# The Split Analysis for Multiple-reader Multiple-case Split-plot Studies

by Jui-Ying Hsieh

M.A. in Economics, July 2015, National Taiwan University

A Thesis submitted to

The Faculty of
The Columbian College of Arts and Sciences
of The George Washington University
in partial fulfillment of the requirements
for the degree of Master of Science

May 20, 2018

Thesis directed by

Huixia Judy Wang
Professor of Statistics

Brandon D. Gallas
Applied Mathematician, U.S. Food and Drug Administration

Abstract of Thesis

The Split Analysis for Multiple-reader Multiple-case Split-plot Studies

One pathway for a new device to gain access to the marketplace requires demonstration that it is equivalent to, or substantially better than, a legally marketed device. To evaluate the equivalence of a medical imaging device, we propose measuring the intra- or inter-reader agreement in a reader study, where the clinicians (readers) make diagnoses on the medical images (cases) using both the new and old imaging devices. Such an endpoint, as well as its variance estimate, enable us to make a statistical inference on the equivalence of two devices. A method for multiple-reader multiple-case agreement analysis was presented in Gallas et al. (2016) for fully-crossed study designs, where every reader reads every case. In practice, having every reader read every case may be impossible when readers have a limited amount of time to participate in the study. One alternative study design is the split-plot study design, where both the readers and the cases are partitioned into a fixed number of groups, and each group of readers reads its own group of cases. In this thesis, we adapt the multiple-reader multiple-case agreement analysis method in Gallas et al. (2016) to analyze split-plot study designs, and propose a new variance estimator based on splitting the analysis across the groups. In each split sub-study, we compute an estimate, and then combine these estimates to obtain the final estimate for the full study. Our numerical studies show that the "split-analysis" variance estimator provides more accurate estimation of the variance of concordance measurements than the full-study-based method for unbalanced split-plot study designs.

Table of Contents

List of Figures

List of Tables

# Chapter 1

# Introduction

## 1.1 Multiple-reader Multiple-case Study

Cross-correlated data arise from experiments in many fields where the measures of interest depend on the combination of factors from different populations. Multiple-Reader Multiple-Case (MRMC) medical imaging studies are one kind of experiment that results in cross-correlated data. In these studies, we have a collection of pathologists (readers) reading the same set of patients' pathology images (cases). Both the readers and the patients' pathology images are samples from the corresponding target populations (e.g., a sample of certified pathologists and a sample of patients' pathology images undergoing evaluations of suspected lesions). Pathologists give ratings to the images. A rating is an ordinal number (e.g., integers from 1 to 100); usually a suspicion score. It indicates the level of suspicion a reader considers the case is diseased. Pathologists give ratings using both the optical microscope and the whole slide imaging (WSI). WSI is a digital scan of an optical image of a real physical object (a section of stained tissue on a glass slide). The analytical goals for this experiment are to measure the agreement of the ratings from the microscope

and those from the WSI. Since human observers naturally have different abilities and experience, there is variability in readers' performance levels. Multiple-reader analyses provide an efficient way to make inferences on the agreement of the two imaging devices accounting for variability due to readers and cases. To highlight the two important sources of variability, these multiple-reader analyses are frequently called Multiple-Reader Multiple-Case (MRMC) analyses.

In medical imaging, modality is a type of equipment used to acquire structural or functional images of the body, such as radiography, ultrasound, computed tomography and magnetic resonance imaging. In this thesis we are comparing pathologists using glass slides on the optical microscope (the reference modality) to pathologists using digital whole slide images (WSIs) on a computer display (the new modality).

The benefits of MRMC studies are addressed in Wagner et al. (2002). One of them is that the results and analyses of the study can be generalized to the population of readers and the population of cases. In other words, the MRMC analysis of agreement studies generalizes to what is expected when that experiment is replicated many times, each time drawing independent samples of readers and cases from their respective populations.

## 1.2 Motivation

In the previous medical imaging example, while the microscope is widely used as the means of reviewing histopathology for anatomic diagnosis, WSI is believed to have the potential to replace microscope. Potential benefits include: (1) improvement of efficiency; (2) cost-effectiveness; (3) accessibility to high-quality pathology review; (4) potential improvement diagnostic accuracy; (5) reproducibility. Pathologists who want to see the widespread adoption of the WSI systems may question whether WSI is adequate for diagnosis.

To address this question, Gallas et al. (2016) presented a method that can test

the non-inferiority between the modalities under comparison. This method gives estimates for agreement, as well as its variance estimate, for conducting statistical inference. However, this method only works for MRMC fully-crossed study designs, where every reader reads every case. In this study, we aim to adapt the method in Gallas et al. (2016) to analyze split-plot study designs, where each reader reads just a subset of the cases. We propose a new variance estimator based on the "split analysis"; the idea of the split analysis is to split the study into sub-studies, and combine the estimates from each sub-study to give the final estimate for the full study. Our numerical studies show that this estimator works well for split-plot study designs.

## 1.3 MRMC Study Designs

Several MRMC study designs for imaging studies have been proposed (Obuchowski et al., 2012), but the most common design to date is the fully-crossed MRMC study design. In fully-crossed MRMC study designs, every reader reads every case. Consider a fully-crossed study design where there are $n_C$ cases, $n_R$ readers, and two modalities $A$ and $B$. Let $X_{mrc}$ denote the rating from a reader $r$ for case $c$ using modality $m$, where $m = A$ or $B$. For example, $X_{A21}$ can be interpreted as the rating from reader 2 for case 1 using modality $A$. Table 1.1 presents the data layout for this fully-crossed study design.

Table 1.1: Data layout for a fully-crossed study design

| | Reader 1 | | Reader 2 | | ... | | Reader $n_R$ | |
|---|---|---|---|---|---|---|---|---|
| Modality | $A$ | $B$ | $A$ | $B$ | ... | ... | $A$ | $B$ |
| Case 1 | $X_{A11}$ | $X_{B11}$ | $X_{A21}$ | $X_{B21}$ | ... | ... | $X_{An_R1}$ | $X_{Bn_R1}$ |
| Case 2 | $X_{A12}$ | $X_{B12}$ | $X_{A22}$ | $X_{B22}$ | ... | ... | $X_{An_R2}$ | $X_{Bn_R2}$ |
| Case 3 | $X_{A13}$ | $X_{B13}$ | $X_{A23}$ | $X_{B23}$ | ... | ... | $X_{An_R3}$ | $X_{Bn_R3}$ |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Case $n_C$ | $X_{A1n_C}$ | $X_{B1n_C}$ | $X_{A2n_C}$ | $X_{B2n_C}$ | ... | ... | $X_{An_Rn_C}$ | $X_{Bn_Rn_C}$ |

The agreement analysis is "most" statistically powerful with fully-crossed study designs (Obuchowski, 2009). However, one drawback of fully-crossed study designs is that a large number of readings is required for each reader. For a study with 400 cases and two modalities, each reader must interpret 800 images. If each image requires an average of three minutes to read, each reader needs an entire week to participate in the study.

The notion that fully-crossed study designs are most powerful is only true when the sample sizes (number of readers and cases) are the only resources considered to achieve statistical power. In practice, the workload of the readers is oftentimes an important cost for the studies.

Alternative MRMC study designs are therefore proposed, and one of them is the split-plot MRMC study design. In split-plot study designs, readers read cases using both the reference and new modalities, but each reader reads just a subset of the cases. Specifically, in split-plot study designs, readers and cases are randomized into one of $n_G$ groups. In each group, all the readers read all of the cases in their group. Given a fixed number of readers and cases, split-plot study designs require much fewer observations than fully-crossed study designs. Chen et al. (2018) showed that when

both sample size and reader's workload are considered, split-plot study designs can be more cost-effective. Split-plot study designs make efficient use of cases, reader's time, and the total number of observations of a study: each case is read by multiple readers to reduce the noise from a single observation, and each case is read by just a proportion of study readers to avoid diminishing returns from adding too many readers.

For split-plot study designs, there are many possible configurations depending on the number of readers $n_R$, number of cases $n_C$, and the number of groups $n_G$. In this work, we will discuss and investigate two types of split-plot study designs: the balanced split-plot study design and the unbalanced split-plot study design. In balanced split-plot designs, we have an equal number of readers and equal number of cases in each group. For instance, a split-plot MRMC study design of six readers and nine cases with three groups is a balanced split-plot MRMC study design; see Table 1.2 for the corresponding reading data layout.

Table 1.2: An example data layout for a balanced split-plot study design

| | Group 1 | | | | Group 2 | | | | Group 3 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Reader 1 | | Reader 2 | | Reader 3 | | Reader 4 | | Reader 5 | | Reader 6 | |
| Modality | $A$ | $B$ | $A$ | $B$ | $A$ | $B$ | $A$ | $B$ | $A$ | $B$ | $A$ | $B$ |
| Case 1 | $X_{A11}$ | $X_{B11}$ | $X_{A21}$ | $X_{B21}$ | | | | | | | | |
| Case 2 | $X_{A12}$ | $X_{B12}$ | $X_{A22}$ | $X_{B22}$ | | | | | | | | |
| Case 3 | $X_{A13}$ | $X_{B13}$ | $X_{A23}$ | $X_{B23}$ | | | | | | | | |
| Case 4 | | | | | $X_{A34}$ | $X_{B34}$ | $X_{A44}$ | $X_{B44}$ | | | | |
| Case 5 | | | | | $X_{A35}$ | $X_{B35}$ | $X_{A45}$ | $X_{B45}$ | | | | |
| Case 6 | | | | | $X_{A36}$ | $X_{B36}$ | $X_{A46}$ | $X_{B46}$ | | | | |
| Case 7 | | | | | | | | | $X_{A57}$ | $X_{B57}$ | $X_{A67}$ | $X_{B67}$ |
| Case 8 | | | | | | | | | $X_{A58}$ | $X_{B58}$ | $X_{A68}$ | $X_{B68}$ |
| Case 9 | | | | | | | | | $X_{A59}$ | $X_{B59}$ | $X_{A69}$ | $X_{B69}$ |

In unbalanced split-plot study designs, we have either an unequal number of readers, or/and cases in each group. Table 1.3 presents an example of the reading data layout for an unbalanced split-plot study design, which involves seven cases, six readers and three groups. In this example, three cases are involved in each of the first two groups, while only one case is involved in the third group, so the design is unbalanced.

Table 1.3: An example data layout for an unbalanced split-plot study design

| | Group 1 | | | | Group 2 | | | | Group 3 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Reader 1 | | Reader 2 | | Reader 3 | | Reader 4 | | Reader 5 | | Reader 6 | |
| Modality | $A$ | $B$ | $A$ | $B$ | $A$ | $B$ | $A$ | $B$ | $A$ | $B$ | $A$ | $B$ |
| Case 1 | $X_{A11}$ | $X_{B11}$ | $X_{A21}$ | $X_{B21}$ | | | | | | | | |
| Case 2 | $X_{A12}$ | $X_{B12}$ | $X_{A22}$ | $X_{B22}$ | | | | | | | | |
| Case 3 | $X_{A13}$ | $X_{B13}$ | $X_{A23}$ | $X_{B23}$ | | | | | | | | |
| Case 4 | | | | | $X_{A34}$ | $X_{B34}$ | $X_{A44}$ | $X_{B44}$ | | | | |
| Case 5 | | | | | $X_{A35}$ | $X_{B35}$ | $X_{A45}$ | $X_{B45}$ | | | | |
| Case 6 | | | | | $X_{A36}$ | $X_{B36}$ | $X_{A46}$ | $X_{B46}$ | | | | |
| Case 7 | | | | | | | | | $X_{A57}$ | $X_{B57}$ | $X_{A67}$ | $X_{B67}$ |

## 1.4 Agreement in MRMC Studies

Table 1.4: Intra-modality agreement for a set of readers

| Intra-modality Agreement | | Reference Modality $A$ | | | | |
|---|---|---|---|---|---|---|
| | | Reader 1 | Reader 2 | Reader 3 | ... | Reader $n_R$ |
| Reference Modality $A$ | Reader 1 | $\widehat{\mathcal{C}}_{1}^{AA}$ | $\widehat{\mathcal{C}}_{12}^{AA}$ | $\widehat{\mathcal{C}}_{13}^{AA}$ | ... | $\widehat{\mathcal{C}}_{1n_R}^{AA}$ |
| | Reader 2 | $\widehat{\mathcal{C}}_{12}^{AA}$ | $\widehat{\mathcal{C}}_{2}^{AA}$ | $\widehat{\mathcal{C}}_{23}^{AA}$ | ... | $\widehat{\mathcal{C}}_{2n_R}^{AA}$ |
| | Reader 3 | $\widehat{\mathcal{C}}_{13}^{AA}$ | $\widehat{\mathcal{C}}_{23}^{AA}$ | $\widehat{\mathcal{C}}_{3}^{AA}$ | ... | $\widehat{\mathcal{C}}_{3n_R}^{AA}$ |
| | ... | ... | ... | ... | ... | ... |
| | Reader $n_R$ | $\widehat{\mathcal{C}}_{1n_R}^{AA}$ | $\widehat{\mathcal{C}}_{2n_R}^{AA}$ | $\widehat{\mathcal{C}}_{3n_R}^{AA}$ | ... | $\widehat{\mathcal{C}}_{n_R}^{AA}$ |

Table 1.5: Inter-modality agreement for a set of readers

| Inter-modality Agreement | | Reference Modality $A$ | | | | |
|---|---|---|---|---|---|---|
| | | Reader 1 | Reader 2 | Reader 3 | ... | Reader $n_R$ |
| New Modality $B$ | Reader 1 | $\widehat{\mathcal{C}}_{1}^{AB}$ | $\widehat{\mathcal{C}}_{12}^{AB}$ | $\widehat{\mathcal{C}}_{13}^{AB}$ | ... | $\widehat{\mathcal{C}}_{1n_R}^{AB}$ |
| | Reader 2 | $\widehat{\mathcal{C}}_{12}^{AB}$ | $\widehat{\mathcal{C}}_{2}^{AB}$ | $\widehat{\mathcal{C}}_{23}^{AB}$ | ... | $\widehat{\mathcal{C}}_{2n_R}^{AB}$ |
| | Reader 3 | $\widehat{\mathcal{C}}_{13}^{AB}$ | $\widehat{\mathcal{C}}_{23}^{AB}$ | $\widehat{\mathcal{C}}_{3}^{AB}$ | ... | $\widehat{\mathcal{C}}_{3n_R}^{AB}$ |
| | ... | ... | ... | ... | ... | ... |
| | Reader $n_R$ | $\widehat{\mathcal{C}}_{1n_R}^{AB}$ | $\widehat{\mathcal{C}}_{2n_R}^{AB}$ | $\widehat{\mathcal{C}}_{3n_R}^{AB}$ | ... | $\widehat{\mathcal{C}}_{n_R}^{AB}$ |

In an MRMC study with two modalities, the following types of agreements may be of interest (Gallas et al., 2016).

(1) **Intra-reader Intra-modality agreement** : the agreement of the ratings of one reader using modality $A$ with the replicate ratings of the same reader using the

same reference modality $A$. These agreement values $\widehat{\mathcal{C}}_r^{AA}$ are depicted as the gray boxes on the diagonal in Table 1.4. We average these over $n_R$ readers to obtain $\widehat{\mathcal{C}}_\cdot^{AA}$. We can similarly define intra-reader intra-modality agreement for the new modality $B$.

(2) **Intra-reader Inter-modality agreement** : the agreement of the ratings of one reader using the reference modality $A$ with the ratings of the same reader using the new modality $B$. These agreement values $\widehat{\mathcal{C}}_r^{AB}$ are depicted as the gray boxes on the diagonal in Table 1.5. We average these over $n_R$ readers to obtain $\widehat{\mathcal{C}}_\cdot^{AB}$.

(3) **Inter-reader Intra-modality agreement** : the agreement of the ratings of one reader using the reference modality $A$ with the ratings of another reader also using the same reference modality $A$. These agreement values $\widehat{\mathcal{C}}_{rr'}^{AA}$ are depicted as the white off-diagonal boxes in Table 1.4. We average these over $n_R(n_R-1)$ pairs of readers to obtain $\widehat{\mathcal{C}}_{\cdot\cdot}^{AA}$. We can similarly define intra-reader intra-modality agreement for the new modality $B$.

