Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Yi Pan Date

A New Permutation-Based Method for Assessing Observer Agreement with Replicated and Repeated Observations

By

Yi Pan

Doctor of Philosophy

Biostatistics

Michael J. Haber, Ph.D.
Advisor

Huiman X. Barnhart, Ph.D.
Committee Member

Lance A. Waller, Ph.D.
Committee Member

Ying Guo, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

A New Permutation-Based Method for Assessing Observer Agreement with Replicated and Repeated Observations

By

Yi Pan
B.S., Nanjing University, 2004
M.S., Auburn University, 2006
M.S., Emory University, 2010

Advisor: Michael J. Haber, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2011

Abstract

The area of observer agreement has rapidly developed over the last half-century. A substantial number of coefficients and approaches have been developed and used to assess the agreement between different observers or methods of measurement. In this dissertation, a new permutation-based coefficient for the evaluation of agreement between two observers making replicated and repeated binary or quantitative measurements is introduced. The new coefficient of individual equivalence (CIE) compares the observed disagreement between the observers to its expected value under the hypothesis of individual equivalence. This hypothesis states that for each subject, the conditional distributions of the readings of the two observers are identical. Therefore, from a statistical viewpoint, it does not matter which observer makes the reading on this subject. In other words, under individual equivalence the observers can be used interchangeably.

Let K and L denote the number of replicated observations that are available from observers X and Y, respectively, on a given subject. Then the expected disagreement under individual equivalence for this subject is based on the $K+L$ choose K possible assignments of X's and Y's to the $K+L$ observations made on this subject. Under individual equivalence, all these assignments have the same probability. Simple methods for nonparametric estimation of the new coefficient and its standard error are derived for both binary and continuous outcomes. Furthermore, model-based approaches are developed for estimation of the CIE for binary and continuous assessments. Model-based methods for estimation of the CIE from repeated binary outcomes are also discussed. Simulation studies confirm the validity of the estimated coefficient and its standard error. Finally, the new coefficient is compared with the coefficient of individual agreement (CIA), Kappa statistic and the concordance correlation coefficient (CCC). Examples with binary and continuous outcomes are used to illustrate the new coefficient. One example involves the evaluation of mammograms by ten radiologists

and another one compares magnetic resonance angiography (MRA) techniques for noninvasive screening of carotid stenosis to an invasive intra-arterial angiogram (IA) method.

A New Method for Assessing Observer Agreement with Replicated and Repeated Observations

By

Yi Pan
B.S., Nanjing University, 2004
M.S., Auburn University, 2006
M.S., Emory University, 2010

Advisor: Michael J. Haber, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2011

Acknowledgement

I would like to express my depest appreciation to my advisor, Dr. Michael, J. Haber for his patient, friendly, and unfailing support over the past years. He is a great supervisor and a close friend!

I would like to thank my committee members, Dr. Huiman X, Barnhart, Dr. Lance A. Waller and Dr. Ying Guo, for their critical help and suggestions on my dissertation along the way. I cannot achieve all those without their help. Appreciation is also due to Dr. Josep L. Carrasco from University of Bacelona for his creative suggestions and I'm looking forward to future collaborations with him. I also want to thank Dr. John J. Halfelt and Dr. Qi Long for their unconditional help on my dissertation work. During my years at Emory, I had great collaboration experience with Dr. John J. Hanfelt and Dr. Robert H. Lyles working on different research projects. Furthermore, I want to thank Kirk A. Easley who trained me in statistical consulting and gave me a lot of consulting opportunities. Special thanks go to Penina Haber who encouraged me so much. Also, I sincerely want to thank many fellow friends Ruosha Li, Jingjing Gao, Ming Wang, Shuo Chen, Li Li and a lot of others who helped and encouraged me during the past years. I cannot make any of those without them.

Finally I am indebted to my parents for encouraging me to pursue this degree and supporting me every step in my life. Last but not the least, I want to thank my husband for his great patience and kindness. The encouragement and support he gives me make this thesis as much his as it is mine.

Contents

List of Figures

List of Tables

Chapter 1

Background and Motivation

1.1 What is Agreement Study?

The topic of agreement among different raters is of importance in many domains of our life. For example, in the Olympic Games, the medals and ranking in gymnastics and diving are based on the scoring of several judges on several different disciplines. Usually, the extreme ratings (highest and lowest) are removed from the pool of scores and used for the ranking (Von Eye and Young Mun 2005).

In the academic area, for example, in Phase II cancer clinical trials, the effect of treatment is measured by imaging devices like computed tomography (CT) or magnetic resonance imaging (MRI), and the success of the treatment is usually decided by a team of radiologists, oncologists and surgeons. Obviously, agreement among the radiologists plays a major role in cancer treatment. Consensus between members of the diagnostic team is necessary in order to judge a treatment as a failure, a success, or worthy of further consideration (Broemeling 2009).

The earliest mention of agreement can be traced back to 1889 and 1901 when Pearson (Liao and Lewis 2000, Haggard 1958, Song 2003) studied fraternal assemblance in genetics by using the correlation coefficient. During that time, people could

not tell the difference between agreement and correlation so that correlation coefficient is used as an index to assess agreement. Later, this question has been corrected by several researchers (Lin 1989, Muller and Petra 1994). The correlation coefficient, which measures the linear association between two variables, only provides partial information about agreement. For example (Haber and Barnhart 2006), if the readings from one rater are exactly twice of the second rater, the correlation of coefficient between the two raters is 1. However, the actual agreement could be low.

Many agreement studies during the early times were in psychological research. Cohen's original kappa, which is the most popular coefficient to measure agreement between two raters with binary or nominal scales of readings, was published in *Educational and Psychological Measurement* in 1960. Furthermore, the weighted kappa which can be extended to ordinal scales of measurements was presented in *Psychological Bulletin* in 1968.

Later on, agreement studies have been heavily used in medical related research. In clinical studies, agreement is often concerned with assessing whether different observers, such as raters, methods or instruments for measuring both continuous and discrete outcomes produce similar results. We are always interested in whether a new rater can replace the standard rater if the new one is less expensive and can reproduce the same or comparable outcome, or whether a new rater and the existing one can be used interchangeably at individual level.

Traditionally, other terminologies such as validity, reliability, repeatability or reproducibility may have been used in studies which are designed to measure agreement. Those terminologies have been explicitly defined in Barnhart et al. (2007b). In general, agreement is defined as an index to evaluate the magnitude of the conformity of readings compared to a true value when the gold standard is available or the consistency of multiple readings when the true value is unavailable (Song 2003). Vangeneugden et al. (2005), Molenberghs et al. (2007), Barnhart et al. (2007b) compared

agreement with reliability: agreement assesses the degree of closeness between readings within a subject, while reliability assesses the degree of differentiation between subjects; i.e., the ability to tell subjects apart from each other within a population.

In this dissertation, we motivate our research through two examples. Next, we introduce a new coefficient, the coefficient of individual equivalence (CIE), which measures agreement for both qualitative and quantitative replicated assessments. The existing methods for measuring agreement for binary, continuous and repeated binary outcomes are reviewed in the following sections. Later, a Bayesian approach and its application to agreement studies are briefly illustrated.

1.2 Motivating Examples

1.2.1 Mammogram Example

In a study described by Elmore et al. (1994), 150 female patients underwent a mammography at the Yale-New Haven Hospital in 1987. Each of ten radiologists read each patient's mammogram and classified it into one of four diagnosis categories: (1) normal, (2) abnormal - probably benign, (3) abnormal - intermediate or (4) abnormal - suggestive of cancer. Four months later, the same films were reviewed again, in a random order, by the same radiologists. We considered the two evaluations as replications. In the present analysis, we considered a radiologist's rating as "positive" if the mammogram was classified into the fourth category, which was abnormal and suggestive of cancer. Otherwise, the rating was considered as "negative". Each of the study participants was followed up for three years, and then a definitive diagnosis was made. The definitive diagnosis was breast cancer if it was histopathologically confirmed within the three years of follow-up. We considered this diagnosis as the patient's "true" breast cancer status. Based on this criterion, 27 of 150 patients (18%) had breast cancer. Ten radiologists were involved. Table 1.1 presents the proportions of positive ratings as well as sensitivity and specificity for the ten radiologist. Since

Table 1.1: Proportions of positive ratings, sensitivity and specificity for each radiologist in the mammography study

Radiologist	Proportion rated positive	Sensitivity	Specificity
A	0.208	0.815	0.927
B	0.120	0.630	0.967
C	0.077	0.333	0.980
D	0.223	0.778	0.898
E	0.180	0.704	0.935
F	0.160	0.722	0.963
G	0.177	0.574	0.911
H	0.107	0.500	0.980
I	0.280	0.796	0.833
J	0.240	0.685	0.858

the total of sensitivity and specificity was highest for radiologist A, we illustrated the new coefficients by estimating the agreement between radiologist A and each of the remaining nine radiologists. Radiologist A was considered as the reference when we used methods for comparing a new observer to a refence observer. A summary table (Table 1.2) which showed the classification of diagnostic interpretation of patients from all ten radiologists. Twenty seven patients were confirmed with cancer; while 123 patients did not present apparent symptoms of breast cancer.

Table 1.2: Diagnostic interpretation for Mammogramphy data

Diagnostic Interpretation	Cancer ($n = 27\,(18\%)$)		Absence of Cancer ($n = 123\,(82\%)$)	
	Reading 1	Reading 2	Reading 1	Reading 2
Normal	22	16	507	480
Abnormal, probably benign	28	21	399	351
Abnormal, indeterminate	46	54	234	303
Abnormal, suggestive of cancer	174	179	90	94

1.2.2 Carotid Stenosis Example

This data set is from a carotid stenosis screening study conducted at Emory University from 1994 to 1996. The study was designed to determine the suitability of magnetic resonance angiography (MRA) for noninvasive screening of carotid artery stenosis, compared to invasive intra-arterial angiogram (IA). The main interest was in comparing two MRA techniques, two-dimensional (MRA-2D) and three-dimensional (MRA-3D) MRA time of flight, to the IA, which was considered as the "gold standard". In this example, the three screening methods were considered as the "observers". Readings were made by each of three raters using each of the three methods to assess carotid stenosis on each of the 55 patients. For this illustration, the three readings made by different raters were considered as replications. Separate readings were made on the left and right carotid arteries. However, in this example, our interest was restricted to the left side.

1.3 Existing Methods for Discrete Outcomes

Qualitative outcomes often occur in both medical and social science. The observers may be physicians who classify patients as having or not having a medical condition, or competing diagnostic devices that classify the extent of disease in patients into either nominal or ordinal categories. Cohen's original Kappa coefficient (Cohen 1960) was introduced half a century ago and it still serves as the most widely employed coefficient to assess rater agreement for discrete outcomes. In this section, existing techniques that are applied to discrete outcomes including binary, nominal and ordinal measurements are discussed. The original Cohen's Kappa is introduced, following by the weighted Kappa. More importantly, the discussion of Kappa coefficients is reviewed.

1.3.1 The Cohen's Kappa Coefficient

The original Kappa is an extension of a chance-corrected measure introduced by Scott (1955) and was extended by Cohen (1960). Originally, it was proposed to measure the agreement between two observers, each of which classifies n subjects into m mutually exclusive categories. Cohen's Kappa coefficient is well known since it takes chance agreement into consideration. It indicated that the observed cases of agreement include some cases for which the agreement was by chance only.

Cohen's Kappa was originally proposed for two observers and two or more nominal classifications. Cohen (1968) later extended his method to multiple ordinal classifications. Fleiss (1971) extended Cohen's Kappa to the case in which each of a sample of subjects is rated on a nominal scale by the same number of raters, but in which the raters rating one subject are not necessarily the same as those rating another (Fleiss et al. 1979).

Let's consider two raters 1 and 2 who made assignments of level of depression on

m patients, each outcome can be represented by 1, 2 or 3 which are considered as nominal levels. The cross-classification of two raters' judgements can be depicted as given in Table 1.3.

Table 1.3: Cross-Classification of Two Raters' Judgements

		Rater 1		
		1	2	3
	1	m_{11}	m_{12}	m_{13}
Rater 2	2	m_{21}	m_{22}	m_{23}
	3	m_{31}	m_{32}	m_{33}

To estimate Cohen's Kappa for two raters, we use the observed frequencies as shown in Table 1.3 to calculate the probabilities of each observer. Let p_{ij} be the probability of cell ij. Then the observed probability of those agreement cells can be defined as $\theta_1 = \sum_{i=1}^{I} p_{ii}$, where I denotes the total number of categories, for instance, I=3 in the above example. We also assume those two raters did not influence each other when they assign scores to the depression level of those patients. Under such assumption, we estimate the proportion when the two raters agreed by chance (under independence) as $\theta_2 = \sum_{i=1}^{I} p_{i.}p_{.j}$, where . indicates the marginal summed across. If $\theta_1 - \theta_2$ is positive, then the two raters agree more often than expected by chance, while if $\theta_1 - \theta_2$ is negative, then the two raters agree less often than expected by chance. Cohen's Kappa extended Scott (1955) index which defined θ_2 using the underlying assumption that the distribution of proportions over m categories for the population is known and is equal for the two raters. Therefore, if the two raters are interchangeable, which is equivalent to say that the marginal distributions are identical, then Cohen's Kappa and Scott's index are the same (Banerjee 1999).

As a result,

$$\kappa = \frac{\theta_1 - \theta_2}{1 - \theta_2}.$$

An estimated κ under a multinomial sampling scheme can be defined as

$$\hat{\kappa} = \frac{N \sum_i m_{ii} - \sum_i m_{i.} m_{.i}}{N^2 - \sum_i m_{i.} m_{.i}},$$

where $i = 1, \ldots, I$ are the categories used by the raters, N is the sample size and m denotes the observed frequencies. $\kappa = 1$ corresponds to perfect agreement while $\kappa = 0$ indicates lack of agreement (i.e. purely random coincidences of observers). A negative value of κ would mean that $\theta_1 - \theta_2$ is negative. In this case, the two raters agree less often than expected by chance.

To make inference and test the hypothesis to determine whether κ is significantly different from zero, one can use the formula presented by Fleiss et al. (1969). Interestingly, it's a long story to achieve the correct variance estimation for Kappa. "*Many human endeavors have been cursed with repeated failures before final success. The scaling of Mount Everest is one example. The discovery of the Northwest Passage is a second. The derivation of a correct standard error for Kappa is a third*" (Fleiss et al. 1979). Fleiss et al. (1969) pointed out the standard error of Kappa published by Cohen (1960), Cohen (1968), Everitt (1968) to be incorrect. Instead, they provided an estimated asymptotic variance for $\hat{\kappa}$, expressed as

$$\begin{aligned} \widehat{\mathrm{Var}}(\kappa) &= \frac{1}{n(1-\theta_2)^2} \left(\sum_{i=1}^{2} \hat{P}_{ii} \left[1 - (\hat{P}_{i.} + \hat{P}_{.i})(1-\hat{\kappa}) \right]^2 \right. \\ &\quad \left. + (1+\hat{\kappa})^2 \sum_{i \neq j}^{2} \hat{P}_{ij} (\hat{P}_{i.} + \hat{P}_{.j})^2 - [\hat{\kappa} - p_c(1-\hat{\kappa})]^2 \right) \end{aligned} \tag{1.1}$$

Von Eye and Young Mun (2005) summarized some interesting characteristics of κ.

- The range of κ is $-\infty < \kappa \leq 1$; the smallest possible value of $\hat{\kappa}$ is $1 - N/(1 - \sum_i m_{ii})$, where N is the sample size and m denotes the frequency in cell ii.

- $\kappa = 1$ only if the probability in the disagreement (off-diagonal) cells is zero.
- κ is defined only if at least two categories are used by both raters, that is, if the probability, p_{ij} is greater than zero for at least two cells.
- Multiplied by 100, κ indicates the percentage by which two raters' agreement exceeds the agreement that could be expected from chance.

Landis and Koch (1977) provided a table for the interpretation of Kappa values (Table 1.4). However, this table was produced mainly based on personal opinions with no statistical evidence to support it. In fact, it has been noted that these guidelines may be misleading, as the number of categories and subjects have impact on the magnitude of the values. The Kappa value increases as the number of categories decreases (Sim and Wright 2005).

Table 1.4: Interpretation of Kappa values by Landis and Koch (1977)

κ	Interpretation
< 0	No agreement
0.0 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

In contrast, Fleiss (1981) suggests the categories (Von Eye and Young Mun 2005).

Table 1.5: Interpretation of Kappa values by Fleiss (1981)

κ	Interpretation
< 0.4	Poor agreement
0.4 – 0.75	good agreement
> 0.75	excellent agreement

Von Eye and Young Mun (2005) suggested researchers that make explicit which guidelines they refer to when report poor, good or excellent agreement using κ. It is well known that even small values of κ can be significant if the sample size is large. Therefore, researchers typically report (1) κ itself, with 95% CI (2) the results from significance tests and (3) other estimates such as the coefficient of raw agreement.

1.3.2 The Weighted Kappa Coefficient

In the above section, we discussed Cohen's Kappa for binary and nominal scales. Cohen (1968) also proposed the weighted Kappa statistic, which provides a measure of agreement between two observers classifying observations into one of $m\,(m > 2)$ ordinal categories. The weighted Kappa coefficient is a generalization of the Kappa statistic to the situations where the categories are weighted by an objective or subjective function.

Recall in Table 1.3, if we redefine the three levels of scores of depression, where 1="not depressed", 2="mildly depressed", and 3="clinically depressed", weighted Kappa needs to be involved to estimate the agreement between two psychiatrists if we wish to assign unequal weights to each of the levels. We redefine θ_1^* as

$$\theta_1^* = \sum_{i=1}^{I}\sum_{j=1}^{J} w_{ij}p_{ii}$$

and θ_2^* as

$$\theta_2^* = \sum_{i=1}^{I}\sum_{j=1}^{J} w_{ij}p_{i.}p_{.j},$$

where w_{ij} are the weights. A weight $w_{ij},\ 0 \leq w_{ij} \leq 1$ is assigned to each cell $(i,\, j)$. The weight w_{ij} quantifies the degree of disagreement between the i^{th} and j^{th} categories. The cells on the diagonal of the table of occurrences corresponding to identical categorizations by both observers, receive weights of one, i.e. $w_{ii} = 1$. The

cells (i, j) with highly different categories i and j are given relatively small weights w_{ij}; whereas large weights w_{ij} are assigned when the respective classes i and j are not far distant. Cohen (1968) requires for the weights that

- $0 \leq w_{ij} \leq 1$, and
- they be ratios.

The second requirement means that if the $w_{ij} = 0.8$ indicates agreement that weights twice that of $w_{ij} = 0.4$. Based on the defined θ_1^* and θ_2^*, the weighted Kappa can be written as

$$\kappa = \frac{\theta_1^* - \theta_2^*}{1 - \theta_2^*}.$$

while the estimation can be defined analogously as

$$\hat{\kappa_w} = \frac{N \sum_i \sum_j w_{ij} m_{ii} - \sum_i \sum_j w_{ij} m_{i.} m_{.i}}{N^2 - \sum_i \sum_j w_{ij} m_{i.} m_{.i}},$$

The maximum value for κ_w is one indicating a complete agreement between two raters; whereas a value of zero corresponds to no agreement better than chance, and negative values show worse than chance agreement. Cohen's Kappa κ is a special case of weighted Cohen's Kappa with weights $w_{ii} = 1$ and $w_{ij} = 0,\ i \neq j$.

Agresti (1989) tried to model agreement with Kappa as parameter. He presented a simple quasi-symmetric model for which Kappa contains all relevant information about the structure of agreement and disagreement. As a measure of agreement, κ and κ_w also have been extended to cases with dependent samples (Williamson et al. 2000, Donner et al. 2000). Barnhart and Williamson (2002) presented an elegant method to estimate and compare correlated Kappa coefficients using weighted least squares (WLS). Their method applies to both Cohen's Kappa and the weighted Kappa. Guo and Manatunga (2005) used local Kappa coefficients to develop a method to assess the agreement between two discrete survival times that are measured on the same subject

by different raters or methods and their method can be extended to multivariate discrete survival distributions.

1.3.3 Critiques on Kappa Coefficients

Besides the development of Kappa coefficients for measuring agreement on binary, nominal and ordinal scales, numerous papers commented on this coefficient and tried to find alternative methods to the study of interrater agreement. Banerjee (1999) provided us with a thorough review of methods beyond Kappa. Some authors (Hutchison 1993) believed that Cohen's Kappa mixes two components of agreement that are inherently different, namely, disagreements which occur due to bias between the raters, and disagreements which occur because the raters rank-order the subjects differently. Thompson and Walter (1988) also stated that one of the limitations of Kappa is that it does not distinguish various types and sources of disagreements.

Furthermore, Kraemer (1979) and Thompson and Walter (1988) showed that besides the sensitivity and specificity for each observer, the prevalence of the characteristic of interest such as the rareness of a disease greatly influence the values of Kappa. Thompson and Walter (1988) showed that under the independence of the errors of the two dichotomous categories, κ can be rephrased as an index of validity using sensitivities, specificities and prevalence, that is

$$\kappa = \frac{2\theta(1-\theta)(1-\alpha_1-\beta_1)(1-\alpha_2-\beta_2)}{\pi_1(1-\pi_2)+\pi_2(1-\pi_1)} \tag{1.2}$$

where

$$\begin{aligned} \theta &= \text{true proportion having the characteristic} \\ 1-\alpha_i &= \text{specificity for } i^{\text{th}} \text{ observer } (i=1,\,2) \\ 1-\beta_i &= \text{sensitivity for } i^{\text{th}} \text{ observer } (i=1,\,2) \\ \pi_i &= \theta(1-\beta_i)+(1-\theta)\alpha_i \ (i=1,\,2) \end{aligned}$$

Equation (1.2) reveals that κ strongly depends on the true prevalence of the condition being diagnosed. As a result, since the true sensitivities, specificities and prevalence are unknown in reality, the heavy dependence of Kappa on the prevalence puts the interpretation and understanding of Kappa as a measurement for agreement in a questionable position. The comparison of two Kappa values may be jeopardized if the underlying prevalence for the situations are far apart. Particularly, it is substantially difficult to attain a high value of Kappa when the disease is considerably rare.

Therefore, a relatively large observed agreement may still result in a relatively small Kappa coefficient after adjusting for chance agreement which is determined by marginal distributions. Feinstein and Cicchetti (1990) defined the term balanced marginals for situation that the proportions of each category are close to 0.5. They further described two paradoxes due to the presence of unbalanced marginals which could produce either unreasonably high or low Kappa statistics, where "unbalance" means that the marginal frequencies significantly differ.

For instance, Tables 1.6, 1.7, 1.8 and 1.9 present four two-by-two contingency tables that illustrate the impact of unbalanced marginals on κ described by Feinstein and Cicchetti (1990). Table 1.6 presents an example of balanced marginals where the percentages of the row and the column are almost equal (52% versus 48% for columns and 49% versus 51% for rows). Table 1.7, on the other hand, shows unbalanced

marginals where the percentages of the row and column are roughly 85% versus 15%. Both of the above tables have the same high percentage of agreement of 89% but the Kappa statistics for the balanced marginals is 0.78, which is much higher than that for the unbalanced marginals (κ=0.56 for Table1.7). Tables 1.8 and 1.9 also present data with unbalanced marginals. When there is symmetric imbalance (70% versus 30% in both the rows and columns), κ is only 0.05. When there is asymmetric imbalance of the rows (45% versus 55%) and for the columns (65% versus 35%) then κ is 0.22.

Therefore, the value of the Kappa statistic not only heavily depends on the prevalence, but also depends on the level of imbalance and asymmetry of the data.

Table 1.6: An example of data with balanced marginals, κ=0.78

		Observer A		
		No	Yes	Total
Observer B	No	45	7	52
	Yes	4	44	48
	Total	49	51	100

Table 1.7: An example of data with unbalanced marginals, κ=0.56

		Observer A		
		No	Yes	Total
Observer B	No	9	5	14
	Yes	6	80	86
	Total	15	85	100

Table 1.8: An example of data with symmetric unbalance, κ=0.05

		Observer A		
		No	Yes	Total
Observer B	No	50	20	70
	Yes	20	10	30
	Total	70	30	100

Table 1.9: An example of data with asymmetric unbalanced, κ=0.22

		Observer A		
		No	Yes	Total
Observer B	No	35	10	45
	Yes	30	25	55
	Total	65	35	100

1.4 Existing Methods for Continuous Outcomes

Numerous methods have been developed to evaluate rater agreement where continuous outcomes are involved. For instance, Bland and Altman (1999) published an interesting example, which includes systolic blood pressure measurements taken by two experienced raters (denoted as raters J and R) using a sphygmomanometer and by a semi-automatic blood pressure monitor (denoted as rater S). Three replicated readings were made in quick succession, resulting in 9 readings for each subject. The research interest is to assess the agreement among raters J, R and S.

