff}}(\hat{\theta})$ in most scenarios and is also applicable to ordinal outcomes. Let F be a CDF, $F(x-) = \lim_{t \uparrow x} F(t)$, and $F^*(x) = \{F(x) + F(x-)\}/2$. If the distribution is continuous, then $F^*(x) = F(x)$. If the distribution is discrete or mixed, $F^*(x)$ corresponds to the population versions of ridits (Bross, 1958). Let γ_I denote the rank ICC, which is a rank-based correlation measuring the degree of within-cluster similarity. It is defined as follows,

$$\gamma_I = corr\{F^*(X_{ij}), F^*(X_{ij'})\},$$

where $(X_{ij}, X_{ij'})$ is a random pair from a random cluster (Tu et al., 2023). γ_I is insensitive to extreme values and skewed distributions, and it does not depend on the scale of data. It is also applicable to ordinal data and easily computed empirically. We analytically show that when $\theta = 1/2$ and sample sizes of the two arms greatly exceed cluster sizes, then

$$D_{\text{eff}}(\hat{\theta}) \approx 1 + \gamma_I(k-1). \tag{4.2}$$

We also show via simulations that in other scenarios, (4.2) is still valid, except for very large γ_I (Figure 4.7). See Supplementary Materials for details. Given the nonparametric nature of γ_I and the approximation between $1+\gamma_I(k-1)$ and $D_{\text{eff}}(\hat{\theta})$, $1+\gamma_I(k-1)$ can serve as a simple inflation factor in sample size calculations with skewed or ordinal outcomes. This inflation factor is as simple as the conventional $1+\rho_I(k-1)$ but it is more robust in the context of skewed data and applicable to ordinal data.

4.3.2 Individual RCTs

For illustrative purposes, we refer to the two arms in cluster RCTs as the control and experiment arms, respectively. Under individual randomization, Whitehead (1993) provided a sample size calculation formula for ordinal outcomes, using the test statistic derived from the likelihood of an unadjusted PO model. Let n_{srs} denote the total sample size for an individual RCT and A denote the allocation ratio of the control arm to the experiment arm. For a two-sided significant level at α and power at $1-\beta$, Whitehead's formula is

$$n_{\text{srs}} = \frac{3(A+1)^2(Z_{1-\alpha/2}+Z_{1-\beta})^2/\delta^2}{A(1-\sum_{i=1}^{l}\bar{\pi}_i^3)}, \tag{4.3}$$

where $\bar{\pi}_i$ is the mean proportion expected in ordinal category i and calculated as $\bar{\pi}_i = (\pi_{1i}+\pi_{2i})/2$, π_{1i} and π_{2i} are the proportions for the control and experiment groups, l is the number of ordered categories, and δ denotes the log OR in the unadjusted PO model.

Continuous outcomes are also ordinal, and the formula (4.3) can also be applied to such outcomes. Since continuous outcomes have no ties, the proportion for each ordinal category is $\pi_{1i} = \pi_{2i} = 1/n_{\text{srs}}$. We plug this proportion into (4.3),

$$n_{\text{srs}} = \frac{3(A+1)^2(Z_{1-\alpha/2}+Z_{1-\beta})^2/\delta^2}{A(1-\sum_{i=1}^{n_{\text{srs}}}1/n_{\text{srs}}^3)},$$

and solve for n_{srs}. Let $S = 3(A+1)^2(Z_{1-\alpha/2}+Z_{1-\beta})^2/(2A\delta^2)$. The sample size for individual RCTs with continuous outcomes is then

$$n_{\text{srs}} = \sqrt{1+S^2}+S. \tag{4.4}$$

The odds ratio δ for continuous outcomes is the relative odds of having a larger outcome. The calculation provided in (4.4) is robust to skewness, extreme values, and any data transformations. There is another

rank-based sample size calculation approach, derived from Wilcoxon rank sum tests and used for continuous outcomes (Rosner and Glynn, 2009). This sample size calculation approach is complex and lacks closed forms. We show via simulation that the sample sizes obtained by (4.4) are very close to those obtained by the calculation approach based on Wilcoxon rank sum tests (Figure 4.8).

4.3.3 Cluster RCTs

We extend Whitehead's sample size calculation for individual RCTs to cluster RCTs by using $1+\gamma_I(k-1)$ to inflate n_{srs} in (4.3). With ordinal outcomes, the total sample size of a cluster RCT for a two-sided significant level at α and power at $1-\beta$ is calculated as

$$n = n_{\text{srs}}\{1+\gamma_I(k-1)\} = \frac{3(A+1)^2(Z_{1-\alpha/2}+Z_{1-\beta})^2/\delta^2}{A(1-\sum_{i=1}^{l}\bar{\pi}_i{}^3)}\{1+\gamma_I(k-1)\}. \tag{4.5}$$

With (4.5), the sample sizes for the experiment and control arms can be easily calculated as $n_E = n/(A+1)$ and $n_C = An/(A+1)$, respectively.

Similar to the extension from (4.3) to (4.4), we can also apply the sample size formula (4.5) to continuous outcomes in cluster RCTs. For continuous outcomes, the proportion at each ordinal category is $\pi_{1i} = \pi_{2i} = 1/n$. We plug this proportion in (4.5) and solve for n,

$$n = \frac{3(A+1)^2(Z_{1-\alpha/2}+Z_{1-\beta})^2/\delta^2}{A(1-\sum_{i=1}^{n}1/n^3)}\{1+\gamma_I(k-1)\}. \tag{4.6}$$

Then the total sample size for cluster RCTs with continuous outcomes is calculated as

$$n = \sqrt{1+S^2\{1+\gamma_I(k-1)\}^2}+S\{1+\gamma_I(k-1)\}. \tag{4.7}$$

where $S = 3(A+1)^2(Z_{1-\alpha/2}+Z_{1-\beta})^2/(2A\delta^2)$. When γ_I is 0 or the cluster size k is 1, (4.5) and (4.7) simplify to (4.3) and (4.4), respectively, which are for individual RCTs. There is an alternative approach for calculating sample sizes for continuous outcomes in cluster RCTs: directly inflating the sample size calculated by (4.4) by $1+\gamma_I(k-1)$. With S^2 much greater than 1, sample sizes calculated by this alternative approach are very close to those calculated by (4.7).

Typically, $1+\rho_I(k-1)$ is used to inflate the sample size for an adequately powered individual RCT based on two-sample t-tests. It is expected that, under the assumption of normality, the sample sizes calculated from this conventional approach would be smaller than those from our calculations. However, we show that under normality, if the allocation ratio is 1 or the outcome variances of both arms are equal, the sample sizes calculated from this conventional approach are very close to those from our calculations. Details on the

derivations are available in the Supplementary Materials (4.7.3). This implies that there is little penalty in terms of additional sample size for not assuming normality and instead using our more robust sample size calculations.

In practical study designs, there are situations where the number of clusters is predetermined and the goal is to calculate the cluster size. In such cases, our method can also calculate cluster sizes with a predetermined number of clusters. Let m denote the total number of clusters, and $m_E = m/(A+1)$ and $m_C = Am/(A+1)$ denote the numbers of clusters in the experiment and control arms, respectively. With ordinal outcomes, the calculation for the cluster sizes is derived from equation (4.5),

$$k = \frac{2S(1-\gamma_I)}{m(1-\sum_{i=1}^{l}\bar{\pi}_i{}^3) - 2S\gamma_I}. \tag{4.8}$$

The calculation for continuous outcomes is derived from equation (4.7),

$$k = \sqrt{\frac{1}{m^2 - 2m\gamma_I S} + \frac{S^2(1-\gamma_I)^2}{(2\gamma_I S - m)^2}} - \frac{S(1-\gamma_I)}{2\gamma_I S - m}. \tag{4.9}$$

4.4 Simulations

We generated cluster RCT data using two additive models; $X_{0ij} = U_{Xi} + R_{Xij}$ and $Y_{0kl} = U_{Yk} + R_{Ykl}$, where $U_{Xi} \overset{i.i.d}{\sim} N(0, \rho_{0I})$, $R_{Xij} \overset{i.i.d}{\sim} N(0, 1-\rho_{0I})$, $U_{Yk} \overset{i.i.d}{\sim} N(\mu, \rho_{0I})$, $R_{Ykl} \overset{i.i.d}{\sim} N(0, 1-\rho_{0I})$, and ρ_{0I} varies over $[0, 0.9]$. Let X_{ij} and Y_{kl} denote observations in the control and experiment groups, respectively. We considered two scenarios for continuous data: (I) $X_{ij} = X_{0ij}$ and $Y_{kl} = Y_{0kl}$; (II) $X_{ij} = exp(X_{0ij})$ and $Y_{kl} = exp(Y_{0kl})$, assuming that the greater the value of the outcome, the more effective the treatment. ρ_{0I} is the ICC on the latent scale. The rank ICC, γ_I, is identical in Scenarios I and II, and $\gamma_I = 6\arcsin(\rho_{0I}/2)/\pi$ (Pearson, 1907). We considered different magnitudes of the treatment effect: $\delta = \{0, 0.1, 0.5, 1, 1.5\}$, where δ is the log OR of the treatment effect in the unadjusted PO model. The value of δ is also the same in both scenarios. As described in Section 4.2, the value of θ can be calculated from δ (De Neve et al., 2019). We then can compute the value of μ from θ by $\mu = \Phi^{-1}(\theta)\sqrt{2}$.

We first performed a comparison between unadjusted cluster PO models and clustered Wilcoxon rank-sum tests regarding type I error rate. Different numbers of clusters were considered, including 20, 50, 100, and 300. For each number of clusters, simulations were conducted with 1,000 replicates under $\delta = 0$ ($\theta = 1/2$) and cluster sizes of 5. The simulation results are summarized in Figure 4.1. Unadjusted PO models had a slightly higher type I error rate than clustered Wilcoxon rank-sum tests, but this difference diminished as γ_I increased or the number of clusters increased. The results under Scenarios I and II are the same.