(4) **Inter-reader Inter-modality agreement** : the agreement of the ratings of one reader using the reference modality $A$ with the ratings of another reader using the new modality $B$. These agreement values $\widehat{\mathcal{C}}_{rr'}^{AB}$ are depicted as the white off-diagonal boxes in Table 1.5. We average these over $n_R(n_R-1)$ pairs of readers to obtain $\widehat{\mathcal{C}}_{\cdot\cdot}^{AB}$.

An analysis that compares intra-reader *inter-modality agreement* with its baseline level agreement, intra-reader *intra-modality agreement*, is called an **intra-reader agreement analysis**. Respectively, an analysis that compares inter-reader *inter-modality agreement* with its baseline level agreement, inter-reader *intra-modality agreement*, is called an **inter-reader agreement analysis**.

From a practical standpoint, the inter-reader analysis is less burdensome than the intra-reader analysis. This is because the inter-reader analysis does not require replicated readings (multiple reading sessions) from the same reader for the same cases while the intra-reader analysis does. From a mathematical standpoint, however, the inter-reader analysis is more challenging than the intra-reader analysis. This is because the inter-reader analysis requires averaging over all pairs of readers while the intra-reader analysis requires only a single average over readers. The averaging over pairs of readers adds complexity to the variance estimation of inter-reader agreement measurements as additional correlations have to be accounted for. In this thesis, we will focus on the inter-reader agreement analysis. The intra-reader agreement analysis can be constructed following similar principles and methods.

Let $\mathcal{C}^{AB}$ be the inter-reader inter-modality agreement for the reference modality $A$ and the new modality $B$, and let $\mathcal{C}^{AA}$ be the inter-reader intra-modality agreement for the reference modality $A$. Gallas et al. (2016) presented a hypothesis test for non-inferiority between two modalities. The corresponding null and alternative hypotheses can be written as:

$$H_0 : \mathcal{C}^{AA} > \mathcal{C}^{AB} + \delta,$$

$$H_1 : \mathcal{C}^{AA} \leq \mathcal{C}^{AB} + \delta,$$

where $\delta$ is a positive non-inferiority margin. Rejecting the null hypothesis implies that there are grounds for believing that the new modality is not much worse than (or non-inferior to) the reference modality. The test statistic is given by:

$$t = \Big(\widehat{\mathcal{C}}^{AA} - \widehat{\mathcal{C}}^{AB} - \delta\Big) \Big/ \sqrt{\widehat{\mathrm{Var}}\big(\widehat{\mathcal{C}}^{AB} - \widehat{\mathcal{C}}^{AA}\big)},$$

where $\widehat{\mathcal{C}}^{AB}$, $\widehat{\mathcal{C}}^{AA}$ and $\widehat{\mathrm{Var}}\big(\widehat{\mathcal{C}}^{AB} - \widehat{\mathcal{C}}^{AA}\big)$ are the estimators for $\mathcal{C}^{AB}$, $\mathcal{C}^{AA}$ and $\mathrm{Var}\big(\widehat{\mathcal{C}}^{AB} - \widehat{\mathcal{C}}^{AA}\big)$, respectively. By Central Limit Theorem, it can be shown that $t$ is asymptotically standard normal when the null hypothesis is true (Randles and Wolfe, 1979). Therefore, for a level $\alpha$ test, we can reject the null hypothesis when $t > Z_\alpha =$

$\Phi^{-1}(1-\alpha)$, where $\Phi$ is the cumulative distribution function of standard normal distribution.

## 1.5 Agreement Measures Based on Paired Comparisons

We use paired comparisons to measure agreement. Let $Y$ and $X$ denote the outcome variables of interest, which can either be continuous or ordinal. In the medical imaging study, the outcome variable is a rating and thus an ordinal variable. Given a pair of observations on the variable $Y$, $Y_i$ and $Y_j$, their relative values can be summarized in one of the three possible "orders": $Y_i > Y_j$, $Y_i = Y_j$ or $Y_i < Y_j$. Similarly, there are three possible "orders" for a given pair of observations on the random variable $X$. In an agreement study for variables $X$ and $Y$, nine combinations $(3 \times 3)$ of the "orders" are possible, and these combinations can be partitioned into five categories, as suggested by Kim (1971). We summarize the five categories and their corresponding probabilities in Table 1.6.

Table 1.6: Paired comparisons at the population level

| Probability | Category | Combination for the orders |
|---|---|---|
| $\Pi_C$ | Concordant | $(X_i > X_j$ and $Y_i > Y_j)$ or $(X_i < X_j$ and $Y_i < Y_j)$ |
| $\Pi_D$ | Discordant | $(X_i < X_j$ and $Y_i > Y_j)$ or $(X_i > X_j$ and $Y_i < Y_j)$ |
| $\Pi_{T_x}$ | Tie on X | $(X_i = X_j$ and $Y_i > Y_j)$ or $(X_i = X_j$ and $Y_i < Y_j)$ |
| $\Pi_{T_y}$ | Tie on Y | $(X_i > X_j$ and $Y_i = Y_j)$ or $(X_i < X_j$ and $Y_i = Y_j)$ |
| $\Pi_{T_{xy}}$ | Tie on both | $X_i = X_j$ and $Y_i = Y_j$ |

The probabilities corresponding to the five categories summed to one: $\Pi_C + \Pi_D + \Pi_{T_x} + \Pi_{T_y} + \Pi_{T_{xy}} = 1$. For continuous variables, $\Pi_{T_x} = \Pi_{T_y} = \Pi_{T_{xy}} = 0$.

In this thesis, we measure agreement using the concordance. Concordance is defined as the probability that the two pairs are in the same order (the sort order by $X$ and by $Y$):

$$\mathcal{C} \doteq \Pi_C = \Pr\big((X_i - X_j)(Y_i - Y_j) > 0\big), \tag{1.1}$$

where $i \neq j$. The range of $\mathcal{C}$ is between 0 and 1, and a higher value of $\mathcal{C}$ means higher agreement between $X$ and $Y$.

There are other well-known agreement measures, including Kendall's $\tau_a$, Kendall's $\tau_b$, Goodman and Kruskal's $\gamma$ and Somer's $D$. Kendall (1938) introduced a rank-based correlation measure $\tau_a$, which is defined as the probability that the pairs are in concordance against the probability that the pairs are in discordance; that is,

$$\tau_a = \Pi_C - \Pi_D.$$

Unlike Kendall's $\tau_a$, Kendall's $\tau_b$ makes adjustments for ties (Agresti, 2010). Kendall's $\tau_b$ is defined as

$$\tau_b = \frac{\Pi_C - \Pi_D}{\sqrt{(1 - \Pi_{T_x} - \Pi_{T_{xy}})(1 - \Pi_{T_y} - \Pi_{T_{xy}})}}.$$

Goodman and Kruskal (1954) introduced an alternative measure that only considers paired observations without ties:

$$\gamma = \frac{\Pi_C - \Pi_D}{1 - \Pi_{T_x} - \Pi_{T_y} - \Pi_{T_{xy}}}. \tag{1.2}$$

Somers (1962) introduced a pair of asymmetric coefficients, which are appropriate for measuring the association in ordered contingency tables:

$$\begin{aligned} D_{xy}\ (\text{x dependent}) &= \frac{\Pi_C - \Pi_D}{1 - \Pi_{T_x} - \Pi_{T_{xy}}}, \\ D_{yx}\ (\text{y dependent}) &= \frac{\Pi_C - \Pi_D}{1 - \Pi_{T_y} - \Pi_{T_{xy}}}. \end{aligned} \tag{1.3}$$

In the following section, we will discuss the estimators for these agreement measures.

## 1.6 Agreement Estimator

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be independent replicates of $(X, Y)$. Table 1.7 summarizes the possible combinations for the orders of pairs $(x_i, x_j)$ and $(y_i, y_j)$.

Table 1.7: Paired comparisons at the sample level

| Number of Pairs | Category | Combination for the orders |
|---|---|---|
| $N_C$ | Concordant | $(x_i > x_j$ and $y_i > y_j)$ or $(x_i < x_j$ and $y_i < y_j)$ |
| $N_D$ | Discordant | $(x_i < x_j$ and $y_i > y_j)$ or $(x_i < x_j$ and $y_i > y_j)$ |
| $N_{T_x}$ | Tie on X | $(x_i = x_j$ and $y_i > y_j)$ or $(x_i = x_j$ and $y_i < y_j)$ |
| $N_{T_y}$ | Tie on Y | $(x_i > x_j$ and $y_i = y_j)$ or $(x_i < x_j$ and $y_i = y_j)$ |
| $N_{T_{xy}}$ | Tie on both | $x_i = x_j$ and $y_i = y_j$ |

The summation of the numbers of pairs corresponding to the five categories equals the total number of pairs from the sample: $N = N_C + N_D + N_{T_x} + N_{T_y} + N_{T_{xy}} = n(n-1)$.

Table 1.8 gives the empirical estimator of the probability parameter for each of the five categories in Table 1.6.

Table 1.8: Empirical estimators of the probability parameters in Table 1.6

| Parameter | Estimator |
|---|---|
| $\Pi_C$ | $N_C/N$ |
| $\Pi_D$ | $N_D/N$ |
| $\Pi_{T_x}$ | $N_{T_x}/N$ |
| $\Pi_{T_y}$ | $N_{T_y}/N$ |
| $\Pi_{T_{xy}}$ | $N_{T_{xy}}/N$ |

The concordance estimator can be written as

$$\widehat{\mathcal{C}} = \widehat{\Pi}_C = \frac{N_C}{N}.$$

By the Law of Large Numbers, we have $\widehat{\mathcal{C}} \xrightarrow{p} \mathcal{C}$ as $n \to \infty$; that is, $\widehat{\mathcal{C}}$ is a consistent estimator for $\mathcal{C}$.

The estimators for Kendall's $\tau_a$, Kendall's $\tau_b$, Goodman and Kruscal's $\gamma$ and Somer's $D$ have the same numerator $N_C - N_D$, but have different normalizing denominators when ties are present:

$$\begin{aligned}
\widehat{\tau_a} &= \frac{N_C - N_D}{N}, \\
\widehat{\tau_b} &= \frac{N_C - N_D}{\sqrt{(N - N_{T_x} N_{T_{xy}})(N - N_{T_y} N_{T_{xy}})}}, \\
\widehat{\gamma} &= \frac{N_C - N_D}{N - N_{T_x} - N_{T_y} - N_{T_{xy}}}, \\
\widehat{D}_{xy} &= \frac{N_C - N_D}{N - N_{T_y} - N_{T_{xy}}}, \\
\widehat{D}_{yx} &= \frac{N_C - N_D}{N - N_{T_x} - N_{T_{xy}}}.
\end{aligned} \tag{1.4}$$

These estimators differ in the ways of treating ties. Given that the number of ties is finite and as $n \to \infty$, we have $\widehat{\tau_a} \to \tau_a$, $\widehat{\tau_b} \to \tau_b$, $\widehat{\gamma} \to \gamma$ and $\widehat{D} \to D$, so they are consistent estimators.

## 1.7 Introduction to $U$-statistics

In this section, we briefly introduce one-sample and two-sample $U$-statistics following Randles and Wolfe (1979).

### 1.7.1 One-sample $U$-statistics

**Definition 1.7.1.** A parameter $\theta$ is said to be estimable of degree $r$ for the family of distributions $\mathcal{F}$ if $r$ is the smallest sample size for which there exists a function

$h^*(x_1, \ldots, x_r)$ such that

$$\mathrm{E}_F[h^*(X_1, \ldots, X_r)] = \theta$$

for every distribution $F(\cdot) \in \mathcal{F}$, where $X_1, \ldots, X_r$ denotes a random sample from $F(\cdot)$.

The function $h^*(\cdot)$ in Definition 1.7.1 is called a kernel for the parameter $\theta$. We can assume that a kernel is symmetric in its arguments; that is,

$$h^*(x_1, \ldots, x_r) = h^*(x_{\alpha_1}, \ldots, x_{\alpha_r})$$

for every permutation $(\alpha_1, \ldots, \alpha_r)$ of the integers $1, \ldots, r$. Otherwise, for any kernel $h^*(x_1, \ldots, x_r)$ we can always define a symmetric version,

$$h(x_1, \ldots, x_r) = \frac{1}{r!} \sum_{\boldsymbol{\alpha} \in A} h^*(x_{\alpha_1}, \ldots, x_{\alpha_r}),$$

where the summation is over $A = \{\alpha | \alpha \text{ is a permutation of the integers } 1, \ldots, r\}$. It is easy to see that $h(\cdot)$ is symmetric in its arguments and is an unbiased estimator of $\theta$ for each $F(\cdot) \in \mathcal{F}$.

Suppose that we have a random sample $X_1, \ldots, X_n$, $n \geq r$, from a distribution with c.d.f. $F(\cdot) \in \mathcal{F}$. Naturally, we want to use all $n$ observations in constructing an unbiased estimator of $\theta$.

**Definition 1.7.2.** A $U$-statistic for the parameter $\theta$ of degree $r$ is created with the symmetric kernel $h(\cdot)$ by forming

$$U(X_1, \ldots, X_n) = \frac{1}{\binom{n}{r}} \sum_{\boldsymbol{\beta} \in B} h(X_{\beta_1}, \ldots, X_{\beta_r}),$$

where $B = \{\boldsymbol{\beta} | \boldsymbol{\beta}$ is one of the $\binom{n}{r}$ unordered subsets of $r$ integers chosen without replacement from the set $\{1, \ldots, n\}\}$.

Note that a $U$-statistic is an unbiased estimator for $\theta$ for every $F(\cdot) \in \mathcal{F}$ and is symmetric in its $n$ arguments. In fact, when $\mathcal{F}$ includes all continuous distributions,

it can be shown that such a $U$-statistic is the unique minimum-variance-unbiased estimator of $\theta$.

**Example 1.7.3.** Let $\mathcal{F}$ denote the class of all distributions with finite first moment $\theta$. Then

$$\theta = \mathrm{E}[X_1].$$

Thus, the mean is an estimable parameter of degree one for $\mathcal{F}$. Here $h(x) = x$ is the kernel, and it is symmetric in its arguments. The $U$-statistic estimator of the mean is

$$U(X_1, \ldots, X_n) = \frac{1}{\binom{n}{1}} \sum_{i=1}^{n} X_i = \bar{X}.$$

We next introduce a general expression for the variance of a $U$-statistic. First, for a symmetric kernel $h(\cdot)$ consider the random variables

$$h(X_1, \ldots, X_c, X_{c+1}, \ldots, X_r) \text{ and } h(X_1, \ldots, X_c, X_{r+1}, \ldots, X_{2r-c}),$$

which have exactly $c$ variables in common. The covariance of these two variables is given by

$$\begin{aligned} \xi_c &= \mathrm{Cov}[h(X_1, \ldots, X_c, X_{c+1}, \ldots, X_r), h(X_1, \ldots, X_c, X_{r+1}, \ldots, X_{2r-c})] \\ &= \mathrm{E}[h(X_1, \ldots, X_c, X_{c+1}, \ldots, X_r) h(X_1, \ldots, X_c, X_{r+1}, \ldots, X_{2r-c})] - \theta^2, \end{aligned} \tag{1.5}$$

since both of the variables have expected values equal to $\theta$. Furthermore, since $h(\cdot)$ is symmetric in its arguments and the variables $X_1, \ldots, X_n$ are i.i.d., it follows that

$$\xi_c = \mathrm{Cov}[h(X_{\beta_1}, \ldots, X_{\beta_r}), h(X_{\beta'_1}, \ldots, X_{\beta'_r})]$$

whenever $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_r)$ and $\boldsymbol{\beta}' = (\beta'_1, \ldots, \beta'_r)$ are subsets of the integers $\{1, \ldots, n\}$ having exactly $c$ integers in common. Note also that if $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$ have no integers in common, then $h(X_{\beta_1}, \ldots, X_{\beta_r})$ and $h(X_{\beta'_1}, \ldots, X_{\beta'_r})$ are independent. Hence, we set

$$\xi_0 = 0.$$

The variance of a $U$-statistic is

$$\begin{aligned}\mathrm{Var}(U) =& \mathrm{E}\left[\left(\frac{1}{\binom{n}{r}}\sum_{\boldsymbol{\beta}\in B}[h(X_{\beta_1},\ldots,X_{\beta_r})-\theta]\right)^2\right] \\ =& \frac{1}{\binom{n}{r}^2}\sum_{\boldsymbol{\beta}\in B}\sum_{\boldsymbol{\beta}'\in B}\mathrm{E}[(h(X_{\beta_1},\ldots,X_{\beta_r}-\theta)(h(X_{\beta'_1},\ldots,X_{\beta'_r}-\theta)] \\ =& \frac{1}{\binom{n}{r}^2}\sum_{\boldsymbol{\beta}\in B}\sum_{\boldsymbol{\beta}'\in B}\mathrm{Cov}[h(X_{\beta_1},\ldots,X_{\beta_r}),h(X_{\beta'_1},\ldots,X_{\beta'_r})].\end{aligned} \tag{1.6}$$

All terms in (1.6) for which $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$ have exactly $c$ integers in common have the same covariance, namely, $\xi_c$. The number of such terms is $\binom{n}{r}\binom{r}{c}\binom{n-r}{r-c}$, which is simply: (the number of ways to choose the $r$ integers in $\boldsymbol{\beta}$) times (the number of ways to select $c$ integers among those in $\boldsymbol{\beta}$ to be in common between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$) times (the number of ways to choose the remaining $r-c$ integers in $\boldsymbol{\beta}'$ so that they are not in $\boldsymbol{\beta}$). It then follows that

$$\begin{aligned}\mathrm{Var}(U) &= \frac{1}{\binom{n}{r}^2}\sum_{c=0}^{r}\binom{n}{r}\binom{r}{c}\binom{n-r}{r-c}\xi_c \\ &= \frac{1}{\binom{n}{r}}\sum_{c=0}^{r}\binom{r}{c}\binom{n-r}{r-c}\xi_c.\end{aligned} \tag{1.7}$$

This provides a general expression for the variance of a $U$-statistic, where $\xi_c$ is derived via the formula in (1.5).