Barnhart et al. (2007b) defined measures of agreement as either scaled or unscaled. Unscaled measures require that in agreement problems we decide what is an "acceptable disagreement" in terms of the variable being measured. Scaled measures compare the interobserver disagreement to some value that can be derived from the data and does not depend on the variable of interest. In this dissertation, I will focus on scaled measures. In this section, unscaled measures such as limits of agreement (LOA) (Bland and Altman 1999, 2007), and scaled measures including the intraclass correlation coefficient (ICC) (Bartko 1966, 1974, Shrout and Fleiss 1979, Eliasziw et al. 1994, Muller and Petra 1994, McGraw and Wong 1996), the concordance correlation coefficient (CCC) (Lin 1989, 1992, 2000a, Lin et al. 2002, 2007, King and Chinchilli 2001a,b, King et al. 2007b, Barnhart et al. 2002, 2005, 2007c), and the coefficient of individual agreement (CIA) (Haber et al. 2007, Haber and Barnhart 2006, Barnhart et al. 2007a, Haber and Barnhart 2008, Pan et al. 2010). All those methods were reviewed carefully by Barnhart et al. (2007b) and the review in this section mainly follows the notation used in Barnhart et al. (2007b).

Except for the coefficient of individual agreement (CIA), all the rest of the methods for continuous measurements are originally designed for data without replication. For example, in the different types of intraclass correlation coefficient (ICCs) presented by McGraw and Wong (1996), and Shrout and Fleiss (1979), no replication is involved.

This is the same for the original concordance correlation coefficient (CCC) introduced by Lin (1989). Later, those methods are extended to accommodate replicated data. Having replicated data is very important when designing the agreement studies. The discussion will be continued when we introduce our new coefficient to measure the agreement of replicated observations.

1.4.1 Limits of Agreement

In medical research, the Bland and Altman plot is very popular due to its simplicity. The associated limits of agreement (LOA) by Bland and Altman (1999, 2007) are widely used for assessing agreement between two observers but can be extended for pairwise comparison of J observers. The original LOA method estimates the difference of single observations between two observers as well as the corresponding $(1-\alpha)100\%$ probability interval (PI) that contains the middle $1-\alpha$ probability of the distribution of difference. The limits of agreement are defined as the estimates for the limits of this PI. The key implicit assumption for the method of estimation is that the difference between the two observers is reasonably stable across the range of measurements. Let $D_i = Y_{i1} - Y_{i2}$ be the difference between the single observations of the two observers. The 95% LOA are $\mu_D \pm 1.96\sigma_D$ where $\mu_D = E(D_i)$ and $\sigma_D^2 = Var(D_i)$, under a normality assumption on D_i. For data without replication, the LOA can be estimated by replacing μ_D with the sample mean $D_\bullet$ and σ_D^2 with the sample variance S_D^2. If the absolute limit is less than an acceptable difference, d_0, then the agreement between the two observers is deemed satisfactory. But as in the nature of unscaled measures of agreement, in practice, how to define the acceptable difference is subjective. Usually, we think the range depends on the clinical justification.

The LOA is often displayed in the popular Bland and Altman plot (average, $(Y_{i1} + Y_{i2})/2$, versus difference, $Y_{i1} - Y_{i2}$) with two horizontal lines of the estimated

LOA: $D_\bullet \pm 1.96 S_D$ and two horizontal lines of the 95% lower bound of the lower limit and 95% upper bound of the upper limit:

$$D_\bullet - 1.96 S_D - 1.96\sqrt{\left(\frac{1}{n} - \frac{1.96^2}{2(n-1)}\right) S_D^2}, \quad D_\bullet + 1.96 S_D + 1.96\sqrt{\left(\frac{1}{n} + \frac{1.96^2}{2(n-1)}\right) S_D^2}.$$

The LOA method has also been extended to data with repeated measures (Bland and Altman 2007). Bland and Altman (2007) include two situations: (1) multiple time-matched observations per individual by two observers where the true value of the subject may or may not change over time; (2) multiple observations per individual (not time-matched) by two observers where the true value of the subject is constant at a prespecified length of time when the two observers take measurements. Furthermore, Bland and Altman (1999, 2007) described a method of moments approach to estimating μ_D and σ_D^2 for data with multiple observations in both situations (Barnhart et al. 2007b) and a variance of σ_ϵ^2. The effect of observer β_j is considered a fixed factor with $\sum_j \beta_j = 0$, and $\sigma_\beta^2 = \sum_j \sum_{j'} (\beta_j - \beta_{j'})^2 / [J(J-1)]$.

1.4.2 Intraclass Correlation Coefficient

For a very long time, the agreement of continuous outcomes has been assessed with intraclass correlation coefficient (ICC). Under the assumptions of different types of ANOVA models, ICC compares the variability of different ratings of the same subject with the total variation across all ratings and all subjects. We consider three versions of ICCs, under three types of ANOVA models. In the first model, we only include the random subject effects; in the second model we add a fixed or random observer effect; and the third model is a full model, where an interaction between observer and subject effects is added.

As in Barnhart et al. (2007b), notations are unified for both cases when observer is treated as an either fixed or random effect. Each observer j ($j = 1, \ldots, J$) is assumed

to have $k\,(k = 1, \ldots, K)$ readings on each subject $i\,(i = 1, \ldots, n)$. $K = 1$ means no replications and $K \geq 2$ indicates the number of replications for each observer (Eliasziw et al. 1994). The estimates for the variance components are derived based on the expected mean sums of squares (MSS) from the specified ANOVA model. The definitions of the three kinds of ICCs and their corresponding estimates are shown below:

- ICC_1 is based on a one-way random effect model without observer effect

$$Y_{ijk} = \mu + \alpha_i + \epsilon_{ijk}$$

with assumptions: $\alpha_i \sim N(0, \sigma_\alpha^2)$; $\epsilon_{ijk} \sim N(0, \sigma_\epsilon^2)$; and ϵ_{ijk} is independent of α_i.

$$ICC_1 = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\epsilon^2}, \quad \widehat{ICC_1} = \frac{MS_\alpha - MS_\epsilon}{MS_\alpha + (JK-1)MS_\epsilon},$$

where MS_α and MS_ϵ are the mean sums of squares from the one-way ANOVA model for between and within subjects, respectively.

- ICC_2 is based on a two-way mixed or random (depending on whether the observers are fixed or random) effect model without the observer-subject interaction:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$$

with assumptions: $\alpha_i \sim N(0, \sigma_\alpha^2)$, $\epsilon_{ijk} \sim N(0, \sigma_\epsilon^2)$, and ϵ_{ijk} is independent of α_i. β_j is treated as either as a fixed or a random effect, depending on whether the observers are fixed or random. If observers are fixed, notation $\sigma_\beta^2 = \sum_{j=1}^{J} \beta_j^2/(J-1)$ is used with constraint of $\sum_{j=1}^{J} \beta_j = 0$. If observers are random, additional assumptions are $\beta_j \sim N(0, \sigma_\beta^2)$ and $\alpha_i, \beta_j, \epsilon_{ijk}$ are mutually

independent.

$$ICC_2 = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\epsilon^2}, \quad \widehat{ICC_2} = \frac{MS_\alpha - MS_\epsilon}{MS_\alpha + (JK-1)MS_\epsilon + J(MS_\beta - MS_\epsilon)/n}.$$

- ICC_3 is based on a two-way mixed or random effect model (depending on whether the observers are fixed or random) with observer-subject interaction.

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

with assumptions: $\alpha_i \sim N(0, \sigma_\alpha^2)$, $\epsilon_{ijk} \sim N(0, \sigma_\epsilon^2)$, and ϵ_{ijk} is independent of α_i. If observers are fixed, notation $\sigma_\beta^2 = \sum_{j=1}^J \beta_j^2/(J-1)$ is used with constraint of $\sum_{j=1}^J \beta_j = 0$ and $\gamma_{ij} \sim N(0, \sigma_\gamma^2)$. If the observers are random, additional assumptions are $\beta_j \sim N(0, \sigma_\beta^2), \gamma_{ij} \sim N(0, \sigma_\gamma^2)$ and $\alpha_i, \beta_j, \gamma_{ij}, \epsilon_{ijk}$ are mutually independent. ICC_3 can only be estimated when $K > 1$.

$$ICC_3(\text{fixed } \beta_j) = \frac{\sigma_\alpha^2 - \sigma_\gamma^2/(J-1)}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2 + \sigma_e^2}, \quad ICC_3(\text{random } \beta_j) = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2 + \sigma_e^2}$$
$$\widehat{ICC_3} = \frac{MS_\alpha - MS_\gamma}{MS_\alpha + J(K-1)MS_\epsilon + (J-1)MS_\gamma + J(MS_\beta - MS_\gamma)/n}.$$

Above are the "agreement" ICCs (McGraw and Wong 1996) and the ICCs from Bartko (1966) and Shrout and Fleiss (1979) for fixed observers do not include σ_β^2 (consistency ICCs).

$$ICC_{3c}(\text{fixed } \beta_j) = \frac{\sigma_\alpha^2 - \sigma_\gamma^2/(J-1)}{\sigma_\alpha^2 + \sigma_\gamma^2 + \sigma_e^2},$$

McGraw and Wong (1996) provided a very detailed summary for the inference about ICCs in terms of consistency and agreement when we don't have replications. The assumptions used to define ICC are the main disadvantages to using ICCs to as-

sess agreement. Of note is the fact that all ICCs are increasing functions of between-subject variability (represented here by σ_α^2). Thus, it would attain a high value for a population with substantial heterogeneity. Vangeneugden et al. (2004), Vangeneugden et al. (2005) and Molenberghs et al. (2007) argued that the ICC should be treated as a reliability measure that assesses the degree of differentiation of subjects from a population, rather than agreement. When we don't have replicated data, in ICC_3, $MS_\epsilon = 0$ and MS_γ is reduced to the same value of MS_ϵ in ICC_2. Therefore, it does not make sense to estimate ICC_3 if the data do not have replication.

The ICCs presented here may be used for data with repeated measures where k denotes the time of the measurement. However, these ICCs may not be very useful unless one modifies the assumptions on ϵ_{ijk} in order to take into account the time structure. A linear mixed-model approach to estimate reliability for repeated measures has been proposed by Vangeneugden et al. (2004) and Molenberghs et al. (2007). The ICC has also been extended for repeated measurements with multivariate observer (Konishi et al. 1991). Chen and Barnhart (2008) compute the expected value of the ICC estimator under a very general model to get a sense of the population parameter that the ICC estimator provides.

1.4.3 Concordance Correlation Coefficient

The concordance correlation coefficient (CCC) is very popular for assessing agreement of continuous outcomes. It was first published by Lin (1989) for the simpliest case where there are two raters and each make one reading per subject. Lin's CCC is shown as follows: assume that the observations are from a bivariate distribution with mean vector (μ_1, μ_2) and variance-covariance matrix $\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$, Lin's CCC

between two observers Y_1 and Y_2 is proposed as

$$\begin{aligned} \text{CCC}_{\text{Lin}} &= 1 - \frac{E(Y_2 - Y_1)^2}{E[(Y_2 - Y_1)^2 | \rho = 0]} \\ &= \frac{2\rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2} \end{aligned}$$

where ρ is the Pearson correlation coefficient between two observers.

Later CCC was extended to J observers for data without replications (Lin 1989, King and Chinchilli 2001a, Lin et al. 2002, Barnhart et al. 2002). Barnhart et al. (2005), and Lin et al. (2007) extended CCC for data with replications where none of the observers is treated as reference. Barnhart et al. (2007b) extended the CCC to the situation where one of the multiple observers is treated as reference. Other extensions include CCC for repeated measures for two or more observers (King et al. 2007b, Quiroz 2005) and for multivariate observers (Jason and Olsson 2001, 2004). Recently, Guo (2004), Guo and Manatunga (2009) introduced CCC to survival outcomes.

Since Barnhart et al. (2005) introduced intra, inter and total CCC with the replicated data, we will introduce those three versions without reference. Note that total CCC is the original CCC when there is no replication. Let Y_{ijk} be the kth replicated measurements for the ith subject by the jth method and write $Y_{ijk} = \mu_{ij} + \epsilon_{ijk}$ with the similar assumptions as shown in defining ICCs.

The intra-CCC for observer j is

$$\begin{aligned} CCC_{j,intra} &= \rho_j^I = 1 - \frac{\sum_{k=1}^{K-1}\sum_{k'=k+1}^{J} E(Y_{ijk} - Y_{ijk'})^2}{\sum_{k=1}^{K-1}\sum_{k'=k+1}^{J} E_I(Y_{ijk} - Y_{ijk'})^2} \\ &= ICC_{1j} = \frac{\sigma_{Bj}^2}{\sigma_{Bj}^2 + \sigma_{Wj}^2}, \end{aligned}$$

$$
\begin{aligned}
CCC_{inter} = \rho_c(\mu) &= 1 - \frac{\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J} E(\mu_{ij} - \mu_{ij'})^2}{\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J} E_I(\mu_{ij} - \mu_{ij'})^2} \\
&= \frac{2\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J}\sigma_{Bj}\sigma_{Bj'}\rho_{\mu jj'}}{\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J}[2\sigma_{Bj}\sigma_{Bj'} + (\mu_j - \mu_{j'})^2 + (\sigma_{Bj} - \sigma_{Bj'})^2]}
\end{aligned}
$$

$$
\begin{aligned}
CCC_{total} = \rho_c &= 1 - \frac{\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J} E(Y_{ijk} - Y_{ij'k'})^2}{\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J} E_I(Y_{ijk} - Y_{ij'k'})^2} \\
&= \frac{2\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J}\sigma_j\sigma_{j'}\rho_{jj'}}{(J-1)\sum_{j=1}^{J}\sigma_j^2 + \sum_{j=1}^{J-1}\sum_{j'=j+1}^{J}(\mu_j - \mu_{j'})^2} \\
&= \frac{2\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J}\sigma_{Bj}\sigma_{Bj'}\rho_{\mu jj'}}{\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J}[2\sigma_{Bj}\sigma_{Bj'} + (\mu_j - \mu_{j'})^2 + (\sigma_{Bj} - \sigma_{Bj'})^2 + \sigma_{Wj}^2 + \sigma_{Wj'}^2]},
\end{aligned}
$$

where $\sigma_{Bj'}^2$ denotes the between subject variance and $\sigma_{Wj'}^2$ denotes the within subject variance. But the CCC is known to depend on between-subject variability (Atkinson and Nevill 1997) that may result from that fact that it is scaled relative to the maximum disagreement defined as the expected squared difference under independence. Note that $CCC = ICC_2$ under two-way model without interaction (without replication) (Carrasco and Jover 2003)and $CCC_{total} = ICC3$ under two-way model with interactions for replicated data (Barnhart et al. 2005, Song 2003).

1.4.4 Coefficient of Individual Agreement

Barnhart and colleagues(Haber and Barnhart 2008, Barnhart et al. 2007a, Haber et al. 2007, Wiener 2009, Gao 2010, Pan et al. 2010) began looking for a scaled agreement index, the coefficient of individual agreement (CIA), which is scaled relative to an acceptable disagreement, with the goal of establishing interchangeability of observers. An acceptable disagreement requires that the differences between measurements of

different observers are similar to the differences between replicated measurements of the same observer. The concept of individual agreement is derived from the idea of individual bioequivalence in bioequivalence studies (Anderson and Hauck 1990, Schall and Luus 1993). Similar agreement indices have been proposed by Haber et al. (2005) and Shao and Zhong (2004). The CIAs compare differences between measurements from different observers to the differences of replicated measurements of the same observer. Therefore, they require replications which allow us to estimate the within-observer variability. The numbers of replications can be different across subjects and observers. Before considering observers for comparison, one must assume that the replication errors of the observers are acceptable. Replicated measurements by observers are required for the CIAs for the purpose of estimation and inference.

Barnhart et al. (2007a) defined the CIAs for cases of no reference observer (ψ^N) and for the case where the Jth observer is a reference (ψ^R), respectively, as follows:

$$\begin{aligned}
\psi^N &= \frac{\sum_{j=1}^{J} E(Y_{ijk} - Y_{ijk'})^2/2}{\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J} E[(Y_{ijk} - Y_{ij'k'})^2]/(J-1)} \quad (\text{where } k \neq k') \\
&= \frac{\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J}(\sigma_{Wj}^2 + \sigma_{Wj'}^2)}{\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J}[2(1-\rho_{\mu jj'})\sigma_{Bj}\sigma_{Bj'} + (\mu_j - \mu_{j'})^2 + (\sigma_{Bj} - \sigma_{Bj'})^2 + \sigma_{Wj}^2 + \sigma_{Wj'}^2]}, \\
\psi^R &= \frac{E(Y_{iJk} - Y_{iJk'})^2/2}{\sum_{j=1}^{J-1} E[(Y_{ijk} - Y_{iJk'})^2]/(J-1)} \quad (\text{where } k \neq k') \\
&= \frac{\sigma_{WJ}^2}{\sum_{j=1}^{J-1}[2(1-\rho_{\mu jJ}\sigma_{Bj}\sigma_{BJ} + (\mu_j - \mu_J)^2 + (\sigma_{Bj} - \sigma_{BJ})^2 + \sigma_{Wj}^2 + \sigma_{WJ}^2]},
\end{aligned}$$

where as usual, $\sigma_{Bj'}^2$ denotes the between subject variance and $\sigma_{Wj'}^2$ denotes the within subject variance.

In order to estimate the CIAs, one has to use data with replicated measurements. The numbers of replications for observer X and Y, which do not have to be equal, are denoted by K and L, respectively.

A more general definition of the CIAs uses the concept of a disagreement function $G(X, Y)$. A disagreement function $G(X, Y)$ must satisfy (a) $G(X, Y) \geq 0$ and (b)

$G(X, Y)$ increases as the disagreement between X and Y. The disagreement function can be defined and estimated for each subject. Let $\hat{G}_i(X, Y)$, $\hat{G}_i(X, X')$ and $\hat{G}_i(Y, Y')$ be the estimated values of the disagreement function for subject i.

$$\hat{G}_i(X, Y) = \frac{1}{KL} \sum_{k=1}^{K} \sum_{l=1}^{L} G(X_{ik}, Y_{il})$$

$$\hat{G}_i(X, X') = \frac{2}{K(K-1)} \sum_{k=1}^{K-1} \sum_{k'=k+1}^{K} G(X_{ik}, X_{ik'})$$

$$\hat{G}_i(Y, Y') = \frac{2}{L(L-1)} \sum_{l=1}^{L-1} \sum_{l'=l+1}^{L} G(Y_{il}, Y_{il'})$$

Therefore $\hat{G}_i(X, Y)$ is the averaged disagreement between observations between X and Y for subject i; $\hat{G}_i(X, X')$ is the averaged disagreement between replications by X for subject i and similarly $\hat{G}_i(Y, Y')$ is the averaged disagreement between replications by Y for subject i. Let $\bar{\hat{G}}(X, Y)$, $\bar{\hat{G}}(X, X')$, $\bar{\hat{G}}(Y, Y')$ be the sample means of the $\hat{G}_i$'s. Then ψ_G^N and ψ_G^R are estimated as follows:

$$\psi_G^N = \frac{[\bar{\hat{G}}(X, X') + \bar{\hat{G}}(Y, Y')]/2}{\bar{\hat{G}}(X, Y)}$$

$$\psi_G^R = \frac{\bar{\hat{G}}(X, X')}{\bar{\hat{G}}(X, Y)}$$

The CIAs can be extended to multiple raters. Barnhart et al. (2007a) proposed a non-parametric approach for estimating the standard errors of the CIAs when $MSD(X, Y) = E(X - Y)^2$ is used as the disagreement function. Haber et al. (2010) generalized CIAs to data with matched repeated measurements.

1.5 Existing Methods for Repeated Binary Outcomes

In previous sections, agreement measurements of both continuous and categorical cross-sectional data. Nowadays, we more often have repeated measurements. For replicated measurements there is no change in the true values, while for repeated measurements, true values may change according to different conditions. Here we review some existing methods to assess agreement for repeated binary outcomes.

1.5.1 Logistic Regression Modeling Agreement Proportion

Coughlin et al. (1992) introduced logistic modeling of inter-observer agreement. The dependent variable is defined to be 1 if the two raters agree and 0 otherwise. Covariates may be included in the regression equation in order to obtain adjusted or subgroup-specific estimates of percent agreement. Besides the logistic regression with covariates, repeated measurements on the same subject can be analyzed by generalized estimating equations (GEE). Model-based percent agreement is estimated.

If each of N subjects are assigned independently by two raters to one of the categories, then the cell frequencies (n_{ii}) along the main diagonal of the two-way contingency table represent the agreement between the raters. The crude agreement is estimated as follows:

$$p = \frac{1}{N}\sum_{i=1}^{I} n_{ii}.$$

By applying the logistic regression, the proportion agreement for particular subgroups can be estimated. Suppose K explanatory variables $(X_1, X_2, ..., X_k)$, the model based agreement is:

$$E(p|X_1, X_2, ..., X_k) = \frac{1}{1 + exp([-(\alpha + \sum_{k=1}^{K} \beta_k X_k)])}.$$

The variance of the logit of the proportion agreement can be estimated using the sandwich estimation.

1.5.2 Extended Kappa Coefficient

Klar et al. (2000) proposed an approach for identifying covariates that are predictive of agreement is to consider regression models for kappa. The authors considered models for kappa that allow this chance-corrected measure of agreement to depend on covariates. Klar et al. (2000) model kappa directly as a function of covariates, and estimated the regression parameters using a generalized estimating equations (GEE) approach. However, a drawback of their method is that it cannot be implemented using existing statistical software.

Lipsitz et al. (2003) used two stage logistic regression to take chance agreement into consideration. Let Y_{ir} denote the measurement of ith subject and rth rater. Redefine $Y_i = Y_{i1}Y_{i2} + (1 - Y_{i1})(1 - Y_{i2})$. The following two steps can be considered:
(1) Use the standard logistic regression of Y_{ir} on (x_i, x_{ir}) , r=1, 2, to obtain $\hat{p}_{ir}$.
(2) Form the estimated offset, $\hat{\eta}_i = logit[\hat{p}_{i1}\hat{p}_{i2} + (1 - \hat{p}_{i1})(1 - \hat{p}_{i2})]$. Finally the model $logit(p_i) = \eta_i + x_{i1}^{'}\beta_1 + x_{i2}^{'}\beta_2 + x_i^{'}\beta_3$. The model based kappa statistic is shown as follows:

$$\hat{k}_i = \frac{\hat{p}_i - e^{\hat{\eta}}/[1 + e^{\hat{\eta}}]}{1 - e^{\hat{\eta}}/[1 + e^{\hat{\eta}}]} = \frac{\hat{p}_i - [\hat{p}_{i1}\hat{p}_{i2} + (1 - \hat{p}_{i1})(1 - \hat{p}_{i2})]}{1 - [\hat{p}_{i1}\hat{p}_{i2} + (1 - \hat{p}_{i1})(1 - \hat{p}_{i2})]}$$

A bootstrap method is applied to estimate the standard error (SE).

Ma et al. (2008) introduced a novel approach based on a new class of kappa estimates to tackle the complexities involved in addressing missing data and other related issues arising from a general multi-rater and longitudinal data setting. In classic approach, kappa is estimated by substituting the sample proportions in place of the respective parameters. The asymptotic distribution of the kappa estimate is derived by first considering the joint asymptotic distribution of $\hat{\pi}(g_1, g_2)$, then followed

by an application of the Delta method. The limitations are as follows: first of all, it involves estimating a large number of parameters. Second, it is quite complex to extend this classic approach to address missing data. An alternative approach based on the theory of U-statistics to address these limitations has been proposed (Ma et al. (2008)).