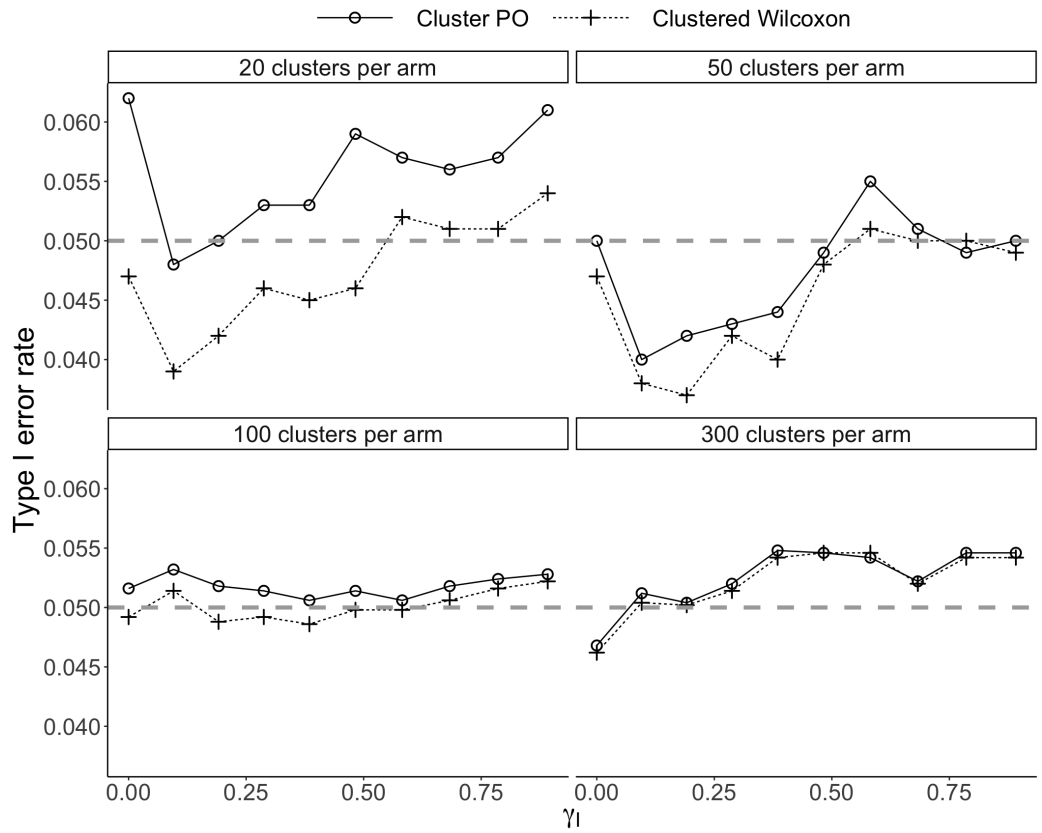
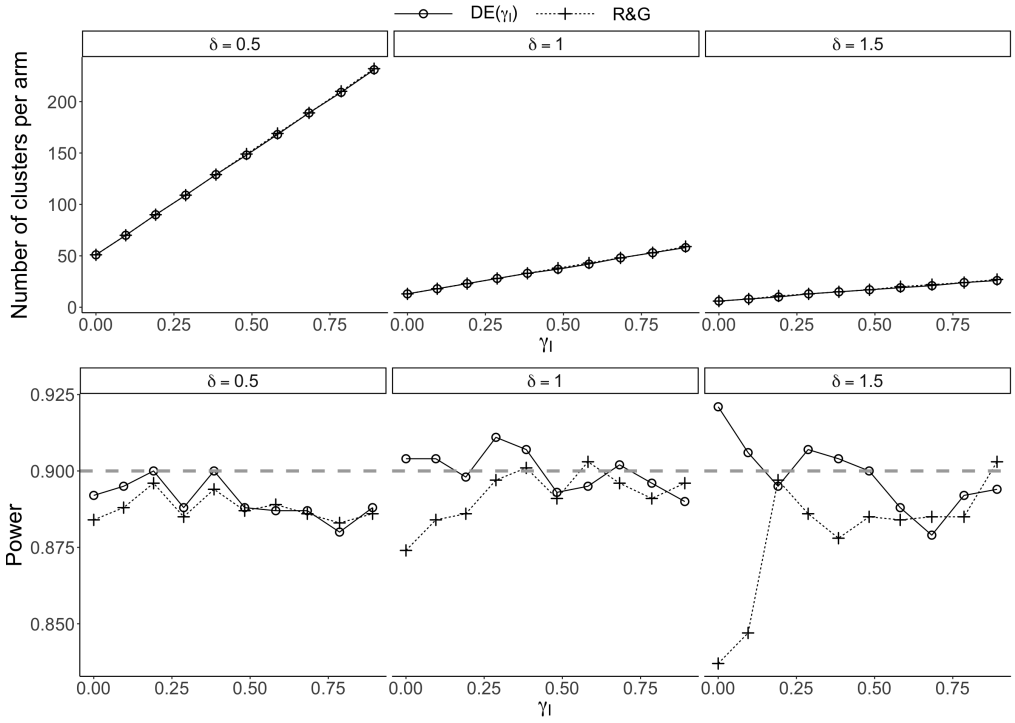


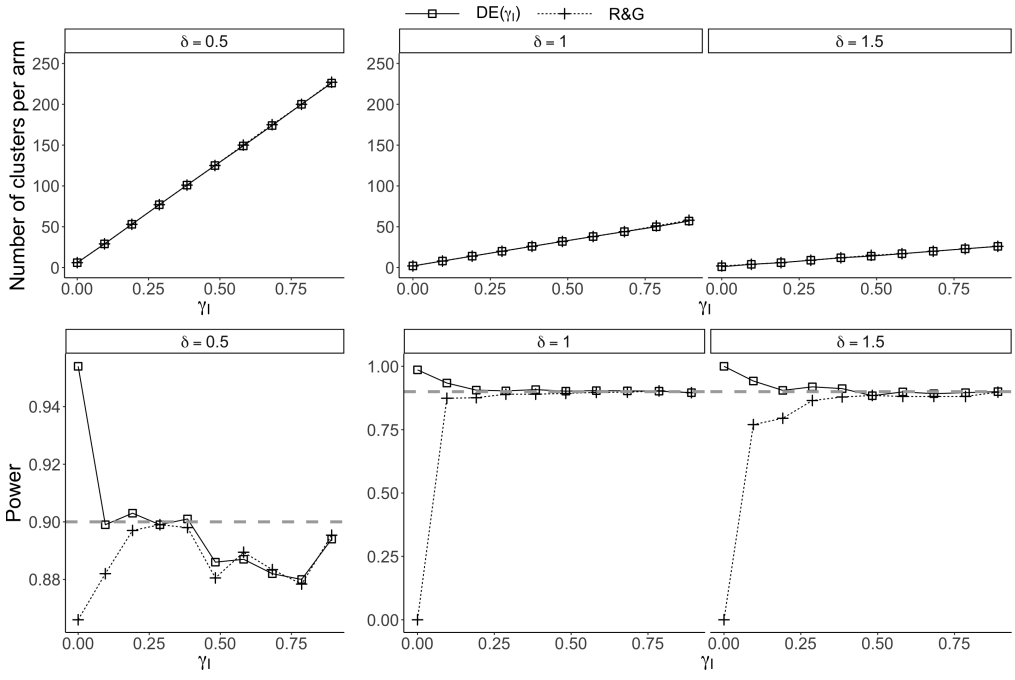
Figure 4.1: Type I error rates of clustered Wilcoxon rank-sum tests and unadjusted cluster PO models with cluster sizes of 5 and the rank ICC γ_I varying between 0 and 0.9.

We then compared our sample size calculations for continuous outcomes with Rosner and Glynn's sample size calculations with respect to power. These two calculation approaches have been developed from unadjusted PO models and clustered Wilcoxon rank-sum tests, respectively. The power of each calculation approach was obtained based on the test from which it was developed; unadjusted cluster PO models for our calculations, and clustered Wilcoxon rank-sum tests for Rosner and Glynn's calculations. The two-sided significant level was set to 0.05 and the required power was set to 0.9. In addition, we explored two design scenarios with predetermined equal cluster sizes: one with a small cluster size of 5, and another with a large cluster size of 50. The calculated number of clusters was rounded up to the nearest integer. The allocation ratio of the control arm to the treatment arm was set to 1. The simulations for power were conducted 1,000 times for each cluster size.

In summary, our sample size calculations had good power in most scenarios we considered for continuous outcomes. The sample sizes obtained by the two calculation approaches were close and both increased approximately linearly with γ_I (Figure 4.2). This suggests that, for continuous outcomes, both calculation approaches have an approximately linear relationship with γ_I, even though Rosner and Glynn's calculations are more complex and lack closed forms. When δ was small, the powers of both approaches were slightly below 0.9. The reason for this in our calculations could be link function misspecification. Since the simulated data were continuous with normally distributed latent variables, the correct link function was probit, but the link function of cluster PO models is logit. We show via additional simulations that in this setting, the power of the probit link was slightly greater than the power of the logit link (Figure 4.9). Furthermore, the power of our calculations was slightly higher than that of Rosner and Glynn's calculations in most scenarios. This difference in the power might be mainly attributed to the difference between unadjusted cluster PO models and clustered Wilcoxon rank-sum tests, as the former tends to have a slightly greater type I error rate when the number of clusters is not large. When the cluster size was large and γ_I was small, the power of our calculations was greater than 0.9 (Figure 4.2b). This is because this predetermined cluster size exceeded the required number of individuals. Rosner and Glynn's calculations had poor power in situations with small γ_I and large cluster sizes. This may be due to the poor performance of the clustered Wilcoxon rank-sum test when dealing with a small number of large clusters, as this test was proposed as a large-sample approach (Rosner et al., 2003).



(a) cluster sizes = 5



(b) cluster sizes = 50

Figure 4.2: Simulation results for numbers of clusters and powers of our calculations and Rosner and Glynn's calculations for continuous data with predetermined cluster sizes of 5 and 50. "DE(γ_I)" represents our calculation and "R&G" represents Rosner and Glynn's calculations. The simulation results under Scenarios I and II are the same.

In practice, when designing a cluster RCT, it is common to predetermine equal cluster sizes to calculate sample sizes, but the cluster sizes after accrual may be unequal. Therefore, we conducted simulations to evaluate the performance of our calculations under such cases. The data generation process involved first computing the numbers of clusters with predetermined equal cluster sizes, and then generating data with the computed numbers of clusters but unequal cluster sizes. The predetermined cluster size for sample size calculations was set to 20. We explored various configurations of cluster sizes of actual sample data: (a) equal cluster sizes of 20; (b) uniformly ranging from 15 to 25; (b) half each of 15 and 25; (d) half each of 10 and 30. The power was obtained via simulations based on cluster PO models. Simulations were performed 1,000 times for each actual cluster size. The results are summarized in Figure 4.3. The powers of (a) and (b) were very close, while the power of (c) was slightly smaller. The power of (d) was much smaller than others. It suggests that if the unequal cluster sizes in actual sample data do not differ much from the predetermined equal cluster sizes, our calculations remain robust, but if the difference is very large, our calculations might have low power. When clustered Wilcoxon rank-sum tests were fit to data generated in a similar manner, power was also low with extreme cluster size imbalance (Figure 4.10). Interestingly, the clustered Wilcoxon rank-sum test had especially low power when cluster sizes were uniformly distributed between 15 to 25; this test appears to have challenges when there are few clusters of the same cluster size because the algorithm performs computations within equal-sized clusters.