**Example 1.7.4.** Consider the $U$-statistic estimator $\bar{X}$ of the population mean, as in Example 1.7.3. In this case, the kernel, $h(x) = x$, is the identity function, so the degree is one ($r = 1$). Assuming that the $X_1,\ldots,X_n$ are independently identically distributed with $\mathrm{Var}(X) = \sigma^2$, we have

$$\xi_0 = \mathrm{Cov}(X_i, X_j) = 0,$$

and

$$\xi_1 = \mathrm{Cov}(X_i, X_i) = \sigma^2,$$

where $i \neq j$. The general expression for the variance of a $U$-statistic then reduces correspondingly to the well known variance formula

$$\mathrm{Var}(\bar{X}) = \frac{1}{\binom{n}{1}} \binom{1}{1} \binom{n-1}{0} \xi_1 = \frac{\sigma^2}{n}.$$

### 1.7.2 Two-sample $U$-statistic

Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be independent random variables from distributions with c.d.f.'s $F(x)$ and $G(y)$, respectively. A parameter $\theta$ is said to be estimable of degree $(r, s)$, for distributions $(F, G)$ in a family $\mathcal{F}$ if $r$ and $s$ are the smallest sample sizes for which there exists an estimator of $\theta$ that is unbiased for every $(F, G) \in \mathcal{F}$. That is, there is a function $h^*(\cdot)$ such that

$$\mathrm{E}_{(F,G)}[h^*(X_1, \ldots, X_r; Y_1, \ldots, Y_s)] = \theta,$$

for every $(F, G) \in \mathcal{F}$. Letting $h(\cdot)$ denote such a symmetric two-sample kernel, we have the following direct extension of one-sample $U$-statistics to this two-sample setting.

**Definition 1.7.5.** For an estimable parameter $\theta$ of degree $(r, s)$, with a symmetric kernel $h(\cdot)$, a two-sample $U$-statistic (estimate) has the form

$$U(X_1, \ldots, X_m; Y_1, \ldots, Y_n) = \frac{1}{\binom{m}{r}\binom{n}{s}} \sum_{\boldsymbol{\alpha} \in A} \sum_{\boldsymbol{\beta} \in B} h(X_1, \ldots, X_m; Y_1, \ldots, Y_n),$$

where $m \geq r$, $n \geq s$, and A (B) is the collection of all subsets of r (s) integers chosen without replacement from the integers $\{1, \ldots, m\}$ ($\{1, \ldots, n\}$).

**Example 1.7.6.** Let $\mathcal{F}$ denote the collection of all pairs of distributions $(F, G)$ such that each has a finite first moment. The difference of the two population means $\mu_Y$ and $\mu_X$ is an estimable parameter of degree (1,1), since

$$\mathrm{E}[Y_1 - X_1] = \mu_Y - \mu_X$$

for every $(F, G) \in \mathcal{F}$. The corresponding two-sample $U$-statistic is

$$U(X_1, \ldots, X_m; Y_1, \ldots, Y_m) = \bar{Y} - \bar{X}.$$

The variance expressions for two-sample $U$-statistics are analogous to those for one-sample $U$-statistics, but they are often considerably more complex. For integers $c$ and $d$ such that $0 \le c \le r$ and $0 \le d \le s$, let $\xi_{c,d}$ denote the covariance between two kernel random variables with exactly $c$ $X_i$s and $d$ $Y_j$s in common. That is, we define

$$\begin{aligned} \xi_{c,d} = \mathrm{Cov}[&h(X_1, \ldots, X_c, X_{c+1}, \ldots, X_r; Y_1, \ldots, Y_d, Y_{d+1}, \ldots, Y_s), \\ &h(X_1, \ldots, X_c, X_{r+1}, \ldots, X_{2r-c}; Y_1, \ldots, Y_d, Y_{s+1}, \ldots, Y_{2s-d})], \end{aligned} \tag{1.8}$$

and set

$$\xi_{0,0} = 0.$$

Similar to the one-sample case, it follows that

$$\begin{aligned} &\mathrm{Var}[U(X_1, \ldots, X_m; Y_1, \ldots, Y_m)] \\ &\qquad = \frac{1}{\binom{m}{r}\binom{n}{s}} \sum_{c=0}^{r} \sum_{d=0}^{s} \binom{r}{c} \binom{m-r}{r-c} \binom{s}{d} \binom{n-s}{s-d} \xi_{c,d}. \end{aligned} \tag{1.9}$$

**Example 1.7.7.** Continue Example 1.7.6, in which the kernel is $h(x, y) = y - x$. By (1.9), we can get

$$\begin{aligned} \mathrm{Var}(\bar{Y} - \bar{X}) = \frac{1}{\binom{m}{1}\binom{n}{1}} \Bigg[ &\binom{1}{1} \binom{m-1}{0} \binom{1}{0} \binom{n-1}{1} \xi_{1,0} \\ &+ \binom{1}{0} \binom{m-1}{1} \binom{1}{1} \binom{n-1}{0} \xi_{0,1} \\ &+ \binom{1}{1} \binom{m-1}{0} \binom{1}{1} \binom{n-1}{0} \xi_{1,1} \Bigg]. \end{aligned} \tag{1.10}$$

Using the expected value of the product of the kernel $h(x, y)$, we can find the covariances

$$\begin{aligned} \xi_{c=1,d=0} &= \mathrm{E}[(Y_1 - X_1)(Y_2 - X_1)] - (\mu_Y - \mu_X)^2 \\ &= \mu_Y^2 - 2\mu_X \mu_Y + \sigma_X^2 + \mu_X^2 - (\mu_Y - \mu_X)^2 \\ &= \sigma_X^2, \end{aligned} \tag{1.11}$$

$$\begin{aligned}\xi_{c=0,d=1} &= \mathrm{E}[(Y_1 - X_1)(Y_1 - X_2)] - (\mu_Y - \mu_X)^2 \\ &= \mu_Y^2 + \sigma_Y^2 - 2\mu_X\mu_Y + \mu_X^2 - (\mu_Y - \mu_X)^2 \\ &= \sigma_Y^2,\end{aligned} \tag{1.12}$$

$$\begin{aligned}\xi_{c=1,d=1} &= \mathrm{E}[(Y_1 - X_1)(Y_1 - X_1)] - (\mu_Y - \mu_X)^2 \\ &= \sigma_Y^2 + \sigma_X^2.\end{aligned} \tag{1.13}$$

Therefore, the variance of the $U$-statistic is

$$\begin{aligned}\mathrm{Var}(\bar{Y} - \bar{X}) &= \frac{1}{mn}\Big((n-1)\sigma_X^2 + (m-1)\sigma_Y^2 + \sigma_X^2 + \sigma_Y^2\Big) \\ &= \frac{1}{m}\sigma_X^2 + \frac{1}{n}\sigma_Y^2.\end{aligned} \tag{1.14}$$

## 1.8 Inter-reader Concordance Estimator in Fully-Crossed Studies

In what follows, we make the assumptions: (1) readers are independent and identically distributed (i.i.d.); (2) cases are i.i.d; (3) readers are independent of cases.

We begin with the inter-reader inter-modality concordance between a given pair of readers $(r_1, r_2)$:

$$\mathcal{C}_{r_1 r_2}^{AB} = \Pr\big((X_{Ar_1c} - X_{Ar_1c'})(X_{Br_2c} - X_{Br_2c'}) > 0\big), \tag{1.15}$$

where $(c \neq c')$ stands for a pair of cases. The concordance $\mathcal{C}_{r_1 r_2}^{AB}$ measures the inter-modality agreement of the two readers $r_1$ and $r_2$ for the modalities $A$ and $B$. We consider the following estimator for this concordance

$$\widetilde{\mathcal{C}}_{r_1 r_2}^{AB} = \binom{n_C}{2}^{-1} \sum_{c_1 < c_2} S(X_{Ar_1c_1}, X_{Ar_1c_2}; X_{Br_2c_1}, X_{Br_2c_2}),$$

where $n_C$ is the number of cases in the study, and the success function $S(\cdot)$ is defined as

$$S(X, X'; Y, Y') = I\Big((X - X')(Y - Y') > 0\Big), \tag{1.16}$$

where

$$I\Big((X-X')(Y-Y')>0\Big)=\begin{cases}1, & (X-X')(Y-Y')>0\\ 0, & (X-X')(Y-Y')\leq 0\end{cases}.$$

Next we define the inter-reader inter-modality concordance for any pair of readers as

$$\mathcal{C}^{AB}=\Pr\big((X_{Arc}-X_{Arc'})(X_{Br'c}-X_{Br'c'})>0\big). \tag{1.17}$$

We can estimate $\mathcal{C}^{AB}$ by averaging the readers-specific concordance estimates across all unique pairs of readers (Gallas et al., 2016):

$$\begin{aligned}\widetilde{\mathcal{C}}^{AB}_{\cdot\cdot} &= \binom{n_R}{2}^{-1}\sum_{r_1=1}^{n_R-1}\sum_{r_2=r_1+1}^{n_R}\widetilde{\mathcal{C}}^{AB}_{r_1r_2}\\ &= \binom{n_R}{2}^{-1}\binom{n_C}{2}^{-1}\sum_{r_1=1}^{n_R-1}\sum_{r_2=r_1+1}^{n_R}\sum_{c_1=1}^{n_C-1}\sum_{c_2=c_1+1}^{n_C}S(X_{Ar_1c_1},X_{Ar_1c_2};X_{Br_2c_1},X_{Br_2c_2}),\end{aligned} \tag{1.18}$$

where we assume that there are $n_R$ readers and $n_C$ cases in the study.

We would like to adapt $\widetilde{\mathcal{C}}^{AB}_{\cdot\cdot}$ to a two-sample $U$-statistic, because $U$-statistics are the minimum-variance-unbiased estimators. By Definition (1.7.5), a two-sample $U$-statistic kernel should be symmetric in the arguments from the first sample and separately symmetric in the arguments from the second sample. Note that the success function $S(\cdot)$ is symmetric in its cases since

$$\begin{aligned}S(X_{Ar_1c_1},X_{Ar_1c_2};X_{Br_2c_1},X_{Br_2c_2}) &= I\Big((X_{Ar_1c_1}-X_{Ar_1c_2})(X_{Br_2c_1}-X_{Br_2c_2})>0\Big)\\ &= I\Big((X_{Ar_1c_2}-X_{Ar_1c_1})(X_{Br_2c_2}-X_{Br_2c_1})>0\Big)\\ &= S(X_{Ar_1c_2},X_{Ar_1c_1};X_{Br_2c_2},X_{Br_2c_1}),\end{aligned} \tag{1.19}$$

but asymmetric in its readers since

$$\begin{aligned} S(X_{Ar_1c_1}, X_{Ar_1c_2}; X_{Br_2c_1}, X_{Br_2c_2}) &= I\Big((X_{Ar_1c_1} - X_{Ar_1c_2})(X_{Br_2c_1} - X_{Br_2c_2}) > 0\Big) \\ &\neq I\Big((X_{Ar_2c_1} - X_{Ar_2c_2})(X_{Br_1c_1} - X_{Br_1c_2}) > 0\Big) \\ &= S(X_{Ar_2c_1}, X_{Ar_2c_2}; X_{Br_1c_1}, X_{Br_1c_2}). \end{aligned} \tag{1.20}$$

Note that the asymmetry in the readers is because the readers are using different modalities; $r_1$ is using modality $A$, $r_2$ is using modality $B$. Otherwise, if both readers are using the same modality, the success function is still symmetric in the readers.

Gallas et al. (2016) defined a kernel that is symmetric in both its $c_j$ and $r_i$ components as follows

$$\begin{aligned} \Phi^{AB}_{c_1c_2r_1r_2} &= \frac{1}{2}S(X_{Ar_1c_1}, X_{Ar_1c_2}; X_{Br_2c_1}, X_{Br_2c_2}) \\ &\quad + \frac{1}{2}S(X_{Br_1c_1}, X_{Br_1c_2}; X_{Ar_2c_1}, X_{Ar_2c_2}). \end{aligned} \tag{1.21}$$

The kernel $\Phi^{AB}_{c_1c_2r_1r_2}$ is unbiased for the inter-reader inter-modality concordance since

$$\begin{aligned} \mathrm{E}(\Phi^{AB}_{c_1c_2r_1r_2}) &= \frac{1}{2}\mathrm{E}\Big(S(X_{Ar_1c_1}, X_{Ar_1c_2}; X_{Br_2c_1}, X_{Br_2c_2})\Big) \\ &\quad + \frac{1}{2}\mathrm{E}\Big(S(X_{Br_1c_1}, X_{Br_1c_2}; X_{Ar_2c_1}, X_{Ar_2c_2})\Big) \\ &= \frac{1}{2}\mathcal{C}^{AB} + \frac{1}{2}\mathcal{C}^{AB} \\ &= \mathcal{C}^{AB}. \end{aligned} \tag{1.22}$$

The unbiasedness of the kernel is another premise for a two-sample $U$-statistic kernel.

With the two-sample $U$-statistic kernel $\Phi^{AB}_{c_1c_2r_1r_2}$, we can construct the following estimator for the inter-reader inter-modality concordance

$$\widehat{\mathcal{C}}^{AB}_{\cdot\cdot} = \binom{n_R}{2}^{-1}\binom{n_C}{2}^{-1}\sum_{r_1=1}^{n_R-1}\sum_{r_2=r_1+1}^{n_R}\sum_{c_1=1}^{n_C-1}\sum_{c_2=c_1+1}^{n_C}\Phi^{AB}_{c_1c_2r_1r_2}. \tag{1.23}$$

The estimator can also be referred to as the reader-averaged concordance estimator (Gallas et al., 2016). By the properties of $U$-statistics, the reader-averaged concordance estimator $\widehat{\mathcal{C}}^{AB}_{\cdot\cdot}$ is a minimum-variance-unbiased estimator for the inter-reader inter-modality concordance.

The analysis for the inter-reader intra-modality concordance follows the same concept so we skip the discussion for the intra-modality concordance here.