1.5.3 Other Agreement Measurement for Repeated Binary Measurements

Nelson and Edwards (2007) introduced a generalized linear mixed model for agreement with covariates and probit link function has been used.

$$\phi^{-1}(p_{ij}) = \eta + \beta x_{ij} + d_i\mu_i + d_j v_j,$$

where x_{ij} is the vector of covariates associated with the ith item classified by the jth rater. The vectors d_i and d_j represent the design vectors of the random effects for the ith item and jth rater respectively, and the vectors and contain the associated random effects for the items and raters. Under the probit model, the authors proposed an alternative measurement which can be compared to kappa statistic: $\rho = \frac{\sigma_u^2}{\sigma_v^2+\sigma_u^2+\sigma_w^2}$ where σ_u^2=1 is the variance of the probit function. Furthermore, authors present the model-based estimation of kappa statistics and its variance. The performance of the model-based estimation of kappa with covariates was examined through the simulation studies. Computationally, the datasets were individually fitted using a Monte-Carlo expectation-maximization algorithm (MCEM) developed by McCulloch (1997) to obtain almost-exact maximum likelihood estimates of the parameters in the generalized linear mixed model. And the entire algorithm was implemented using a program written in C programming language.

Furthermore, the intraclass correlation coefficient (ICC) can also be applied to

binary data (Ridout et al. (1999)) and can be extended to repeated binary data. King et al. (2007b) proposed a class of repeated measures concordance correlation coefficient (CIA) including both continuous and categorical measurements.

Last but not the least, Gao (2010) applied three logistic models to estimate coefficient of individual agreement (CIA) for binary data. Here, the same three logistic models will be applied to estimate coefficient of individual equivalence (CIEA) in Chapter 4.

1.5.4 Agreement Measurement for Repeated Continuous Measurements

King et al. (2007a) proposed a repeated measures concordance correlation coefficient between two raters or two methods of measuring a response in the presence of repeated continuous measurements. Carrasco et al. (2009) furthermore introduced the concordance correlation coefficient for repeated measures estimated by variance components. Haber et al. (2010) extended the coefficient of individual agreement using a variance components approach. The variance components idea will be applied to our estimation of coefficient of individual equivalence when we have replicated and repeated binary measurements.

1.6 Bayesian Approaches for Evaluating Agreement

In medicine and biology and even the social science, Bayesian methods are being used frequently. For example, Bayesian sequential stopping rules are implemented in some designs of clinical trials. As we know, Bayes' theorem allows to incorporate previous information with the current data. This is considered attractive for many areas in medicine since Bayesian methods are based on prior information which is usually available in the related previous studies. Gelman et al. (2004) shared that the primary motivation for believing Bayesian thinking important is that it facilitates a common-sense interpretation of statistical conclusions. They use interval estimation as an example. The Bayesian (probability) interval for an unknown quantity of interest can be directly regarded as having a high probability of containing the unknown quantity, in contrast to a frequentist (confidence) interval, which may strictly be interpreted only in relation to a sequence of similar inferences that might be made in repeated practice (Gelman et al. 2004).

The application of a Bayesian approach to observer agreement studies is fairly recent. Broemeling (2009) well summarized the existing Bayesian methods for measures of agreement on both discrete and continuous outcomes. In this section, the basic foundation of Bayesian methods is reviewed as well as the existing application to Kappa, limits of agreement (LOA) and intraclass correlation coefficient (ICC).

All Bayesian approaches are based on Bayes' theorem, which provides the posterior distribution as the basis for statistical inference, which includes estimation of parameters and tests of hypothesis. Suppose X is a continuous observable random vector and $\theta \in \Omega \subset R^m$ is an unknown parameter vector, and we define $f(x|\theta)$ as the conditional density of X given θ. The the conditional density of θ given $X = x$ is

$$p(\theta|x) = cf(x|\theta)p(\theta), \theta \in \Omega, x \in R^m.$$

In the above equation, $c > 0$ is the normalizing constant, which is chosen so that the integral of $f(x|\theta)p(\theta)$ with respect to θ is unity. If X is discrete, $f(x|\theta)$ is the probability mass function of X. The density $p(\theta)$ is the prior density of θ and represents the knowledge one possesses about the parameter before one observes the data X. Here θ is considered a random variable so that Bayes theorem transforms one's prior knowledge of θ, which is the prior density, to the posterior density. Such transformation combines the prior information about θ with the sample information represented by the likelihood $f(x|\theta)$.

Using the Bayesian approach, the posterior distribution and analysis of some measures of agreement, including Kappa, limits of agreement and intraclass correlation coefficient have been presented recently. Basu et al. (2000), one of the few Bayesian investigations of Kappa, determined the posterior density of Kappa by MCMC techniques with WinBUGS and developed a test for the homogeneity of Kappa across several different experiments. A very simple comparison between conventional and Bayesian methods to estimate Kappa is given by Broemeling (2009). One study examines the agreement between two psychiatrists who are assigning degrees of depression to 129 patients. Table 1.10 contains the summary of measures between two observers, where 1="not depressed", 2="mildly depressed", and 3="clinically depressed". We notice this set up is similar to Table 1.3.

Table 1.10: Example: Agreement for Depression

		Psychiatrist 1			
		1	2	3	Total
Psychiatrist 2	1	11	2	19	32
	2	1	3	3	7
	3	0	8	82	90
	Total	12	13	104	129

By conventional methods, Kappa=0.375 with 0.079 as its standard error. If one adopts a Bayesian approach with a uniform prior density for the θ_{ij}, where

$i = j = 1, 2, 3$, then the parameters have a Dirichlet (12,3,20,2,4,4,1,9,83) posterior distribution and the posterior mean and standard error of Kappa are 0.358 and 0.072, respectively. The raw agreement by itself indicates a strong association between psychiatrists, however the chance agreement if fairly strong also, which reduces the overall agreement to only fair agreement. In this study, the Bayesian and conventional analysis of Kappa agree quite well. Similarly, Bayesian methods can be extended to Bland and Altman's limits of agreement (LOA) (Broemeling 2009).

The application of Bayesian approaches to the research of intraclass correlation coefficient (ICC) started around 2000. Turner et al. (2001) introduced Bayesian hierarchical modeling for the analysis of randomized trials with binary outcome data. This approach provides us with a credible interval for the ICC for binary outcome data. Several approaches to constructing informative priors from empirical ICC values are described in their paper. The authors pointed out that a Bayesian approach allows us to assume distributions other than normality for the random effects used to model the clustering and this enables us to gain insight into the robustness of our parameter estimates to the classical normality assumption. Those authors extended their research in 2006 and they described Bayesian Markov chain Monte Carlo (MCMC) modeling approaches to interval estimation of the ICC, which offered greater flexibility than existing approaches (Turner et al. 2006). It is known that the ANOVA estimator underestimated the true ICC when the diagnostic test was imperfect. Branscum et al. (2005) examined the effect of substituting diagnostic test outcomes for true infection status on an ANOVA estimator of ICC through Monte Carlo simulation. They also proposed a Bayesian model for estimating the ICC that incorporated imperfect sensitivity and specificity. A similar Bayesian method has been extended to estimate ICC where ordinal scales measurements are involved (Gajewski et al. 2007). Ahmed and Shoukri (2010) extended ICC to correlated binary outcomes using Bayesian methods.

1.7 Discussion

As discussed in the previous sections, there are many different measures of agreement in terms of qualitative and quantitative outcomes. However, the ICC and the CCC are originally defined for quantitative data and these coefficients have been shown to be equivalent to the weighted kappa for categorical data (Fleiss and Cohen 1973, Lin et al. 2002). In addition, Shoukri (2004) and King and Chinchilli (2001a) also defined ICC and CCC for qualitative data. Those are also carefully reviewed by Gao (2010).

While both ICC and CCC are popularly used for continuous outcomes with or without replications, there are literatures comparing the similarity and the difference between them. In some special cases, for instance, when there are no replications and the observer is treated as fixed effect in ICC, Carrasco and Jover (2003) showed the equality of ICC_2 and the total CCC without the ANOVA model assumptions. Also, when there are replications, the total CCC is equivalent to ICC_3 (Barnhart et al. 2005, Song 2003). The differences between ICCs and CCCs were summarized in Barnhart et al. (2007b) as: (1) the ICC has been proposed for fixed and random observers, while the CCC usually treats the observers fixed; (2) the ICC requires ANOVA model assumptions, while the CCC does not. Chen and Barnhart (2008) compared ICC and CCC when there are data without and with replications. They computed the expected value of ICC estimator under a very general model to get a sense of the population parameter of the ICC and then compared the expected value to CCC, which is defined without ANOVA assumptions. They reported their results for data without replication and with replication for three types of ICCs as defined in section 1.4.2, and recommended to use ICC_3 as its estimate is similar to the estimate of CCC regardless whether the ANOVA assumptions are met or not.

It's also of interest to compare CCC and CIA when we want to compare agreement with continuous measurements. Barnhart et al. (2007b) did a brief review. The CCC and the CIA were compared for quantitative data with replications when none

of the observers is considered as the reference (Barnhart et al. 2007c). When two observers are involved and the between-subject and within-subject variabilities across two observers are assumed equal, i.e. $\sigma^2_{B_1} = \sigma^2_{B_2} = \sigma^2_B$; and $\sigma^2_{W_1} = \sigma^2_{W_2} = \sigma^2_W$. Then the total CCC and the CIA are restated as

$$\rho_c = \frac{2\sigma^2_B \rho_{\mu_{12}}}{(\mu_1 - \mu_2)^2 + 2(\sigma^2_B + \sigma^2_W)}$$

and

$$\psi^N = \frac{2\sigma^2_W}{(\mu_1 - \mu_2)^2 + 2(1 - \rho_{\mu_{12}})\sigma^2_B + 2\sigma^2_W}$$

Therefore, both coefficients increase as the correlation ($\rho_{\mu_{12}}$) increases and decrease as the difference of the means ($\mu_1 - \mu_2$) increases. Due to the difference in the numerators (σ^2_B for ρ_c and σ^2_W for ψ^N), the CCC increases when the between-subject variability (σ^2_B) increases and the within-subject (σ^2_W) variability decreases. However, the CIA, on the other hand, increases when the within-subjects variability increases and when the between-subjects variability decreases. Furthermore, Barnhart et al. (2007c) found that the CCC is more dependent on the relative magnitude of the between- and within-subject variabilities, σ^2_B/σ^2_W, than the CIA.

Finally, since Bayesian approaches are pretty new to the area of observer agreement, the literature is fairly limited. Further discussion including the possible Bayesian modeling of the coefficient of individual equivalence (CIE) which is introduced since next chapter, is coming in the later Chapter.

Here is the brief outline of my dissertation. In Chapter 2, the basic idea of the coefficient individual equivalence (CIE) as well as its nonparametric estimation, are introduced. In Chaper 3, the application of this new coefficients to replicated binary measurements is discussed. In Chapter 4, we extend this approach to replicated quantitative data. Comparisons between the new coefficient and the existing Kappa, CIA and CCC follow in Chapter 5. Models for replicated and repeated binary data

are investigated in Chapter 6. This dissertation is summarized in Chapter 7, where we also present ideas for future work.

Chapter 2

Introduction of the Coefficient of Individual Equivalence

2.1 Motivation for introducing the Coefficient of Individual Equivalence (CIE)

The number of coefficients proposed to assess agreement is large, but the number used in practise is small. For binary, nominal and ordinal measurements, Cohen's kappa (Cohen 1960) and weighted kappa (Cohen 1968) are the most commonly used method. However, kappa has been criticized since its value heavily depends on the prevalence (Kraemer 1979, Thompson and Walter 1988) and is also affected by the data structure, i.e., imbalance or asymmetry (Feinstein and Cicchetti 1990). Among the scaled agreement indices for quantitative data, the intraclass correlation coefficient (ICC) and the concordance correlation coefficient (CCC) are the most popular. Under certain conditions, the CCC is equivalent to one version of the ICC. Specifically, if the ANOVA model assumptions are satisfied, the CCC reduces to the agreement ICC defined by this ANOVA model (Barnhart et al. 2007b, McGraw and Wong 1996, Barnhart et al. 2002). The CCC is based on comparing the mean squared deviation

(MSD) to its value under independence. However, independence and disagreement are two different concepts (Haber and Barnhart 2006). Furthermore, the CCC depends on the between-subject variability. Atkinson and Nevill (1997) pointed out that an increase in the between-subject variability results in a larger value of CCC even if the individual differences between measurements by the two methods remain the same. Barnhart et al. (2007c) also showed that the CCC depends on the between-subject variability due to the fact that it is scaled relative to the maximum disagreement defined as the expected MSD under independence.

The introduction of coefficients of individual agreement (CIAs), which are scaled relative to an acceptable disagreement, was motivated by the desire to establish interchangeability of observers. However, these coefficients are not suitable in case one or both observer have an unacceptably large within-observer disagreement (repeatability). Furthermore, ψ_N is not estimable when we have only one observation on either observer and ψ_R is not defined when there is no replicated reading for the reference observer, for example when the reference observer is a perfect gold standard.

In this dissertation we propose a new criterion, coefficient of individual equivalence (CIE), for "reasonable" or "acceptable" agreement with replicated data. The importance of having replicated data has been presented before by Bland and Altman (1999) when they introduced limits of agreement (LOA). If we have only one measurement using each method on each subject we cannot tell which method is more repeatable (precise). Lack of repeatability can interfere with the comparison of two methods because if one method has poor repeatability, in the sense that there is considerable variation in repeated measurements on the same subject, the agreement between the two methods is bound to be poor. Even if the measurements by the two methods agreed very closely on average, poor repeatability of one method would lead to poor agreement between the methods for individuals. When the old method has poor repeatability even a new method which was perfect would not agree with it.

Lack of agreement in unreplicated studies may suggest that the new method cannot be used, but it might be caused by poor repeatability of the standard method. If both methods have poor repeatability, then poor agreement is highly likely. Bland and Altman (1999) recommended the simultaneous estimation of repeatability and agreement by collecting replicated data. Further justifications for the use of CIE can be found in Section 2.5.

2.2 Overview and Definition of the Coefficient of Individual Equivalence

For simplicity, suppose that there are only two observers, and let the readings of the two observers be represented by the random variables X and Y. Further, let $f_X(u)$ and $f_Y(u)$, denote the respective probability mass functions (for categorical observations) or the probability density functions (for quantitative observations). Hawkins (2002) called two observers *equivalent* if for each study subject, the conditional distribution of X and Y, given each subject, are identical: i.e., $f_X(u|i) = f_Y(u|i)$, for every u and every $i = 1, \ldots N$. In our opinion, if observers are equivalent then we can argue that their agreement is at least 'acceptable' because from a statistical point of view, *it does not matter whether the next observation on a given subject will be made by observer X or by observer Y.* In other words, the two observers can be used interchangeably, or one observer can be replaced by the other at any time during the study.

We assume that the magnitude of disagreement between two readings, x and y, on the same subject is quantified via a disagreement function $G(x, y)$. A disagreement function $G(X, Y)$ must satisfy (a) $G(X, Y) \geq 0$ and (b) $G(X, Y)$ increases as the disagreement between X and Y, according to a specific criterion, increases. The most commonly used disagreement function is the mean squared deviation (MSD): $G(X, Y) = E(X - Y)^2$. A more meaningful disagreement function is $G(X, Y) =$

$E|X-Y|$, the mean absolute difference, which can be expressed in the same units as the individual observations. Other possible choices are the mean relative difference: $MRD = E[|X-Y|/X]$ or the mean of the Winsorized squared distance:

$$d(x-y) = \begin{cases} (x-y)^2 & \text{when } |x-y| \leq a; \\ a^2 & \text{when } |x-y| > a. \end{cases}$$

for a pre-selected positive constant a (King and Chinchilli 2001a). The last disagreement function is more robust to the effects of outliers.

The objective of this dissertation is to introduce a method to determine whether the disagreement between observers is in line with or is substantially larger than "acceptable" disagreement, which we define as the disagreement that can be expected if the observers are equivalent. Therefore we define a new *coefficient of individual equivalence* based on comparing the observed value of the disagreement function to its expected value under the hypothesis of equivalence. This expected disagreement under equivalence can be viewed as *disagreement by chance*, i.e., the magnitude of the disagreement function that would be observed if, in fact, all the readings on the same subject were made by the same observer and later someone randomly assigned the letter X to some of the observations and the letter Y to the remaining observations on this subject. This concept of disagreement by chance is very different from the concept of agreement by chance, which is often used to obtain scaled agreement measures (e.g., kappa or the CCC). Agreement by chance is the expected agreement under independence of the observers.

The coefficient of individual equivalence (CIE) for subject i is then defined as $\eta_i = G_i^E/G_i(X,Y)$, where G_i^E denotes the disagreement under equivalence in the individual level and $G_i(X,Y)$ is the observed disagreement for subject i. The overall CIE is obtained as the ratio of the means (over i) of the numerators and denominators

in the subject-specific coefficients:

$$\eta = CIE = \frac{E_i(G_i(X,Y) \text{ under equivalence})}{E_i(G_i(X,Y))} = \frac{G^E}{G(X,Y)}.$$

In order to assess the magnitude of departure from equivalence, replicated readings of at least one observer on each subject should be available. Suppose that for a given subject there are K replicated readings by observer X and L replicated readings by observer Y ($K, L \geq 1$ and $K + L \geq 3$). Under the hypothesis of equivalence, all the C_K^{K+L} allocations of the $K+L$ readings, so that K readings are assigned to observer X and the remaining L readings are assigned to observer Y, are equally likely. Therefore, the expected disagreement under equivalence will be estimated as the mean of the values of the disagreement function over all the possible C_K^{K+L} allocations. Values close to 1 indicate that the observed disagreement is similar to the disagreement that can be expected if the two observers are "equivalent" in the sense that for each subject, the (conditional) distributions of the values they would report are identical. In other words, under equivalence one would not expect any systematic difference if observer Y would replace X for some or all the subject. Small values of η indicate that there is some systematic difference between the two observers.

We believe that it makes much more sense to compare the observed disagreement to its expected value under good (or acceptable) agreement because we want the coefficient to quantify the deviation (if any) from good agreement. We also believe that equivalence of observers is an appropriate model for good agreement because under this model, the distribution of an observation made by observer X is the same as that of an observation made by observer Y on the same subject. With CIA, the acceptable inter-observer disagreement is the average intra-observer disagreement or the intra-observer disagreement of a reference. With CIE it is the expected disagreement by chance, i.e. under the hypothesis of equivalence.

2.3 An Alternative Expression for the CIE

As described before, the CIE compares the observed disagreement to the expected disagreement under individual equivalence, G^E i.e., to disagreement by chance. We now define an explicit expression for G^E, which can be used to estimate the CIE and to explore its properties. If we denote the $K + L$ measurements on a given subject as $Z_1, Z_2, Z_3, \ldots, Z_{K+L}$, then the value of G^E for the subject is the mean of all C_2^{K+L} terms $G(Z_h, Z_{h'})$, $1 \leq h < h' \leq K + L$. We define G(X,X') as the disagreement function between two replicated measurements made by observer X, $G(Y, Y')$ is analogously defined for observer Y, and $G(X, Y)$ as the disagreement between X and Y. Then among all the pairs $(Z_h, Z_{h'})$, there are C_2^K pairs (X, X'), C_2^L pairs (Y, Y') and $K \cdot L$ pairs (X, Y). Hence

$$\begin{aligned} G^E &= \frac{\sum_{h<h'} G(Z_h, Z_{h'})}{C_2^{K+L}} \\ &= \frac{C_2^K G(X, X') + C_2^L G(Y, Y') + K \cdot L \cdot G(X, Y)}{C_2^{K+L}}. \end{aligned} \tag{2.1}$$

Thus CIE can be written as:

$$\begin{aligned} \eta = CIE &= \frac{C_2^K G(X, X') + C_2^L G(Y, Y') + K \cdot L \cdot G(X, Y)}{C_2^{K+L} G(X, Y)} \\ &= \frac{C_2^K G(X, X') + C_2^L G(Y, Y')}{C_2^{K+L} G(X, Y)} + \frac{K \cdot L}{C_2^{K+L}} \end{aligned} \tag{2.2}$$

The CIE is defined and estimable if $K \geq 1$, $L \geq 1$ and $K + L \geq 3$.

2.4 Nonparametric Estimation of the CIE

In order to estimate the CIE we need replicated observations of at least one of the observers on each subject. Suppose that on subject i, there are K and L replicated readings by X and Y, respectively, where K, $L \geq 1$ and $K + L \geq 3$. Denote the ob-

served values for subject i by $\mathbf{X}_i = (X_{i1}, \ldots, X_{iK})$, $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iL})$. Then the estimated disagreement for this subject is $\hat{G}_i(X, Y) = G(X_i, Y_i) = mean_{k,l}[G(X_{ik}, Y_{il})]$ (the mean over all $K \cdot L$ pairs of an observation from X and an observation from Y). The average observed disagreement is $\hat{G}(X, Y) = mean_i[\hat{G}_i(X, Y)]$. The estimated disagreement between two observations of observer X on subject i is $\hat{G}_i(X, X') = mean_{k<k'}[G(X_{ik}, X_{ik'}]$ and the overall disagreement is $\hat{G}(X, X') = mean_i[\hat{G}_i(X, X')]$. $\hat{G}(Y, Y')$ is defined in a similar way. Turning now to the estimation of the numerator of η we again begin with a single subject i. In order to estimate the expected disagreement for this subject under equivalence, consider all the C_K^{K+L} possible assignments of K $X's$ and L $Y's$ to the $K + L$ observations made on this subject. Under equivalence, all these assignments are equally likely, and hence the expected value of the disagreement function for this subject is the mean of $G_i(X, Y)$ over all C_K^{K+L} assignments. Thus, we estimate the expected disagreement for subject i under equivalence as $\hat{G}_i^{\,E} = mean_A[G(X_i^A, Y_i^A)]$, where A is one of the C_K^{K+L} assignments, and (X_i^A, Y_i^A) is the array of K $X's$ and L $Y's$ corresponding to this assignment. Then the estimate of G^E is $\hat{G}^E = mean_i(\hat{G}_i^{\,E})$ and the estimate of the CIE is

$$\widehat{CIE} = \hat{\eta} = \frac{\hat{G}^E}{\hat{G}(X, Y)}.$$

Alternatively, we plug these estimates into (6.5) to obtain

$$\hat{\eta} = \widehat{CIE} \;=\; \frac{C_2^K \hat{G}(X, X') + C_2^L \hat{G}(Y, Y')}{C_2^{K+L} \hat{G}(X, Y)} + \frac{K \cdot L}{C_2^{K+L}} \tag{2.3}$$

Note that the definition and estimation of CIE can be easily extended when K and L vary across subjects.

2.5 Discussion

In general, there are two types of coefficients (indices) used to assess agreement between two observers or measurement methods. *Unscaled* coefficients are usually based on the distribution of the differences ($D = X - Y$) of the readings of both observers on the same subject. For example, the coverage probability (CP, Lin et al. (2002)) is defined as $CP(d) = Pr(|D| \leq d)$, i.e., the probability that the absolute difference does not exceed a pre-determined value d. Hence, a value of d that represents "acceptable" or "good" agreement has to be determined in advance, and the selection of the appropriate value is usually subjective. All unscaled coefficients share a similar limitation, in the sense that they require an a-priori definition of what constitutes "acceptable agreement" in terms of the magnitude of the difference D.

Scaled coefficients of agreement, on the other hand, compare the observed value of a disagreement function to a baseline value that is obtained from some assumption or model that represents either "poor agreement" or "good agreement". Most of the existing scaled coefficients, such as kappa for categorical data and CCC for quantitative data, compare the observed agreement to the expected agreement when the observers are independent. Here, the expected agreement under independence represents a worst-case scenario, or simply a case of poor agreement. We have two reservations related to this approach. (a) Independence and lack of agreement are different concepts (see Haber and Barnhart (2006)), hence the expected agreement under independence does not necessarily corresponds to poor agreement. (b) Comparing the observed agreement to the expected under poor agreement does not provide any information regarding how close or far we are from good agreement. For example, suppose that for a given dataset $CCC = 0.8$. This means that the observed value of the disagreement function, $MSD = E(X - Y)^2$, is five times smaller than the expected MSD under independence. While this may be considered an important finding, it does not contain information on whether or not the two observers are in

good agreement.