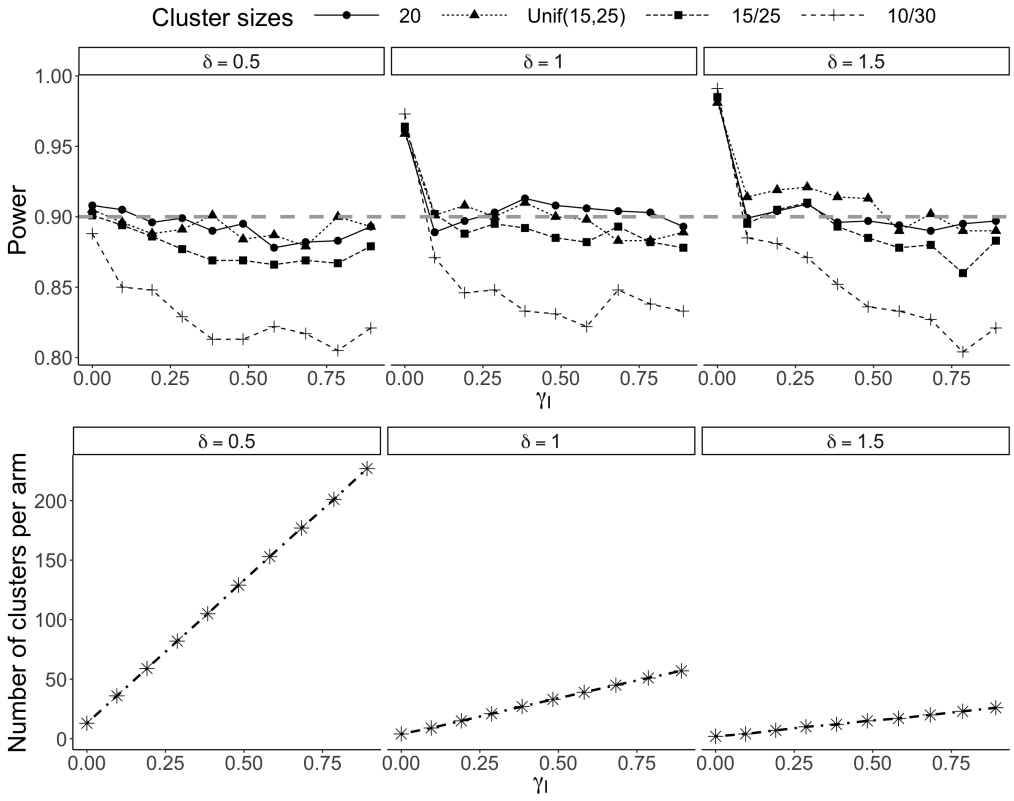
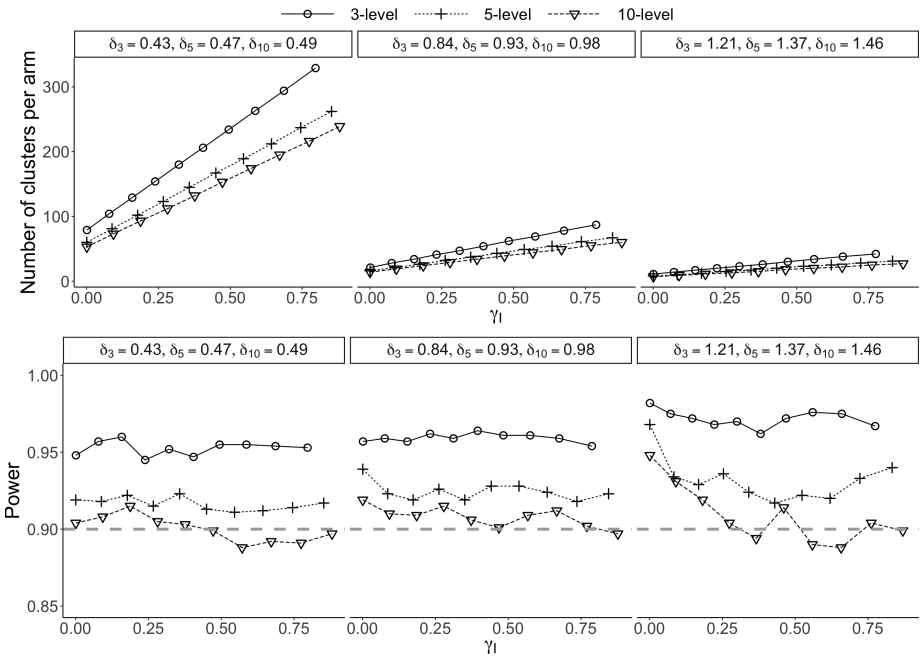
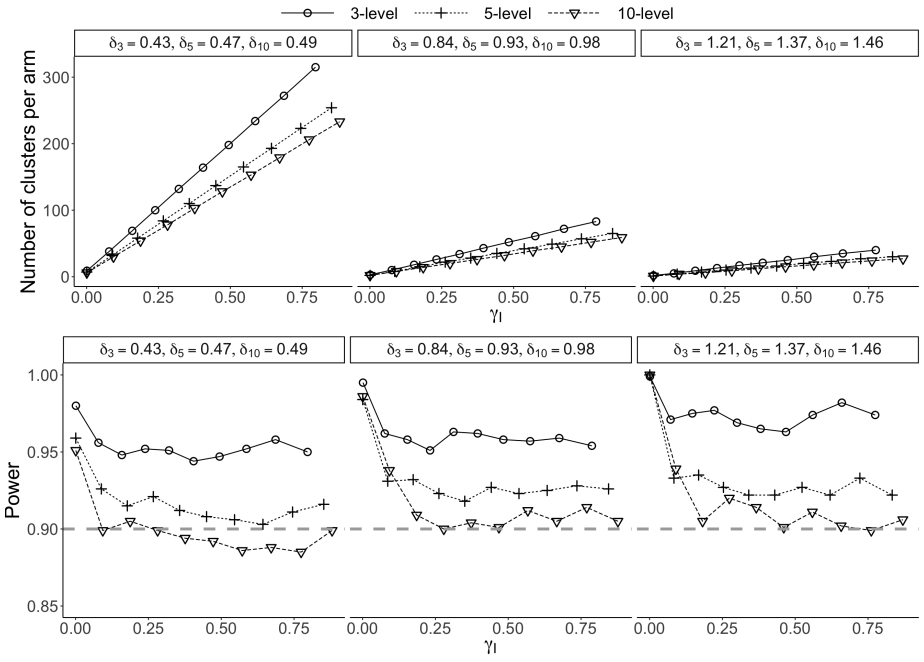


Figure 4.3: Simulation results for number of clusters per arm obtained by our calculations with predetermined equal cluster sizes of 20 and power under equal or unequal cluster sizes in actual sample data. “Unif(15,25)”, “15/25”, and “10/30” represent cluster sizes in actual sample data uniformly ranging from 15 to 25, half each of 15 and 25, and half each of 10 and 30, respectively.

Furthermore, we evaluated the performance of our sample size calculations for ordinal data. We generated clustered data of 3-level, 5-level, and 10-level ordinal variables by discretizing X_{0ij} and Y_{0kl} with cut-offs at quantiles of a standard normal distribution (i.e., using the 1/3 and 2/3 quantiles for 3 levels; the 0.2, 0.4, 0.6, 0.8 quantiles for 5 levels; and the 0.1, 0.2, ..., 0.8, 0.9 quantiles for 10 levels). The proportion of each ordinal category was analytically derived. To calculate sample sizes for each ordinal variable, we empirically computed γ_l and δ by generating a million clusters and 100 observations per cluster. The empirical values of γ_l and δ of the 3-level ordinal outcome are slightly smaller than those of the other two ordinal outcomes. In summary, our calculations had good power for ordinal data in most scenarios. The calculated number of clusters for the three ordinal variables all increased as the γ_l increased (Figure 4.4). The numbers of clusters calculated for the three ordinal variables, in descending order, are as follows: 3-level $>$ 5-level $>$ 10-level. The powers are in the same order from largest to smallest. The power of the 3-level ordinal variable was around 0.95, indicating slight overestimation in the number of clusters for this variable.



(a) cluster sizes = 5



(b) cluster sizes = 50

Figure 4.4: Simulation results for the number of clusters and power for ordinal outcomes with predetermined cluster sizes of 5 and 50. δ_3, δ_5, δ_{10} are the log ORs of the 3-level, 5-level, and 10-level ordinal outcomes, respectively.

4.5 Applications

4.5.1 A cluster randomized trial with a skewed continuous outcome

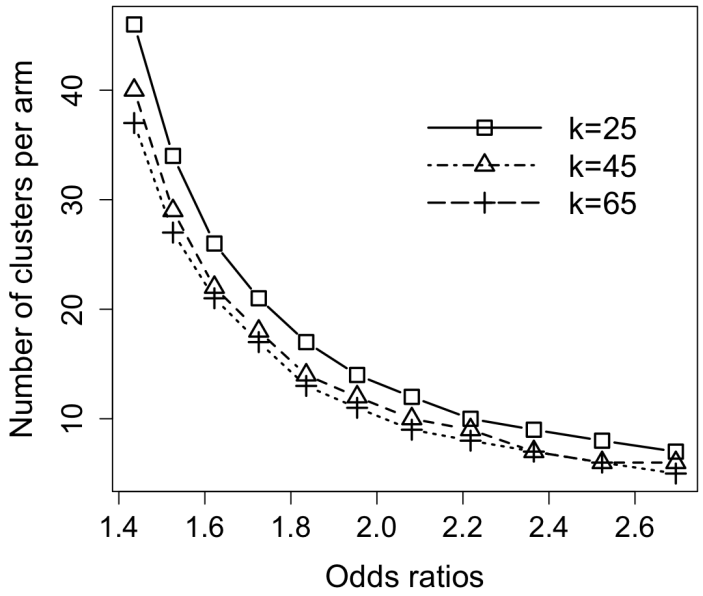
The Homens para Saúde Mais (HoPS+) study conducted a cluster randomized trial to measure the impact of a multi-component intervention on adherence to ART of pregnant women living with HIV in Zambézia Province, Mozambique (Audet et al., 2018, 2024). The primary outcome was adherence to ART, quantified as the proportion of medications taken within 1 year. This measure of adherence is pseudo-continuous ranging from 0 to 1 with a left-skewed distribution; many participants were highly adherent, while others had moderate to poor adherence. Sample size calculations for this trial were based on a simplified binary outcome (i.e., retention at 6 months) and an assumed ICC of 0.07 for this binary outcome. The study design anticipated having approximately 85% power to detect an improvement in 6-month retention from 31% to 48% (equivalent to the log OR $\delta = 2.05$) under a type I error rate of 0.05 and cluster sizes of 45. The number of clinics under this design was 24 with 12 per arm. This simplified sample size calculation, while fairly standard, was conservative, likely resulting in a larger sample size than needed.