## 1.9 Variance of Reader-averaged Concordance Estimator for Fully-crossed Study Designs

Recall the expression given for the variance of the two-sample $U$-statistic in (1.9). The variance of the reader-averaged concordance estimator can be written as

$$\mathrm{Var}(\widehat{\mathcal{C}}_{\cdot\cdot}^{AB}) = \binom{n_R}{2}^{-1}\binom{n_R}{2}^{-1}\sum_{k=0}^{2}\sum_{k=0}^{2}\binom{2}{k}\binom{2}{k'}\binom{n_R-2}{2-k}\binom{n_C-2}{2-k'}\xi_{k,k'}, \quad (1.24)$$

where the covariances, $\xi$'s, are related to the second order moments:

$$\xi_{k,k'} = M_{kk'} - M_{00}. \quad (1.25)$$

There are nine types of second order moments depending on the number of cases and readers in common, and these moments can be expressed in the expected value of two kernels $\Phi^{AB}_{cc'rr'}$ and $\Phi^{AB}_{c^*c^{**}r^*r^{**}}$ as follows

$$M_{kk'} = \mathrm{E}\Big(\Phi^{AB}_{cc'rr'}\Phi^{AB}_{c^*c^{**}r^*r^{**}}\Big|k \text{ cases and } k' \text{ readers in common}\Big), \quad (1.26)$$

where $k, k' = 0, 1, 2$. For example, for the moment with 0 cases and 0 readers in common, we consider

$$\begin{aligned} M_{00} &= \mathrm{E}\Big(\Phi^{AB}_{cc'rr'}\Phi^{AB}_{c^*c^{**}r^*r^{**}}\Big|0 \text{ cases and } 0 \text{ readers in common}\Big) \\ &= \mathrm{E}\Big(\Phi^{AB}_{c_1c_2r_1r_2}\Phi^{AB}_{c_3c_4r_3r_4}\Big) \\ &= \mathrm{E}\Big(\Phi^{AB}_{c_1c_2r_1r_2}\Big)\mathrm{E}\Big(\Phi^{AB}_{c_3c_4r_3r_4}\Big) \\ &= \Big(\mathcal{C}^{AB}\Big)^2. \end{aligned} \quad (1.27)$$

Notice that we will consider the kernels with no common components in the subscripts as independent random variables.

Similarly for the rest of the second order moments, we can write the expressions as follows

$$\begin{aligned} M_{10} &= \mathrm{E}\Big(\Phi^{AB}_{cc'rr'}\Phi^{AB}_{c^*c^{**}r^*r^{**}}\big|1 \text{ case and } 0 \text{ readers in common}\Big) \\ &= \mathrm{E}\Big(\Phi^{AB}_{c_1c_2r_1r_2}\Phi^{AB}_{c_1c_3r_3r_4}\Big), \end{aligned} \tag{1.28}$$

$$\begin{aligned} M_{20} &= \mathrm{E}\Big(\Phi^{AB}_{cc'rr'}\Phi^{AB}_{c^*c^{**}r^*r^{**}}\big|2 \text{ cases and } 0 \text{ readers in common}\Big) \\ &= \mathrm{E}\Big(\Phi^{AB}_{c_1c_2r_1r_2}\Phi^{AB}_{c_1c_2r_3r_4}\Big), \end{aligned} \tag{1.29}$$

$$\begin{aligned} M_{01} &= \mathrm{E}\Big(\Phi^{AB}_{cc'rr'}\Phi^{AB}_{c^*c^{**}r^*r^{**}}\big|0 \text{ cases and } 1 \text{ reader in common}\Big) \\ &= \mathrm{E}\Big(\Phi^{AB}_{c_1c_2r_1r_2}\Phi^{AB}_{c_3c_4r_1r_3}\Big), \end{aligned} \tag{1.30}$$

$$\begin{aligned} M_{11} &= \mathrm{E}\Big(\Phi^{AB}_{cc'rr'}\Phi^{AB}_{c^*c^{**}r^*r^{**}}\big|1 \text{ case and } 1 \text{ reader in common}\Big) \\ &= \mathrm{E}\Big(\Phi^{AB}_{c_1c_2r_1r_2}\Phi^{AB}_{c_1c_3r_1r_3}\Big), \end{aligned} \tag{1.31}$$

$$\begin{aligned} M_{21} &= \mathrm{E}\Big(\Phi^{AB}_{cc'rr'}\Phi^{AB}_{c^*c^{**}r^*r^{**}}\big|2 \text{ cases and } 1 \text{ reader in common}\Big) \\ &= \mathrm{E}\Big(\Phi^{AB}_{c_1c_2r_1r_2}\Phi^{AB}_{c_1c_2r_1r_3}\Big), \end{aligned} \tag{1.32}$$

$$\begin{aligned} M_{02} &= \mathrm{E}\Big(\Phi^{AB}_{cc'rr'}\Phi^{AB}_{c^*c^{**}r^*r^{**}}\big|0 \text{ cases and } 2 \text{ readers in common}\Big) \\ &= \mathrm{E}\Big(\Phi^{AB}_{c_1c_2r_1r_2}\Phi^{AB}_{c_3c_4r_1r_2}\Big), \end{aligned} \tag{1.33}$$

$$\begin{aligned} M_{12} &= \mathrm{E}\Big(\Phi^{AB}_{cc'rr'}\Phi^{AB}_{c^*c^{**}r^*r^{**}}\big|1 \text{ case and } 2 \text{ readers in common}\Big) \\ &= \mathrm{E}\Big(\Phi^{AB}_{c_1c_2r_1r_2}\Phi^{AB}_{c_1c_3r_1r_2}\Big), \end{aligned} \tag{1.34}$$

$$\begin{aligned} M_{22} &= \mathrm{E}\Big(\Phi^{AB}_{cc'rr'}\Phi^{AB}_{c^*c^{**}r^*r^{**}}\big|2 \text{ cases and } 2 \text{ readers in common}\Big) \\ &= \mathrm{E}\Big(\Phi^{AB}_{c_1c_2r_1r_2}\Phi^{AB}_{c_1c_2r_1r_2}\Big). \end{aligned} \tag{1.35}$$

The variance of the reader-averaged concordance estimator in (1.24) can be rewritten as the linear combination of the covariances $\xi$'s;

$$\mathrm{Var}(\widehat{\mathcal{C}}^{AB}) = \boldsymbol{b}^T \boldsymbol{\xi}. \tag{1.36}$$

The vectors $\boldsymbol{b}$ and the $\boldsymbol{\xi}$ are defined as

$$\boldsymbol{b} = \begin{bmatrix} b_{00} \\ b_{10} \\ b_{20} \\ b_{01} \\ b_{11} \\ b_{21} \\ b_{02} \\ b_{12} \\ b_{22} \end{bmatrix}, \boldsymbol{\xi} = \begin{bmatrix} \xi_{00} \\ \xi_{10} \\ \xi_{20} \\ \xi_{01} \\ \xi_{11} \\ \xi_{21} \\ \xi_{02} \\ \xi_{12} \\ \xi_{22} \end{bmatrix} = \begin{bmatrix} M_{00} - M_{00} \\ M_{10} - M_{00} \\ M_{20} - M_{00} \\ M_{01} - M_{00} \\ M_{11} - M_{00} \\ M_{21} - M_{00} \\ M_{02} - M_{00} \\ M_{12} - M_{00} \\ M_{22} - M_{00} \end{bmatrix},$$

where

$$\begin{aligned}
b_{00} &= \binom{n_R}{2}^{-1} \binom{n_C}{2}^{-1} \binom{2}{0} \binom{n_R - 2}{2} \binom{2}{0} \binom{n_C - 2}{2}, \\
b_{10} &= \binom{n_R}{2}^{-1} \binom{n_C}{2}^{-1} \binom{2}{0} \binom{n_R - 2}{2} \binom{2}{1} \binom{n_C - 2}{1}, \\
b_{20} &= \binom{n_R}{2}^{-1} \binom{n_C}{2}^{-1} \binom{2}{0} \binom{n_R - 2}{2} \binom{2}{2} \binom{n_C - 2}{0}, \\
b_{01} &= \binom{n_R}{2}^{-1} \binom{n_C}{2}^{-1} \binom{2}{1} \binom{n_R - 2}{1} \binom{2}{0} \binom{n_C - 2}{2}, \\
b_{11} &= \binom{n_R}{2}^{-1} \binom{n_C}{2}^{-1} \binom{2}{1} \binom{n_R - 2}{1} \binom{2}{1} \binom{n_C - 2}{1}, \\
b_{21} &= \binom{n_R}{2}^{-1} \binom{n_C}{2}^{-1} \binom{2}{1} \binom{n_R - 2}{1} \binom{2}{2} \binom{n_C - 2}{0},
\end{aligned} \tag{1.37}$$

$$
\begin{aligned}
b_{02} &= \binom{n_R}{2}^{-1}\binom{n_C}{2}^{-1}\binom{2}{2}\binom{n_R-2}{0}\binom{2}{0}\binom{n_C-2}{2},\\
b_{12} &= \binom{n_R}{2}^{-1}\binom{n_C}{2}^{-1}\binom{2}{2}\binom{n_R-2}{0}\binom{2}{1}\binom{n_C-2}{1},\\
b_{12} &= \binom{n_R}{2}^{-1}\binom{n_C}{2}^{-1}\binom{2}{2}\binom{n_R-2}{0}\binom{2}{2}\binom{n_C-2}{0}.
\end{aligned}
\tag{1.38}
$$

The population moments $M_{kk'}$ can be estimated by the corresponding sample moments:

$$
\begin{aligned}
\widehat{M}_{00} &= \binom{n_R}{4}^{-1}\frac{1}{24}\binom{n_C}{4}^{-1}\frac{1}{24}\sum_{\{r_1\neq r_2\neq r_3\neq r_4\}}\sum_{\{c_1\neq c_2\neq c_3\neq c_4\}}\Phi^{AB}_{c_1c_2r_1r_2}\Phi^{AB}_{c_3c_4r_3r_4},\\
\widehat{M}_{10} &= \binom{n_R}{4}^{-1}\frac{1}{24}\binom{n_C}{3}^{-1}\frac{1}{6}\sum_{\{r_1\neq r_2\neq r_3\neq r_4\}}\sum_{\{c_1\neq c_2\neq c_3\}}\Phi^{AB}_{c_1c_2r_1r_2}\Phi^{AB}_{c_1c_3r_3r_4},\\
\widehat{M}_{20} &= \binom{n_R}{4}^{-1}\frac{1}{24}\binom{n_C}{2}^{-1}\frac{1}{2}\sum_{\{r_1\neq r_2\neq r_3\neq r_4\}}\sum_{\{c_1\neq c_2\}}\Phi^{AB}_{c_1c_2r_1r_2}\Phi^{AB}_{c_1c_2r_3r_4},\\
\widehat{M}_{01} &= \binom{n_R}{3}^{-1}\frac{1}{6}\binom{n_C}{4}^{-1}\frac{1}{24}\sum_{\{r_1\neq r_2\neq r_3\}}\sum_{\{c_1\neq c_2\neq c_3\neq c_4\}}\Phi^{AB}_{c_1c_2r_1r_2}\Phi^{AB}_{c_3c_4r_1r_3},\\
\widehat{M}_{11} &= \binom{n_R}{3}^{-1}\frac{1}{6}\binom{n_C}{3}^{-1}\frac{1}{6}\sum_{\{r_1\neq r_2\neq r_3\}}\sum_{\{c_1\neq c_2\neq c_3\}}\Phi^{AB}_{c_1c_2r_1r_2}\Phi^{AB}_{c_1c_3r_1r_3},\\
\widehat{M}_{21} &= \binom{n_R}{3}^{-1}\frac{1}{6}\binom{n_C}{2}^{-1}\frac{1}{2}\sum_{\{r_1\neq r_2\neq r_3\}}\sum_{\{c_1\neq c_2\}}\Phi^{AB}_{c_1c_2r_1r_2}\Phi^{AB}_{c_1c_2r_1r_3},\\
\widehat{M}_{02} &= \binom{n_R}{2}^{-1}\frac{1}{2}\binom{n_C}{4}^{-1}\frac{1}{24}\sum_{\{r_1\neq r_2\}}\sum_{\{c_1\neq c_2\neq c_3\neq c_4\}}\Phi^{AB}_{c_1c_2r_1r_2}\Phi^{AB}_{c_3c_4r_1r_2},\\
\widehat{M}_{12} &= \binom{n_R}{2}^{-1}\frac{1}{2}\binom{n_C}{3}^{-1}\frac{1}{6}\sum_{\{r_1\neq r_2\}}\sum_{\{c_1\neq c_2\neq c_3\}}\Phi^{AB}_{c_1c_2r_1r_2}\Phi^{AB}_{c_1c_3r_1r_2},\ \text{and}\\
\widehat{M}_{22} &= \binom{n_R}{2}^{-1}\frac{1}{2}\binom{n_C}{2}^{-1}\frac{1}{2}\sum_{\{r_1\neq r_2\}}\sum_{\{c_1\neq c_2\}}\Phi^{AB}_{c_1c_2r_1r_2}\Phi^{AB}_{c_1c_2r_1r_2}.
\end{aligned}
\tag{1.39}
$$

The sample moment $\widehat{M}_{00}$ can be shown to be unbiased for the population moment

$M_{00}$:

$$\begin{aligned}
\mathrm{E}\Big(\widehat{M}_{00}\Big) &= \mathrm{E}\left(\binom{n_R}{4}^{-1}\frac{1}{24}\binom{n_C}{4}^{-1}\frac{1}{24}\sum_{\{r_1\neq r_2\neq r_3\neq r_4\}}\sum_{\{c_1\neq c_2\neq c_3\neq c_4\}}\Phi^{AB}_{c_1c_2r_1r_2}\Phi^{AB}_{c_3c_4r_3r_4}\right)\\
&= \binom{n_R}{4}^{-1}\frac{1}{24}\binom{n_C}{4}^{-1}\frac{1}{24}\sum_{\{r_1\neq r_2\neq r_3\neq r_4\}}\sum_{\{c_1\neq c_2\neq c_3\neq c_4\}}\mathrm{E}\Big(\Phi^{AB}_{c_1c_2r_1r_2}\Phi^{AB}_{c_3c_4r_3r_4}\Big)\\
&= \binom{n_R}{4}^{-1}\frac{1}{24}\binom{n_C}{4}^{-1}\frac{1}{24}\sum_{\{r_1\neq r_2\neq r_3\neq r_4\}}\sum_{\{c_1\neq c_2\neq c_3\neq c_4\}}M_{00}\\
&= M_{00}.
\end{aligned} \tag{1.40}$$

Similarly, we can show that the rest of the sample moments in (1.39) are unbiased; that is,

$$\mathrm{E}\big(\widehat{M}_{kk'}\big) = M_{kk'}, \tag{1.41}$$

for $k, k' = 0, 1, 2$. Therefore, we can replace the population moments $M_{kk'}$ in (1.36) with the corresponding sample moments $\widehat{M}_{kk'}$ in (1.39) to obtain the variance estimator:

$$\widehat{\mathrm{Var}}(\widehat{C}^{AB}_{\cdot\cdot}) = \boldsymbol{b}^T\widehat{\boldsymbol{\xi}}, \tag{1.42}$$

where

$$\widehat{\boldsymbol{\xi}} = \begin{bmatrix}\widehat{\xi}_{00}\\ \widehat{\xi}_{10}\\ \widehat{\xi}_{20}\\ \widehat{\xi}_{01}\\ \widehat{\xi}_{11}\\ \widehat{\xi}_{21}\\ \widehat{\xi}_{02}\\ \widehat{\xi}_{12}\\ \widehat{\xi}_{22}\end{bmatrix} = \begin{bmatrix}\widehat{M}_{00}\\ \widehat{M}_{10}-\widehat{M}_{00}\\ \widehat{M}_{20}-\widehat{M}_{00}\\ \widehat{M}_{01}-\widehat{M}_{00}\\ \widehat{M}_{11}-\widehat{M}_{00}\\ \widehat{M}_{21}-\widehat{M}_{00}\\ \widehat{M}_{02}-\widehat{M}_{00}\\ \widehat{M}_{12}-\widehat{M}_{00}\\ \widehat{M}_{22}-\widehat{M}_{00}\end{bmatrix}.$$

We can show that this variance estimator is unbiased for $\mathrm{Var}(\widehat{C}_{\cdot\cdot}^{AB})$ since

$$\mathrm{E}\Big(\widehat{\mathrm{Var}}(\widehat{C}_{\cdot\cdot}^{AB})\Big) = \mathrm{E}\Big(\boldsymbol{b}^T\widehat{\boldsymbol{\xi}}\Big) = \boldsymbol{b}^T\mathrm{E}\Big(\widehat{\boldsymbol{\xi}}\Big) = \boldsymbol{b}^T\boldsymbol{\xi} = \mathrm{Var}(\widehat{C}_{\cdot\cdot}^{AB}), \tag{1.43}$$

due to the unbiasedness of the sample moments $\widehat{M}_{kk'}$ as stated in (1.40).