We believe that it makes much more sense to compare the observed disagreement to its expected value under good (or acceptable) agreement because we want the coefficient to quantify the deviation (if any) from good agreement. We also believe that individual equivalence of observers, as defined in Section 2.2, is an appropriate model for good agreement because under this model, the distribution of an observation made by observer X is the same as that of an observation made by observer Y on the same subject. Therefore we propose the coefficient of individual equivalence, which compares the observed disagreement to its expected values under individual equivalence, as an alternative to kappa (in the binary case) or the CCC (in the continuous case).

The coefficients of individual agreement (CIA's, Haber et al. (2007), Barnhart et al. (2007a), Haber and Barnhart (2008), Pan et al. (2010)) are also based on comparing the observed value of a disagreement function to its expected value under "good agreement", where the latter value is based on the notion that under good agreement, the disagreement between the observers is similar to the disagreement between replicated readings of the same observer. The approach underlying CIA's have two limitations: (a) they require that the within-observer disagreement, which is used as a baseline for "good agreement" is acceptable, and (b) the CIA comparing an observer to a reference or a gold standard, i.e. ψ^R defined in Chapter 1.4.4, cannot be estimated when replicated readings on the reference observer are unavailable (as is the case when the reference value of each subject can be determined without an error). CIE and the adjusted CIE do not share these limitations.

Both CIE and CIA's require replicated readings on the same subject by the same observer. In addition, CIE can be applied when we don't have replication from one of the observers. We believe it's important to have replicated measurements. The importance of replications has been emphasized by Bland and Altman (1999). Even if the

measurements by the two methods agreed very closely on average, poor repeatability of one method would lead to poor agreement between the methods. Furthermore, when comparing a new method and a standard method, lack of agreement in un-replicated studies may suggest that the new method cannot be used, but it might be caused by poor repeatability of the standard method. Therefore, we recommend having at least one replication for the new method. In some cases replicated readings cannot be obtained, or involve logistic problems. We believe that when replications are unavailable then one should use an unscaled coefficient for evaluating agreement between two observers.

Chapter 3

Coefficient of Individual Equivalence for Replicated Binary Measurements

3.1 Introduction

The coefficient of individual equivalence (CIE), which compares the observed value of disagreement function to its expected value under the hypothesis of equivalence, was introduced in Chapter 2 as a tool to assess agreement for replicated data. While the idea of using equivalence as a criterion for acceptable agreement is valid for any measurement scale (binary, ordinal, nominal and quantitative), this chapter focuses on binary observations. In Section 3.2, we present an alternative way to define and estimate CIE for binary outcomes when mean squared error (MSD) is used as the disagreement function. An adjusted CIE, as well as its estimation, is discussed in Section 3.3. Simulation results are presented in Section 3.4, while in Section 3.5 we apply the new concepts and methods to data from a study designed to assess the agreement between ten radiologists' readings of mammograms. A discussion in Section 3.6 concludes this Chapter.

3.2 The Coefficient of Individual Equivalence for Binary Observation

3.2.1 Non-Parametric Estimation of CIE for $G(X,Y) = E(X-Y)^2 = P(X \neq Y)$

While the general expression (2.3) for $\widehat{CIE}$ can be used for any disagreement function G, we now propose a computationally simpler method for the most commonly used disagreement function for binary readings, $G(X,Y) = E(X-Y)^2 = P(X \neq Y)$. An alternative expression for estimating the numerator in equation (2.1) is described below and this expression is shown in appendix B to be equivalent to the non-parametric estimator of the numerator of CIE, when MSD(X,Y) is chosen for the disagreement function G in (2.2). The denominator, $\hat{G}(X,Y)$, is estimated as described in Section

2.3. Note that $\hat{G}_i(X,Y) = 1$ when (X,Y)=(1,0) or (X,Y)=(0,1). Define T_i as the number of "1"s for observer X and U_i as the number of "1"s for observer Y. The numbers of pairs (1,0) and (0,1) are $T_i(L-U_i)$ and $U_i(K-T_i)$, respectively. While averaging across all the $K \cdot L$ pairs, $\hat{G}_i(X,Y) = [T_i(L-U_i) + (K-T_i)U_i]/KL$. Let's define T_i^A as the number of "1"s in the permutation assignment for observer X and U_i^A as the number of "1"s in the same permutation assigment for observer Y. If we consider all the assignments of K $X's$ and the L $Y's$ to the $K+L$ observations made on subject i, we note that $W_i = T_i^A + U_i^A$ is fixed in all the assignments and the conditional distribution of T_i^A is hypergeometric given K and L. $\hat{G_i^E}$ can be written as:

$$\hat{G_i}^E = E_H[2(T_i^A)^2 + (L-K-2W_i)T_i^A + KW_i]/(KL). \tag{3.1}$$

where E_H stands for the expectation under the hypergeometric distribution. We can calculate the last expression for $\hat{G_i}^E$ from the first two moments of the hypergeometric distribution:

$$E_H(T_i^A) = E(T_i^A|W_i) = W_iK/M. \tag{3.2}$$

$$\begin{aligned} E_H(T_i^A)^2 &= E[(T_i^A)^2|W_i] \\ &= [(W_iKL(M-W_i) + W_i^2K^2(M-1)]/[M^2(M-1)]. \end{aligned} \tag{3.3}$$

where $M = K+L$ is the total number of observations on this subject. Therefore, by (3.1), (3.2) and (3.3), $\hat{G}_i^E$ can be simplified as follows:

$$\hat{G}_i^E = \frac{2W_i(M-W_i)}{M(M-1)} = \frac{2W_i(K+L) - 2W_i^2}{(K+L)(K+L-1)}. \tag{3.4}$$

The average value of equation (3.4) provides an estimate for the numerator of CIE. Thus, the non-parametric estimate of CIE can be written as

$$\widehat{CIE} = \frac{\sum_{i=1}^{N} \hat{G}_i^E}{\sum_{i=1}^{N} \hat{G}_i(X,Y)}. \tag{3.5}$$

3.2.2 Minimum and Maximum Values of CIE

3.2.2.1 Minimum Value of CIE

Denote $\pi_i = P(X_{ik} = 1)$ and $\lambda_i = P(Y_{il} = 1)$, CIE achieves its minimum when $\pi_i = 0$ and $\lambda_i = 1$ or $\pi_i = 1, \lambda_i = 0$ for each subject i. Then for subject i, assuming $\pi_i = 0$ and $\lambda_i = 1$:

$$G_i^E = \frac{2KL}{(K+L)(K+L-1)},$$

and

$$G^E = \frac{\sum_i^N G_i^E}{N} = \frac{2KL}{(K+L)(K+L-1)}.$$

Obviously, in the denominator of CIE, $G_i(X,Y) = \pi_i + \lambda_i - 2\pi_i\lambda_i = 1$ when $\pi_i = 0$ and $\lambda_i = 1$. Therefore $G(X,Y) = 1$ and hence

$$CIE_{\min} = \frac{2KL}{(K+L)(K+L-1)}.$$

For example, when K=L=3, $CIE_{\min} = 0.6$; when K=4 and L=2, $CIE_{\min} = 0.533$; when K=1 and L=2, $CIE_{\min} = 0.667$.

3.2.2.2 Maximum Value of CIE

Let's take a look at the individual level first. By the general expression,

$$G_i^E = \frac{C_2^K G_i(X,X') + C_2^L G_i(Y,Y') + K \cdot L \cdot G_i(X,Y)}{C_2^{K+L}}.$$

We know

$$\begin{aligned} G_i(X, X') &= 2\pi_i(1-\pi_i) \\ G_i(Y, Y') &= 2\lambda_i(1-\lambda_i) \\ G_i(X, Y) &= \pi_i + \lambda_i - 2\pi_i\lambda_i \end{aligned}$$

Therefore, we can rewrite the general expression as

$$\begin{aligned} G_i^E &= \frac{2K(K-1)\pi_i(1-\pi_i) + 2L(1-L)\lambda_i(1-\lambda_i) + 2KL(\pi_i + \lambda_i - 2\pi_i\lambda_i)}{(K+L)(K+L-1)} \\ &= \frac{2K^2\pi_i(1-\pi_i) + 2L^2\lambda_i(1-\lambda_i) - 2[K\pi_i(1-\pi_i) + L\lambda_i(1-\lambda_i)] + 2KL(\pi_i + \lambda_i - 2\pi_i\lambda_i)}{(K+L)(K+L-1)} \end{aligned}$$

CIE achieves its maximum when $\pi_i = \lambda_i$ for all subjects, which indicates perfect agreement. Then

$$G_i^E = \frac{2(K^2+L^2)\pi_i(1-\pi_i) - 2(K+L)\pi_i(1-\pi_i) + 2KL \cdot 2\pi_i(1-\pi_i)}{(K+L)(K+L-1)}.$$

Also, when $\pi_i = \lambda_i$,

$$G_i(X, Y) = 2\pi_i(1-\pi_i).$$

$$\begin{aligned} G^E &= \frac{1}{n}\sum_{i=1}^{N} G_i^E \\ &= \frac{2(K^2+L^2)\sum_{i=1}^{N}\pi_i(1-\pi_i) - 2(K+L)\sum_{i=1}^{N}\pi_i(1-\pi_i) + 2KL \cdot 2\sum_{i=1}^{N}\pi_i(1-\pi_i)}{n(K+L)(K+L-1)} \end{aligned}$$

and $G(X, Y) = \sum_{i=1}^{N} G_i(X, Y) = 2\sum_{i=1}^{N}\pi_i(1-\pi_i)/n$. Therefore,

$$CIE_{max} = \frac{\sum_{i=1}^{N} G^E}{\sum_{i=1}^{N} G(X, Y)} = \frac{K^2+L^2-K-L+2KL}{(K+L)(K+L-1)} = 1.$$

3.3 The Adjusted CIE (Definition and Estimation)

3.3.1 General Definition

As discussed in the previous section, the minimum of CIE isn't equal to 0 and in almost all the cases it's greater than 0.5. To force the index to range from 0 to 1, we define the adjusted CIE. The adjusted CIE is denoted as CIEA and defined as:

$$CIEA = \frac{CIE - CIE_{\min}}{1 - CIE_{\min}}. \tag{3.6}$$

Thus the true value of CIEA ranges from 0 to 1. The estimate of CIEA can be obtained from $\widehat{CIE}$ because $CIEA$ is a linear function of CIE.

3.3.2 The Standard Error of the Adjusted CIE

Let's first consider the standard error of CIE. Assume that for each i, the pair (π_i, λ_i) is randomly drawn from a large hypothetical population. Then $\hat{\pi}_i$ is the maximum likelihood estimator (MLE) of π_i and similarly, $\hat{\lambda}_i$ is the MLE of λ_i. Since $\hat{G}_i^{\,E}$ and $\hat{G}_i(X,Y)$ are functions of $\hat{\pi}_i$ and $\hat{\lambda}_i$, the asymptotic normality of the estimate of CIE can be established by applying the multivariate central limit theorem and multivariate delta method. As CIEA is just a linear transformation of CIE, the asymptotic property holds for $\widehat{CIEA}$.

Let $A = \hat{G}^E = \frac{1}{N}\sum_{i=1}^{N}\hat{G}_i^E$ and $B = \hat{G}_i(X,Y) = \frac{1}{N}\sum_{i=1}^{N}\hat{G}_i(X,Y)$.Then $\widehat{CIE} = A/B$. The sample variance of A is $S^2(A)$=$\frac{1}{N-1}\sum_{i=1}^{N}(\hat{G}_i^E - \hat{G}^E)^2$ and then $\widehat{Var}(A) = S^2(A)/N$. Similarly, $\widehat{Var}(B) = S^2(B)/N$, where $S^2(B)$=$\frac{1}{N-1}\sum_{i=1}^{N}(\hat{G}_i(X,Y)-\hat{G}(X,Y))^2$. Also, $\hat{Cov}(A,B) = [\sum_{i=1}^{N}(\hat{G}_i^E - \hat{G}^E)(\hat{G}_i(X,Y) - \hat{G}(X,Y))]/N(N-1)$. Finally

$$Var(\widehat{CIE}) = \widehat{Var}(\frac{A}{B}) \approx \frac{A^2}{B^2}[\frac{\widehat{Var}(A)}{A^2} + \frac{\widehat{Var}(B)}{B^2} - \frac{\widehat{Cov}(A,B)}{A \cdot B}].$$

The standard error of the adjusted CIE is obtained as:

$$SE(\widehat{CIEA}) = \frac{1}{(1 - CIE_{\min})}\sqrt{Var(\widehat{CIE})}.$$

3.3.3 Asymptotic Property of CIE

Denote $\pi_i = Pr(X_{ik} = 1)$ and $\lambda_i = Pr(Y_{il} = 1)$. Assume

$$\begin{pmatrix} \pi_i \\ \lambda_i \end{pmatrix} \overset{iid}{\sim} \text{P, where P denotes a known population.}$$

Denote T_i is the number of "1"s for observer X on subject i and K is the total number of replications observer X made per subject. Thus, $\hat{\pi}_i = \frac{T_i}{K}$ is the maximum likelihood estimator of π_i. Similarly, $\hat{\lambda}_i = \frac{U_i}{L}$ is the MLE of λ_i, where U_i is the number of "1"s for observer Y on subject i and L is the total number of replications over observer Y per subject. Therefore,

$$\begin{pmatrix} \hat{\pi}_i \\ \hat{\lambda}_i \end{pmatrix} \overset{iid}{\sim} \text{distributed.}$$

$$CIE = \frac{E_i(G_i^E)}{E_i(G_i(X,Y))}.$$

$$\widehat{CIE} = \frac{\frac{1}{N}\sum_{i=1}^{N} \hat{G_i}^E}{\frac{1}{N}\sum_{i=1}^{N} \hat{G}_i(X,Y)}.$$

Assume $A = \bar{\hat{G}}_i^E$ and $B = \bar{\hat{G}}_i(X,Y)$. Let

$$Z = \begin{pmatrix} A \\ B \end{pmatrix}$$

Since $\hat{G}_i^{\,E}$ and $\hat{G}_i(X,Y)$ are functions of the joint distribution of $(\hat{\pi}_i, \hat{\lambda}_i)$ which are *iid* distributed, by multivariate central limit theorem,

$$Z \to N\left(\begin{pmatrix} E(A) \\ E(B) \end{pmatrix}, \quad \begin{pmatrix} Var(A) & Cov(A,B) \\ Cov(A,B) & Var(B) \end{pmatrix}\right).$$

Assume $f(Z) = \frac{A}{B}$, where $f(Z) = \widehat{CIE}$. The asymptotic normality of $\widehat{CIE}$ will be established by multivariate Delta method.

$$\frac{\partial f(Z)}{\partial A} = \frac{1}{B}$$

$$\frac{\partial f(Z)}{\partial B} = -\frac{A}{B^2}$$

$$f'(Z) = \begin{pmatrix} \frac{1}{B} \\ -\frac{A}{B^2} \end{pmatrix}.$$

Therefore,

$$\begin{aligned} Var(\frac{A}{B}) &= \begin{pmatrix} \frac{1}{B} & -\frac{A}{B^2} \end{pmatrix} \begin{pmatrix} Var(A) & Cov(A,B) \\ Cov(A,B) & Var(B) \end{pmatrix} \begin{pmatrix} \frac{1}{B} \\ -\frac{A}{B^2} \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{B}Var(A) - \frac{A}{B^2}Cov(A,B) & \frac{1}{B}Cov(A,B) - \frac{A}{B^2}Var(B) \end{pmatrix} \begin{pmatrix} \frac{1}{B} \\ -\frac{A}{B^2} \end{pmatrix} \\ &= \frac{1}{B}Var(A) - \frac{A}{B^3}Cov(A,B) + \frac{A^2}{B^4}VarB \\ &= \frac{A^2}{B}[\frac{Var(A)}{A^2} + \frac{Var(B)}{B^2} - \frac{Cov(A,B)}{A.B}]. \end{aligned}$$

3.3.4 Interpretation of CIEA

As with every new agreement coefficient, it is necessary to specify criteria for "good" or "acceptable" agreement in a given application. Intuitively CIE compares the observed disagreement to disagreement under equivalence and CIEA is the adjusted CIE using the minimum value of CIE. Based on our experience, we suggest the rule of thumb that the true CIEA value should be greater than and equal to 0.8 in order to have good agreement. This implies that the lower confidence limit of CIEA needs to be greater than or equal to 0.8 when analyzing data in order to claim good agreement.

3.4 Simulation Studies

3.4.1 Data Generation

We assume that observers' binary readings are based on the value of a "true" variable T (observed or unobserved), where $T \sim N(\mu_T, \sigma_T^2)$ and σ_T^2 is the between subjects variability. Let A_i and B_i be the biases for observer X and Y, respectively. We assume that $A_i \sim N(\mu_A, \sigma_A^2), B_i \sim N(\mu_B, \sigma_B^2)$ and T, A and B are mutually independent.

For subject i, let $T_i = t_i$, $A_i = a_i$, $B_i = b_i$. We use U_i and V_i to denote the observers' readings on the subject's true value, and we assume $U_i \sim N(t_i+a_i, \sigma_U^2), V_i \sim N(t_i + b_i, \sigma_V^2)$, where σ_U^2 and σ_V^2 are the within-observer error variances for observers X and Y, respectively. Denote by U_{ik} the kth replicated reading of U_i, $i = 1, \ldots K$. Similarly, V_{il} is the lth replicated reading of V_i, $l = 1, \ldots L$. Finally, for a fixed common threshold C, we generate binary readings $X_{i1}, \ldots, X_{iK}$ and $Y_{i1}, \ldots, Y_{iL}$ by the criterion that if $U_{ik} > C$ then $X_{ik} = 1$, else then $X_{ik} = 0$, and if $V_{il} > C$ then $Y_{il} = 1$, else then $Y_{il} = 0$. Note that X's and Y's are conditionally independent given the subject.

For given t_i, a_i, b_i, the values of $P(X_{ik} = 1)$ and $P(Y_{il} = 1)$ are π_i and λ_i, as

follows:

$$\pi_i = P(X_{ik} = 1) = P(U_{ik} > C) = 1 - \Phi\left(\frac{C - (t_i + a_i)}{\sigma_U}\right)$$

$$\lambda_i = P(Y_{il} = 1) = P(V_{il} > C) = 1 - \Phi\left(\frac{C - (t_i + b_i)}{\sigma_V}\right),$$

where Φ is the CDF of a standard normal.

3.4.2 Simulation Process

Suppose T stands for the systolic blood pressure and we consider a hypothetical population of obese type II diabetes, i.e., persons with $\mu_T = 138mmHg$ and a between-subjects $\sigma_T = 5mmHg$. The usual cut-off point for hypertension is 140mmHg for systolic blood pressure, which is the threshold C in our model. We set the within-observers standard errors to 3 ($\sigma_U = \sigma_V = 3$). Furthermore, we set the average bias of observer X, μ_A, to 0 and gradually increase the average bias of observer Y, μ_B, from 1 to 3 to 5 in order to accommodate good, moderate and poor agreement, respectively. The closer μ_A and μ_B, the better the agreement when other conditions remain unchanged. Standard deviation for biases in observers X and Y were set to 1 ($\sigma_A = \sigma_B = 1$).

For each value of μ_B, we used three values of the sample size (N=50, 100 and 200) along with four combinations of the numbers of replications (K,L)=(1,2), (3,3), (4,2) and (10,10). The $K = 1$ and $L = 2$ combination was used because (a) $K + L = 3$ is the smallest number of replicates that is allowed for the CIE and adjusted CIE (CIEA); (b) the CIA without a reference, ψ^N in Chapter 1.4.4, is not defined when $K = 1$ or $L = 1$; (c) The CIA with observer X as the reference, ψ^R in Chapter 1.4.4, cannot be estimated when there are no replicated observation by X ($K = 1$). Hence only the CIE or CIEA can be estimated when $K = 1, L = 2$.

In our simulation study, we assessed the accuracy and precision of the estimated CIEA relative to the true CIEA. We estimated the CIEA and its standard error from

each sample. The average of the bias, the average estimated standard error and the coverage probabilities of the 95% confidence interval were reported for each set-up. One thousand simulations were conducted for each parameter set.

3.4.3 Simulation Results

Table 3.1 presents the true values, biases, standard errors and the coverage probability of the estimates for all the combinations of (N, K, L) for poor, moderate and good agreement cases defined earlier. As shown in Table 3.1, the bias was minimal for all combinations and it decreased with the increase of sample size. Both standard errors based on simulations of CIE and the mean of estimated standard errors calculated from SE estimator are presented. The similarity between those two standard errors confirms the robustness of our standard error estimates. For moderate sample sizes, the coverage probability was very close to the nominal 95% level. Even when sample size was 50, the coverage probabilities were above 91% for almost all the combinations.

Table 3.1: Simulation Results: Estimation of CIEA when (K,L)=(1,2), (3,3), (4,2) and (10,10)

Sample Size	K	L	μ_B	True Value	Bias	SE^a	SE^b	CP
50	1	2	1	0.942	0.022	0.243	0.246	0.943
100	1	2	1	0.942	0.007	0.178	0.173	0.938
200	1	2	1	0.942	0.007	0.120	0.123	0.950
50	3	3	1	0.922	0.006	0.099	0.095	0.912
100	3	3	1	0.922	0.003	0.070	0.070	0.926
200	3	3	1	0.922	0.000	0.051	0.050	0.935
50	4	2	1	0.908	0.007	0.115	0.115	0.923
100	4	2	1	0.908	0.004	0.082	0.083	0.941
200	4	2	1	0.908	0.001	0.061	0.059	0.940
50	10	10	1	0.922	0.001	0.020	0.019	0.892
100	10	10	1	0.922	0.000	0.014	0.013	0.927
200	10	10	1	0.922	0.000	0.010	0.010	0.935
50	1	2	3	0.783	0.020	0.228	0.215	0.918
100	1	2	3	0.783	0.005	0.146	0.151	0.946
200	1	2	3	0.783	0.006	0.102	0.107	0.960
50	3	3	3	0.766	0.005	0.094	0.097	0.947
100	3	3	3	0.766	0.004	0.069	0.069	0.939
200	3	3	3	0.766	0.001	0.050	0.050	0.941
50	4	2	3	0.755	0.012	0.111	0.111	0.934
100	4	2	3	0.755	0.003	0.077	0.079	0.951
200	4	2	3	0.755	0.003	0.057	0.056	0.938
50	10	10	3	0.766	0.002	0.024	0.024	0.920
100	10	10	3	0.766	-0.001	0.017	0.017	0.938
200	10	10	3	0.766	0.000	0.012	0.012	0.944
50	1	2	5	0.554	0.015	0.176	0.170	0.933
100	1	2	5	0.554	0.005	0.119	0.119	0.935
200	1	2	5	0.554	0.005	0.081	0.084	0.963
50	3	3	5	0.576	0.009	0.084	0.085	0.942
100	3	3	5	0.576	0.005	0.059	0.061	0.945
200	3	3	5	0.576	0.002	0.043	0.043	0.943
50	4	2	5	0.592	0.010	0.096	0.099	0.946
100	4	2	5	0.592	0.005	0.069	0.070	0.945
200	4	2	5	0.592	0.003	0.050	0.049	0.938
50	10	10	5	0.576	0.002	0.023	0.023	0.945
100	10	10	5	0.576	0.000	0.017	0.016	0.936
200	10	10	5	0.576	0.001	0.012	0.011	0.942

[a]Standard errors based on simulations of CIEA

[b]Mean of estimated standard errors calculated from SE estimator in Section 3.2

3.5 An Example

Data from a mammography study (Haber et al. (2007), Elmore et al. (1994)) and Section 1.2 in this dissertation was used to illustrate the new coefficient of agreement. Each of the study participants was followed up for three years, and then a definitive diagnosis was made. The definitive diagnosis was breast cancer if it was histopathologically confirmed within the three years of follow-up. We considered this diagnosis as the patient's "true" breast cancer status. Since the total of sensitivity and specificity was highest for radiologist A, we illustrated the new coefficients by estimating the agreement between radiologist A and each of the remaining nine radiologists. One observation with missing data in radiologist A was removed from the analysis.