We recalculated the sample size with our calculations but using the primary outcome without dichotomization. Since the primary outcome ranged from 0 to 1 with a left-skewed distribution and 366 possible values, the outcome can be treated either as a continuous or an ordinal variable with our sample size calculations. If the primary outcome is treated as ordinal, the proportion in each ordered category must be estimated to calculate the sample size. Now the trial is over, these proportions can be roughly estimated post-hoc using published data (Tu et al., 2024). In contrast, if the primary outcome is treated as continuous, the proportion in each ordered category is simply 1 over the total sample size, which gently simplifies calculations because it does not require preliminary estimates for the numbers in each of the 366 categories. It turns out that whether the primary outcome was considered either continuous or ordinal with proportions post-hoc estimated from the published trial, the calculated number of clinics remained the same. This is expected because the outcome is roughly continuous with 366 possible values between 0 and 1, and no single proportion was very large. Therefore, we consider the outcomes to be continuous in all of the following calculations in this subsection.

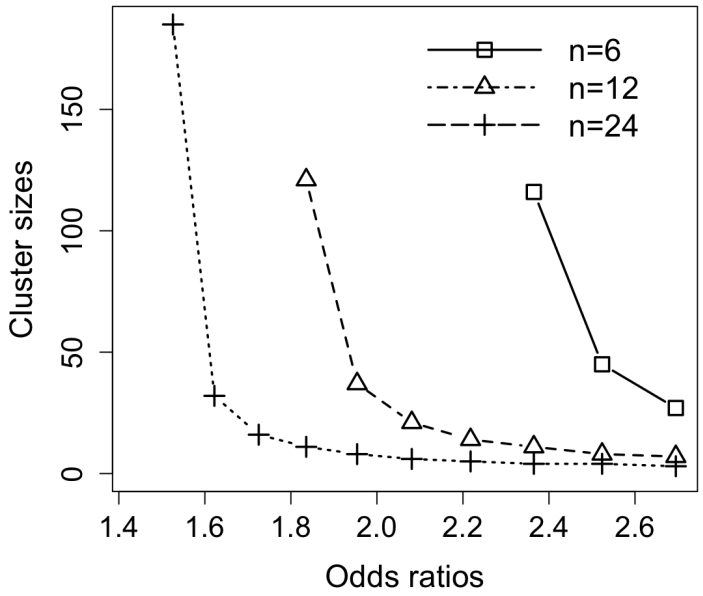
Under the same setting as the original design (power of 85%, type I error rate of 0.05, cluster sizes of 45, rank ICC of 0.07, and $\delta = 2.05$), our calculations yielded a calculated number of clinics of 10 per arm, which is smaller than that of the original design (12 per arm). Alternatively, with 12 clusters per arm to have the same power with the continuous outcome, we only would have needed clusters of size 21.

The rank ICC of 12-month adherence for women in the HoPS+ study was 0.074 (Tu et al., 2024), close to the assumed ICC of 0.07. Using the rank ICC of 0.074, we calculated the sample size across different ORs (Figures 4.5a and 4.5b). As the OR increased, the calculated sample size had an initial rapid decrease

followed by a gradual decrease. Increasing the numbers in each cluster helped to reduce the calculated number of clusters, but at some point the benefits became incremental. For example, the calculated number of clusters for $k = 45$ was slightly smaller than that for $k = 65$, whereas for $k = 25$, it was notably lower compared to $k = 45$. In addition, a limited predetermined number of clusters may hinder the detection of small treatment effects, even in cases when the rank ICC is small.



(a) Number of clinics per arm calculated with predetermined cluster sizes



(b) Cluster sizes calculated with predetermined numbers of clinics per arm

Figure 4.5: Results for the HoPS+ study example, including the number of clinics per arm calculated with predetermined cluster sizes, and cluster sizes calculated with predetermined numbers of clinics per arm across different ORs. The cluster size is denoted by k and the number of clinics per arm is denoted by n. The two-sided significant level and required power were set to 0.05 and 0.85, respectively

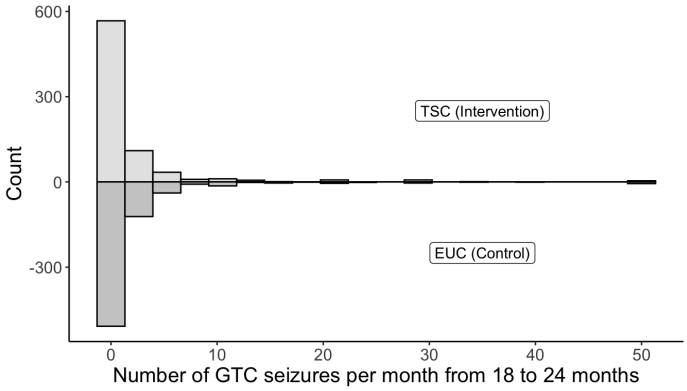
4.5.2 A non-inferiority cluster randomized clinical trial with an ordinal outcome

In the Bridging the Childhood Epilepsy Treatment Gap in Africa (BRIDGE) study, a non-inferiority randomized clinical trial was designed to understand if task-shifting childhood epilepsy treatment by trained community health workers can be as effective at reducing seizures as treatment by trained physicians (Aliyu et al., 2019). This trial recruited children with untreated epilepsy from primary healthcare centers (PHCs) in northern Nigeria. The intervention was task-shifted epilepsy care by trained community health workers (TSC). The control was enhanced usual care (EUC, referral to a physician plus primary care by an epilepsy-trained community healthcare worker). The study had general inclusion criteria which included children with many types of epilepsy. The primary outcome was whether the child had been seizure-free (yes/no) for 6 months or more at the 24-month follow-up visit. The definition of this binary outcome is standard in this setting where children with a wide variety of seizures and seizure frequencies were included. The one-sided null hypothesis was that the seizure-free rate of TSC patients (intervention) was inferior to that of EUC patients (control) by $\geq$10% (equivalent to the log OR ≥ 1.5). The type I and type II errors were set to 0.05 and 0.2, respectively, and the ICC was assumed to be 0.05. With a predetermined number of clusters at 30 per arm, the cluster size was calculated to be 19.

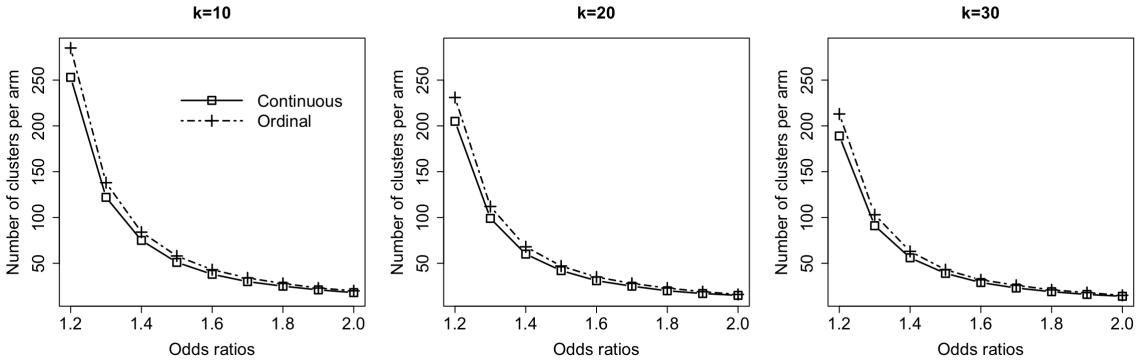
There may be interest in performing a new study to examine interventions on children with generalized tonic-clonic (GTC) seizures. In this more homogeneous population, the number of seizures in the past 6 months at the 24-month follow-up visit is a scientifically meaningful response variable, resulting in higher power than a dichotomized (0 versus > 0) response variable. We consider the same one-sided hypothesis in the design of a new cluster RCT among children with GTC seizures, with the primary outcome being the number of seizures from months 18 to 24. The BRIDGE study data among the subset of children with GTC seizures can be used as preliminary data. A histogram of GTC seizure counts between months 18 and 24 in the BRIDGE trial is given in Figure 4.6a.

Since the new primary outcome is an irregularly distributed count variable, it is reasonable to treat it as an ordinal variable. The proportion of responses for each ordered category can be easily estimated from the BRIDGE study data. One could alternatively treat the number of GTC seizures as a continuous variable for sample size calculations, where the proportion of each outcome is 1 over the total sample size (even though in the preliminary data, it only took integer values from 0 to 50). We compared the sample sizes obtained from our calculations when treating the new outcome as ordinal with the observed proportions in the BRIDGE trial (i.e., equation (4.5)) versus continuous (i.e., equation (4.7)) across different values of ORs. The one-sided significance level and required power were set to 0.05 and 0.8, respectively. The rank ICC was 0.14, estimated from the BRIDGE study. The results are shown in Figures 4.6b and 4.6c. The calculation treating

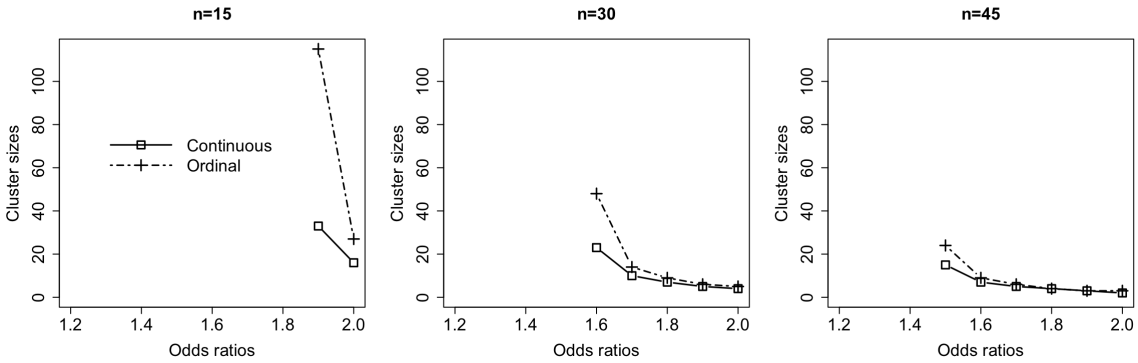
the new outcome as ordinal yielded larger sample sizes than the calculation treating the new outcome as continuous. This difference is expected as the outcome is an integer ranging from 0 to 50 with a left-skewed distribution, and the value of 0 had a very large proportion. This difference decreased as the OR increased or the predetermined number of clusters increased.