We have now presented the inter-reader inter-modality concordance estimator $\widehat{\mathcal{C}}^{AB}$ and its variance estimator $\widehat{\mathrm{Var}}\big(\widehat{\mathcal{C}}^{AB}\big)$ for fully-crossed study designs. The inter-reader intra-modality concordance estimator $\widehat{\mathcal{C}}^{AA}$ and its variance estimator $\widehat{\mathrm{Var}}\big(\widehat{\mathcal{C}}^{AA}\big)$ can be constructed in a similar way, since the intra-modality concordance is just a special case of the inter-modality concordance when the modalities in comparison are the same.

In the following chapter we focus on adapting these estimators for alternative MRMC study designs. The adaptation follows the same concept as in Gallas and Brown (2008). Furthermore, we propose an improved split-analysis variance estimator for split-plot study designs.

# Chapter 2

# Concordance Analysis for Split-plot Study Designs

In this chapter, we will first extend the MRMC agreement analysis methods in Sections 1.8 and 1.9 to treat alternative study designs following the concept in Gallas and Brown (2008). Gallas and Brown (2008) has extended MRMC receiver operating curve analysis for fully-crossed study designs to treat alternative study designs. However, we will show in Section 3.2 that even though the extended method works well for fully-crossed study designs and balanced split-plot study designs, it fails for unbalanced split-plot study designs, for which the resulting variance estimates have too much variability to be useful. To treat unbalanced split-plot study designs, we propose an improved variance estimator through an approach we refer to as the "split analysis". The idea of the split analysis is to treat each group in the split-plot study design as a separate fully-crossed study design. For each separate fully-crossed study, we obtain a reader-averaged concordance estimate and its variance estimate using (1.23) and (1.42). Then we combine the reader-averaged concordance estimator from each group to give the reader-averaged concordance estimator for the full split-plot study; and we combine the corresponding variance estimates of reader-averaged concordance

from each group to give the variance estimate of reader-averaged concordance for the entire split-plot study.

## 2.1 Design Matrix

We first introduce the design matrix $D$ and the reading data matrix $X$. Define $X = (X_{mcr})$, where $X_{mcr}$ denotes the reading from reader $r$ for case $c$ using modality $m$, $m = A, B$, $c = 1, \ldots, n_C$ and $r = 1, \ldots, n_R$. Let $D = (d_{mcr})$, where $d_{mcr}$ takes value one if $X_{mcr}$ is collected and zero otherwise; that is,

$$d_{mcr} = \begin{cases} 1, & \text{if reader } r \text{ reads case } c \text{ using modality } m \\ 0, & \text{otherwise.} \end{cases} \tag{2.1}$$

Table 2.1 gives an example design matrix for a split-plot study design with six readers, nine cases, three groups and two modalities $A$ and $B$. Note that the readers and cases in one group do not overlap with those in another group.

Table 2.1: An example design matrix for a balanced split-plot study design

| | Group 1 | | | | Group 2 | | | | Group 3 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Reader 1 | | Reader 2 | | Reader 3 | | Reader 4 | | Reader 5 | | Reader 6 | |
| Modality | $A$ | $B$ | $A$ | $B$ | $A$ | $B$ | $A$ | $B$ | $A$ | $B$ | $A$ | $B$ |
| Case 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Case 2 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Case 3 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Case 4 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| Case 5 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| Case 6 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| Case 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| Case 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| Case 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |

The design matrix can also be applied to fully-crossed study designs. The design matrix for a fully-crossed study design is a matrix full of ones; since every reader reads every case using all of the modalities.

## 2.2 Full Concordance Analysis for Alternative Study Designs

Gallas and Brown (2008) extended the MRMC receiver operating curve analysis method for fully-crossed study designs to alternative study designs. The idea of the extension is to consider the elements in the design matrix as indicator variables. We modify the $U$-statistic kernel by multiplying it with the indicator variable which

equals zero if any of the arguments in the kernel are missing. In this way, the kernel will give zero when any of the arguments in the kernel is missing. Otherwise, if no argument is missing, the kernel gives the normal output. Following the idea, we adapt the agreement analysis method discussed in Sections 1.8 and 1.9 to accommodate alternative study designs.

We begin with rewriting the success function for alternative study designs as

$$S^*(X_{Arc}, X_{Arc'}; X_{Br'c}, X_{Br'c'}) = d^{AB}_{rr'cc'} S(X_{Arc}, X_{Arc'}; X_{Br'c}, X_{Br'c'}), \tag{2.2}$$

where $d^{AB}_{r_1r_2c_1c_2} = d_{Ar_1c_1} \times d_{Ar_1c_2} \times d_{Br_2c_1} \times d_{Br_2c_1}$. The success function $S(\cdot)$ is now multiplied by $d^{AB}_{rr'cc'}$ so that when any of the arguments in the success function is missing, we will have $S^*(X_{Arc}, X_{Arc'}; X_{Br'c}, X_{Br'c'}) = 0$. Accordingly, the two-sample $U$-statistic kernel for alternative study designs can be rewritten as

$$\begin{aligned} d\Phi^{AB}_{rr'cc'} = &\frac{1}{2} S^*(X_{Arc}, X_{Arc'}; X_{Br'c}, X_{Br'c'}) \\ &+ \frac{1}{2} S^*(X_{Brc}, X_{Brc'}; X_{Ar'c}, X_{Ar'c'}). \end{aligned} \tag{2.3}$$

Analogous to (1.23), we average $d\Phi^{AB}_{rr'cc'}$ over unique pairs of readers and unique pairs of cases to construct the reader-averaged concordance estimator for alternative study designs as

$$\widehat{\mathcal{C}}^{AB}_{\cdot\cdot} = \frac{\sum_{r_1<r_2} \sum_{c_1<c_2} d\Phi^{AB}_{c_1c_2r_1r_2}}{N^*}, \tag{2.4}$$

where

$$N^* = \sum_{c_1<c_2} \sum_{r_1<r_2} \frac{1}{2} \Big( d^{AB}_{r_1r_2c_1c_2} + d^{BA}_{r_1r_2c_1c_2} \Big). \tag{2.5}$$

Note that $N^*$ equals the number of unique observations under comparison in the MRMC study, where an observation includes a pair of readers and a pair of cases.

We may consider (1.23) as a special case of (2.4), in which the design matrix $D$

consists of only ones since replacing $d^{AB}_{rr'cc'}$ with a value of one gives

$$\begin{aligned}\widehat{\mathcal{C}}^{AB}_{\cdot\cdot} &= \frac{\sum_{r_1<r_2}\sum_{c_1<c_2}\Phi^{AB}_{c_1c_2r_1r_2}}{\sum_{r_1^*<r_2^*}\sum_{c_1^*<c_2^*}\left(\frac{1}{2}+\frac{1}{2}\right)} \\ &= \binom{n_R}{2}^{-1}\binom{n_C}{2}^{-1}\sum_{r_1<r_2}\sum_{c_1<c_2}\Phi^{AB}_{c_1c_2r_1r_2},\end{aligned} \tag{2.6}$$

which coincides with (1.23). It can be shown that the reader-averaged concordance estimator $\widehat{\mathcal{C}}^{AB}_{\cdot\cdot}$ is unbiased for $\mathcal{C}^{AB}$:

$$\begin{aligned}\mathrm{E}\left(\widehat{\mathcal{C}}^{AB}_{\cdot\cdot}\right) &= \frac{\sum_{r_1<r_2}\sum_{c_1<c_2}\mathrm{E}\left(d\Phi^{AB}_{c_1c_2r_1r_2}\right)}{\sum_{r_1^*<r_2^*}\sum_{c_1^*<c_2^*}\left(\frac{1}{2}d^{AB}_{r_1^*r_2^*c_1^*c_2^*}+\frac{1}{2}d^{BA}_{r_1^*r_2^*c_1^*c_2^*}\right)} \\ &= \frac{\sum_{r_1<r_2}\sum_{c_1<c_2}\left(\frac{1}{2}d^{AB}_{r_1r_2c_1c_2}\mathcal{C}^{AB}+\frac{1}{2}d^{BA}_{r_1r_2c_1c_2}\mathcal{C}^{AB}\right)}{\sum_{r_1^*<r_2^*}\sum_{c_1^*<c_2^*}\left(\frac{1}{2}d^{AB}_{r_1^*r_2^*c_1^*c_2^*}+\frac{1}{2}d^{BA}_{r_1^*r_2^*c_1^*c_2^*}\right)} \\ &= \frac{\sum_{r_1<r_2}\sum_{c_1<c_2}\left(\frac{1}{2}d^{AB}_{r_1r_2c_1c_2}+\frac{1}{2}d^{BA}_{r_1r_2c_1c_2}\right)\mathcal{C}^{AB}}{\sum_{r_1^*<r_2^*}\sum_{c_1^*<c_2^*}\left(\frac{1}{2}d^{AB}_{r_1^*r_2^*c_1^*c_2^*}+\frac{1}{2}d^{BA}_{r_1^*r_2^*c_1^*c_2^*}\right)} \\ &= \mathcal{C}^{AB}.\end{aligned} \tag{2.7}$$

Now we discuss the variance estimator for $\widehat{\mathcal{C}}^{AB}_{\cdot\cdot}$ for alternative study designs. Recall that from (1.36), the variance of $\widehat{\mathcal{C}}^{AB}_{\cdot\cdot}$ is a linear combination of the covariance vector $\boldsymbol{\xi}$ with the coefficient vector $\boldsymbol{b}$. Analogous to the coefficients for fully-crossed study

designs (1.37), we can determine the coefficients for alternative study designs:

$$
\begin{aligned}
b_{00} &= \frac{\sum_{r_1<r_2}\sum_{c_1<c_2}\left(d^{AB}_{c_1c_2r_1r_2}\sum_{(r_3<r_4)\in H}\sum_{(c_3<c_4)\in I}d^{AB}_{c_3c_4r_3r_4}\right)\times\binom{2}{0}\times\binom{2}{0}}{(N^*)^2},\\
b_{10} &= \frac{\sum_{r_1<r_2}\sum_{c_1<c_2}\left(d^{AB}_{c_1c_2r_1r_2}\sum_{(r_3<r_4)\in H}\sum_{c_3\in I}d^{AB}_{c_1c_3r_3r_4}\right)\times\binom{2}{1}\times\binom{2}{0}}{(N^*)^2},\\
b_{20} &= \frac{\sum_{r_1<r_2}\sum_{c_1<c_2}\left(d^{AB}_{c_1c_2r_1r_2}\sum_{(r_3<r_4)\in H}d^{AB}_{c_1c_2r_3r_4}\right)\times\binom{2}{2}\times\binom{2}{0}}{(N^*)^2},\\
b_{01} &= \frac{\sum_{r_1<r_2}\sum_{c_1<c_2}\left(d^{AB}_{c_1c_2r_1r_2}\sum_{r_3\in H}\sum_{(c_3<c_4)\in I}d^{AB}_{c_3c_4r_1r_3}\right)\times\binom{2}{0}\times\binom{2}{1}}{(N^*)^2},\\
b_{11} &= \frac{\sum_{r_1<r_2}\sum_{c_1<c_2}\left(d^{AB}_{c_1c_2r_1r_2}\sum_{(r_3)\in H}\sum_{(c_3)\in I}d^{AB}_{c_1c_3r_1r_3}\right)\times\binom{2}{1}\times\binom{2}{1}}{(N^*)^2},\\
b_{21} &= \frac{\sum_{r_1<r_2}\sum_{c_1<c_2}\left(d^{AB}_{c_1c_2r_1r_2}\sum_{(r_3)\in H}d^{AB}_{c_1c_2r_1r_3}\right)\times\binom{2}{2}\times\binom{2}{1}}{(N^*)^2},\\
b_{02} &= \frac{\sum_{r_1<r_2}\sum_{c_1<c_2}\left(d^{AB}_{c_1c_2r_1r_2}\sum_{(c_3<c_4)\in I}d^{AB}_{c_3c_4r_1r_2}\right)\times\binom{2}{0}\times\binom{2}{2}}{(N^*)^2},\\
b_{12} &= \frac{\sum_{r_1<r_2}\sum_{c_1<c_2}\left(d^{AB}_{c_1c_2r_1r_2}\sum_{(c_3)\in I}d^{AB}_{c_1c_3r_1r_2}\right)\times\binom{2}{1}\times\binom{2}{2}}{(N^*)^2},\\
b_{22} &= \frac{\sum_{r_1<r_2}\sum_{c_1<c_2}\left(d^{AB}_{c_1c_2r_1r_2}\right)\times\binom{2}{2}\times\binom{2}{2}}{(N^*)^2}.
\end{aligned}
\tag{2.8}
$$

We consider $H$ the set of possible pairs of readers in $\big\{(1,\ldots,n_R)\neq(r_1,r_2)\big\}$, and $I$ the set of possible pairs of cases in $\big\{(1,\ldots,n_C)\neq(c_1,c_2)\big\}$.