The estimates, standard errors and 95% confidence intervals of CIEA are presented in Table 3.2 for the nine comparisons. Among all those pairwise comparisons, the estimated CIEA of radiologists A and F was the highest (0.763) with 95% CI (0.475, 1.051). However, the lower limit of 95% CI is less than 0.8, which means that radiologist A and F failed to achieve a good agreement. In this example, since $K = L = 2$ for every subject, our estimates of CIEA's are almost identical to the estimates of CIA's (ψ^N) reported in Haber et al. (2007) for the same data. (The small differences result from the deletion of a patient who was missing one reading by radiologist A from our analysis.) Since our estimates of the standard error of CIEA are based on the method in Section 3.3.2, while the estimates of the SE of CIA in Haber et al. (2007) were obtained via the bootstrap, the confidence intervals are not the same. The bootstrapped CIs, compared to the one based on the proposed expression in Section 3.3.2, were right shifted a little bit.

As we mentioned earlier, CIEA can be estimated when some of the subjects have only one reading by one of the observers. CIA cannot be estimated in this case. To illustrate this advantage of CIEA over CIA, we randomly select 30 patients (20%), and deleted at random the first or the second reading by radiologist A. We kept the

complete data on the remaining 119 patients. Table 3.3 presents the estimates, SEs and the 95% CIs of CIEA with one of the readings of radiologist A missing in 20% of the patients. Similar estimates and slightly larger SEs of CIEA were obtained for each comparison. As before, the comparison of radiologist A and F had the largest $\widehat{CIEA}$ but the lower limit 95% confidence limit is less than 0.8 that fails to conclude good agreement by CIEA.

Table 3.2: Estimated CIEA's for Nine Pairs of Radiologists

Radiologist	CIEA	SE(CIEA)	95% CI(CIEA)
(A,B)	0.646	0.141	(0.370 0.922)
(A,C)	0.358	0.093	(0.176 0.540)
(A,D)	0.697	0.126	(0.450 0.944)
(A,E)	0.643	0.141	(0.367 0.919)
(A,F)	0.763	0.147	(0.475 1.051)
(A,G)	0.541	0.111	(0.323 0.759)
(A,H)	0.486	0.108	(0.274 0.699)
(A,I)	0.739	0.111	(0.521 0.957)
(A,J)	0.619	0.108	(0.407 0.831)

Table 3.3: Estimated CIEA's for Nine Pairs of Radiologists with 20% of Patients are Missing One of the Readings of Radiologist A

Radiologist	CIEA	SE(CIEA)	95% CI(CIEA)
(A,B)	0.677	0.149	(0.385 0.969)
(A,C)	0.333	0.098	(0.142 0.524)
(A,D)	0.719	0.135	(0.453 0.984)
(A,E)	0.600	0.153	(0.300 0.900)
(A,F)	0.762	0.182	(0.406 1.118)
(A,G)	0.553	0.122	(0.313 0.793)
(A,H)	0.459	0.115	(0.235 0.684)
(A,I)	0.744	0.121	(0.506 0.982)
(A,J)	0.682	0.136	(0.414 0.949)

3.6 Discussion

In this Chapter, we presented an approach to defining and estimating a new index of agreement, CIE, for the comparison of two fixed observers or methods with replicated binary measurements. This coefficient compares the observed disagreement to its expected value under individual equivalence, i.e., when the distribution of the readings of both observers on the same subject are identical. We used the conditional version of CIEA given K and L. The dependence of CIEA on K and L results from the permutation-based approach. Figure 3.1, 3.2 and 3.3 show the relations among CIEA, K and L where K and L ranges from 2 to 50 under good, moderate and poor agreement, respectively, with the same parameter setups in Section 3.4.2. When K and L are small, there are some curvatures on the surface plot. When K and L increase, the surfaces seem to be flat. Furthermore, Table 3.4 presents the true values of CIEA under the simulation model described in Section 3.4.1. We observed the dependency of CIEA on the number of replications seems to depend on the ratio K/L. There is a differences of about 0.02 in CIEA when the ratio of K/L decreases from 1.0 to 0.1. As shown in Figure 3.4 (L=1000), for all the good, moderate and

poor agreement scenarios, CIEA stays approximately constant across different values of K/L.

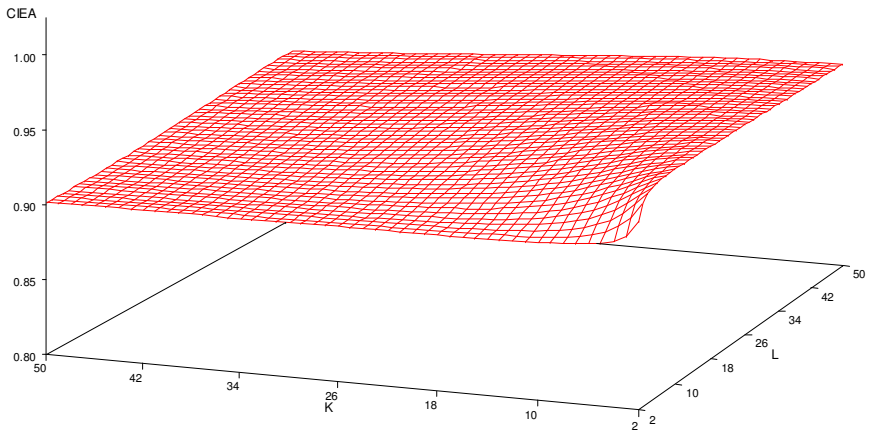

Figure 3.1: Surface Plot of CIEA, K and L for Replicated Binary Measurements under Good Agreement

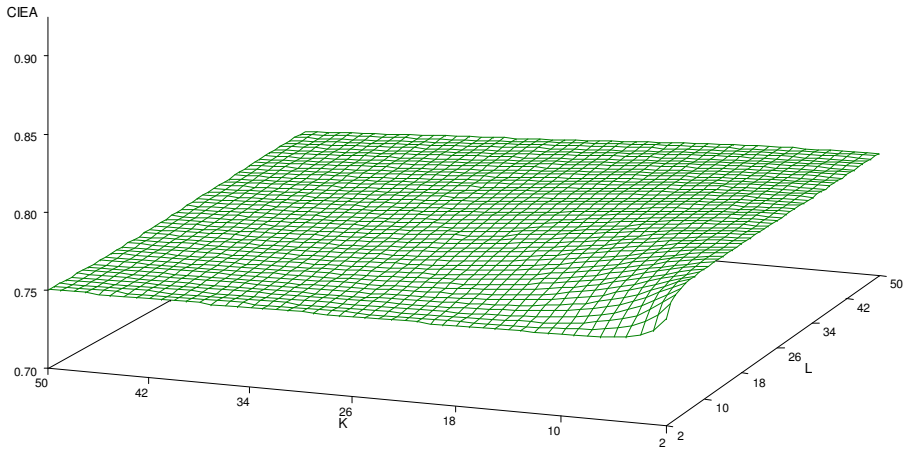

Figure 3.2: Surface Plot of CIEA, K and L for Replicated Binary Measurements under Moderate Agreement

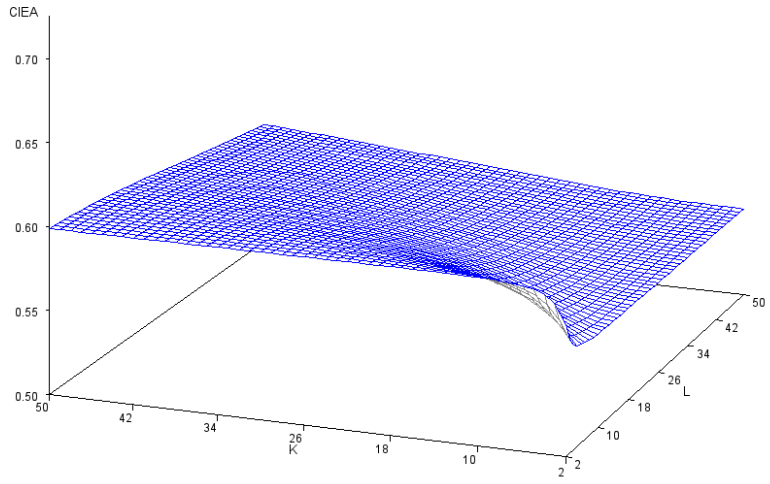

Figure 3.3: Surface Plot of CIEA, K and L for Replicated Binary Measurements under Poor Agreement

Table 3.4: Dependence of CIEA on K and L under Simulation Set up

μ_b	L	K/L					
		0.1	0.3	0.5	0.7	0.9	1.0
1	10	0.9419	0.9394	0.9376	0.9292	0.9241	0.9219
1	100	0.9416	0.9387	0.9340	0.9288	0.9240	0.9219
1	1000	0.9415	0.9386	0.9339	0.9287	0.9240	0.9219
3	10	0.7826	0.7806	0.7767	0.7723	0.7683	0.7665
3	100	0.7823	0.7800	0.7762	0.7720	0.7682	0.7665
3	1000	0.7823	0.7799	0.7762	0.7720	0.7682	0.7665
5	10	0.5537	0.5567	0.5620	0.5681	0.5738	0.5763
5	100	0.5543	0.5575	0.5628	0.5686	0.5739	0.5763
5	1000	0.5543	0.5576	0.5628	0.5686	0.5739	0.5763

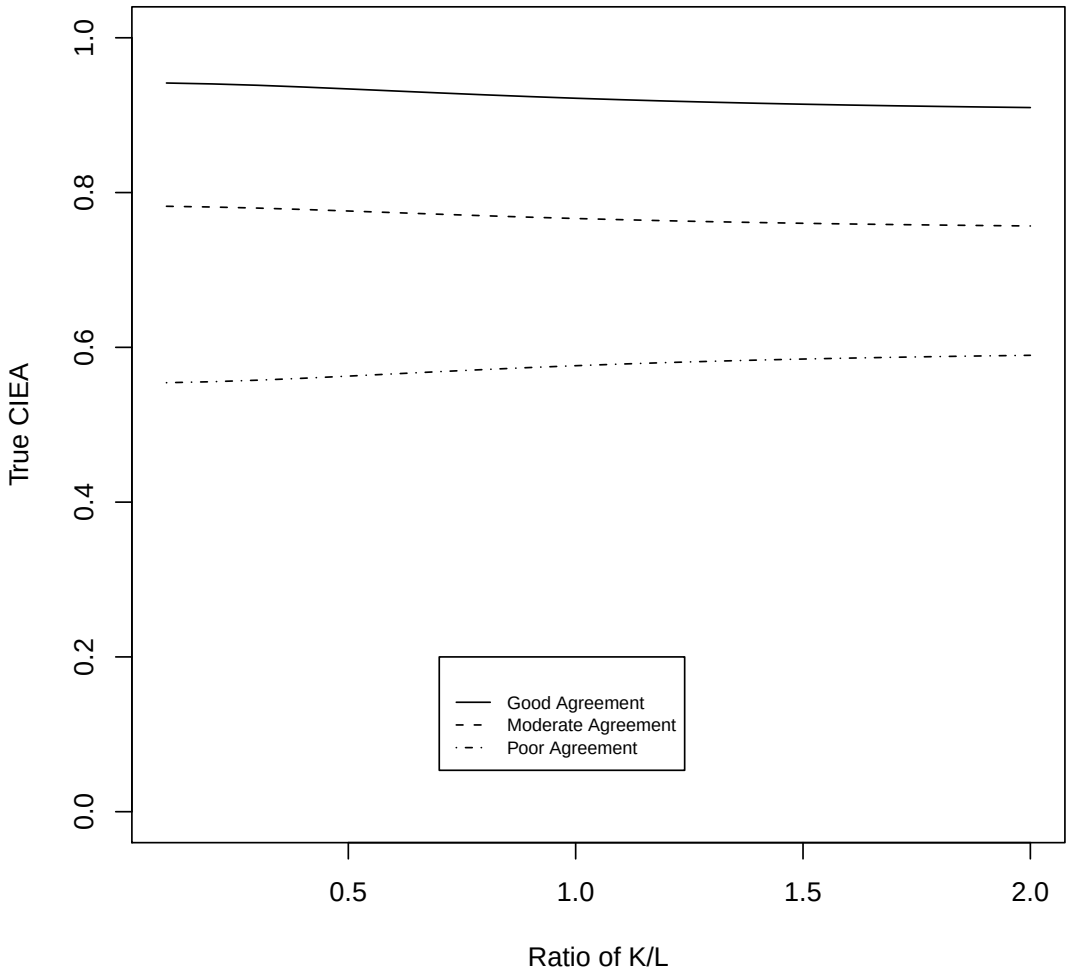

Figure 3.4: Dependence of True CIEA on the ratio of K/L when L=1000

Sample size calculations are essential in agreement studies because it's important to determine the number of subjects and the number of replications needed in order to achieve a desired precision when estimating CIEA. For binary data, Gao (2010) obtained the standard error (SE) of CIA in terms of N,K and L. One can get a similar expression of the SE of CIEA so that the corresponding sample size of interest can be calculated.

In Chapter 6, we will introduce a model-based approach to estimating CIEA from replicated and repeated binary data.

Chapter 4

Coefficient of Individual Equivalence for Quantitative Replicated Measurements

4.1 Introduction

In Chapters 2 and 3 as well as in Pan et al. (2011), CIE has been introduced and discussed when two observers make binary readings on each subject. In this Chapter we focus on the properties of the new approach when the observations are quantitative. In addition to the nonparametric approach, which is described in Chapter 2, we introduce a parametric approach, based on ANOVA mixed effects models, for estimation and inference. The parametric model is discussed in Section 4.2. And adjusted CIE, based on the minimum value of CIE, is introduced in Section 4.3. Simulation results are presented in Section 4.4, while in Section 4.5 we apply the new concepts and methods to data from a study designed to determine the suitability of magnetic resonance angiography (MRA) for noninvasive screening of carotid artery stenosis, compared to invasive intra-arterial angiogram (IA). A discussion in Section 4.6 concludes this Chapter.

4.2 Definition and Estimation of the Coefficient of Individual Equivalence (CIE)

4.2.1 Estimation of CIE with $G(X,Y) = E(X-Y)^2$ using Linear Mixed Effects Models for Normally Distributed Observations

When X and Y are normally distributed and $G(X,Y) = E(X-Y)^2$, a parametric approach for estimation and inference on CIE can be based on a two-way mixed linear model with subject effect α_i, observer effect β_j as well as the interaction between subject and observer γ_{ij}. The residual term, ε_{ijk}, represents the within-observer replication variability. When there are $J \geq 2$ observers ($j = 1, \cdots, J$) we model the

k^{th} replication of observer j on subject i as

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \tag{4.1}$$

where the subject effect is random and the observer effect is fixed. We assume that the random effects distributions are $\alpha_i \sim N(0, \sigma^2_\alpha)$, $\gamma_{ij} \sim N(0, \sigma^2_\gamma)$ and $\varepsilon_{ijk} \sim N(0, \sigma^2_{\varepsilon_j})$. We also assume that α_i, γ_{ij} and ε_{ijk} are independent. Note that each observer may have a different within-observer variance σ_{ε_j}. Since the observer is treated as a fixed effect, define σ^2_β as $\sigma^2_\beta = \frac{1}{J-1}\sum_{j=1}^{J}(\beta_j - \bar{\beta})^2$.

Suppose now that there are only two observers, X and Y, and denote $X = Y_1$ and $Y = Y_2$. Let $G(X,Y) = MSD(X,Y) = E(X-Y)^2$. It is easy to see that

$$\begin{aligned}
MSD(X,X') &= E(Y_{i1k} - Y_{i1k'})^2 = E(\varepsilon_{i1k} - \varepsilon_{i1k'})^2 = 2\sigma^2_{\varepsilon_1} \\
MSD(Y,Y') &= E(Y_{i2k} - Y_{i2k'})^2 = E(\varepsilon_{i2k} - \varepsilon_{i2k'})^2 = 2\sigma^2_{\varepsilon_2} \\
MSD(X,Y) &= E(Y_{i1k} - Y_{i2k'})^2 = E(\beta_1 - \beta_2)^2 + E(\gamma_{i1} - \gamma_{i2})^2 + E(\varepsilon_{i1k} - \varepsilon_{i2k'})^2 \\
&= 2\sigma^2_\beta + 2\sigma^2_\gamma + \sigma^2_{\varepsilon_1} + \sigma^2_{\varepsilon_2}.
\end{aligned}$$

As we have seen, the expression of CIE can be simplified as

$$CIE = \frac{C_2^K G(X,X') + C_2^L G(Y,Y')}{C_2^{K+L} G(X,Y)} + \frac{K \cdot L}{C_2^{K+L}}.$$

When $G = MSD$ is used as the disagreement function, G(X,X'), G(Y,Y') and G(X,Y) can be written in terms of the parameters of the mixed model as shown before. Hence,

$$CIE = \frac{C_2^K \sigma^2_{\varepsilon_1} + C_2^L \sigma^2_{\varepsilon_2}}{C_2^{K+L} \cdot (\sigma^2_\beta + \sigma^2_\gamma + \sigma^2_{\varepsilon_1} + \sigma^2_{\varepsilon_2})} + \frac{K \cdot L}{C_2^{K+L}}. \tag{4.2}$$

The CIE can be estimated by replacing the parameters by their REML estimates.

In the special case $K = L$, the estimate of CIE can be written as

$$\widehat{CIE} = \frac{K}{2K-1} \cdot \frac{\hat{\sigma}_{\varepsilon}^2}{\hat{\sigma}_{\beta}^2 + \hat{\sigma}_{\gamma}^2 + \hat{\sigma}_{\varepsilon}^2} + \frac{K}{(2K-1)}.$$

where $\hat{\sigma}_{\varepsilon}^2 = (\hat{\sigma}_{\varepsilon_1}^2 + \hat{\sigma}_{\varepsilon_2}^2)/2$.

4.2.2 Minimum and Maximum Values of CIE

4.2.2.1 Minimum Value of CIE

In order to derive a coefficient ranging from zero to one, we need to look at the minimum of CIE and make an adjustment if necessary. In (2), if $G(X, X') = G(Y, Y') = 0$ while keeping $G(X, Y) > 0$, CIE achieves its minimum as follows:

$$\begin{aligned} CIE_{\min} &= \frac{K \cdot L}{C_2^{K+L}} \\ &= \frac{2K \cdot L}{(K+L)(K+L-1)}. \end{aligned} \tag{4.3}$$

4.2.2.2 Maximum Value of CIE

CIE can be written as shown in (4.2). Notice that acceptable within-observer disagreement is required when estimating CIE. We assume that the within observer disagreement should not be greater than the between observer disagreement, that is, $G(X, X') \leq G(X, Y)$ and $G(Y, Y') \leq G(X, Y)$. Under these assumptions, CIE achieves its maximum when $G(X, X') = G(X, Y)$ and $G(Y, Y') = G(X, Y)$.

$$\begin{aligned} CIE_{max} &= \frac{C_2^K + C_2^L + K \cdot L}{C_2^{K+L}} = \frac{\frac{K(K-1)}{2} + \frac{L(L-1)}{2} + K \cdot L}{\frac{(K+L)(K+L-1)}{2}} \\ &= \frac{K(K-1) + L(L-1) + 2K \cdot L}{(K+L)(K+L-1)} = \frac{(K+L)^2 - (K+L)}{(K+L)(K+L-1)} = 1 \end{aligned}$$

Therefore, the maximum value of CIE for continuous measurement under reasonable assumptions is 1.

4.3 Adjusted CIE

4.3.1 Definition

As we see, the minimum of CIE is greater than 0 and in almost all cases it's greater than 0.5. To force the index to range from 0 to 1, we define an adjusted CIE. The adjusted CIE is denoted as CIEA and defined as:

$$CIEA = \frac{CIE - CIE_{\min}}{1 - CIE_{\min}}, \tag{4.4}$$

where $CIE_{\min}$ is defined in (4.3).The estimate of CIEA can be easily obtained from $\widehat{CIE}$ because $CIEA$ is a linear function of CIE.

4.3.2 Large Sample Distribution and Standard Error of $\widehat{CIEA}$

Let's first consider the nonparametric estimators of CIE. Since both estimator of G^E and $G(X,Y)$ are means overall all N subjects, the asymptotic normality of the estimate of CIE can be established by applying the multivariate central limit theorem and multivariate delta method when the sample size is large enough.

Let $A = \hat{G}^E = \frac{1}{N}\sum_{i=1}^{N}\hat{G}_i^E$ and $B = \hat{G}_i(X,Y) = \frac{1}{N}\sum_{i=1}^{N}\hat{G}_i(X,Y)$. Then $\widehat{CIE} = A/B$. The sample variance of A is $S^2(A)=\frac{1}{N-1}\sum_{i=1}^{N}(\hat{G}_i^E - \hat{G}^E)^2$ and then $\hat{Var}(A) = S^2(A)/N$. Similarly, $\hat{Var}(B) = S^2(B)/N$, where $S^2(B)=\frac{1}{N-1}\sum_{i=1}^{N}(\hat{G}_i(X,Y)-\hat{G}(X,Y))^2$. Also, $\hat{Cov}(A,B) = [\sum_{i=1}^{N}(\hat{G}_i^E - \hat{G}^E)(\hat{G}_i(X,Y) - \hat{G}(X,Y))]/N(N-1)$. Finally

$$\hat{Var}(\widehat{CIE}) = \hat{Var}(\frac{A}{B}) \approx \frac{A^2}{B^2}[\frac{\hat{Var}(A)}{A^2} + \frac{\hat{Var}(B)}{B^2} - \frac{\hat{Cov}(A,B)}{A \cdot B}].$$

For the parametric estimation of CIE from a linear mixed model, bootstrap SE's and normalized confidence interval based on bootstrap SE's can be used.

As CIEA is just a linear transformation of CIE, the above asymptotic properties hold for $\widehat{CIEA}$. The standard error of $\widehat{CIEA}$ for both nonparametric method and linear mixed model estimation is obtained as:

$$SE(\widehat{CIEA}) = \frac{1}{(1 - CIE_{\min})}\sqrt{Var(\widehat{CIE})}.$$

4.3.3 Interpretation of CIEA

As with every new agreement coefficient, it is necessary to specify criteria for "good" or "acceptable" agreement in a given application. Intuitively CIE compares the observed disagreement to disagreement under individual equivalence and CIEA is the adjusted CIE using the minimum value of CIE. We suggest that the true CIEA value should be greater than or equal to 0.8 in order to have good agreement based on empirical experience. This implies that the lower confidence limit of CIEA needs to be greater than or equal to 0.8 when analyzing data in order to claim good agreement.

4.4 Simulation Studies

Simulation studies were conducted to evaluate the performance of both nonparametric and parametric approaches for estimation and inference on the adjusted coefficient of individual equivalence (CIEA). For nonparametric estimation, simulations were performed for small (n=50), moderate (n=100) and large sample sizes (n=200). For the parametric method, only sample size n=100 was included. In all settings, we look at balanced number of replications for both raters (K=L=3) and unbalanced scenarios (K=1, L=2) and (K=2, L=3). The (K=1, L=2) scenario correspondents to the case where X is a perfect gold standard that does not require replications. The

disagreement function $G = MSD$ was used in all the simulations.