(a) Histogram of GTC seizure counts per month between 18- and 24-month visits by the two arm



(b) Number of clusters per arm calculated with predetermined cluster sizes



(c) Cluster sizes calculated with predetermined numbers of clusters per arm

Figure 4.6: Results for the BRIDGE trial example, including histogram of GTC seizure counts per month between 18- and 24-month visits, the number of clusters per arm calculated with predetermined cluster sizes, and cluster sizes calculated with predetermined numbers of clusters for different ORs. The cluster size is denoted by k and the number of clusters per arm is denoted by n. "Continuous" represents treating the outcome as continuous and "Ordinal" represents treating the outcome as ordinal. The one-sided significant level and required power were set to 0.05 and 0.8, respectively.

4.6 Discussion

In this Chapter, we propose unified and simple sample size calculations for cluster RCTs with skewed or ordinal outcomes. Our calculations involve inflating the sample size for an adequately powered individual RCT for an ordinal outcome with a design effect that incorporates the rank ICC. We show that in most scenarios, our design effect is close to the ratio of the variance of the ordinary Wilcoxon rank-sum statistic to the variance of the clustered Wilcoxon rank-sum statistic. Our calculations can be applied to compute either the number of clusters with predetermined cluster sizes or compute the cluster sizes with a predetermined number of clusters. With continuous data, we show that our calculations closely approximate those more complicated calculations based on clustered Wilcoxon rank-sum tests. Our calculations are simple and applicable to ordinal data, in contrast to those based on clustered Wilcoxon rank-sum tests are not. Furthermore, our calculations had good performance in most simulation scenarios we considered for skewed or ordinal data.

Our sample size calculations can also be applied to binary outcomes, which can be treated as ordinal with two categories. We show via simulation that sample sizes obtained by our calculations are very similar to those obtained by commonly used calculations (Hayes and Moulton, 2009) for binary outcomes. Those commonly used calculations treat $1+\rho_I(k-1)$ as the DE to inflate the sample size of an individual RCT (Hayes and Moulton, 2009). Notably, the ICC ρ_I for binary outcomes is equal to the rank ICC for binary outcomes.

In our application examples, we saw that the sample sizes using our approach for continuous outcomes tended to be lower than the sample sizes based on dichotomizing the continuous outcome. These results illustrate the well-known fact that power decreases when outcomes are dichotomized. Sample sizes based on dichotomization are more conservative (larger) than needed, which is arguably better than having underpowered, too small studies. However, it may be unethical to expose more people than needed to experimental treatments.

In the process of developing sample size formulas for clustered data, we extended the sample size calculations introduced by Whitehead (1993) for individual RCTs with ordinal outcomes to continuous outcomes. We showed via simulation that with continuous outcomes, these ordinal PO-based calculations yield sample sizes that are very close to the sample sizes obtained by more complex calculations based on Wilcoxon rank sum tests. These formulas may be useful for individual RCTs with continuous outcomes because they make minimal assumptions on the unknown distribution of the outcome.

We also conduct a comparative review of Wilcoxon rank-sum tests and PO models, along with their extension to clustered data. With independent data, unadjusted PO models with a single covariate are essentially

Wilcoxon rank-sum tests. With clustered data, unadjusted cluster PO models and clustered Wilcoxon rank-sum tests have slightly different test statistics due to different approaches used to correct for within-cluster correlation, but their p-values are close with a large number of clusters. These findings motivated our development of simple sample size calculations for cluster RCTs using PO models and a design effect, offering an alternative to more complicated calculations based on clustered Wilcoxon rank-sum tests.

Our sample size calculations have some limitations. For ordinal data with a very small number of ordered categories, our calculations may overestimate the required sample size. In such cases, the DE (i.e., $1+\gamma_I(k-1)$) may inflate the sample size of an adequately powered individual RCT beyond what is necessary. In addition, our calculations consider equal cluster sizes or an average of cluster sizes. In unbalanced designs, if the variation in unequal cluster sizes is not extreme, our calculations are applicable; otherwise, they might underestimate sample sizes. Furthermore, our calculations in (4.5) and (4.7) use the rank ICC of the entire population. The rank ICC may differ between the two arms. One may initially use preliminary data to estimate the rank ICCs of the two arms, and then calculate the sample size of each arm with its respective rank ICC. In practice, however, it is difficult to obtain precise rank ICC estimates for both arms because preliminary data are often limited and small. Moreover, our calculations require selecting effect sizes using odds ratios, which may be challenging for continuous outcomes. In such cases, using the probabilistic index to select effect sizes may be more natural. As mentioned in Section 4.2, there is a numerical relationship between the odds ratio and the probabilistic index. With this relationship, we can use the probabilistic index to select effect sizes and then use our sample size calculations with the odds ratio computed from the probabilistic index.

Future work could consider improving the sample size calculations to accommodate unbalanced designs or developing calibration approaches for ordinal outcomes with very few ordered categories.

4.7 Supplementary Materials

4.7.1 Wilcoxon rank-sum tests, Mann-Whitney U tests, and unadjusted PO models

Let n_1 and n_2 be the numbers of individuals in the control and experiment arms, respectively. Let R_i denote the rank of observation i, δ_i denote the indicator of the control arm for observation i (i.e., $\delta_i = 1$ if in the control arm and $\delta_i = 0$ otherwise), and $N = n_1 + n_2$. The Mann-Whitney U statistic is the number of times that an observation in the experiment arm precedes an observation in the control arm in the ranking,

$$W_{\text{MWU}} = n_1 n_2 + \frac{n_1(n_1+1)}{2} - \sum_{i=1}^{N} \delta_i R_i.$$

Under H_0,

$$E_0(W_{\text{MWU}}) = n_1 n_2/2,$$

$$\text{var}_0(W_{\text{MWU}}) = \frac{n_1 n_2}{(n_1+n_2)(n_1+n_2-1)}\left\{\frac{(n_1+n_2)^3-(n_1+n_2)}{12} - \sum_{i=1}^{l}\frac{k_i^3-k_i}{12}\right\},$$

where l denotes the total number of ordered categories of the outcome and k_i is the number of observations tied for a given rank i (Siegel, 1956). We have the test statistic

$$Z_{\text{MWU}} = \frac{W_{\text{MWU}} - n_1 n_2/2}{\sqrt{\text{var}_0(W_{\text{MWU}})}},$$

which is asymptotically normal with $N(0,1)$. With continuous date, $\text{var}_0(W_{\text{MWU}}) = n_1 n_2(n_1+n_2)/12$.

The Wilcoxon rank-sum statistic is

$$W_{\text{WRS}} = \sum_{i=1}^{N}\delta_i R_i.$$

Under H_0,

$$E_0(W_{\text{WRS}}) = n_1(N+1)/2,$$

$$\text{var}_0(W_{\text{WRS}}) = (n_1 n_2/12)[N+1-\sum_{i=1}^{l}(k_i^3-k_l)/[N(N-1)]].$$

We then have the test statistic

$$Z_{\text{WRS}} = \frac{W_{\text{WRS}} - 1/2}{\sqrt{\text{var}_0(W_{\text{WRS}})}} \sim N(0,1).$$

Because $W_{\text{MWU}} = n_1 n_2 + n_1(n_1+1)/2 - W_{\text{WRS}}$, we can simply show that $Z_{\text{MWU}} = Z_{\text{WRS}}$. If the outcome is continuous with no ties, then $\text{var}_0(W_{\text{WRS}})$ can be simplified to $n_1 n_2(N+1)/12$.

Whitehead (1993) introduced a statistic W_{PO} based on an unadjusted PO model for evaluating the treatment effect with ordinal data. It can be shown that

$$W_{\text{PO}} = 2W_{\text{MWU}}/(n_1+n_2+1),$$

and

$$\text{var}_0(W_{\text{PO}}) = 4\text{var}_0(W_{\text{MWU}})/(n_1+n_2+1)^2.$$

Hence, the hypothesis test based on unadjusted PO models is essentially the Mann-Whitney U/Wilcoxon rank-sum test.

The Wilcoxon rank-sum test can also be formulated in terms of the probabilistic index $\theta = P(X_i < Y_j)$, where X_i and Y_i are random variables from the control and experiment populations, respectively (Hollander and Wolfe, 1999). The hypothesis is $H_0 : \theta = 1/2$ vs. $H_1 : \theta \neq 1/2$. On this basis, the test statistic is derived

as

$$Z_{\hat{\theta}} = \frac{\hat{\theta} - 1/2}{\sqrt{\text{var}_0(\hat{\theta})}},$$

where $\hat{\theta} = W_{\text{MWU}}/(n_1 n_2)$ and $\text{var}_0(\hat{\theta}) = (N+1)/(12 n_1 n_2)$ is the variance of $\hat{\theta}$ under the null. Note that $\text{var}_0(\hat{\theta})$ is derived with an assumption that the outcome is continuous with no tie. Hence, this test may not apply to ordinal data. With continuous outcomes, we can simply show that $Z_{\text{MWU}} = Z_{\text{WRS}} = Z_{\hat{\theta}}$.