Analogous to (1.26), the second order moments for alternative study designs can

be written as

$$
\begin{aligned}
\widehat{M}_{00} &= \frac{\sum_{r_1\neq r_2}\sum_{c_1\neq c_2}\sum_{r_3\neq r_4}\sum_{c_3\neq c_4} d\Phi^{AB}_{c_1c_2r_1r_2} d\Phi^{AB}_{c_3c_4r_3r_4}}{\sum_{r_1^*\neq r_2^*}\sum_{c_1^*\neq c_2^*}\sum_{r_3^*\neq r_4^*}\sum_{c_3^*\neq c_4^*}\left(\frac{1}{2}d^{AB}_{c_1^*c_2^*r_1^*r_2^*}+\frac{1}{2}d^{BA}_{c_1^*c_2^*r_1^*r_2^*}\right)\left(\frac{1}{2}d^{AB}_{c_3^*c_4^*r_3^*r_4^*}+\frac{1}{2}d^{BA}_{c_3^*c_4^*r_3^*r_4^*}\right)},\\
\widehat{M}_{10} &= \frac{\sum_{r_1\neq r_2}\sum_{c_1\neq c_2}\sum_{r_3\neq r_4}\sum_{c_3} d\Phi^{AB}_{c_1c_2r_1r_2} d\Phi^{AB}_{c_1c_3r_3r_4}}{\sum_{r_1^*\neq r_2^*}\sum_{c_1^*\neq c_2^*}\sum_{r_3^*\neq r_4^*}\sum_{c_3^*}\left(\frac{1}{2}d^{AB}_{c_1^*c_2^*r_1^*r_2^*}+\frac{1}{2}d^{BA}_{c_1^*c_2^*r_1^*r_2^*}\right)\left(\frac{1}{2}d^{AB}_{c_1^*c_3^*r_3^*r_4^*}+\frac{1}{2}d^{BA}_{c_1^*c_3^*r_3^*r_4^*}\right)},\\
\widehat{M}_{20} &= \frac{\sum_{r_1\neq r_2}\sum_{c_1\neq c_2}\sum_{r_3\neq r_4} d\Phi^{AB}_{c_1c_2r_1r_2} d\Phi^{AB}_{c_1c_2r_3r_4}}{\sum_{r_1^*\neq r_2^*}\sum_{c_1^*\neq c_2^*}\sum_{r_3^*\neq r_4^*}\left(\frac{1}{2}d^{AB}_{c_1^*c_2^*r_1^*r_2^*}+\frac{1}{2}d^{BA}_{c_1^*c_2^*r_1^*r_2^*}\right)\left(\frac{1}{2}d^{AB}_{c_1^*c_2^*r_3^*r_4^*}+\frac{1}{2}d^{BA}_{c_1^*c_2^*r_3^*r_4^*}\right)},\\
\widehat{M}_{01} &= \frac{\sum_{r_1\neq r_2}\sum_{c_1\neq c_2}\sum_{r_3}\sum_{c_3\neq c_4} d\Phi^{AB}_{c_1c_2r_1r_2} d\Phi^{AB}_{c_3c_4r_1r_3}}{\sum_{r_1^*\neq r_2^*}\sum_{c_1^*\neq c_2^*}\sum_{r_3^*}\sum_{c_3^*\neq c_4^*}\left(\frac{1}{2}d^{AB}_{c_1^*c_2^*r_1^*r_2^*}+\frac{1}{2}d^{BA}_{c_1^*c_2^*r_1^*r_2^*}\right)\left(\frac{1}{2}d^{AB}_{c_3^*c_4^*r_1^*r_3^*}+\frac{1}{2}d^{BA}_{c_3^*c_4^*r_1^*r_3^*}\right)},\\
\widehat{M}_{11} &= \frac{\sum_{r_1\neq r_2}\sum_{c_1\neq c_2}\sum_{r_3}\sum_{c_3} d\Phi^{AB}_{c_1c_2r_1r_2} d\Phi^{AB}_{c_1c_3r_1r_3}}{\sum_{r_1^*\neq r_2^*}\sum_{c_1^*\neq c_2^*}\sum_{r_3^*}\sum_{c_3^*}\left(\frac{1}{2}d^{AB}_{c_1^*c_2^*r_1^*r_2^*}+\frac{1}{2}d^{BA}_{c_1^*c_2^*r_1^*r_2^*}\right)\left(\frac{1}{2}d^{AB}_{c_1^*c_3^*r_1^*r_3^*}+\frac{1}{2}d^{BA}_{c_1^*c_3^*r_1^*r_3^*}\right)},\\
\widehat{M}_{21} &= \frac{\sum_{r_1\neq r_2}\sum_{c_1\neq c_2}\sum_{r_3} d\Phi^{AB}_{c_1c_2r_1r_2} d\Phi^{AB}_{c_1c_2r_1r_3}}{\sum_{r_1^*\neq r_2^*}\sum_{c_1^*\neq c_2^*}\sum_{r_3^*}\left(\frac{1}{2}d^{AB}_{c_1^*c_2^*r_1^*r_2^*}+\frac{1}{2}d^{BA}_{c_1^*c_2^*r_1^*r_2^*}\right)\left(\frac{1}{2}d^{AB}_{c_1^*c_2^*r_1^*r_3^*}+\frac{1}{2}d^{BA}_{c_1^*c_2^*r_1^*r_3^*}\right)},\\
\widehat{M}_{02} &= \frac{\sum_{r_1\neq r_2}\sum_{c_1\neq c_2}\sum_{c_3\neq c_4} d\Phi^{AB}_{c_1c_2r_1r_2} d\Phi^{AB}_{c_3c_4r_1r_2}}{\sum_{r_1^*\neq r_2^*}\sum_{c_1^*\neq c_2^*}\sum_{c_3^*\neq c_4^*}\left(\frac{1}{2}d^{AB}_{c_1^*c_2^*r_1^*r_2^*}+\frac{1}{2}d^{BA}_{c_1^*c_2^*r_1^*r_2^*}\right)\left(\frac{1}{2}d^{AB}_{c_3^*c_4^*r_1^*r_2^*}+\frac{1}{2}d^{BA}_{c_3^*c_4^*r_1^*r_2^*}\right)},\\
\widehat{M}_{12} &= \frac{\sum_{r_1\neq r_2}\sum_{c_1\neq c_2}\sum_{c_3} d\Phi^{AB}_{c_1c_2r_1r_2} d\Phi^{AB}_{c_1c_3r_1r_2}}{\sum_{r_1^*\neq r_2^*}\sum_{c_1^*\neq c_2^*}\sum_{c_3^*}\left(\frac{1}{2}d^{AB}_{c_1^*c_2^*r_1^*r_2^*}+\frac{1}{2}d^{BA}_{c_1^*c_2^*r_1^*r_2^*}\right)\left(\frac{1}{2}d^{AB}_{c_1^*c_3^*r_1^*r_2^*}+\frac{1}{2}d^{BA}_{c_1^*c_3^*r_1^*r_2^*}\right)},\\
\widehat{M}_{22} &= \frac{\sum_{r_1\neq r_2}\sum_{c_1\neq c_2} d\Phi^{AB}_{c_1c_2r_1r_2} d\Phi^{AB}_{c_1c_2r_1r_2}}{\sum_{r_1^*\neq r_2^*}\sum_{c_1^*\neq c_2^*}\left(\frac{1}{2}d^{AB}_{c_1^*c_2^*r_1^*r_2^*}+\frac{1}{2}d^{BA}_{c_1^*c_2^*r_1^*r_2^*}\right)\left(\frac{1}{2}d^{AB}_{c_1^*c_2^*r_1^*r_2^*}+\frac{1}{2}d^{BA}_{c_1^*c_2^*r_1^*r_2^*}\right)}.
\end{aligned}
\tag{2.9}
$$

For each of the sample moment above, the normalizing denominator equals the total number of terms involved in the numerators.

Combining (2.8), (2.9) and (1.42), we obtain the variance estimator for the reader-

averaged concordance measurement:

$$\widehat{\mathrm{Var}}\big(\widehat{C}_{\cdot\cdot}^{AB}\big) = \boldsymbol{b}^T\widehat{\boldsymbol{\xi}}, \tag{2.10}$$

where

$$\boldsymbol{b} = \begin{bmatrix} b_{00} \\ b_{10} \\ b_{20} \\ b_{01} \\ b_{11} \\ b_{21} \\ b_{02} \\ b_{12} \\ b_{22} \end{bmatrix}, \widehat{\boldsymbol{\xi}} = \begin{bmatrix} \widehat{\xi}_{00} \\ \widehat{\xi}_{10} \\ \widehat{\xi}_{20} \\ \widehat{\xi}_{01} \\ \widehat{\xi}_{11} \\ \widehat{\xi}_{21} \\ \widehat{\xi}_{02} \\ \widehat{\xi}_{12} \\ \widehat{\xi}_{22} \end{bmatrix} = \begin{bmatrix} \widehat{M}_{00} - \widehat{M}_{00} \\ \widehat{M}_{10} - \widehat{M}_{00} \\ \widehat{M}_{20} - \widehat{M}_{00} \\ \widehat{M}_{01} - \widehat{M}_{00} \\ \widehat{M}_{11} - \widehat{M}_{00} \\ \widehat{M}_{21} - \widehat{M}_{00} \\ \widehat{M}_{02} - \widehat{M}_{00} \\ \widehat{M}_{12} - \widehat{M}_{00} \\ \widehat{M}_{22} - \widehat{M}_{00} \end{bmatrix},$$

and the elements $b_{kk'}$ and $\widehat{M}_{kk'}$ are defined in (2.8) and (2.9).

From now on, we refer to the variance estimator in (2.10) as the full-analysis variance estimator (FAVE). The FAVE works well for fully-crossed and balanced split-plot study designs. However, our numerical studies show that the variability of this estimator is often large, and it may even produce negative variance estimates for the unbalanced split-plot studies.

## 2.3 Split Analysis

To overcome the problem of negative variance estimates for unbalanced split-plot studies, we propose a new variance estimator based on "split analysis". The idea of split analysis is to treat each group in the split-plot study as a separate fully-crossed study design. We illustrate the idea using an example study design in Table 1.2: a split-plot study design with six readers, nine cases and three groups. For this study

design, we can treat the three groups as three separate fully-crossed sub-studies as shown in Tables 2.2-2.4.

Table 2.2: Group 1 for the split-plot study in Table 1.2

| | Group 1 | | | |
|---|---|---|---|---|
| | Reader 1 | | Reader 2 | |
| Modality | $A$ | $B$ | $A$ | $B$ |
| Case 1 | $X_{A11}$ | $X_{B11}$ | $X_{A21}$ | $X_{B21}$ |
| Case 2 | $X_{A12}$ | $X_{B12}$ | $X_{A22}$ | $X_{B22}$ |
| Case 3 | $X_{A13}$ | $X_{B13}$ | $X_{A23}$ | $X_{B23}$ |

Table 2.3: Group 2 for the split-plot study in Table 1.2

| | Group 2 | | | |
|---|---|---|---|---|
| | Reader 3 | | Reader 4 | |
| Modality | $A$ | $B$ | $A$ | $B$ |
| Case 4 | $X_{A34}$ | $X_{B34}$ | $X_{A44}$ | $X_{B44}$ |
| Case 5 | $X_{A35}$ | $X_{B35}$ | $X_{A45}$ | $X_{B45}$ |
| Case 6 | $X_{A36}$ | $X_{B36}$ | $X_{A46}$ | $X_{B46}$ |

Table 2.4: Group 3 for the split-plot study in Table 1.2

| | Group 3 | | | |
|---|---|---|---|---|
| | Reader 5 | | Reader 6 | |
| Modality | $A$ | $B$ | $A$ | $B$ |
| Case 7 | $X_{A57}$ | $X_{B57}$ | $X_{A67}$ | $X_{B67}$ |
| Case 8 | $X_{A58}$ | $X_{B58}$ | $X_{A68}$ | $X_{B68}$ |
| Case 9 | $X_{A59}$ | $X_{B59}$ | $X_{A69}$ | $X_{B69}$ |

We will show in the following that the reader-averaged concordance estimator for the full split-plot study design equals the weighted average of the reader-averaged concordance estimator for each sub-study.

We begin with the reader-averaged concordance estimator for the full split-plot study design in Table 1.2. By (2.4), we can write the reader-averaged concordance estimator as

$$\widehat{\mathcal{C}}_{\cdot\cdot}^{AB} = \frac{\sum_{r_1=1}^{5}\sum_{r_2=r_1+1}^{6}\sum_{c_1=1}^{8}\sum_{c_1=c_1+1}^{9} d\Phi_{c_1c_2r_1r_2}^{AB}}{N^*}, \tag{2.11}$$

where

$$N^* = \sum_{r_1=1}^{5}\sum_{r_2=r_1+1}^{6}\sum_{c_1=1}^{8}\sum_{c_2=c_1+1}^{9}\frac{1}{2}\Big(d_{r_1r_2c_1c_2}^{AB} + d_{r_1r_2c_1c_2}^{BA}\Big). \tag{2.12}$$

We can decompose $N^*$ by discarding the terms for which both $d_{r_1r_2c_1c_2}^{AB}$ and $d_{r_1r_2c_1c_2}^{BA}$

equal zero as follows

$$
\begin{aligned}
N^* = & \sum_{r_1=1}^{1} \sum_{r_2=r_1+1}^{2} \sum_{c_1=1}^{2} \sum_{c_2=c_1+1}^{3} \left( \frac{1}{2} d^{AB}_{r_1 r_2 c_1 c_2} + \frac{1}{2} d^{BA}_{r_1 r_2 c_1 c_2} \right) \\
& + \sum_{r_1=3}^{3} \sum_{r_2=r_1+1}^{4} \sum_{c_1=4}^{5} \sum_{c_2=c_1+1}^{6} \left( \frac{1}{2} d^{AB}_{r_1 r_2 c_1 c_2} + \frac{1}{2} d^{BA}_{r_1 r_2 c_1 c_2} \right) \\
& + \sum_{r_1=5}^{5} \sum_{r_2=r_1+1}^{6} \sum_{c_1=7}^{8} \sum_{c_2=c_1+1}^{9} \left( \frac{1}{2} d^{AB}_{r_1 r_2 c_1 c_2} + \frac{1}{2} d^{BA}_{r_1 r_2 c_1 c_2} \right) \\
= & N_1^* + N_2^* + N_3^*,
\end{aligned}
\tag{2.13}
$$

where $N_g^*$ is the number of unique pairs under comparison in group $g$, $g = 1, 2, 3$. Similarly, we can decompose the summand in the numerator in (2.11) by discarding the terms for which $d\Phi^{AB}_{c_1 c_2 r_1 r_2} = 0$. Then we can rewrite (2.11) as follows

$$
\begin{aligned}
\widehat{\mathcal{C}}^{AB}_{\cdot\cdot} = & \frac{\sum_{r_1=1}^{1} \sum_{r_2=r_1+1}^{2} \sum_{c_1=1}^{2} \sum_{c_2=c_1+1}^{3} d\Phi^{AB}_{c_1 c_2 r_1 r_2}}{N_1^* + N_2^* + N_3^*} \\
& + \frac{\sum_{r_1=3}^{3} \sum_{r_2=r_1+1}^{4} \sum_{c_1=4}^{5} \sum_{c_2=c_1+1}^{6} d\Phi^{AB}_{c_1 c_2 r_1 r_2}}{N_1^* + N_2^* + N_3^*} \\
& + \frac{\sum_{r_1=5}^{5} \sum_{r_2=r_1+1}^{6} \sum_{c_1=7}^{8} \sum_{c_2=c_1+1}^{9} d\Phi^{AB}_{c_1 c_2 r_1 r_2}}{N_1^* + N_2^* + N_3^*},
\end{aligned}
\tag{2.14}
$$

where in each of the three fractions above, $d^{AB}_{c_1 c_2 r_1 r_2}$ always equals one; replacing all

$d^{AB}_{c_1c_2r_1r_2}$ with a value of one gives

$$
\begin{aligned}
\widehat{\mathcal{C}}^{AB}_{\cdot\cdot} &= (N_1^* + N_2^* + N_3^*)^{-1} \sum_{r_1=1}^{1} \sum_{r_2=r_1+1}^{2} \sum_{c_1=1}^{2} \sum_{c_2=c_1+1}^{3} \Phi^{AB}_{c_1c_2r_1r_2} \\
&\quad + (N_1^* + N_2^* + N_3^*)^{-1} \sum_{r_1=3}^{3} \sum_{r_2=r_1+1}^{4} \sum_{c_1=4}^{5} \sum_{c_2=c_1+1}^{6} \Phi^{AB}_{c_1c_2r_1r_2} \\
&\quad + (N_1^* + N_2^* + N_3^*)^{-1} \sum_{r_1=5}^{5} \sum_{r_2=r_1+1}^{6} \sum_{c_1=7}^{8} \sum_{c_2=c_1+1}^{9} \Phi^{AB}_{c_1c_2r_1r_2} \\
&= \frac{N_1^*}{N^*} \left( \frac{1}{N_1^*} \sum_{r_1=1}^{1} \sum_{r_2=r_1+1}^{2} \sum_{c_1=1}^{2} \sum_{c_2=c_1+1}^{3} \Phi^{AB}_{c_1c_2r_1r_2} \right) \\
&\quad + \frac{N_2^*}{N^*} \left( \frac{1}{N_2^*} \sum_{r_1=3}^{3} \sum_{r_2=r_1+1}^{4} \sum_{c_1=4}^{5} \sum_{c_2=c_1+1}^{6} \Phi^{AB}_{c_1c_2r_1r_2} \right) \\
&\quad + \frac{N_3^*}{N^*} \left( \frac{1}{N_3^*} \sum_{r_1=5}^{5} \sum_{r_2=r_1+1}^{6} \sum_{c_1=7}^{8} \sum_{c_2=c_1+1}^{9} \Phi^{AB}_{c_1c_2r_1r_2} \right) \\
&= \frac{N_1^*}{N^*} \left( \widehat{\mathcal{C}}^{AB}_{\cdot\cdot} \middle| \text{Group} = 1 \right) \\
&\quad + \frac{N_2^*}{N^*} \left( \widehat{\mathcal{C}}^{AB}_{\cdot\cdot} \middle| \text{Group} = 2 \right) \\
&\quad + \frac{N_3^*}{N^*} \left( \widehat{\mathcal{C}}^{AB}_{\cdot\cdot} \middle| \text{Group} = 3 \right),
\end{aligned}
\tag{2.15}
$$

where $\left( \widehat{\mathcal{C}}^{AB}_{\cdot\cdot} \middle| \text{Group} = g \right)$ is the reader-averaged concordance estimator for the group $g$, $g = 1, 2, 3$. We learn from (2.15) that the reader-averaged concordance estimator for the full split-plot study design (2.11) is a weighted average of the reader-averaged concordances for each group, where the weights are the number of unique pairs under comparison for group $g$ over the number of unique pairs under comparison for the full study.