In the first set of simulations, we assume that the true value $T_i \sim N(\mu_T, \sigma_T^2)$. Furthermore, we assume the conditional means and standard deviations of the observers' readings given subjects' true value are linear functions of t: $\mu_{X|t} = a + bt, \mu_{Y|t} = c + dt, \sigma_{X|t} = e + ft, \sigma_{Y|t} = g + ht$. Then for subject i, K replicated measurements of observer X were generated from $N(\mu_{X|t_i}, \sigma^2_{X|t_i})$ and L replicated measurements of observer Y were from $N(\mu_{Y|t_i}, \sigma^2_{Y|t_i})$. The three MSD functions can be expressed as Haber and Barnhart (2008):

$$\begin{aligned} MSD(X,Y) &= (a-c)^2 + e^2 + g^2 + 2[(a-c)(b-d) + ef + gh]\mu_T \\ &\quad + [(b-d)^2 + f^2 + h^2](\mu_T^2 + \sigma_T^2) \\ MSD(X,X') &= 2e^2 + 4ef\mu_T + 2f^2(\mu_T^2 + \sigma_T^2) \\ MSD(Y,Y') &= 2g^2 + 4gh\mu_T + 2h^2(\mu_T^2 + \sigma_T^2) \end{aligned}$$

Data from a carotid stenosis study (see Section 1.2.2) were used to investigate the behavior of the new coefficient. The distribution of T was defined using the sample moments from the data, $\mu_T = 43.29, \sigma_T = 29.87$. We used two sets of parameters for the conditional means and variance, in order to explore the effect of difference in variances of X and Y given T: (1) $b = 1, d = 1, e = g = 1.5, f = h = 0.3$; (2) $b = 1, d = 1, e = 1.5, g = 1, f = h = 0.3$. To accommodate good, moderate and poor agreement, we used $a = 0$ and let $c = 3.8, 16.3$ and 28.1. To investigate the performance of nonparametric and parametric approaches when the distribution of the true value is skewed, we conducted a second set of simulation as the true value T followed an exponential distribution $T_i \sim EXP(\lambda_T)$ with the mean and standard deviation parameter as 29.87. The numbers of simulations when we have equal variances was 1000. For scenarios with unequal variances, the number of simulations was reduced to 100 since we had to fit the ANOVA mixed model with heteroscedastic error terms.

The number of bootstrap samples in estimating the standard error for the parametric approach was 100.

Tables 4.1, 4.2, 4.3 and 4.4 present the true values, biases, standard errors, root mean squared errors (RMSE) and the coverage probabilities of the estimates of CIEA for all the combinations of (N, K, L) for the poor, moderate and good agreement cases defined earlier when T was normally distributed. We considered nonparametric estimation with equal variance, nonparametric estimation with unequal variance, parametric estimation with equal variance and parametric estimation with unequal variance. As shown in Tables 1 to 4, the bias was minimal for all combinations and it decreased when the sample size increased. Both standard errors based on simulations of CIE and the mean of estimated standard errors are presented. The similarity between those two standard errors confirms the robustness of our standard error estimates. For moderate and large sample sizes, the coverage probabilities were very close to the nominal 95% level.

Furthermore, we noticed that when equal variances were used, the true values of CIEA were not affected by the choice of K and L. When the variance parameters were not the same, the true values of CIEA varied in a limited amount. In addition, root mean square error (RMSE) was used for comparing the efficiency of the nonparametric and parametric estimates. For all the scenarios we considered, the RMSE of the parametric estimates was consistently smaller than that of the nonparametric method. This indicates that when the data were generated from normal distributions, linear mixed models produced more efficient estimates than the nonparametric method. Similar simulation results were obtained when T was exponentially distributed (results not shown) but the coverage probabilities under all scenarios were smaller than when T was normally distributed.

Table 4.1: Nonparametric Simulation results: estimation of CIEA with (K,L)=(1,2), (2,3) and (3,3) when T is normal and variances are equal (e=g=1.5)

Sample size	**k**	**l**	**c**	**True**	**Bias**	SE^1	SE^2	$RMSE^3$	CP^4
50	1	2	0.0	1.000	0.04	0.300	0.262	0.264	0.884
100	1	2	0.0	1.000	0.015	0.205	0.195	0.195	0.924
200	1	2	0.0	1.000	0.006	0.146	0.141	0.141	0.931
50	2	3	0.0	1.000	0.008	0.141	0.125	0.125	0.894
100	2	3	0.0	1.000	0.008	0.100	0.093	0.094	0.906
200	2	3	0.0	1.000	0.002	0.072	0.068	0.068	0.928
50	3	3	0.0	1.000	0.000	0.099	0.092	0.092	0.905
100	3	3	0.0	1.000	0.004	0.068	0.068	0.068	0.934
200	3	3	0.0	1.000	0.004	0.051	0.05	0.05	0.936
50	1	2	3.8	0.976	0.038	0.293	0.255	0.258	0.879
100	1	2	3.8	0.976	0.014	0.2	0.19	0.191	0.926
200	1	2	3.8	0.976	0.005	0.143	0.138	0.138	0.931
50	2	3	3.8	0.976	0.008	0.139	0.124	0.124	0.896
100	2	3	3.8	0.976	0.007	0.099	0.092	0.093	0.909
200	2	3	3.8	0.976	0.002	0.071	0.067	0.067	0.931
50	3	3	3.8	0.976	0.000	0.099	0.092	0.092	0.899
100	3	3	3.8	0.976	0.004	0.069	0.068	0.068	0.93
200	3	3	3.8	0.976	0.004	0.051	0.05	0.05	0.932
50	1	2	16.3	0.686	0.017	0.204	0.185	0.186	0.904
100	1	2	16.3	0.686	0.005	0.142	0.136	0.136	0.922
200	1	2	16.3	0.686	0.000	0.101	0.098	0.098	0.93
50	2	3	16.3	0.686	0.003	0.111	0.104	0.104	0.914
100	2	3	16.3	0.686	0.003	0.079	0.075	0.076	0.932
200	2	3	16.3	0.686	0.000	0.057	0.054	0.054	0.925
50	3	3	16.3	0.686	-0.002	0.088	0.084	0.084	0.922
100	3	3	16.3	0.686	0.001	0.062	0.062	0.062	0.934
200	3	3	16.3	0.686	0.001	0.046	0.044	0.044	0.933
50	1	2	28.1	0.424	0.006	0.126	0.117	0.117	0.899
100	1	2	28.1	0.424	0.001	0.089	0.086	0.086	0.921
200	1	2	28.1	0.424	-0.001	0.063	0.061	0.061	0.931
50	2	3	28.1	0.424	0.001	0.076	0.073	0.073	0.91
100	2	3	28.1	0.424	0.001	0.055	0.053	0.053	0.923
200	2	3	28.1	0.424	-0.001	0.039	0.038	0.038	0.925
50	3	3	28.1	0.424	-0.001	0.064	0.063	0.063	0.93
100	3	3	28.1	0.424	0.000	0.046	0.046	0.046	0.928
200	3	3	28.1	0.424	0.000	0.034	0.033	0.033	0.936

[1]Standard errors based on simulations of CIE

[2]Mean of estimated standard errors calculated from SE estimator

[3]Root mean squared error

[4]coverage probability of 95% confidence interval

Table 4.2: Nonparametric Simulation results: estimation of CIEA when (K, L)=(1,2), (2,3) and (3,3) when T is normal and variances are unequal (e=1.5, g=1)

Sample size	**k**	**l**	**c**	**TRUE**	**Bias**	SE^1	SE^2	$RMSE^3$	CP^4
50	1	2	3.8	0.951	0.038	0.290	0.252	0.255	0.878
100	1	2	3.8	0.951	0.014	0.198	0.188	0.188	0.926
200	1	2	3.8	0.951	0.005	0.142	0.136	0.136	0.931
50	2	3	3.8	0.963	0.008	0.138	0.123	0.124	0.894
100	2	3	3.8	0.963	0.007	0.098	0.092	0.092	0.909
200	2	3	3.8	0.963	0.002	0.070	0.067	0.067	0.934
50	3	3	3.8	0.975	0.000	0.100	0.092	0.092	0.9
100	3	3	3.8	0.975	0.004	0.069	0.068	0.068	0.928
200	3	3	3.8	0.975	0.004	0.051	0.05	0.05	0.933
50	1	2	16.3	0.663	0.017	0.201	0.181	0.182	0.901
100	1	2	16.3	0.663	0.005	0.140	0.134	0.134	0.921
200	1	2	16.3	0.663	0.000	0.099	0.096	0.096	0.931
50	2	3	16.3	0.672	0.003	0.110	0.102	0.103	0.914
100	2	3	16.3	0.672	0.003	0.078	0.075	0.075	0.93
200	2	3	16.3	0.672	-0.001	0.056	0.054	0.054	0.924
50	3	3	16.3	0.681	-0.002	0.088	0.084	0.084	0.921
100	3	3	16.3	0.681	0.001	0.062	0.062	0.062	0.934
200	3	3	16.3	0.681	0.001	0.046	0.044	0.044	0.933
50	1	2	28.1	0.407	0.006	0.123	0.114	0.114	0.901
100	1	2	28.1	0.407	0.001	0.087	0.083	0.083	0.921
200	1	2	28.1	0.407	-0.001	0.061	0.06	0.06	0.931
50	2	3	28.1	0.412	0.001	0.075	0.072	0.072	0.905
100	2	3	28.1	0.412	0.001	0.054	0.052	0.052	0.923
200	2	3	28.1	0.412	-0.001	0.039	0.037	0.037	0.923
50	3	3	28.1	0.418	-0.001	0.064	0.063	0.063	0.931
100	3	3	28.1	0.418	0.000	0.046	0.045	0.045	0.928
200	3	3	28.1	0.418	0.000	0.034	0.033	0.033	0.934

[1]Standard errors based on simulations of CIE

[2]Mean of estimated standard errors calculated from SE estimator

[3]Root mean squared error

[4]coverage probability of 95% confidence interval

Table 4.3: Parametric Simulation results: estimation of CIEA when (K, L) = (1, 2), (2, 3) and (3, 3) when T is normal and variances are equal (e=g=1.5)

Sample size	**K**	**L**	**c**	**TRUE**	**Bias**	SE^1	SE^2	$RMSE^3$	CP^4
100	1	2	0	1.000	-0.08	0.101	0.093	0.122	0.946
100	2	3	0	1.000	-0.037	0.054	0.047	0.06	0.964
100	3	3	0	1.000	-0.028	0.040	0.038	0.048	0.977
100	1	2	3.8	0.975	-0.078	0.102	0.095	0.122	0.936
100	2	3	3.8	0.975	-0.036	0.056	0.050	0.062	0.953
100	3	3	3.8	0.975	-0.028	0.042	0.041	0.05	0.964
100	1	2	16.3	0.686	-0.050	0.088	0.086	0.100	0.899
100	2	3	16.3	0.686	-0.022	0.059	0.057	0.061	0.910
100	3	3	16.3	0.686	-0.018	0.051	0.050	0.053	0.933
100	1	2	28.1	0.424	-0.029	0.062	0.061	0.067	0.889
100	2	3	28.1	0.424	-0.012	0.046	0.044	0.046	0.898
100	3	3	28.1	0.424	-0.010	0.041	0.040	0.041	0.918

[1]Standard errors based on simulations of CIE

[2]Mean of estimated standard errors calculated from bootstrap SE estimator

[3]Root mean squared error

[4]coverage probability of 95% confidence interval

Table 4.4: Parametric Simulation results: estimation of CIEA when (K, L) = (1, 2), (2, 3) and (3, 3) when T is normal and variances are unequal (e=1.5, g=1)

Sample size	**K**	**L**	**c**	**TRUE**	**Bias**	SE^1	SE^2	$RMSE^3$	CP^4
100	1	2	3.8	0.951	-0.006	0.206	0.171	0.171	0.920
100	2	3	3.8	0.963	-0.031	0.061	0.068	0.075	0.940
100	3	3	3.8	0.975	-0.036	0.041	0.046	0.058	0.960
100	1	2	16.3	0.663	-0.005	0.135	0.118	0.118	0.940
100	2	3	16.3	0.672	-0.022	0.064	0.061	0.065	0.900
100	3	3	16.3	0.681	-0.026	0.051	0.051	0.058	0.880
100	1	2	28.1	0.407	-0.004	0.083	0.074	0.074	0.940
100	2	3	28.1	0.412	-0.010	0.050	0.046	0.047	0.900
100	3	3	28.1	0.418	-0.013	0.044	0.040	0.042	0.900

[1]Standard errors based on simulations of CIE

[2]Mean of estimated standard errors calculated from bootstrap SE estimator

[3]Root mean squared error

[4]coverage probability of 95% confidence interval

4.5 Carotid Stenosis Example

The carotid stenosis example, introduced in Section 1.2.2, comparing two MRA techniques, two-dimensional (MRA-2D) and three-dimensional (MRA-3D) MRA time of flight, to the IA, which was considered as the "gold standard". We treated three readings from three raters as the replication and 55 patients were recruited.

The estimates, standard errors and 95% confidence intervals of the CIEA by nonparametric and parametric estimation methods are presented in Tables 4.5 for all the pairwise comparisons of the three screening methods. For nonparametric estimation, the estimated CIEA of MRA-2D and the gold standard IA method was 0.592 with the 95% CI (0.348, 0.835), which indicated moderate agreement. Similarly, a moderate agreement was obtained when compare MRA-3D to IA with CIEA=0.452 (95%=(0.242, 0.661)). Even the comparison between MRA-2D and MRA-3D with CIEA=0.881 and 95% CI (0.688, 1.000) failed to show good agreement since the 95% lower limit is below 0.8. The parametric approach produced compatible estimations of CIEA.

Table 4.5: Estimation of Agreement in the Carotid Stenosis Study using CIEA

Methods Compared	Estimate	SE of CIEA	95% CI for CIEA
Nonparametric Estimation			
(IA, MRA-2D)	0.592	0.124	(0.348,0.835)
(IA, MRA-3D)	0.452	0.107	(0.242,0.661)
(MRA-2D, MRA-3D)	0.881	0.099	(0.688,1.000)
Parametric Estimation			
(IA, MRA-2D)	0.585	0.133	(0.373,0.874)
(IA, MRA-3D)	0.447	0.100	(0.290,0.671)
(MRA-2D, MRA-3D)	0.875	0.091	(0.634,0.998)

4.6 Discussion

In this chapter we introduced the coefficient of individual equivalence (CIE) for replicated quantitative measurements. CIE compares the observed disagreement to its expected value under individual equivalence, i.e., when the probability density function of the readings of both observers on the same subject are identical. Both nonparametric and parametric estimators were introduced and validated through a simulation study and applied in the carotid stenosis example. We concluded that a nonparametric approach always gives us robust estimation. The parametric method also works well and gives us smaller RMSE when the true values are normally distributed (i.e. when the parametric assumptions are valid).

As a scaled index, CIE has the advantage of judging the degree of agreement based on standardized value. For other scaled agreement coefficients, such as ICCs and CCCs, the values of the coefficients are not comparable across different populations, and sometimes artificially high or low agreement values may be obtained due to the dependence of those indices on the population heterogeneity. Both CIA and CIE are fundamentally different from ICCs and CCCs because CIA uses the within-subject, rather than between-subject, variability as the scaling factor. CIE, on the other hand, introduces the idea of permutations to estimate the disagreement under individual equivalence while sharing the same denominator as CIA for the observed disagreement.

The CIA with MSD as the disagreement functions compares the observed disagreement to the expected when $E(X_i|T) = E(Y_i|T)$ for every subject i Haber and Barnhart (2008). The CIE compares the observed disagreement to the expected when X and Y have the same conditional distribution (given T) for every subject, which is a stronger requirement. When we compare the definition of CIA to the expression for CIE involving the between and within observer disagreement (1), we see that the approaches differ by the weights they assign to $G(X, X')$ and $G(Y, Y')$. Basically, in

CIA equal weights were assigned to $G(X, X')$ and $G(Y, Y')$, while in CIE, where we applied the permutation-based method, C_2^K and C_2^L were used as weights for $G(X, X')$ and $G(Y, Y')$ respectively. The comparison between CIEA and CIA and the relation among CIEA, CIA and CCC will be discussed in Chapter 5.

In general we do not expect that either $G(X, X')$ or $G(Y, Y')$ to exceed G(X,Y) and thus CIE or CIEA is generally less or equal to 1. However, in practice it is possible for the estimated value of $G(X, X')$ or $G(Y, Y')$ (or both) to exceed the estimated value of G(X,Y). Therefore, in practice it is possible to have estimated CIE or CIEA greater than 1. In this case, we will set the estimates of CIE and CIEA to one.

Furthermore, in this Chapter we used the conditional version of CIEA given K and L. The dependence of CIEA on K and L results from the permutation-based approach. Figure 4.1, 4.2 and 4.3 show the relations among CIEA, K and L under good, moderate and poor agreement, respectively, with the setups in Section 4.4 with unequal variances. Furthermore, we observed the dependency of CIEA on the number of replications seems to be a function of the ratio K/L. CIEA does not depend on K, L when the conditional variances of both observers are the same (Figure 4.4 (a)). In general, there is a difference of about 0.02 in CIEA when the ratio of K/L increases from 0.1 to 1.0 when we have different variances. As shown in Figure 4.4 (b) (L=1000), for all the good, moderate and poor agreement scenarios, CIEA stays approximately constant across different values of K/L.

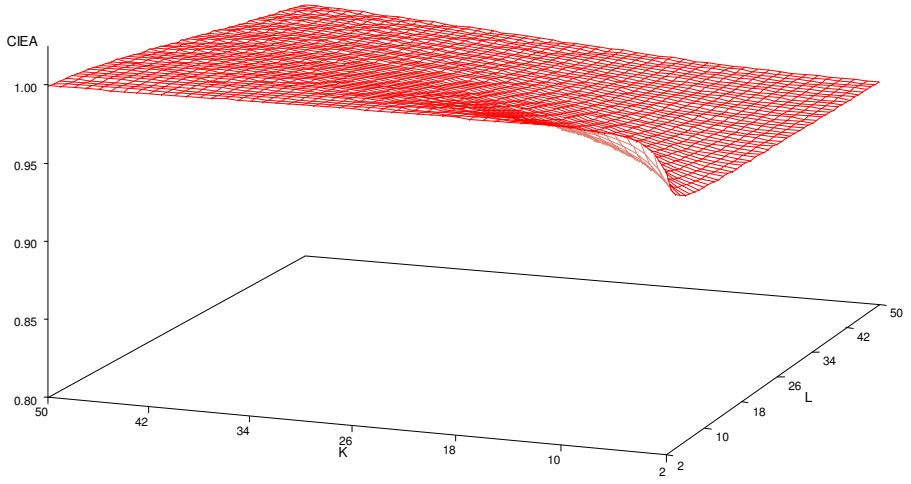

Figure 4.1: Surface Plot of CIEA, K and L for Continuous Measurements with Unequal Variances under Good Agreement

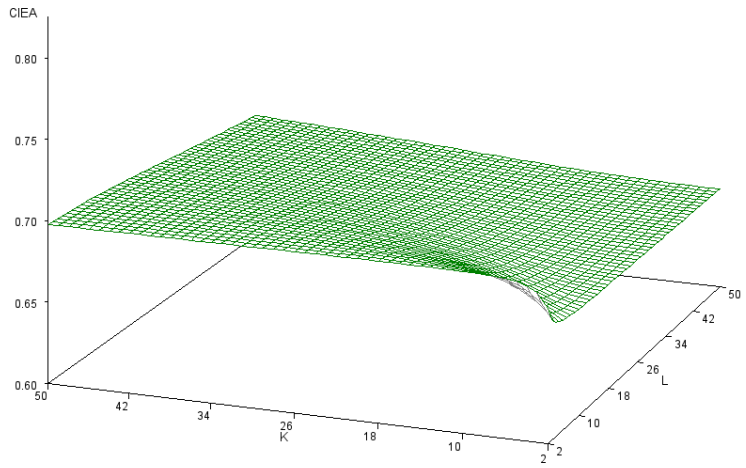

Figure 4.2: Surface Plot of CIEA, K and L for Continuous Measurements with Unequal Variances under Moderate Agreement

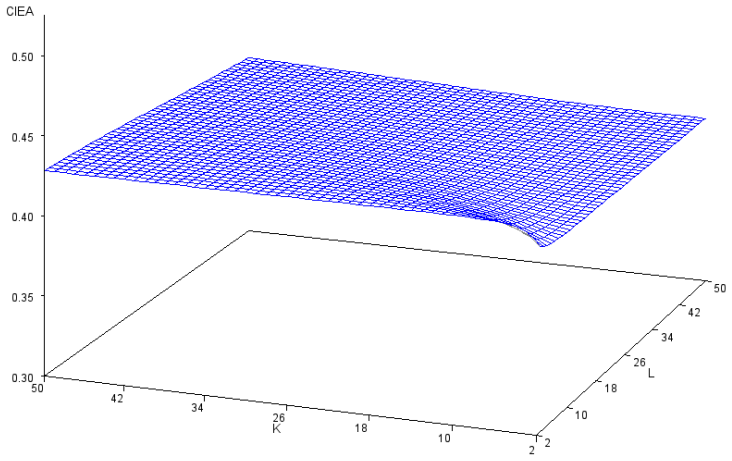

Figure 4.3: Surface Plot of CIEA, K and L for Continuous Measurements with Unequal Variances under Poor Agreement

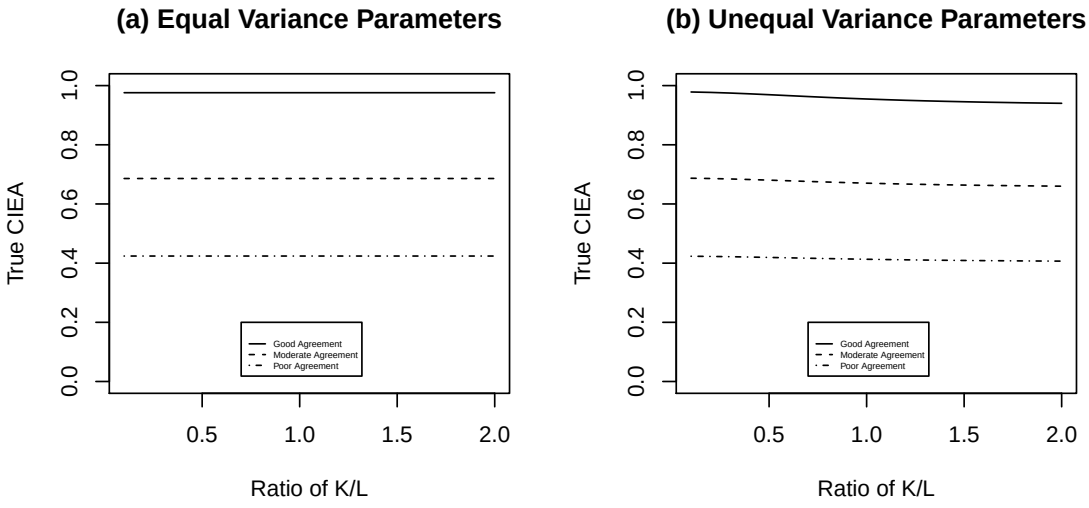

Figure 4.4: Dependence of True CIEA on the ratio of K/L when L=1000

Sample size calculations are essential in agreement studies because it is important to determine the number of subjects and the number of replications needed in order to achieve a desired precision when estimating CIEA. Gao (2010) obtained the standard error (SE) of CIA in terms of N, K and L. One can get a similar expression for the SE of CIEA, so that the corresponding sample size of interest can be calculated. In addition, the definition and nonparametric estimation of CIE can be easily extended when the number of replications on each subject by the same observer is not fixed.

Chapter 5

Comparison of CIE/CIEA to CIA, Kappa and CCC

5.1 Introduction

As stated in Chapter 1, Kappa and weighted Kappa are the most popular agreement coefficients for binary and categorical measurements. When we have continuous outcomes, the concordance correlation coefficients (CCC) is often applied. Besides those coefficients, the coefficient of individual agreement (CIA) can also be used for both categorical and continuous measurements when replications are available. In this chapter, the comparison between CIEA and CIA, Kappa and CCC are discussed.