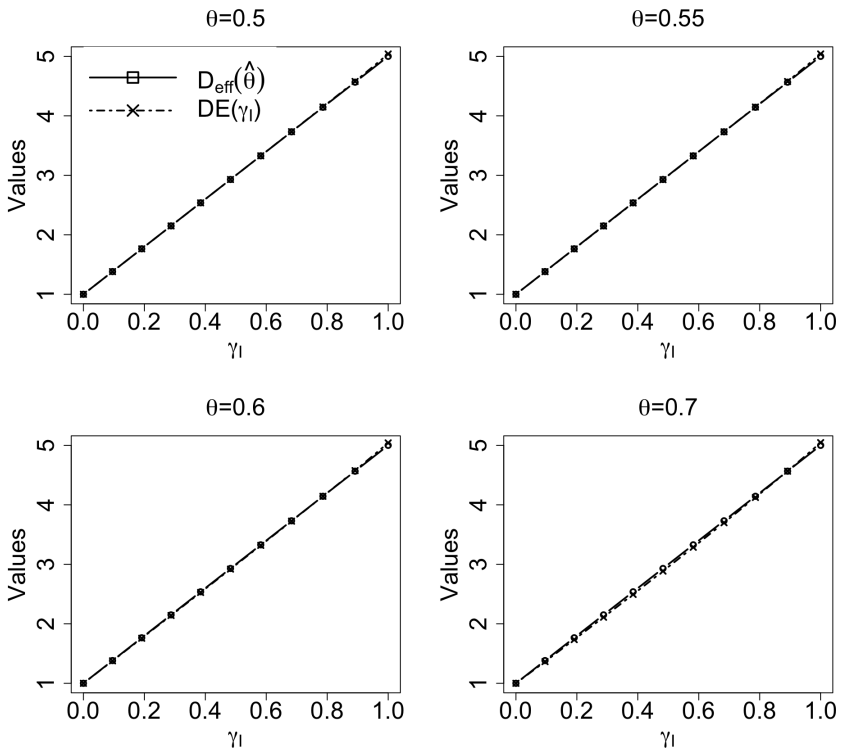
4.7.2 The design effect of cluster RCTs associated with the clustered Wilcoxon rank-sum test statistic

With continuous outcomes, Rosner and Glynn (2011) derived the variances of $\hat{\theta}$ for cluster and independent data. Then we can derive the DE of cluster RCTs associated with $\hat{\theta}$ for continuous outcomes. Let ρ denote the intraclass correlation coefficient (ICC) after the probit transformation on the cumulative distribution function. Let $Q(\theta,\rho) = \Phi_2(\Phi^{-1}(\theta),\Phi^{-1}(\theta),\rho) - \theta^2$, where $\Phi_2(\Phi^{-1}(\theta),\Phi^{-1}(\theta),\rho) = P(Z_1 \leq \Phi^{-1}(\theta), Z_2 \leq \Phi^{-1}(\theta) | (Z_1,Z_2) \sim N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix})$. The DE of cluster RCTs associated with $\hat{\theta}$ for continuous outcomes is

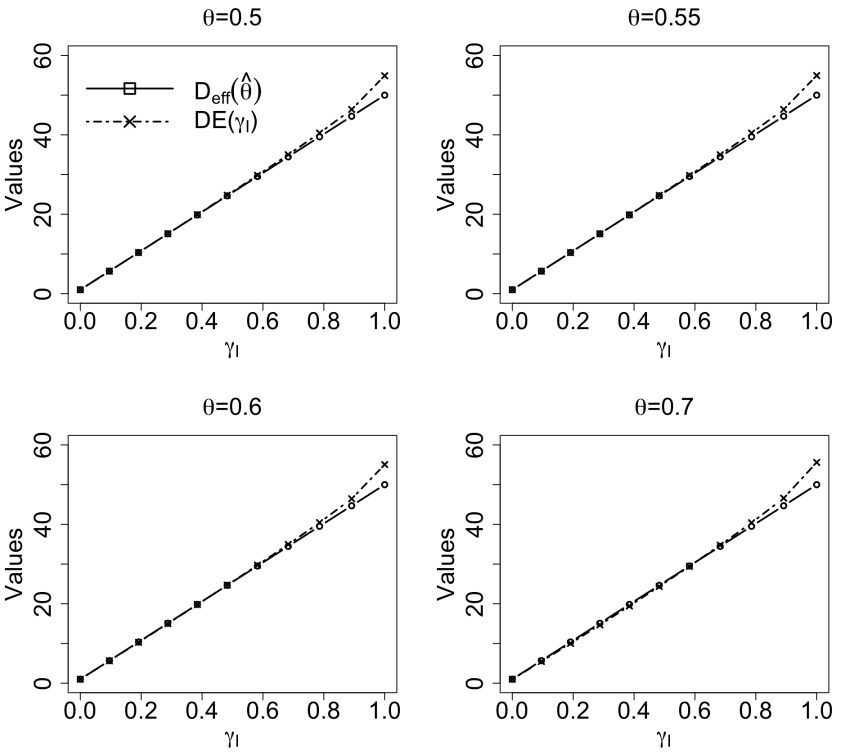
$$\begin{aligned} D_{\text{eff}}(\hat{\theta}) &= \frac{\text{var}(\hat{\theta})}{\text{var}_{\text{srs}}(\hat{\theta})} \\ &= \frac{\left\{\theta(1-\theta) + 2(k-1)Q(\theta,\frac{1+\rho}{2}) + (k-1)^2 Q(\theta,\rho)\right\}}{\left\{\theta(1-\theta) + (n_1 k + n_2 k - 2)Q(\theta,\frac{1}{2})\right\}} \\ &+ \frac{\left\{k(n_1+n_2-2)[Q(\theta,\frac{1}{2}) - (k-1)Q(\theta,\frac{\rho}{2})]\right\}}{\left\{\theta(1-\theta) + (n_1 k + n_2 k - 2)Q(\theta,\frac{1}{2})\right\}}. \end{aligned} \tag{4.10}$$

When $n_1, n_2 >> k$, $D_{\text{eff}}(\hat{\theta}) \approx 1 + (k-1)Q(\theta,\rho/2)/Q(\theta,1/2)$. When $\theta = 1/2$, $Q(\theta,\rho/2)/Q(\theta,1/2) = 6\sin^{-1}(\rho/2)/\pi$. After the probit transformation, since the distribution is normal, we have $6\sin^{-1}(\rho/2)/\pi = \gamma_I$ (Pearson, 1907), where γ_I is the rank ICC (Tu et al., 2023). Hence, under the null, if $n_1, n_2 >> k$, $D_{\text{eff}}(\hat{\theta}) \approx 1 + (k-1)\gamma_I$.

We conducted simulations to compare $D_{\text{eff}}(\hat{\theta})$ and $1 + \gamma_I(k-1)$ under different values of θ, considering scenarios where the number of clusters was smaller than the cluster size (Figure 4.7). When $\theta = 1/2$, $D_{\text{eff}}(\hat{\theta})$ is close to $1 + \gamma_I(k-1)$ even when the cluster size is larger than the number of clusters. When $\theta > 1/2$ and the number of clusters is greater than the cluster size, they are also close, except for large γ_I.



(a) 50 clusters and 5 per cluster



(b) 5 clusters and 50 per cluster

Figure 4.7: The values of $D_{\text{eff}}(\hat{\theta})$ and $1+\gamma_I(k-1)$ over different values of γ_I and θ. "DE(γ_I)" represents $1+\gamma_I(k-1)$.

4.7.3 Comparison between the conventional calculations and our calculations under normality

In this section, we analytically compare the conventional approach using the DE based on the ICC with our method under normality. Let X_{ij} denote a random variable of the jth individual in the ith cluster from the control population and Y_{kl} denote a random variable of the lth individual in the kth cluster from the experiment population, and $X_{ij} \sim N(\mu_C, \sigma_C^2)$ and $Y_{kl} \sim N(\mu_E, \sigma_E^2)$. Let n denote the total number of individuals for a cluster RCT, and $n_E = n/(A+1)$ and $n_C = An/(A+1)$ denote the number of individuals in the experiment and control arms, respectively, where A is the allocation ratio. The conventional approach based on t-tests and the ICC calculates the total sample size for cluster RCTs with a two-sided significant level at α and power at $1-\beta$ as

$$n^* = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2 (A\sigma_E^2 + \sigma_C^2)(A+1)}{A(\mu_E - \mu_C)^2}(1 + (k-1)\rho_I),$$

where ρ_I is the ICC and k is the cluster size. There is a relationship between θ and $\mu_E - \mu_C$ under normality. Because $X_{ij} - Y_{kl} \sim N(\mu_C - \mu_E, \sigma_C^2 + \sigma_E^2)$, then we have

$$\begin{aligned}
\theta &= P(X_{ij} < Y_{kl}) \\
&= P(X_{ij} - Y_{kl} < 0) \\
&= P\left(\frac{X_{ij} - Y_{kl} - (\mu_C - \mu_E)}{\sqrt{\sigma_C^2 + \sigma_E^2}} < \frac{-(\mu_C - \mu_E)}{\sqrt{\sigma_C^2 + \sigma_E^2}}\right) \\
&= P\left(Z < \frac{-(\mu_C - \mu_E)}{\sqrt{\sigma_C^2 + \sigma_E^2}}\right)
\end{aligned}$$

That is, $\mu_E - \mu_C = \Phi^{-1}(\theta)\sqrt{\sigma_C^2 + \sigma_E^2}$. As mentioned previously, $\theta = exp(\delta)[exp(\delta) - \delta - 1]/(exp(\delta) - 1)^2$ (De Neve et al., 2019). For simplicity, we denote $\theta = h(\delta)$. Then we have $\mu_E - \mu_C = \Phi^{-1}[h(\delta)]\sqrt{\sigma_C^2 + \sigma_E^2}$, and

$$n^* = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2 (A\sigma_E^2 + \sigma_C^2)(A+1)}{A\{\Phi^{-1}[h(\delta)]\}^2 (\sigma_E^2 + \sigma_C^2)}(1 + (k-1)\rho_I).$$

As described previously, with continuous outcomes, our method is

$$n = \frac{3(A+1)^2 (Z_{1-\alpha/2} + Z_{1-\beta})^2 / \delta^2}{A(1 - 1/n^2)}\{1 + \gamma_I(k-1)\}.$$

Then we have $n(1-1/n^2) = 3(A+1)^2(Z_{1-\alpha/2} + Z_{1-\beta})^2/(A\delta^2)\{1 + \gamma_I(k-1)\}$. We compare the conventional approach with our method,

$$\frac{n - 1/n}{n*} = \frac{3(A+1)^2 (Z_{1-\alpha/2} + Z_{1-\beta})^2 / (A\delta^2)\{1 + \gamma_I(k-1)\}}{\frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2 (A\sigma_E^2 + \sigma_C^2)(A+1)}{A\{\Phi^{-1}[h(\delta)]\}^2 (\sigma_E^2 + \sigma_C^2)}\{1 + (k-1)\rho_I\}}$$

$$= \frac{3(A+1)\{1+\gamma_I(k-1)\}/\delta^2}{\frac{A\sigma_E^2+\sigma_C^2}{\sigma_E^2+\sigma_C^2}\{1+(k-1)\rho_I\}/\{\Phi^{-1}[h(\delta)]\}^2}.$$

If A=1 (or $\sigma_E = \sigma_C$), then $\frac{n-1/n}{n*} = \frac{6\{1+\gamma_I(k-1)\}/\delta^2}{\{1+(k-1)\rho_I\}/\{\Phi^{-1}[h(\delta)]\}^2}$. We show via simulations that $\frac{6\{1+\gamma_I(k-1)\}/\delta^2}{\{1+(k-1)\rho_I\}/\{\Phi^{-1}[h(\delta)]\}^2} \approx 1$, except for very large δ. Hence, when $A = 1$ and δ is not very large, we can have $n \approx n^*$ since $1/n < 1$, except $n = 1$.