Furthermore, the variance of the reader-averaged concordance estimator for the

split-plot study in this example can be decomposed as

$$\begin{aligned}
\mathrm{Var}\Big(\widehat{\mathcal{C}}_{\cdot\cdot}^{AB}\Big) &= \mathrm{Var}\bigg(\frac{N_1^*}{N^*}\Big(\widehat{\mathcal{C}}_{\cdot\cdot}^{AB}\big|\mathrm{Group}=1\Big) + \frac{N_2^*}{N^*}\Big(\widehat{\mathcal{C}}_{\cdot\cdot}^{AB}\big|\mathrm{Group}=2\Big) \\
&\quad + \frac{N_3^*}{N^*}\Big(\widehat{\mathcal{C}}_{\cdot\cdot}^{AB}\big|\mathrm{Group}=3\Big)\bigg) \\
&= \frac{(N_1^*)^2}{(N^*)^2}\mathrm{Var}\Big(\widehat{\mathcal{C}}_{\cdot\cdot}^{AB}\big|\mathrm{Group}=1\Big) + \frac{(N_2^*)^2}{(N^*)^2}\mathrm{Var}\Big(\widehat{\mathcal{C}}_{\cdot\cdot}^{AB}\big|\mathrm{Group}=2\Big) \\
&\quad + \frac{(N_3^*)^2}{(N^*)^2}\mathrm{Var}\Big(\widehat{\mathcal{C}}_{\cdot\cdot}^{AB}\big|\mathrm{Group}=3\Big) \\
&\quad + 2\frac{N_1^*N_2^*}{(N^*)^2}\mathrm{Cov}\bigg(\Big(\widehat{\mathcal{C}}_{\cdot\cdot}^{AB}\big|\mathrm{Group}=1\Big), \Big(\widehat{\mathcal{C}}_{\cdot\cdot}^{AB}\big|\mathrm{Group}=2\Big)\bigg) \\
&\quad + 2\frac{N_1^*N_3^*}{(N^*)^2}\mathrm{Cov}\bigg(\Big(\widehat{\mathcal{C}}_{\cdot\cdot}^{AB}\big|\mathrm{Group}=1\Big), \Big(\widehat{\mathcal{C}}_{\cdot\cdot}^{AB}\big|\mathrm{Group}=3\Big)\bigg) \\
&\quad + 2\frac{N_2^*N_3^*}{(N^*)^2}\mathrm{Cov}\bigg(\Big(\widehat{\mathcal{C}}_{\cdot\cdot}^{AB}\big|\mathrm{Group}=2\Big), \Big(\widehat{\mathcal{C}}_{\cdot\cdot}^{AB}\big|\mathrm{Group}=3\Big)\bigg).
\end{aligned} \tag{2.16}$$

In MRMC studies, we make the assumptions: (1) readers are independent and identically distributed (i.i.d.); (2) cases are i.i.d; (3) readers are independent of cases. The assumptions imply that the reader-averaged concordance estimates for different groups are independent. Hence,

$$\mathrm{Cov}\bigg(\Big(\widehat{\mathcal{C}}_{\cdot\cdot}^{AB}\big|\mathrm{Group}=g_1\Big), \Big(\widehat{\mathcal{C}}_{\cdot\cdot}^{AB}\big|\mathrm{Group}=g_2\Big)\bigg) = 0, \tag{2.17}$$

for $g_1 \neq g_2$. The variance in (2.16) can thus be rewritten as

$$\begin{aligned}
\mathrm{Var}\Big(\widehat{\mathcal{C}}_{\cdot\cdot}^{AB}\Big) &= \frac{(N_1^*)^2}{(N^*)^2}\mathrm{Var}\Big(\widehat{\mathcal{C}}_{\cdot\cdot}^{AB}\big|\mathrm{Group}=1\Big) + \frac{(N_2^*)^2}{(N^*)^2}\mathrm{Var}\Big(\widehat{\mathcal{C}}_{\cdot\cdot}^{AB}\big|\mathrm{Group}=2\Big) \\
&\quad + \frac{(N_3^*)^2}{(N^*)^2}\mathrm{Var}\Big(\widehat{\mathcal{C}}_{\cdot\cdot}^{AB}\big|\mathrm{Group}=3\Big).
\end{aligned} \tag{2.18}$$

Since we consider each group as a fully-crossed study design, the variance of the reader-averaged concordance estimator for each group $g$ can be estimated with (1.42). Therefore, we propose the following split-analysis variance estimator (SAVE):

$$\begin{aligned}
\widetilde{\mathrm{Var}}\Big(\widehat{\mathcal{C}}_{\cdot\cdot}^{AB}\Big) &= \frac{(N_1^*)^2}{(N^*)^2}\widehat{\mathrm{Var}}\Big(\widehat{\mathcal{C}}_{\cdot\cdot}^{AB}\big|\mathrm{Group}=1\Big) + \frac{(N_3^*)^2}{(N^*)^2}\widehat{\mathrm{Var}}\Big(\widehat{\mathcal{C}}_{\cdot\cdot}^{AB}\big|\mathrm{Group}=2\Big) \\
&\quad + \frac{(N_3^*)^2}{(N^*)^2}\widehat{\mathrm{Var}}\Big(\widehat{\mathcal{C}}_{\cdot\cdot}^{AB}\big|\mathrm{Group}=3\Big),
\end{aligned} \tag{2.19}$$

where $\widehat{\mathrm{Var}}\Big(\widehat{\mathcal{C}}_{\cdot\cdot}^{AB}\big|\mathrm{Group}=\mathrm{g}\Big)$ is the variance estimator for the reader-averaged concordance estimator for group $g$, $g=1,2,3$. In other words, the proposed split-analysis estimator $\widetilde{\mathrm{Var}}\Big(\widehat{\mathcal{C}}_{\cdot\cdot}^{AB}\Big)$ is a linear combination of the variance estimators, each one of which is obtained by using (1.42).

We can show that SAVE is unbiased since

$$\begin{aligned}
\mathrm{E}\Big(\widetilde{\mathrm{Var}}\Big(\widehat{\mathcal{C}}_{\cdot\cdot}^{AB}\Big)\Big) &= \frac{(N_1^*)^2}{(N^*)^2}\mathrm{E}\Big(\widehat{\mathrm{Var}}\Big(\widehat{\mathcal{C}}_{\cdot\cdot}^{AB}\big|\mathrm{Group}=1\Big)\Big) + \frac{(N_2^*)^2}{(N^*)^2}\mathrm{E}\Big(\widehat{\mathrm{Var}}\Big(\widehat{\mathcal{C}}_{\cdot\cdot}^{AB}\big|\mathrm{Group}=2\Big)\Big) \\
&\quad + \frac{(N_3^*)^2}{(N^*)^2}\mathrm{E}\Big(\widehat{\mathrm{Var}}\Big(\widehat{\mathcal{C}}_{\cdot\cdot}^{AB}\big|\mathrm{Group}=3\Big)\Big) \\
&= \frac{(N_1^*)^2}{(N^*)^2}\mathrm{Var}\Big(\widehat{\mathcal{C}}_{\cdot\cdot}^{AB}\big|\mathrm{Group}=1\Big) + \frac{(N_2^*)^2}{(N^*)^2}\mathrm{Var}\Big(\widehat{\mathcal{C}}_{\cdot\cdot}^{AB}\big|\mathrm{Group}=2\Big) \\
&\quad + \frac{(N_3^*)^2}{(N^*)^2}\mathrm{Var}\Big(\widehat{\mathcal{C}}_{\cdot\cdot}^{AB}\big|\mathrm{Group}=3\Big) \\
&= \mathrm{Var}\Big(\widehat{\mathcal{C}}_{\cdot\cdot}^{AB}\Big).
\end{aligned} \tag{2.20}$$

# Chapter 3

# Simulation Study

## 3.1 Simulation Design

An MRMC simulation model was first introduced by Roe and Metz (1997) for MRMC receiver operating characteristic analysis. The model is a linear mixed effect model given by

$$X_{mrc} = \mu + \tau_m + R_r + C_c + RC_{rc} + \tau R_{mr} + \tau C_{mc} + \tau RC_{mrc} + E_{mrc}, \tag{3.1}$$

where $X_{mrc}$ denotes the rating (e.g. suspicion score) from reader $r$ for case $c$ using modality $m$. It is built on several fixed (Greek letters) and random effects: $\mu$ denotes the overall mean (fixed effect); $\tau_m$ denotes the modality-specific means (fixed effect); $R_r$ denotes the random effect for reader $r$; $C_c$ denotes the random effect for case $c$; $RC_{rc}$ denotes the random effect for reader $r$ on case $c$; $\tau R_{mr}$ denotes a random effect for reader $r$ using modality $m$; $\tau C_{mc}$ denotes the random effect for case $c$ in modality $m$; $\tau RC_{mrc}$ denotes the random effect for reader $r$ on case $c$ using modality $m$; and $E_{mrc}$ corresponds to the measurement error, within-reader variability (internal noise) that represents a readers' inability to reproduce their scores when the experiment is repeated under identical conditions. The original model also has a $t$ subscript on all

terms (except $\mu$). This subscript indicates the true state of the case: $t = 1$ indicates the diseased case, and $t = 0$ indicates the non-diseased case. We have left the $t$ subscript off this model since we are not splitting the cases by truth.

Except for the overall mean $\mu$ and the diagnostic capacity $\tau_m$, the remaining terms $R_r$, $C_c$, $RC_{rc}$, $\tau R_{mr}$, $\tau C_{mc}$, $\tau RC_{mrc}$, $E_{mrc}$ are assumed to be independent normal random variables with mean zero and variances $\sigma_R^2, \sigma_C^2, \sigma_{RC}^2, \sigma_{MR}^2, \sigma_{MC}^2, \sigma_{MRC}^2, \sigma_E^2$, respectively. Thus, the variance of the ratings is

$$\mathrm{Var}\Big(X_{mrc}\Big) = \sigma_R^2 + \sigma_C^2 + \sigma_{RC}^2 + \sigma_{MR}^2 + \sigma_{MC}^2 + \sigma_{MRC}^2 + \sigma_E^2. \tag{3.2}$$

Unfortunately, the simulation model in (3.1) cannot be used in MRMC agreement analysis since it does not yield different levels of inter-reader agreement for different pairs of readers, which is expected in MRMC agreement analysis. We briefly explore the reason as follows. We begin with the success function $S$ that compares two pairs of observations:

$$S(X_{mrc}, X_{mrc'}; X_{m'r'c}, X_{m'r'c'}) = \begin{cases} 1, & \text{if } (X_{mrc} - X_{mrc'})(X_{m'r'c} - X_{m'r'c'}) > 0 \\ 0, & \text{if } (X_{mrc} - X_{mrc'})(X_{m'r'c} - X_{m'r'c'}) \leq 0 \end{cases}.$$

The difference of the ratings for reader $r$ can be expressed as follows

$$\begin{aligned} W_r &= X_{mrc} - X_{mrc'} \\ &= \mu + \tau_m + R_r + C_c + RC_{rc} + \tau R_{mr} + \tau C_{mc} + \tau RC_{mrc} + E_{mrc} \\ &\quad - \mu - \tau_m - R_r - C_{c'} - RC_{rc'} - \tau R_{mr} - \tau C_{mc'} - \tau RC_{mrc'} - E_{mrc'} \\ &= C_c + RC_{rc} + \tau C_{mc} + \tau RC_{mrc} + E_{mrc} \\ &\quad - C_{c'} - RC_{rc'} - \tau C_{mc'} - \tau RC_{mrc'} - E_{mrc'}. \end{aligned} \tag{3.3}$$

Moreover,

$$W_r \sim N\Big(0, 2\sigma_C^2 + 2\sigma_{RC}^2 + 2\sigma_{MC}^2 + 2\sigma_{MRC}^2 + 2\sigma_E^2\Big). \tag{3.4}$$

Since $\sigma_C^2, \ldots, \sigma_E^2$ are fixed parameters, $W_r$ does not vary in distribution for different readers. Hence, we do not expect different levels of inter-reader agreement for different pairs of readers.

Gallas et al. (2016) adapted the model in (3.1) to better represent variables in an MRMC agreement study. The model is given by

$$X_{mrc} = \Lambda_c + RC_{rc} + \tau RC_{mrc} + RCE_{rc} + \tau RCE_{mrc}, \tag{3.5}$$

where $\Lambda_c$ denotes the mean rating for case $c$, $RC_{rc}$ denotes the random effect for reader $r$ on case $c$, $\tau RC_{mrc}$ denotes the random effect for reader $r$ on case $c$ using modality $m$, $RCE_{rc}$ denotes the replication error for reader $r$ on case $c$, $\tau RCE_{mrc}$, denotes the replication error for reader $r$ on case $c$ using modality $m$.

In Gallas et al. (2016), $\Lambda_c$ is sampled from a standard normal distribution, and $RC_{rc}$, $\tau RC_{mrc}$, $\tau RC_{mrc}$ and $\tau RCE_{mrc}$ are sampled from normal distributions:

$$\begin{aligned} RC_{rc} &\sim N\Big(0, (\sigma_r + \sigma_c)^2\Big), \\ \tau RC_{mrc} &\sim N\Big(0, (\sigma_{mr} + \sigma_{mc})^2\Big), \\ RCE_{rc} &\sim N\Big(0, (\sigma_{re} + \sigma_{ce})^2\Big), \\ \tau RCE_{mrc} &\sim N\Big(0, (\sigma_{mre} + \sigma_{mce})^2\Big), \end{aligned} \tag{3.6}$$

where $\sigma_r$, $\sigma_c$, $\sigma_{mr}$, $\sigma_{mc}$, $\sigma_{re}$, $\sigma_{ce}$, $\sigma_{mre}$ and $\sigma_{mce}$ are themselves random variables that are sampled from the exponential distributions with means $\mu_r$, $\mu_c$, $\mu_{mr}$, $\mu_{mc}$, $\mu_{re}$, $\mu_{ce}$, $\mu_{mre}$, $\mu_{mce}$, respectively.

We consider $\sigma_r, \sigma_{mr}$ to represent the variability of a reader; larger values of $\sigma_r, \sigma_{mr}$ correspond to a reader that gives readings with more variability. We consider the case terms $\sigma_c, \sigma_{mc}$ to reflect the variability a case evokes from readers; larger values of $\sigma_c, \sigma_{mc}$ correspond to cases that are more challenging to rate. The parameters $\sigma_{re}, \sigma_{mre}, \sigma_{ce}$, $\sigma_{mce}$ are similar to $\sigma_r, \sigma_{mr}, \sigma_c$, $\sigma_{mc}$, but they represent replication variability. Model (3.5) contains total eight parameters: the four parameters $\mu_r, \mu_{\tau r}, \mu_{re}, \mu_{\tau re}$ are interpreted as the mean variability of a reader, and the four parameters $\mu_c, \mu_{\tau c}, \mu_{ce}, \mu_{\tau ce}$ are interpreted as the mean variability evoked by a case.

## 3.2 Simulation Results

The goal for the simulation study is to assess and compare the performance of the two variance estimators FAVE and SAVE. We run $N_{MC} = 10000$ trials of simulation, and use the Monte Carlo average and variances as surrogates for the true means and variances of the concordance measures.

Following Gallas et al. (2016), we generate data from the model in (3.5) with parameters

$$\begin{aligned} \mu_r &= \mu_{\tau r} = \mu_{re} = \mu_{\tau re}, \\ \mu_c &= \mu_{\tau c} = \mu_{ce} = \mu_{\tau ce}, \end{aligned} \tag{3.7}$$

where $\mu_r$ measures the mean reader variability, and $\mu_c$ captures the mean case variability. We choose to vary the parameters in a factorial way by exploring two levels of variabilities: $\mu_r = \{0.05, 0.4\}$ and $\mu_c = \{0.05, 0.4\}$. We investigate three study designs in our simulation: (1) a fully-crossed study design with 15 readers, 50 cases and two modalities; (2) a balanced split-plot study design with 15 readers, 150 cases, two modalities and three groups; and (3) an unbalanced split-plot study design with 14 readers, 164 cases, two modalities and three groups. Specifically, in the balanced split-plot study design, there are five readers and 50 cases in each group. In the unbalanced split-plot study design, there are five readers and 47 cases for the first two groups, and four readers and 70 cases for the third group. Note that in each study design above, there are total 750 data points, and we will expect that the "cost" of each study design is the same.