5.2 CIEA vs. CIA

5.2.1 Equality of $\widehat{CIEA}$ and $\widehat{CIA}$ when K=L

5.2.1.1 Binary Measurements

When we have binary outcomes, the equality of $\widehat{CIEA}$ and $\widehat{CIA}$ when $K = L$ can be shown as follows. Here we consider CIA=ψ^N.

$$\widehat{CIEA} = \frac{\widehat{CIE} - CIE_{\min}}{1 - CIE_{min}} = \frac{\frac{\hat{G}^E}{\hat{G}(X,Y)} - CIE_{min}}{1 - CIE_{min}} = \frac{\hat{G}^E - \hat{G}(X,Y)CIE_{\min}}{(1 - CIE_{\min})} \cdot \frac{1}{\hat{G}(X,Y)}$$

Since $K = L$, $CIE_{\min} = \frac{K}{2K-1}$, which is a constant. By the estimation of CIA:

$$\widehat{CIA} = \frac{[\hat{G}(X,X') + \hat{G}(Y,Y')]}{2\hat{G}(X,Y)}.$$

Therefore we need to show:

$$\frac{\hat{G}^E - \hat{G}(X,Y)CIE_{\min}}{(1 - CIE_{\min})} \equiv \frac{(\hat{G}(X,X') + \hat{G}(Y,Y'))}{2} \tag{5.1}$$

Since all the estimates $\hat{G}$ in (5.1) are means over all study subjects, it is sufficient to show that the contributions of every subject to the two sides of (5.1) are euqal, i.e. for every i:

$$\frac{\hat{G}_i^E - \hat{G}_i(X,Y)CIE_{\min}}{(1 - CIE_{\min})} \equiv \frac{(\hat{G}_i(X,X') + \hat{G}_i(Y,Y'))}{2} \tag{5.2}$$

For subject i, we first look at the r.h.s. of (5.1). Assuming $K = L$,

$$\frac{(\hat{G}_i(X,X') + \hat{G}_i(Y,Y'))}{2} = \frac{K\hat{\pi}_i(1-\hat{\pi}_i)}{(K-1)} + \frac{K\hat{\lambda}_i(1-\hat{\lambda}_i)}{(K-1)} = \frac{K[\hat{\pi}_i(1-\hat{\pi}_i) + \hat{\lambda}_i(1-\hat{\lambda}_i)]}{K-1}.$$

Regarding l.h.s. of (5.1), when $K = L$:

$$\hat{G}_i^E = \frac{2KW_i - W_i^2}{K(2K-1)}.$$

We know that T_i is the number of "1"s for observer X and U_i is the number of "1"s for observer Y. $\hat{\pi}_i = \frac{T_i}{K_i}$ and $\hat{\lambda}_i = \frac{U_i}{L_i}$. Since $K_i = L_i = K$ and $W_i = T_i + U_i$, then $W_i = K(\hat{\pi}_i + \hat{\lambda}_i)$.

Therefore,

$$\hat{G}_i^E = \frac{2K(\hat{\pi}_i + \hat{\lambda}_i) - K(\hat{\pi}_i + \hat{\lambda}_i)^2}{2K-1}.$$

$$\frac{\hat{G}_i^E - \hat{G}_i(X,Y)CIE_{\min}}{1 - CIE_{\min}} = \frac{2K(\hat{\pi}_i + \hat{\lambda}_i) - K(\hat{\pi}_i + \hat{\lambda}_i)^2}{K-1} - (\hat{\pi}_i - \hat{\lambda}_i + 2\hat{\pi}_i\hat{\lambda}_i)K$$
$$= \frac{K[\hat{\pi}_i(1-\hat{\pi}_i) + \hat{\lambda}_i(1-\hat{\lambda}_i)]}{K-1}$$

Hence we established the equality (5.1) for every subject, which is sufficient for the equality of $\widehat{CIEA}$ and $\widehat{CIA}$.

5.2.1.2 Continuous Measurements: Nonparametric Estimation

When $K = L$, the mean over i of $\hat{G}^E$ is $[C_2^K(G(X,X')+G(Y,Y'))+K^2G(X,Y)]/C_2^{2K}$. Compare this to the numerator of ψ^N: $[G(X,X') + G(Y,Y')]/2 = W$, the numerator of CIE is $[2C_2^K W + K^2G(X,Y)]/C_2^{2K}$. Divided by $G(X,Y)$ to get

$$\eta = CIE = [K(K-1)\psi^N + K^2]/C_2^{2K}.$$

Note when $K = L$, $CIE_{\min} = K/(2K-1)$. Then it's straightforward to show that CIEA=CIA. Similarly, we establish the equality between $\widehat{CIEA}$ and $\widehat{CIA}$.

5.2.1.3 Continuous Measurements: Parametric Estimation using Full Model Only

To estimate CIAs using MSD as the disagreement function, either ψ_N or ψ_R, we can use a two-way effects model. Applying the same two-way effects ANOVA model in Chapter 4, when there is no reference,

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}.$$

we assume those random effects $\alpha_i \sim N(0, \sigma_\alpha^2)$, $\gamma_{ij} \sim N(0, \sigma_\gamma^2)$ and $\varepsilon_{ijk} \sim N(0, \sigma_{\varepsilon_j}^2)$. Also we assume that α_i, γ_{ij} and ε_{ijk} are independent. The overall error variance is defined as $\sigma_\varepsilon^2 =$ mean of $\sigma_{\varepsilon_1}^2, \cdots, \sigma_{\varepsilon_J}^2$. When observer is treated as the fixed effect, then σ_β^2 is a sum of squares, $\sigma_\beta^2 = \frac{1}{J-1}\sum_{j=1}^{J}(\beta_j - \bar{\beta})^2$. Therefore,

$$\psi_N = \frac{\sigma_{\varepsilon_1}^2 + \sigma_{\varepsilon_2}^2}{2\sigma_\beta^2 + 2\sigma_\gamma^2 + \sigma_{\varepsilon_1}^2 + \sigma_{\varepsilon_2}^2}.$$

$\hat{\sigma}_\beta^2$, $\hat{\sigma}_\gamma^2$, $\hat{\sigma}_{\varepsilon_1}^2$ and $\hat{\sigma}_{\varepsilon_2}^2$ are used to estimate ψ_N. The SE estimator of $\hat{\psi}_N$ was obtained with delta method, by establishing the variances and covariances matrix of random effects first (Wiener 2009).

Recall under the same model set up, we estimate CIEA using estimations from the full model,

$$CIE = \frac{C_2^K \sigma_{\varepsilon_1}^2 + C_2^L \sigma_{\varepsilon_2}^2}{C_2^{K+L} \cdot (\sigma_\beta^2 + \sigma_\gamma^2 + \sigma_{\varepsilon_1}^2 + \sigma_{\varepsilon_2}^2)} + \frac{K \cdot L}{C_2^{K+L}}.$$

when $K = L$, $CIE = [K(K-1)CIA + K^2]/C_2^{2K}$. Then CIE can be obtained through

CIA. It's not hard to show that under this case, $CIEA$ is identical to CIA.

$$\begin{aligned} CIEA &= \frac{CIE - CIE_{\min}}{1 - CIE_{\min}} \\ &= \frac{[K(K-1)CIA + K^2]/C_2^{2K} - CIE_{\min}}{1 - CIE_{\min}} \\ &= \frac{K(K-1)CIA + K^2]/(K(2K-1)) - K/(2K-1)}{1 - K/(2K-1)} \\ &= CIA. \end{aligned}$$

The equality between $\widehat{CIEA}$ and $\widehat{CIA}$ follows the same way.

We found that CIEA=CIA when K=L. However, unlike CIA, CIEA can be used when K=1 or L=1, i.e., the scenario when we do not have replicated measurements on a gold standard that is measured without error.

5.2.2 Comparison between $\widehat{CIEA}$ and $\widehat{CIA}$ when $K \neq L$

After establishing the equality between $\widehat{CIEA}$ and $\widehat{CIA}$ under the assumption $K = L$, it's also interesting to compare $\widehat{CIEA}$ to $\widehat{CIA}$ when $K \neq L$, especially when sample sizes are small. CIEA is based on a permutation method which is a nonparametric approach and could result in better estimation for small sample size. Tables 5.1 and 5.2 present the simulation results when K=4 and L=2 for good, moderate and poor agreement with sample sizes 10, 20 and 50 for binary outcomes, following the same simulation set up in Chapter 3. And Tables 5.3 and 5.4 showed the same comparison when we have continuous outcomes with a normally distributed true value. Generally, when $K \neq L$, CIA experienced smaller bias than CIEA while CIEA archived better coverage probabilities under many scenarios. Furthermore, CIA (ψ^N) is not estimable when there is no replication by either observer and ψ^R is not estimable when there is no replicated data for the gold standard, whereas CIEA is always estimable when at least one of K, L is greater than 1.

Table 5.1: Simulation Results: Estimation of CIEA when (K,L)=(4,2) for Binary Outcomes

Sample Size	K	L	μ_B	Bias	SE^a	SE^b	CP
10	4	2	1	0.041	0.280	0.214	0.767
10	4	2	3	0.046	0.276	0.219	0.819
10	4	2	5	0.031	0.241	0.207	0.853
20	4	2	1	0.019	0.192	0.171	0.878
20	4	2	3	0.023	0.187	0.170	0.890
20	4	2	5	0.019	0.168	0.153	0.884
50	4	2	1	0.007	0.115	0.115	0.923
50	4	2	3	0.012	0.111	0.111	0.934
50	4	2	5	0.010	0.096	0.099	0.946

[a]Standard errors based on simulations of CIEA

[b]Mean of estimated standard errors calculated from SE estimator

Table 5.2: Simulation Results: Estimation of CIA when (K,L)=(4,2) for Binary Outcomes

Sample Size	K	L	μ_B	Bias	SE^a	SE^b	CP
10	4	2	1	0.034	0.265	0.196	0.744
10	4	2	3	0.042	0.264	0.205	0.759
10	4	2	5	0.025	0.230	0.191	0.810
20	4	2	1	0.019	0.187	0.166	0.837
20	4	2	3	0.027	0.184	0.168	0.873
20	4	2	5	0.021	0.164	0.147	0.880
50	4	2	1	0.006	0.116	0.115	0.906
50	4	2	3	0.010	0.117	0.112	0.911
50	4	2	5	0.009	0.094	0.097	0.944

[a]Standard errors based on simulations of CIA

[b]Mean of estimated standard errors calculated from SE estimator

Table 5.3: Simulation Results: Estimation of CIEA when (K,L)=(2,3) for Continuous Outcomes

Sample Size	K	L	c	Bias	SE^a	SE^b	CP
10	2	3	3.8	0.038	0.253	0.197	0.828
10	2	3	16.3	0.022	0.209	0.173	0.847
10	2	3	28.1	0.008	0.139	0.121	0.860
20	2	3	3.8	0.024	0.183	0.159	0.890
20	2	3	16.3	0.014	0.149	0.133	0.900
20	2	3	28.1	0.006	0.100	0.091	0.910
50	2	3	3.8	0.008	0.138	0.123	0.894
50	2	3	16.3	0.003	0.110	0.102	0.914
50	2	3	28.1	0.001	0.075	0.072	0.905

[a]Standard errors based on simulations of CIEA

[b]Mean of estimated standard errors calculated from SE estimator

Table 5.4: Simulation Results: Estimation of CIA when (K,L)=(2,3) for Continuous Outcomes

Sample Size	K	L	c	Bias	SE^a	SE^b	CP
10	2	3	3.8	0.031	0.235	0.181	0.808
10	2	3	16.3	0.020	0.205	0.171	0.837
10	2	3	28.1	0.008	0.143	0.125	0.859
20	2	3	3.8	0.018	0.174	0.150	0.868
20	2	3	16.3	0.011	0.148	0.133	0.881
20	2	3	28.1	0.005	0.102	0.094	0.891
50	2	3	3.8	0.000	0.100	0.092	0.900
50	2	3	16.3	-0.002	0.088	0.084	0.921
50	2	3	28.1	-0.001	0.064	0.063	0.931

[a]Standard errors based on simulations of CIA

[b]Mean of estimated standard errors calculated from SE estimator

5.3 CIEA vs. Kappa for Binary Observations

For this comparison, a latent class model for dialogistic agreement is introduced and comparisons between CIEA and Cohen's kappa are considered in the context of this

model. Let 1 indicate presence of illness while 0 indicates absence of illness. Let X and Y be two observers or disgnostic tests and T be the true binary illness status:

$$\omega = P(T = 1), \;\; \pi_t = P(X = 1|T = t), \;\; \lambda_t = P(Y = 1|T = t), \;\; t = 0, 1$$

where ω is the disease prevalence, π_1, λ_1 are the sensitivities of X and Y respectively, and π_0, λ_0 are the complement of their specificities. Under this latent model, introduced by Dawid and Skene (1979), the disagreement functions were derived as in Haber et al. (2007):

$$\begin{aligned} G(X, Y) &= \omega(\pi_1 + \lambda_1 - 2\pi_1\lambda_1) + (1 - \omega)(\pi_0 + \lambda_0 - 2\pi_0\lambda_0) \\ G(X, X') &= 2\omega\pi_1(1 - \pi_1) + 2(1 - \omega)\pi_0(1 - \pi_0) \\ G(Y, Y') &= 2\omega\lambda_1(1 - \lambda_1) + 2(1 - \omega)\lambda_0(1 - \lambda_0) \end{aligned}$$

Thus, CIEA can be obtained from (3.5) and (3.6) when plugging in the above G functions. On the other hand, by assuming the errors of the two bianry classifications to be independent, kappa can be written as follows:

$$\kappa = \frac{2\omega(1 - \omega)(\pi_1 - \pi_0)(\lambda_1 - \lambda_0)}{\pi^*(1 - \lambda^*) + \pi^*(1 - \lambda^*)},$$

where $\pi^* = \omega\pi_1 + (1 - \omega)\pi_0$ and $\lambda^* = \omega\lambda_1 + (1 - \omega)\lambda_0$ (Kraemer (1979), Thompson and Walter (1988) and Shoukri (2004)). For example, assume $K = L = 3$, $\pi_1 = 0.95$, π_0=0.15, $\lambda_1 = 0.9$ and $\lambda_0 = 0.1$. In other words, the sensitivities and specificities of both observers are high and close to each other. Figure 5.1 shows the CIEA and κ when the prevalence, ω, varies from 0 to 1. We see that CIEA is close to 1 and remains constant across $0 < \omega < 1$ while κ varies dramatically with the changes in ω. This is not surprising since we know that κ heavily depends on the prevalence (Kraemer (1979), Thompson and Walter (1988)).

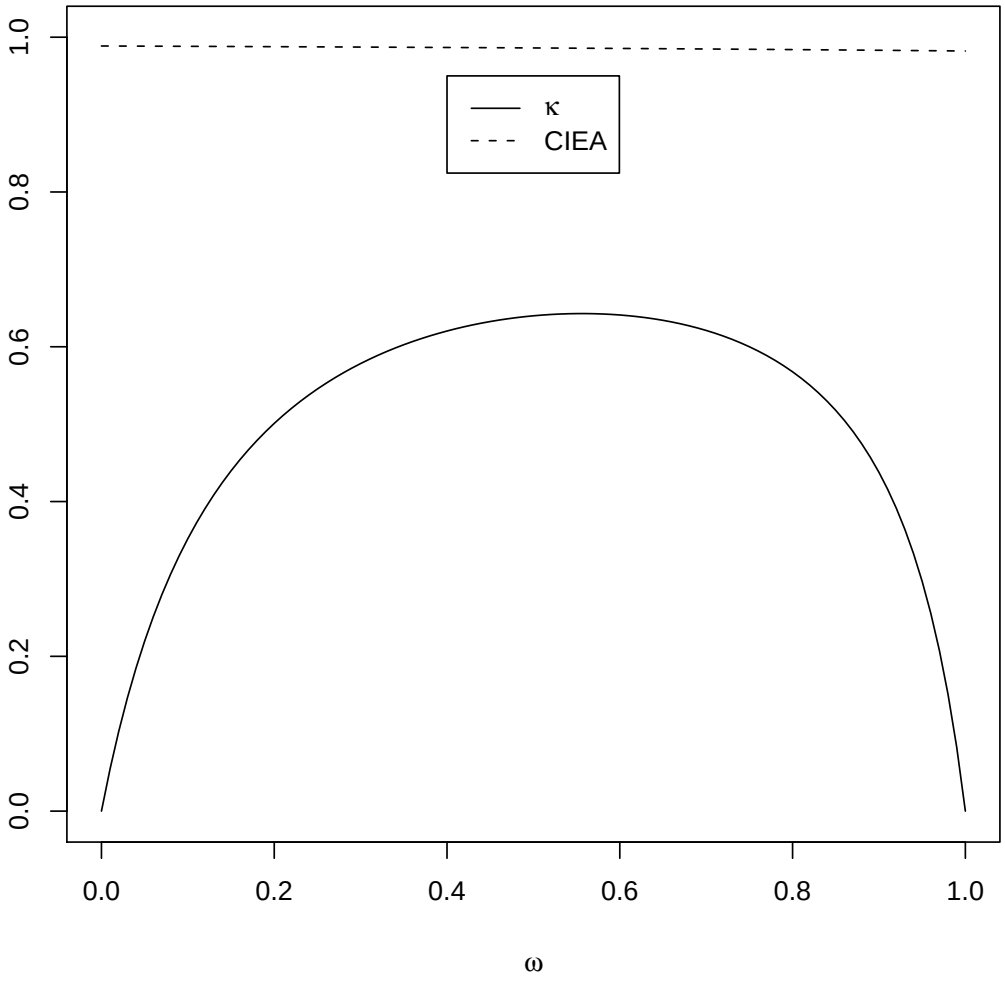

Figure 5.1: CIEA and κ as functions of prevalence ω.

5.4 CIEA, CIA vs. CCC for Quantitative Data

The concordance correlation coefficient (CCC) is commonly used for assessing agreement for continuous outcomes. It was first published by Lin Lin (1989) for the simplest case where there are two raters and each make one reading per subject. Lin's CCC is defined as follows: assume that the observations (X, Y) are from a bivariate distribution with mean vector (μ_x, μ_y) and variance-covariance matrix $\begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$. Lin's CCC is defined as

$$\begin{aligned} \text{CCC}_{\text{Lin}} &= 1 - \frac{E(X-Y)^2}{E[(X-Y)^2|\rho=0]} \\ &= \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2+(\mu_x-\mu_y)^2}, \end{aligned}$$

where ρ is the Pearson correlation coefficient between two observers.

Following the introduction of the adjusted coefficient of individual equivalence (CIEA) for quantitative measurements, it is of interest to compare this coefficient to CCC, which shares the same denominator as CIE with G=MSD. Since we have replicated observations from each observer, we compare the CIEA to the total CCC Barnhart et al. (2005) defined as

$$CCC_{total} = 1 - \frac{E(X_{ik}-Y_{il})^2}{E_I(X_{ik}-Y_{il})^2},$$

where E_I is the expectation given independence of X, Y. Barnhart et al. (2007c) compared the total CCC and CIA when the between-subject variability is increased and found that the CCC is inflated when the between-subject variability is large. We will follow the same strategy to compare the CIEA with the total CCC.

Let's use the simple latent class model introduced in Section 4.4. Obviously, all three MSDs increase with the between-subject variability σ_T^2. We're going to

explore the dependence of total CCC, CIA and CIEA on σ_T^2. In Section 4.4, the distribution of T was defined using the sample moments from the stenosis data, $\mu_T = 43.29$, $\sigma_T = 29.87$. To investigate the dependence of the coefficients on the between-subjects variability, we now keep μ_T fixed at 43.29 but let σ_T vary from 0 to 60. The parameters defining the conditional means and standard deviations of the observers' measurements given T are: $b = d = 1, e = 1.5, g = 1, f = h = 0.3$. We keep $a = 0$ and let $c = 3.8, 16.3, 28.1$ to account for good, moderate and poor agreement. From the previous section, we know that when $K = L$, CIA and CIEA are identical. Therefore, we consider K=L=3 and K=2, L=3 as the two scenarios in our comparison.

In Figure 5.2, the total CCC, CIA and CIEA were plotted while varying σ_T. With K=L=3, we consider good, moderate and poor agreement (cases ((a), (b) and (c) in Figure 5.2, respectively). As we showed earlier when K=L, CIEA and CIA coincide. Therefore we also consider case (d) where K=2, L=3 for moderate agreement. In (a), where we have good agreement, both CIEA and CIA were constant with the increment in σ_T, while total CCC increased rapidly with σ_T. Similarly, when we have moderate and poor agreement (cases (b) and (c)), CIEA and CIA increased at a modest rate with σ_T while total CCC increased very rapidly. Furthermore, when $\sigma_T \geq 20$, total CCC was much higher than CIEA and CIA. When K and L are not equal, similar trend was observed except that CIA was a little bit higher than CIEA (Figure 5.2(d)). Therefore, we believe that CCC is inflated with the between subject variability is large, while CIEA and CIA are more stable.

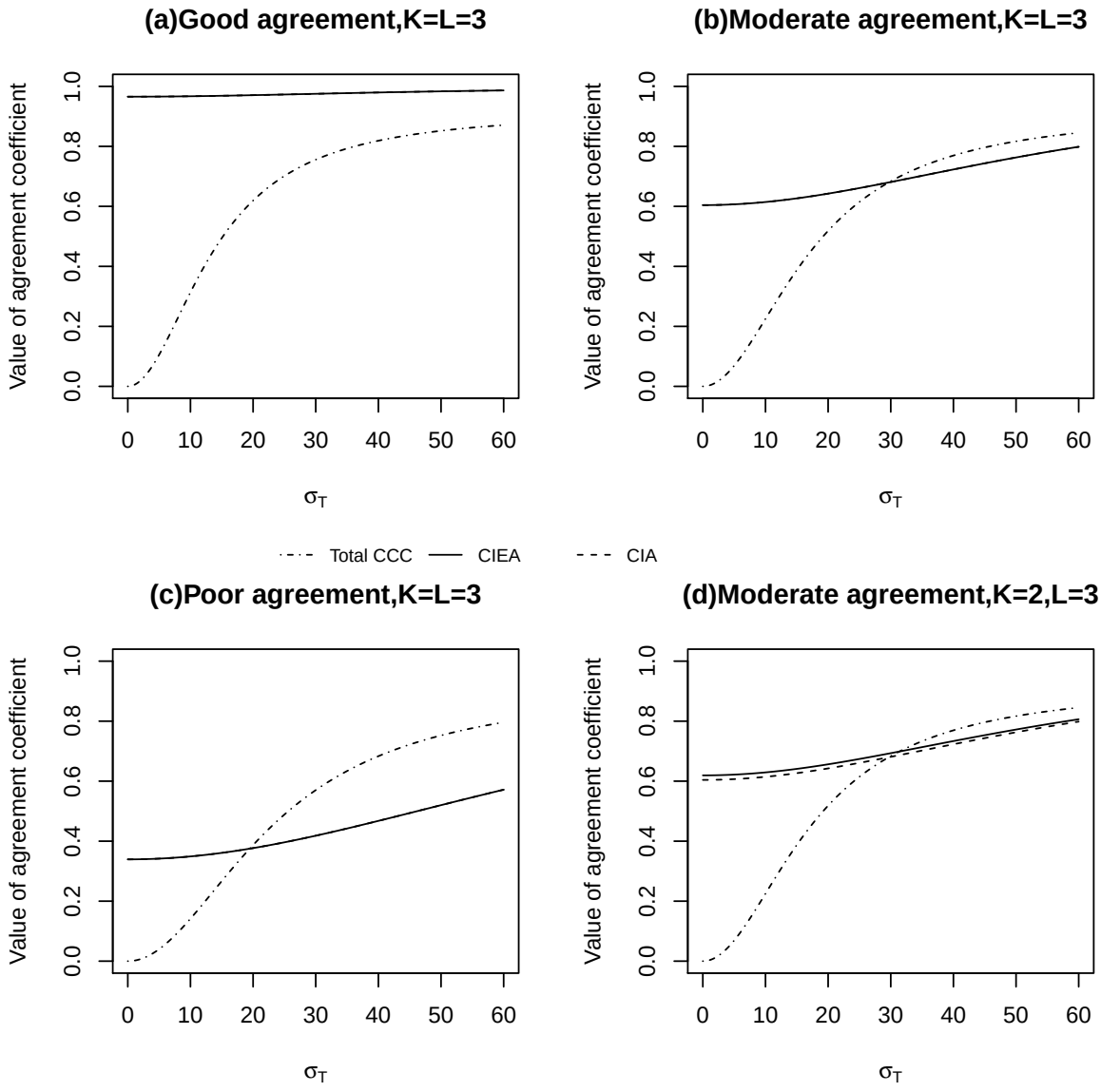

Figure 5.2: Agreement coefficients with varying σ_T

Chapter 6

Model-based Estimation of the Coefficient of Individual Equivalence for Replicated and Repeated Binary Measurements

A nonparametric approach for estimating CIE/CIEA from replicated binary outcomes was presented in Chapter 3 , while in Chapter 4, a parametric method has been proposed for estimating CIE/CIEA with replicated continuous measurements. There are several advantages for using a parametric random effects model to estimate CIE/CIEA. First, it's quite convenient to generate all the estimations from random effects models although usually modeling requires more assumptions. Second, it will be easier to extend this coefficient to the scenario with multiple raters. Last but not the least, in this chapter, we extend the coefficient of individual equivalence (CIE) to repeated binary outcomes. It is then easy to use model-based estimation when we have more complex data structure. In this exploratory chapter, several different modeling approaches will be introduced and investigated to estimate the adjusted

coefficient of individual equivalence (CIEA) for binary data.