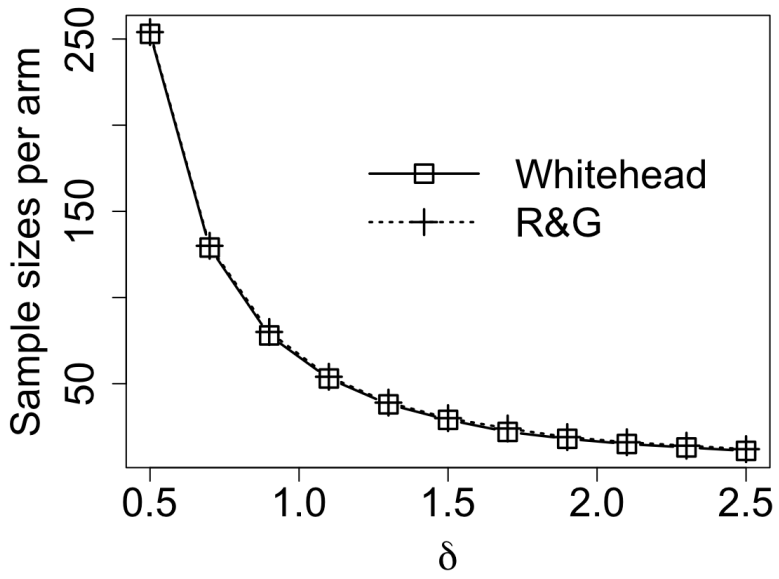


Figure 4.8: Calculated sample sizes per arm for continuous data under individual randomization across different odds ratios. "Whitehead" represents the sample size calculation (the formula (4.4)) adapted from Whitehead's sample size calculation. "R&G" represents Rosner and Glynn's sample size calculation based on ordinary Wilcoxon rank-sum tests.

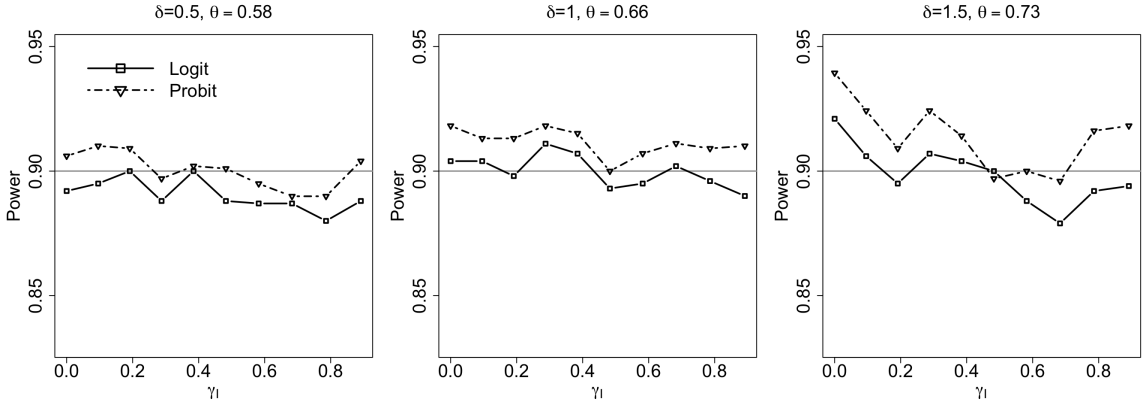


Figure 4.9: Comparison between the logit and probit links with respect to power. The cluster size was set to 5. The number of clusters was calculated based on the logit link (i.e., PO models), but data were generated such that the probit link function is correct and logit is misspecified.

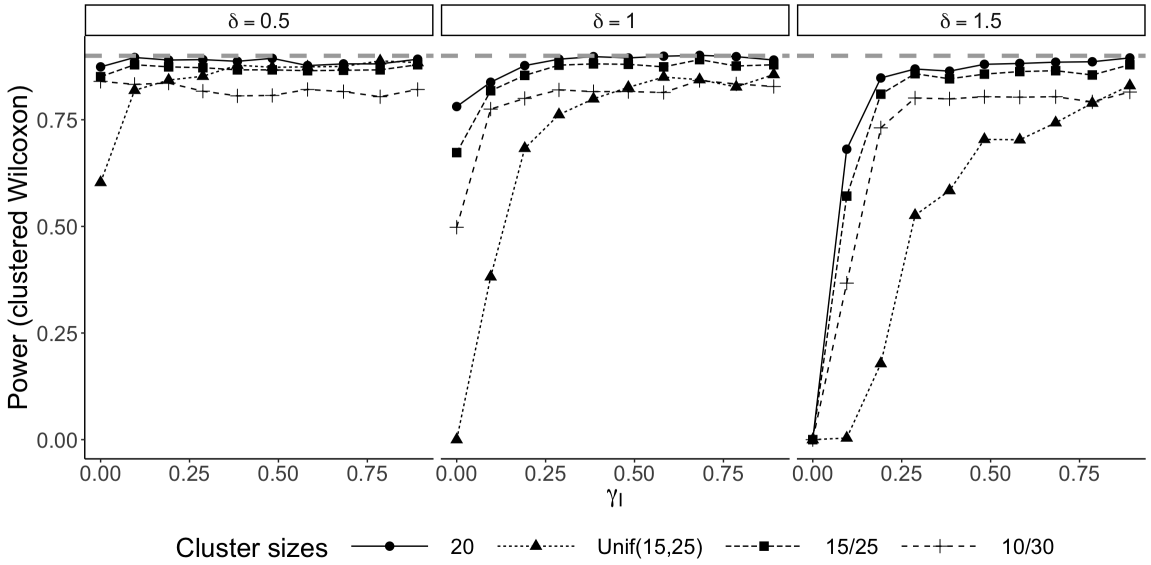


Figure 4.10: Number of clusters per arm calculated with predetermined equal cluster sizes of 20 and power of clustered Wilcoxon rank-sum tests under equal or unequal cluster sizes of the actual sample data.

CHAPTER 5

Conclusion

5.1 Summary

Cluster data are common in practice and often are skewed, ordinal, or mixtures of the two. Conventional approaches (i.e., the ICC and Pearson correlations), while frequently used, are inadequate to analyze data of these types. The overall goal of this dissertation has been to develop innovative rank-based methods that provide more robust and accurate analyses and designs for clustered data.

In Chapter 2, we defined the rank ICC as a natural extension of Fisher's ICC to the rank scale, and described its population parameter. Our approach maintains the spirit of Fisher's ICC while creating a non-parametric rank ICC measure analogous to Spearman's rank correlation. The rank ICC is simply interpreted as the rank correlation between a pair of observations from the same cluster. We also extended the rank ICC for distributions with more than two hierarchies. Our estimator of the rank ICC is insensitive to extreme values and skewed distributions, and does not depend on the scale of the data. It is also consistent and asymptotically normal, with low bias and good coverage in our simulations. Furthermore, we discussed assigning weights to clusters and observations under different cases when estimating the rank ICC for two-level data with heterogeneous cluster sizes. This work is accompanied by a new R package, `rankICC`, available on CRAN.

Chapter 3 introduces the population parameters of the between- and within-cluster Spearman rank correlations, which are natural extensions of the between- and within-cluster Pearson correlations to the rank scale. We also show their approximated relationship with the total Spearman rank correlation and rank ICCs of the two variables. Compared with traditional Pearson correlations, our method is insensitive to extreme values and skewed distributions, and does not depend on the scale of the data. Our framework is general, and is applicable to any orderable variables. Our estimators are asymptotically normal, with generally low bias and good coverage in our simulations. A developed R package, `rankCorr`, available on CRAN, facilitates the implementation of the new method.

In Chapter 4, we extend the use of the rank ICC to designing cluster RCTs with skewed or ordinal outcomes, proposing unified and simple sample size calculations. Our calculations involve inflating the sample size for an adequately powered individual RCT for an ordinal outcome with a design effect that incorporates the rank ICC. For continuous outcomes, our calculations set the number of distinct ordinal levels to the sample size. We show that in most scenarios, our design effect is close to the ratio of the

variance of the ordinary Wilcoxon rank-sum statistic to the variance of the clustered Wilcoxon rank-sum statistic. We also show that with continuous data, our calculations closely approximate more complicated calculations based on clustered Wilcoxon rank-sum tests. In addition, our calculations can be applied to compute either the number of clusters with predetermined cluster sizes or compute the cluster sizes with a predetermined number of clusters. Our calculations are simple and applicable to ordinal data, whereas those based on clustered Wilcoxon rank-sum tests are not. Furthermore, our calculations had good performance in most scenarios of the simulations for skewed or ordinal data.

We hope that our work provides new directions and tools for researchers in the analyses and designs for clustered data.

5.2 Future Research

Future work could consider covariate-adjusted Spearman rank correlations. For example, in the application of CD4 and CD8 data in Chapter 3, there may be interest in measuring the rank correlations after adjusting for age. Our methods proposed in Chapter 3 could be extended to allow for covariate adjustment by fitting CPMs that include the covariate, in addition to the cluster indicators. We suspect that the correlation between probability-scale residuals from these fitted models could be used to estimate covariate-adjusted within-cluster Spearman rank correlations and that the correlation between cluster indicator coefficients from these fitted models could be used to estimate covariate-adjusted between-cluster Spearman rank correlations. This approach is somewhat similar to random effects approaches used for estimating covariate-adjusted within- and between-cluster Pearson correlations (Ferrari et al., 2005). Further investigation is warranted into such an approach, as well as Spearman rank correlation as a function of time with longitudinal data.