Table 3.1-3.3 summarize the following measurements:

1. $\widehat{C}_{\cdot\cdot}^{AB}$: the Monte Carlo average (MC-average) of all the reader-averaged concordance estimates across simulation;

2. $\mathrm{Var}\big(\widehat{C}^{AB}\big)$: the Monte Carlo variance (MC-variance) of the reader-averaged concordance estimates across simulation;

3. $\widehat{\text{Var}}\big(\widehat{C}^{AB}\big)$: the MC-average of the variance estimates based on the FAVE;

4. $\widetilde{\text{Var}}\big(\widehat{C}^{AB}\big)$: the MC-average of the variance estimates based on the SAVE;

5. $\text{CV}\Big(\widehat{\text{Var}}\big(\widehat{C}^{AB}_{\cdot\cdot}\big)\Big)$: the square root of MC-variance of the variance estimates based on FAVE divided by item 2, that is,

$$\text{CV}\Big(\widehat{\text{Var}}\big(\widehat{C}^{AB}_{\cdot\cdot}\big)\Big) = \frac{\sqrt{\text{MC-Variance}\Big(\widehat{\text{Var}}\big(\widehat{C}^{AB}_{\cdot\cdot}\big)\Big)}}{\text{Var}\big(\widehat{C}^{AB}\big)};$$

6. $\text{CV}\Big(\widetilde{\text{Var}}\big(\widehat{C}^{AB}_{\cdot\cdot}\big)\Big)$: the square root of MC-variance of the variance estimates based on SAVE divided by item 2, that is,

$$\text{CV}\Big(\widetilde{\text{Var}}\big(\widehat{C}^{AB}_{\cdot\cdot}\big)\Big) = \frac{\sqrt{\text{MC-Variance}\Big(\widetilde{\text{Var}}\big(\widehat{C}^{AB}_{\cdot\cdot}\big)\Big)}}{\text{Var}\big(\widehat{C}^{AB}\big)}.$$

Table 3.1: Simulation results for the fully-crossed study design

| $\mu_R$ | $\mu_C$ | $\widehat{\mathcal{C}}^{AB}_{\cdot\cdot}$ | $\text{Var}\big(\widehat{C}^{AB}_{\cdot\cdot}\big)$ | $\widehat{\text{Var}}\big(\widehat{C}^{AB}_{\cdot\cdot}\big)$ | $\widetilde{\text{Var}}\big(\widehat{C}^{AB}_{\cdot\cdot}\big)$ | $\text{CV}\Big(\widehat{\text{Var}}\big(\widehat{C}^{AB}_{\cdot\cdot}\big)\Big)$ | $\text{CV}\Big(\widetilde{\text{Var}}\big(\widehat{C}^{AB}_{\cdot\cdot}\big)\Big)$ |
|---|---|---|---|---|---|---|---|
| 0.05 | 0.05 | 0.896 | 1.911e-4 | 1.861e-4 | 1.861e-4 | 0.441 | 0.441 |
| 0.05 | 0.4 | 0.654 | 4.637e-4 | 4.571e-4 | 4.571e-4 | 0.214 | 0.214 |
| 0.4 | 0.05 | 0.664 | 6.649e-4 | 6.513e-4 | 6.513e-4 | 0.231 | 0.231 |
| 0.4 | 0.4 | 0.579 | 2.584e-4 | 2.540e-4 | 2.540e-4 | 0.320 | 0.320 |

As expected, for the fully-crossed study design, the FAVE and SAVE are identical as we can show that

$$\widehat{\text{Var}}\big(\widehat{C}^{AB}_{\cdot\cdot}\big) = \widetilde{\text{Var}}\big(\widehat{C}^{AB}_{\cdot\cdot}\big)$$

and both estimators are unbiased as

$$\widehat{\text{Var}}\big(\widehat{C}^{AB}_{\cdot\cdot}\big) = \widetilde{\text{Var}}\big(\widehat{C}^{AB}_{\cdot\cdot}\big) \approx \text{Var}\big(\widehat{C}^{AB}_{\cdot\cdot}\big).$$

Table 3.2: Simulation results for the balanced split-plot study design

| $\mu_R$ | $\mu_C$ | $\widehat{\mathcal{C}}_{\cdot\cdot}^{AB}$ | $\mathrm{Var}\big(\widehat{C}_{\cdot\cdot}^{AB}\big)$ | $\widehat{\mathrm{Var}}\big(\widehat{C}_{\cdot\cdot}^{AB}\big)$ | $\widetilde{\mathrm{Var}}\big(\widehat{C}_{\cdot\cdot}^{AB}\big)$ | $\mathrm{CV}\Big(\widehat{\mathrm{Var}}\big(\widehat{C}_{\cdot\cdot}^{AB}\big)\Big)$ | $\mathrm{CV}\Big(\widetilde{\mathrm{Var}}\big(\widehat{C}_{\cdot\cdot}^{AB}\big)\Big)$ |
|---|---|---|---|---|---|---|---|
| 0.05 | 0.05 | 0.896 | 8.988e-5 | 8.926e-5 | 8.957e-5 | 0.728 | 0.370 |
| 0.05 | 0.4 | 0.654 | 2.088e-4 | 2.069e-4 | 2.096e-4 | 0.631 | 0.250 |
| 0.4 | 0.05 | 0.664 | 4.738e-4 | 4.746e-4 | 4.709e-4 | 0.625 | 0.359 |
| 0.4 | 0.4 | 0.579 | 1.721e-4 | 1.714e-4 | 1.720e-4 | 0.650 | 0.431 |

For the balanced study design, while the variance estimates from both methods are unbiased as

$$\widehat{\mathrm{Var}}\big(\widehat{C}_{\cdot\cdot}^{AB}\big) = \widetilde{\mathrm{Var}}\big(\widehat{C}_{\cdot\cdot}^{AB}\big) \approx \mathrm{Var}\big(\widehat{C}_{\cdot\cdot}^{AB}\big),$$

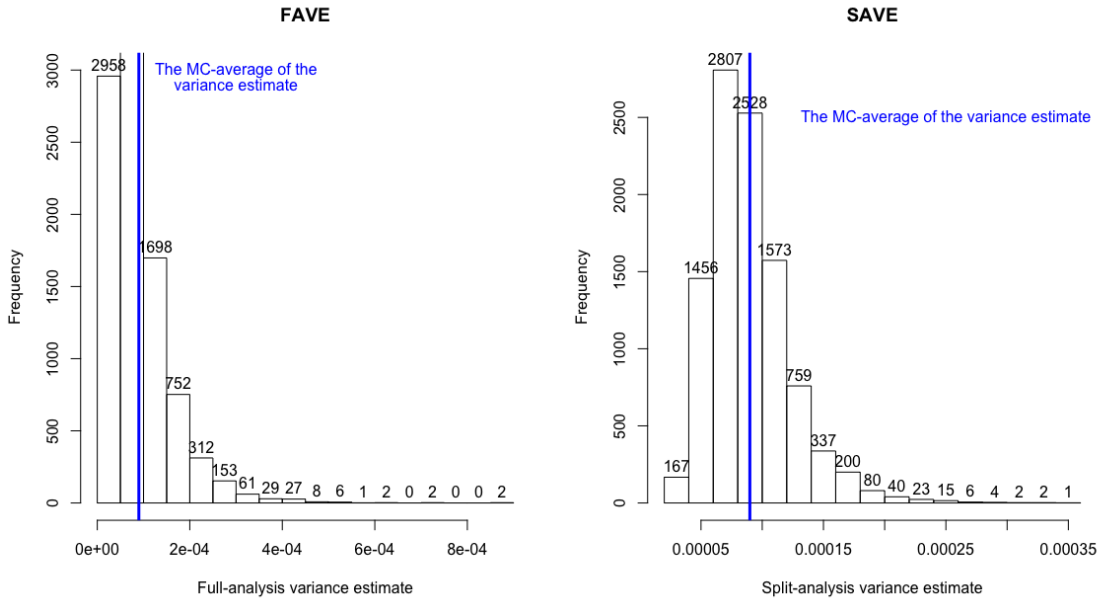
there is a noticeable amount of difference between the coefficient of variation (CV) for the variance estimates. Compared to the CV for the estimates from FAVE, the CV for the estimates from SAVE is lower. That means the SAVE gives a variance estimate with lower variability. This fact can also be observed from the histograms in Figure 3.1. As we can see in the histograms, the full-analysis variance estimates are more spread than the split-analysis variance estimates; most of the variance estimates from FAVE are between 0 and 0.0002, while most of the estimates from SAVE are between 0.00005 and 0.00015.

Table 3.3: Simulation results for the unbalanced split-plot study design

| $\mu_R$ | $\mu_C$ | $\widehat{\mathcal{C}}_{\cdot\cdot}^{AB}$ | $\mathrm{Var}\big(\widehat{C}_{\cdot\cdot}^{AB}\big)$ | $\widehat{\mathrm{Var}}\big(\widehat{C}_{\cdot\cdot}^{AB}\big)$ | $\widetilde{\mathrm{Var}}\big(\widehat{C}_{\cdot\cdot}^{AB}\big)$ | $\mathrm{CV}\Big(\widehat{\mathrm{Var}}\big(\widehat{C}_{\cdot\cdot}^{AB}\big)\Big)$ | $\mathrm{CV}\Big(\widetilde{\mathrm{Var}}\big(\widehat{C}_{\cdot\cdot}^{AB}\big)\Big)$ |
|---|---|---|---|---|---|---|---|
| 0.05 | 0.05 | 0.896 | 9.008e-5 | 8.130e-5 | 9.048e-5 | 17.098 | 0.480 |
| 0.05 | 0.4 | 0.654 | 2.003e-4 | 2.033e-4 | 2.001e-4 | 7.358 | 0.338 |
| 0.4 | 0.05 | 0.664 | 4.986e-4 | 4.326e-4 | 5.054e-4 | 5.812 | 0.460 |
| 0.4 | 0.4 | 0.579 | 1.743e-4 | 1.564e-4 | 1.760e-4 | 8.041 | 0.573 |

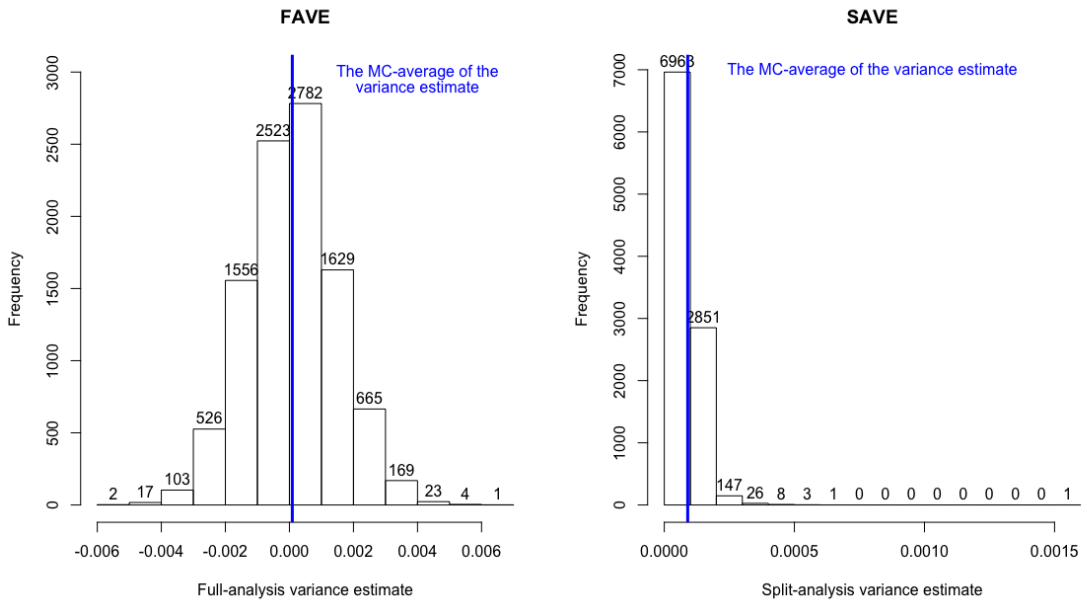
For the unbalanced split-plot study design, the major difference between the two

Figure 3.1: Variance estimates for the balanced split-plot study design with low reader variability and low case variability




methods is the CV for the variance estimates. Even though the variance estimates from both methods are centered around the truth, $\mathrm{Var}\left(\widehat{C}_{\cdot\cdot}^{AB}\right)$, the variance estimates from the split-analysis method have much smaller variability than those from the full-analysis method. The large variability of the FAVE may lead to negative variance estimates. We can observe from the histograms in Figure 3.2 that in this simulation study, almost half of the variance estimates from the FAVE are negative, whereas none of the variance estimates from the SAVE is negative. This suggests that the split-analysis provides more reasonable variance estimates and thus should be preferred for unbalanced split-plot designs.

Figure 3.2: Variance estimates for the unbalanced split-plot study design with low reader variability and low case variability



# Chapter 4

# Conclusion and Future Work

In this study, we propose SAVE, the split-analysis variance estimator, for split-plot study designs. SAVE works well for split-plot study designs, particularly for unbalanced split-plot study designs compared to the FAVE. Our numerical studies show significant improvement in the precision from FAVE to SAVE.

There are several directions for future work in this area. One direction is to adapt the current SAVE to analyze alternative study designs, where the readers and/or cases overlap across groups. This requires the derivation of an efficient estimator for the covariances of concordances between groups. Another direction for future work is to analytically derive reader-averaged concordance and its variance for the simulation model provided here. In other words, one could derive the underlying distribution for the reader-averaged concordance given the simulation model.

# Bibliography


[1] Alan Agresti. *Analysis of ordinal categorical data*, volume 656. John Wiley & Sons, 2010.

[2] Weijie Chen, Qi Gong, and Brandon D Gallas. Efficiency gain of paired split-plot designs in mrmc roc studies. volume 10577, pages 10577 – 10577 – 8, 2018. doi: 10.1117/12.2293741.

[3] Brandon D Gallas and David G Brown. Reader studies for validation of cad systems. *Neural Networks*, 21(2):387–397, 2008.

[4] Brandon D Gallas, Amrita Anam, Weijie Chen, Adam Wunderlich, and Zhiwei Zhang. Mrmc analysis of agreement studies. In *SPIE Medical Imaging*, pages 97870F–97870F. International Society for Optics and Photonics, 2016.

[5] Leo A Goodman and William H Kruskal. Measures of association for cross classifications. *Journal of the American statistical association*, 49(268):732–764, 1954.

[6] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2): 81–93, 1938.

[7] Jae-On Kim. Predictive measures of ordinal association. *American Journal of Sociology*, 76(5):891–907, 1971.

[8] Nancy A Obuchowski. Reducing the number of reader interpretations in mrmc studies. *Academic radiology*, 16(2):209–217, 2009.

[9] Nancy A Obuchowski, Brandon D Gallas, and Stephen L Hillis. Multi-reader roc studies with split-plot designs: a comparison of statistical methods. *Academic radiology*, 19(12):1508–1517, 2012.

[10] Ronald H. Randles and Douglas A. Wolfe. *Introduction to the Theory of Nonparametric Statistics*. John Wiley and Sons, New York, 1979.

[11] Cheryl A Roe and Charles E Metz. Dorfman-berbaum-metz method for statistical analysis of multireader, multimodality receiver operating characteristic data: validation with computer simulation. *Academic radiology*, 4(4):298–303, 1997.

[12] Robert H Somers. A new asymmetric measure of association for ordinal variables. *American sociological review*, pages 799–811, 1962.

[13] Robert F Wagner, Sergey V Beiden, Gregory Campbell, Charles E Metz, and William M Sacks. Assessment of medical imaging and computer-assist systems: lessons from recent experience. *Academic radiology*, 9(11):1264–1277, 2002.