Unbalanced data are common in cluster RCTs. If the variation in unequal cluster sizes is not extreme, our sample size calculations proposed in Chapter 4 are applicable; otherwise, they might underestimate sample sizes. Future work could focus on improving our sample size calculations to accommodate unbalanced designs.

Our sample size calculations may overestimate sample sizes for ordinal data with a very small number of ordered categories. In such cases, our design effect may inflate the sample size of an adequately powered individual RCT beyond what is necessary. We could develop calibration approaches for calculating sample sizes for ordinal outcomes with a very small number of ordered categories.

References

Agresti, A. and Natarajan, R. (2001). Modeling clustered ordered categorical data: A survey. *Int Stat Rev*, 69(3):345–371.

Aliyu, M. H., Abdullahi, A. T., Iliyasu, Z., Salihu, A. S., Adamu, H., Sabo, U., et al. (2019). Bridging the childhood epilepsy treatment gap in northern nigeria (BRIDGE): Rationale and design of pre-clinical trial studies. *Contemp Clin Trials Commun*, 15:100362.

Audet, C. M., Graves, E., Barreto, E., De Schacht, C., Gong, W., Shepherd, B. E., et al. (2018). Partners-based HIV treatment for seroconcordant couples attending antenatal and postnatal care in rural mozambique: A cluster randomized trial protocol. *Contemp Clin Trials Commun*, 71:63–69.

Audet, C. M., Graves, E., Shepherd, B. E., Prigmore, H. L., Brooks, H. L., Emilio, A., Matino, A., Paulo, P., Diemer, M. A., Frisby, M., Sack, D. E., Aboobacar, A., Barreto, E., Van Rompaey, S., and De Shacht, C. (2024). Partner-based hiv treatment for seroconcordant couples attending antenatal and postnatal care in rural mozambique: a cluster randomized controlled trial. *Journal of Acquired Immune Deficiency Syndromes*, page (in press).

Bross, I. D. J. (1958). How to use ridit analysis. *Biometrics*, 14:18–38.

Campbell, M. and Walters, S. (2014). *How to Design, Analyse and Report Cluster Randomised Trials in Medicine and Health Related Research*. Wiley, Chichester.

Castilho, J., Shepherd, B., Koethe, J., Turner, M., Bebawy, S., Logan, J., Rogers, W., Raffanti, S., and Sterling, T. (2016). Cd4+/cd8+ ratio, age, and risk of serious noncommunicable diseases in hiv-infected adults on antiretroviral therapy. *AIDS*, 30(6):899–908.

Chakraborty, H., Solomon, N., and Anstrom, K. J. (2021). A method to estimate intra-cluster correlation for clustered categorical data. *Commun Stat Theory Methods*, 0(0):1–18.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

De Neve, J. D., Thas, O., and Gerds, T. A. (2019). Semiparametric linear transformation models: Effect measures, estimators, and applications. *Statistics in Medicine*, 38(8):1484–1501.

Denham, B. E. (2016). *Categorical Statistics for Communication Research*. John Wiley and Sons, Ltd.

Donner, A. (1986). A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. *Int Stat Rev*, 54(1):67–82.

Donner, A., Birkett, N., and Buck, C. (1981). Randomization by cluster-sample size requirements and analysis. *American Journal of Epidemiology*, 114(6):906–914.

Donner, A. and Klar, N. (2000). *Design and Analysis of Cluster Randomization Trials in Health Research*. John Wiley & Sons.

Eldridge, S. M. and Kerry, S. M. (2012). *A Practical Guide to Cluster Randomised Trials in Health Services Research*. John Wiley & Sons.

Ferrari, P., Al-Delaimy, W. K., Slimani, N., Boshuizen, H. C., Roddam, A., Orfanos, P., Skeie, G., Rodríguez-Barranco, M., Thiebaut, A., Johansson, G., Palli, D., Boeing, H., Overvad, K., and Riboli, E. (2005). An Approach to Estimate Between- and Within-Group Correlation Coefficients in Multicenter Studies: Plasma Carotenoids as Biomarkers of Intake of Fruits and Vegetables. *American Journal of Epidemiology*, 162(6):591–598.

Field, C. A. and Welsh, A. H. (2007). Bootstrapping clustered data. *J R Stat Soc Series B Stat Methodol*, 69(3):369–390.

Fieller, E. and Smith, C. (1951). Note on the analysis of variance and intraclass correlation. *Ann Eugen*, 16:97–104.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd: Edinburgh.

Hallgren, K. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutor Quant Methods Psychol*, 8:23–34.

Harrell, F. E. (2015). *rms: Regression Modeling Strategies*. R package version 4.2–1.

Harris, J. A. (1913). 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.

Hayes, R. J. and Moulton, L. H. (2009). *Cluster Randomised Trials*. Chapman & Hall/CRC.

Heagerty, P. J. and Zeger, S. L. (1996). Marginal regression models for clustered ordinal measurements. *J Am Stat Assoc*, 91(435):1024–1036.

Hedges, L. V. and Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educ Eval Policy Anal*, 29(1):60–87.

Hollander, M. and Wolfe, D. A. (1999). *Nonparametric Statistical Methods*. New York: John Wiley & Sons, Inc.

Hunsberger, S., Long, L., Reese, S. E., Hong, G. H., Myles, I. A., Zerbe, C. S., Chetchotisakd, P., and Shih, J. H. (2022). Rank correlation inferences for clustered data with small sample size. *Statistica Neerlandica*, 76(3):309–330.

Karlin, S., Cameron, E. C., and Williams, P. T. (1981). Sibling and parent–offspring correlation estimation with variable family size. *Proc Natl Acad Sci U S A*, 78(5):2664–2668.

Kendall, M. (1970). *Rank Correlation Methods*. Charles Griffin.

Kim, H. Y., Williamson, J. M., and Lyles, C. M. (2005). Sample-size calculations for studies with correlated ordinal outcomes. *Statistics in Medicine*, 24:2977–2987.

Kish, L. (1965). *Survey Sampling*. Wiley.

Kish, L. (1987). Weighting in deft. *The Survey Statistician*.

Koo, T. K. and Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*, 15(2):155–163.

Kruskal, W. H. (1958). Ordinal measures of association. *J Am Stat Assoc*, 53(284):814–861.

Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. San Francisco: Holden-Day.

Li, C. and Shepherd, B. E. (2012). A new residual for ordinal outcomes. *Biometrika*, 99(2):473–480.

Liu, Q., Li, C., Wanga, V., and Shepherd, B. E. (2018). Covariate-adjusted spearman's rank correlation with probability-scale residuals. *Biometrics*, 74:595–605.

Liu, Q., Shepherd, B. E., Li, C., and Harrell Jr., F. E. (2017). Modeling continuous response variables using ordinal regression. *Statistics in Medicine*, 36(27):4316–4335.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18(1):50–60.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B*, 43:109–142.

Murray, D. M., Varnell, S. P., and Blitstein, J. L. (2004). Design and analysis of group-randomized trials: a review of recent methodological developments. *Am J Public Health*, 94(3):423–432.

Nakagawa, S., Johnson, P. C. D., and Schielzeth, H. (2017). The coefficient of determination r2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *J R Soc Interface*, 14:20170213.

Parsons, N. R., Edmondson, R., and Gilmour, S. (2006). A generalized estimating equation method for fitting autocorrelated ordinal score data with an application in horticultural research. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 55:507–524.

Pearson, K. G. (1907). *On Further Methods of Determining Correlation*. Cambridge University Press.

Rosner, B. and Glynn, R. (2011). Power and sample size estimation for the clustered wilcoxon test. *Biometrics*, 67:646–653.

Rosner, B. and Glynn, R. J. (2009). Power and sample size estimation for the wilcoxon rank sum test with application to comparisons of c statistics from alternative prediction models. *Biometrics*, 65(1):188–197.

Rosner, B. and Glynn, R. J. (2017). Estimation of rank correlation for clustered data. *Statistics in Medicine*, 36(14):2163–2186.

Rosner, B., Glynn, R. J., and Lee, M.-L. T. (2003). Incorporation of clustering effects for the wilcoxon rank sum test: A large sample approach. *Biometrics*, 59:1089–1098.

Rothery, P. (1979). A nonparametric measure of intraclass correlation. *Biometrika*, 66(3):629–639.

Rutterford, C., Copas, A., and Eldridge, S. (2015). Methods for sample size determination in cluster randomized trials. *International Journal of Epidemiology*, 44(3):1051–1067.

Shepherd, B., Li, C., and Liu, Q. (2016). Probability-scale residuals for continuous, discrete, and censored data. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 44(4):463–479.

Shih, J. H. and Fay, M. P. (2017). Pearson's chi-square test and rank correlation inferences for clustered data. *Biometrics*, 73(3):822–834.

Shirahata, S. (1981). Intraclass rank tests for independence. *Biometrika*, 68(2):451–456.

Shirahata, S. (1982). Nonparametric measures of intraclass correlation. *Commun Stat Theory Methods*, 11:1707–1721.

Shrout, P. and Fleiss, J. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*, 86(2):420—428.

Siddiqui, O., Hedeker, D., Flay, B. R., and Hu, F. B. (1996). Intraclass correlation estimates in a school-based smoking prevention study. outcome and mediating variables, by sex and ethnicity. *Am J Epidemiol*, 144 4:425–33.

Siegel, S. (1956). *Nonparametric Statistics for the Behavioural Sciences*. New York: McGraw-Hill.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*. Chapman & Hall/CRC.

Snijders, T. and Bosker, R. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. London: Sage Publishers.

Stefanski, L. A. and Boos, D. D. (2002). The calculus of m-estimation. *The American Statistician*, 56(1):29–38.

Tian, Y., Shepherd, B. E., Li, C., Zeng, D., and Schildcrout, J. S. (2023). Analyzing clustered continuous response variables with ordinal regression models. *Biometrics*, 79(4):3764–3777.

Tu, S., Li, C., and Shepherd, B. E. (2024). Between- and within-cluster spearman rank correlations.

Tu, S., Li, C., Zeng, D., and Shepherd, B. E. (2023). Rank intraclass correlation for clustered data. *Statistics in Medicine*, 42(24):4333–4348.

Whitehead, J. (1993). Sample size calculations for ordered categorical data. *Statistics in Medicine*, 12(24):2257–2271.

Wudil, U., Aliyu, M., Prigmore, H., Ingles, D., Ahonkhai, A., Musa, B., et al. (2021). Apolipoprotein-1 risk variants and associated kidney phenotypes in an adult hiv cohort in nigeria. *Kidney Int*, 100:146–154.