6.1 Model-Based Estimation of CIEA for Replicated Binary Outcomes

The identity link is not the canonical link function when we have binary outcomes. However, the identity link has been widely used for binary outcome in agreement studies when the proportion of agreement does not approach to the extreme values, such as 0 or 1. Anderson and Aitkin (1985) discussed using variance components to estimate intraclass correlation for binary response. In estimating the variance components, the categorical or binary nature of the response is usually ignored, and the analysis is carried out using analysis of variance. This is not a maximum likelihood method, but analysis of variance has some desirable properties when the error variables are non-normal. The rule of thumb generally applied is that analysis of variance is reasonably accurate as long as the proportions in each of the response categories are between 0.1 and 0.9.

Dunn (2004) and Ridout et al. (1999) introduced an ANOVA model with identity link to estimate the intraclass correlation coefficient (ICC) with binary or even categorical measurements. Lin et al. (2007) used the GEE method with identity link to extend the concordance correlation coefficient (CCC) to replicated categorical measurements. In the following sections, random effects models with both identity and the canonical logit links are discussed. Variance components linear and generalized linear mixed models are investigated. When the logit link is applied, different methods, including first and second order Taylor expansions, delta method, cumulative Gaussian approximation and Adaptive Gauss-Hermite Quadrature were explored. However, only the cumulative Gaussian approximation and Adaptive Gauss-Hermite Quadrature are included in this dissertation to approximate the marginal likelihood

in order to estimate the variance components estimators. Furthermore, a Bayesian approach is applied to estimate CIEA when we have replicated binary outcomes.

6.1.1 Identity Link

A parametric approach to estimating and making inference on CIE can be based on a mixed two-way linear model with subject effect α_i, observer effect β_j as well as the interaction between subject and observer γ_{ij}. Finally, the residual term is ε_{ijk}, assuming the replication index is k. Since we're focusing on applying the identity link, we define the model as

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \tag{6.1}$$

We assume that the subject effect is random, the observer effect is fixed, and that the random effects $\alpha_i \sim N(0, \sigma_\alpha^2)$, $\gamma_{ij} \sim N(0, \sigma_\gamma^2)$ and $\varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2)$ are independent. Since the observer is treated as a fixed effect, we define σ_β^2 as $\sigma_\beta^2 = \frac{1}{(J-1)} \sum_{j=1}^{J} (\beta_j - \bar{\beta})^2$, $\beta_j = \mu_j - \mu$ is the difference between the mean of observer j with respect to the overall mean, and J is the number of observers (Carrasco 2003).

6.1.1.1 Estimation of CIEA using identity link

The estimation of the CIEA using the identity link for binary outcomes is similar to the case of continuous measurements in Chapter 4. To simplify our notation, we consider two observers X and Y and denote $X = Y_1$ and $Y = Y_2$. It is easy to see

that

$$\begin{aligned} MSD(X, X') &= E(Y_{i1k} - Y_{i1k'})^2 = E(\varepsilon_{i1k} - \varepsilon_{i1k'})^2 = 2\sigma_\varepsilon^2 \\ MSD(Y, Y') &= E(Y_{i2k} - Y_{i2k'})^2 = E(\varepsilon_{i2k} - \varepsilon_{i2k'})^2 = 2\sigma_\varepsilon^2 \\ MSD(X, Y) &= E(Y_{i1k} - Y_{i2k'})^2 = E(\beta_1 - \beta_2)^2 + E(\gamma_{i1} - \gamma_{i2})^2 + E(\varepsilon_{i1k} - \varepsilon_{i2k'})^2 \\ &= 2\sigma_\beta^2 + 2\sigma_\gamma^2 + 2\sigma_\varepsilon^2. \end{aligned}$$

As we have seen, the expression of CIE can be simplified as

$$\frac{C_2^K \cdot G(X, X') + C_2^L \cdot G(Y, Y') + K \cdot L \cdot G(X, Y)}{C_2^{K+L} \cdot G(X, Y)}.$$

When MSD is used as the G function, $MSD(X, X')$, $MSD(Y, Y')$ and $MSD(X, Y)$ can be estimated from the saturated parametric mixed model as shown above.

In general,

$$CIE = \frac{2C_2^K \sigma_\varepsilon^2}{C_2^{K+L} \cdot (\sigma_\beta^2 + \sigma_\gamma^2 + 2\sigma_\varepsilon^2}) + \frac{K \cdot L}{C_2^{K+L}}.$$

As before, $\hat{\sigma}_\beta^2$, $\hat{\sigma}_\gamma^2$, $\hat{\sigma}_\varepsilon^2$ are used to estimate $\widehat{CIE}$.

6.1.1.2 Variance Components Approach with Identity Link

Still under the model 6.1, assume $\mathbf{u}$ is a vector containing all the random effects, including random subject effect and the interaction between subject and observer. Therefore, $\mathbf{u} \sim MVN(0, G)$ where G is a block diagonal matrix with components σ_α^2 and σ_γ^2 in the diagonal and 0 otherwise. Furthermore,

$$E[Y_{ijk}|\mathbf{u}] = \pi_{ijk} = g^{-1}(\mu + \alpha_i + \beta_j + \gamma_{ij}) = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

and $Var[Y_{ijk}|\mathbf{u}] = \phi h(\pi_{ijk})$. ϕ denotes the dispersion parameter and h is the variance function associated with the conditional probability distribution of $Y|\mathbf{u}$. Then we

can derive the following equations:

$$\begin{aligned}
E(Y_{ijk}) &= E(E(Y_{ijk}|\mathbf{u})) = E(\pi_{ij}) = \mu \\
Var(Y_{ijk}) &= Var_u(E(Y_{ijk}|\mathbf{u})) + E_u(Var_u(Y_{ijk}|\mathbf{u})) \\
&= Var_u(\pi_{ij}) + E_u(\phi h(\pi_{ij})) \\
Cov(Y_{ijk}, Y_{ij'k}) &= Cov_u(E(Y_{ijk}|\mathbf{u}), E(Y_{ij'k}|\mathbf{u})) + E_u(Cov(Y_{ijk}, Y_{ij'k}|\mathbf{u})) \\
&= Cov_u(E(Y_{ijk}|\mathbf{u}), E(Y_{ij'k}|\mathbf{u})) = Cov_u(\pi_{ij}, \pi_{ij'}) = \sigma^2_\alpha \\
Cov(Y_{ijk}, Y_{ijk'}) &= Cov_u(E(Y_{ijk}|\mathbf{u}), E(Y_{ijk'}|\mathbf{u})) + E_u(Cov(Y_{ijk}, Y_{ijk'}|\mathbf{u})) \\
&= Cov_u(E(Y_{ijk}|\mathbf{u}), E(Y_{ijk'}|\mathbf{u})) = Cov_u(\pi_{ij}, \pi_{ij}) \\
&= Var(\mu_{ij}) = \sigma^2_\alpha + \sigma^2_\beta + \sigma^2_\gamma
\end{aligned}$$

Therefore, when constructing three MSDs,

$$\begin{aligned}
MSD(Y_{ijk}, Y_{ijk'}) = E(Y_{ijk} - Y_{ijk'})^2 &= 2E(Y_{ijk})^2 - 2[Cov_u(\pi_{ij}, \pi_{ij}) + E(Y_{ijk})E(Y_{ijk'})] \\
&= 2\sigma^2_\alpha + 2\sigma^2_\beta + 2\sigma^2_\gamma + 2\sigma^2_\varepsilon - 2\sigma^2_\alpha - 2\sigma^2_\beta - 2\sigma^2_\gamma \\
&= 2\sigma^2_\varepsilon
\end{aligned}$$

On the other hand,

$$\begin{aligned}
MSD(Y_{ijk}, Y_{ij'k}) &= E(Y_{ijk} - Y_{ij'k})^2 \\
&= E(Y_{ijk})^2 + E(Y_{ij'k})^2 + E(-2[Cov_u(\pi_{ij}, \pi_{ij'}) + E(Y_{ijk})E(Y_{ij'k})] \\
&= 2\sigma^2_\alpha + 2\sigma^2_\beta + 2\sigma^2_\gamma + 2\sigma^2_\varepsilon - 2\sigma^2_\alpha \\
&= 2\sigma^2_\beta + 2\sigma^2_\gamma + 2\sigma^2_\varepsilon
\end{aligned}$$

Therefore, we achieve the same estimation as in Section 6.1.1.1. In the following section, the variance components approach will be applied when we have a logit link.

6.1.2 Logit Link

The above section demonstrated the scenario when the error term is considered normally distributed and the identity link is specified. However, in practice this is problematic when the proportions for any category are approaching the extreme values such as 0 or 1.

Alternatively, we can try to derive the coefficient under the logit link, which is the canonical link for binary outcomes. Under the logit link, we want to follow the same strategy to derive the three MSD functions. We know that $Y_{ijk}|\mathbf{u} \sim Bernoulli(\pi_{ijk})$ and $logit(\pi_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij}$. Therefore,

$$\pi_{ijk} = \frac{1}{1 + \exp(-(\mu + \alpha_i + \beta_j + \gamma_{ij}))},$$

in addition, $E(Y_{ijk}|\mathbf{u}) = \pi_{ijk}$ and $Var(Y_{ijk}|\mathbf{u}) = \pi_{ijk}(1-\pi_{ijk})$, where $\mathbf{u} \sim MVN(0, G)$ and G is a block diagonal matrix with components σ^2_α and σ^2_γ in the diagonal and 0 otherwise. In order to construct three MSD functions, we need to get the analytic forms of $E(Y_{ijk})$, $Var(Y_{ijk})$, $Cov(Y_{ijk}, Y_{ijk'})$ and $Cov(Y_{ijk}, Y_{ij'k})$. How to approximate the marginal moments in order to estimate the three MSD functions and CIE is of interest.

For instance,

$$E(Y_{ijk}) \quad = \quad E(E(Y_{ijk}|\mathbf{u})) = E(\pi_{ijk}) = E(\frac{1}{1 + \exp(-(\mu + \alpha_i + \beta_j + \gamma_{ij}))}) \quad (6.2)$$

and generally,

$$\begin{aligned} Var(\pi_{ijk}) &= E(\pi_{ijk}^2) - (E(\pi_{ijk}))^2 \\ Var(Y_{ijk}) &= Var(\pi_{ijk}) + E((\pi_{ijk})(1-\pi_{ijk})) = E(\pi_{ijk}) - (E(\pi_{ijk}))^2 \\ Cov(Y_{ijk}, Y_{ijk'}) &= Cov(\pi_{ijk}, \pi_{ijk}) = Var(\pi_{ijk}) \\ Cov(Y_{ijk}, Y_{ij'k}) &= Cov(\pi_{ijk}, \pi_{ij,k}) = E(\pi_{ijk} \cdot \pi_{ij'k}) - E(\pi_{ijk})E(\pi_{ij'k}) \end{aligned}$$

Finally the above components can be used to construct the MSD functions as follows:

$$\begin{aligned} MSD(Y_{ijk}, Y_{ijk'}) &= 2E(Y_{ijk})^2 - 2[Cov(Y_{ijk}, Y_{ijk'}) + E(Y_{ijk})E(Y_{ijk'})] \\ MSD(Y_{ij'k}, Y_{ij'k}) &= 2E(Y_{ij'k})^2 - 2[Cov(Y_{ij'k}, Y_{ij'k}) + E(Y_{ij'k})E(Y_{ij'k})] \\ MSD(Y_{ijk}, Y_{ij'k'}) &= E(Y_{ijk})^2 + (E(Y_{ij'k'}))^2 - 2[Cov(Y_{ijk}, Y_{ij'k'}) + E(Y_{ijk})E(Y_{ij'k'}) \\ &= Var(Y_{ijk}) + (E(\pi_{ijk}))^2 + Var(Y_{ij'k'}) + (E(\pi_{ij'k'}))^2 \\ &\quad -2[Cov(\pi_{ijk}, \pi_{ij'k'}) + E(\pi_{ijk})E(\pi_{ij'k'})] \end{aligned}$$

The following methods to estimate the marginal moments of this generalized linear mixed model are under consideration: cumulative Gaussian approximation to logistic function and Adaptive Gauss-Hermite Quadrature with SAS PROC GLIMMIX.

6.1.2.1 Cumulative Gaussian Approximation

Let's redefine our problem under a generalized linear mixed model instead of ANOVA.

$$logit(\pi_{ijk}) = \beta_0 + \beta_1 X_{ij} + b_{0i} + b_{1i} X_{ij}.$$

Here, X_{ij} denotes the dichotomous fixed (method) effect with $X_{ij} = 1$ and $X_{ij'} = 0$. Using a cumulative Gaussian approximation to the logistic function (Johnson and

Kotz (1970), Zeger et al. (1988) and Lin and Breslow (1996)), we have the expression $logit(\pi_{ijk}) \approx a_l(D) \cdot (\mu_{ij} + \beta_{ij})$ where $a_l(D) = |c^2 D z_{ij} z'_{ij} + I|^{-q/2}$ and $c = 16\sqrt{3}/(15\pi)$. The details of such approximation is presented as follows.

In the above notation, D is the variance matrix of random effects. Under our set up,

$$D = \begin{pmatrix} \sigma_\alpha^2 & 0 \\ 0 & \sigma_\gamma^2 \end{pmatrix}$$

q is the dimension of random effects, here $q = 2$. z_{ij} denotes the design vector of random effects $z_{ij} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and z'_{ij} is the corresponding transpose. I is the identity matrix, x_{ij} is design matrix of fixed effects and $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$.

Therefore,

$$\begin{aligned} logit(\pi_{ij}) &\approx |c^2 D z_{ij} z'_{ij} + I|^{-q/2} \cdot (\mu_{ij} + \beta_{ij}) \\ &= \left| c^2 \begin{pmatrix} \sigma_\alpha^2 & 0 \\ 0 & \sigma_\gamma^2 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right|^{(-1)} \cdot \begin{pmatrix} 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \\ &= \frac{\beta_0 + \beta_1}{c^2\sigma_\alpha^2 + c^2\sigma_\alpha^2} \end{aligned}$$

Similarly,

$$logit(\pi_{ij'k}) \approx \frac{\beta_0}{c^2\sigma_\alpha^2 + 1}.$$

In terms of the variance covariance matrix, we still can follow Zeger et al. 1988,

$$\begin{aligned} cov(Y_i) &= \begin{pmatrix} Var(Y_{ijk}) & Cov(Y_{ijk}, Y_{ij'k}) \\ Cov(Y_{ijk}, Y_{ij'k}) & Var(Y_{ij'k}) \end{pmatrix} \\ &= \begin{pmatrix} Var(\pi_{ijk}) & Cov(\pi_{ijk}, \pi_{ij'k}) \\ Cov(\pi_{ijk}, \pi_{ij'k}) & Var(\pi_{ij'k}) \end{pmatrix} + \begin{pmatrix} E((\pi_{ijk})(1-\pi_{ijk})) & 0 \\ 0 & E((\pi_{ij'k})(1-\pi_{ij'k})) \end{pmatrix} \\ &\approx L_i Z_i D Z_i' L_i + \phi A_i \end{aligned}$$

where $L_i = diag\frac{\partial h(\pi)}{\partial(\pi)}$ and $h(\pi)$ is the logit link, $\pi = x_{ij}'\beta$; $Z_i = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$; D has been defined as above. ϕ is the overdispersion parameter and for simplicity we assume it's 1 here. $A_i = diag(g^*(\pi_{ijk}), g^*(\pi_{ij'k}))$ and $g^*(\pi_{ijk}) = \pi_{ijk}(1-\pi_{ijk})$.

Therefore,

$$\begin{aligned} & Cov(Y_i) \\ = & \begin{pmatrix} \pi_{ijk}(1-\pi_{ijk}) & 0 \\ 0 & \pi_{ij'k}(1-\pi_{ij'k}) \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} \sigma_\alpha^2 & 0 \\ 0 & \sigma_\gamma^2 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \\ & \cdot \begin{pmatrix} \pi_{ijk}(1-\pi_{ijk}) & 0 \\ 0 & \pi_{ij'k}(1-\pi_{ij'k}) \end{pmatrix} + \begin{pmatrix} \pi_{ijk}(1-\pi_{ijk}) & 0 \\ 0 & \pi_{ij'k}(1-\pi_{ij'k}) \end{pmatrix} \\ = & \begin{pmatrix} \pi_{ijk}^2(1-\pi_{ijk})^2(\sigma_\alpha^2+\sigma_\gamma^2) & \pi_{ijk}(1-\pi_{ijk})\pi_{ij'k}(1-\pi_{ij'k})\sigma_\alpha^2 \\ \pi_{ijk}(1-\pi_{ijk})\pi_{ij'k}(1-\pi_{ij'k})\sigma_\alpha^2 & \pi_{ij'k}^2(1-\pi_{ij'k})^2(\sigma_\alpha^2) \end{pmatrix} \\ & + \begin{pmatrix} \pi_{ijk}(1-\pi_{ijk}) & 0 \\ 0 & \pi_{ij'k}(1-\pi_{ij'k}) \end{pmatrix} \end{aligned}$$

Finally we have

$$\begin{aligned}
Var(\pi_{ijk}) &= \pi_{ijk}^2(1-\pi_{ijk})^2(\sigma_\alpha^2+\sigma_\gamma^2) \\
Var(\pi_{ij'k}) &= \pi_{ij'k}^2(1-\pi_{ij'k})^2\sigma_\alpha^2 \\
Var(Y_{ijk}) &= Var(\pi_{ijk}) + \pi_{ijk}(1-\pi_{ijk}) \\
Var(Y_{ij'k}) &= Var(\pi_{ij'k}) + \pi_{ij'k}(1-\pi_{ij'k}) \\
Cov(Y_{ijk}, Y_{ijk'}) &= Cov(\pi_{ijk}, \pi_{ijk}) = Var(\pi_{ijk}) \\
Cov(Y_{ij'k}, Y_{ij'k}) &= Cov(\pi_{ij'k}, \pi_{ij'k}) = Var(\pi_{ij'k}) \\
Cov(Y_{ijk}, Y_{ij'k'}) &= Cov(\pi_{ijk}, \pi_{ij'k'}) = \pi_{ijk}(1-\pi_{ijk})\pi_{ij'k}(1-\pi_{ij'k})\sigma_\alpha^2
\end{aligned}$$

where π_{ijk} and $\pi_{ij'k}$ can be estimated using the Gaussian approximation.

6.1.2.2 Adaptive Gauss-Hermite Quadrature

The cumulative Gaussian approximation was not satisfactory even when we have big sample size (results were shown through simulations for cumulative Gaussian approximation). In this section, we introduce Adaptive Gauss-Hermite Quadrature to estimate the marginal likelihood in order to estimate CIE using logistic random effects models.

Gauss-Hermite Quadrature (GHQ) is often used for numerical approximation of integrals with Gaussian kernels. In generalized linear mixed models, random effects are assumed to have Gaussian distributions, but often the marginal likelihood, which has the key role in parameter estimation and inference, is analytically intractable. Gaussian quadrature is particularly well suited to numerically evaluate integrals against probability measures. And Gauss-Hermite quadrature is appropriate when the density has kernel $exp(-x^2)$ and integration extends over the real line, as is the case for the normal distribution. Suppose that $p(x)$ is a probability density function and the function $f(x)$ is to be integrated against it. Then the quadrature

rule is

$$\int_{-\infty}^{\infty} f(x)p(x)dx = \sum_{i=1}^{N} w_i f(x_i).$$

where N denotes the number of quadrature points. The Gaussian quadrature chooses abscissas in areas of high density, and if $p(x)$ is continuous, the quadrature rule is exact if $f(x)$ is a polynomial of up to degree $2N-1$. In the generalized linear mixed model the roles of $f(x)$ and $p(x)$ are played by the conditional distribution of the data given the random effects, and the random-effects distribution, respectively. Quadrature abscissas and weights are those of the standard Gauss-Hermite quadrature (SAS,2009).

However, with the complex of adaptive Gauss-Hermite Quadrature approximation, three simpler logistic random effects models were applied to estimate $MSD(Y_1, Y_1')$, $MSD(Y_2, Y_2')$ and $MSD(Y_1, Y_2)$. Let Y_{ijk} be the observation made on the i^{th} subject by the j^{th} observer under the k^{th} replication. To estimate $MSD_i(Y_1, Y_2)$, we consider subject as a random factor; while observer is fixed factor. We construct three generalized linear mixed models

$$\eta_{i1} = \mu_1 + \alpha_{1i} \tag{6.3}$$

$$\eta_{i2} = \mu_2 + \alpha_{2i} \tag{6.4}$$

$$\eta_{ijk} = \mu + \alpha_i + \beta_j \tag{6.5}$$

where η_{ijk} is the linear predictor under logit link, i.e. $logit(\pi_{ijk}) = \eta_{ijk}$; μ, μ_1 and μ_2 are constants; β_j is fixed observer effect in (6.5) and α_{1i}, α_{2i} and α_i stand for random subject effects in the above three models. All three models contribute to estimate $MSD_i(Y_1, Y_1')$, $MSD_i(Y_2, Y_2')$ and $MSD_i(Y_1, Y_2)$. π_{i1} can be estimated using the measurement of the first observer as shown in (6.4). Similarly, π_{i2} can be obtained

by (6.5). Furthermore, $\pi_{i1'}$ and $\pi_{i2'}$ can be estimated by (6.5), such as:

$$\pi_{i1} = \frac{\exp(\mu_1 + \alpha_{1i})}{1 + \exp(\mu_1 + \alpha_{1i})}$$
$$\pi_{i2} = \frac{\exp(\mu_2 + \alpha_{2i})}{1 + \exp(\mu_2 + \alpha_{2i})}$$
$$\pi_{i1'} = \frac{\exp(\mu + \alpha_i + \beta_1)}{1 + \exp(\mu + \alpha_i + \beta_1)}$$
$$\pi_{i2'} = \frac{\exp(\mu + \alpha_i + \beta_2)}{1 + \exp(\mu + \alpha_i + \beta_2)}$$

Therefore

$$MSD_i(Y_1, Y_1') = 2\pi_{i1}(1 - \pi_{i1}) \tag{6.6}$$

$$MSD_i(Y_2, Y_2') = 2\pi_{i2}(1 - \pi_{i2}) \tag{6.7}$$

$$MSD_i(Y_1, Y_2) = Pr(Y_{i1} \neq Y_{i2}|i) = \pi_{i1'} + \pi_{i2'} - 2\pi_{i1'}\pi_{i2'} \tag{6.8}$$

In PROC GLIMMIX procedure of SAS 9.2, "ilink" option in "output" statement gives the inverse of logit function and "blup" option also in "output" statement gives the predicted values based on both fixed and random effects. Besides that, METHOD=QUAD in a generalized linear mixed model approximates the marginal log likelihood with an adaptive Gauss-Hermite quadrature. And points=1000 were specified with "QUAD" method to improve estimation precision. A numerical integration rule is called adaptive when it uses a variable step size to control the error of the approximation. For example, an adaptive trapezoidal rule uses serial splitting of intervals at midpoints until a desired tolerance is achieved. Furthermore, the GLIMMIX procedure centers and scales the quadrature points by using the empirical Bayes estimates (EBEs) of the random effects and the Hessian (second derivative) matrix from the EBE suboptimization. This centering and scaling improves the like-

lihood approximation by placing the abscissas according to the density function of the random effects (SAS,2009).

6.1.3 Bayesian Method

Following the estimating CIEA with three logit models (6.3-6.5), Bayesian method can be alternatively used for estimation, besides the mixed effects models. The corresponding bayesian set up for equation(6.3) is shown as follows:

- Likelihood: $Y_{i1k} \sim Bernoulli(\pi_{i1})$

  $\text{logit}(\pi_{i1}) = \beta_0 + b_0$

- priors: $\beta_0 \sim Normal(0, 0.0001)$

  $b_{0i} \sim Normal(0, s)$

  $s \sim Gamma(0.001, 0.001)$

Similarly we can derive the set up for Y_{i2k}. In equation (6.5), we have

- Likelihood: $Y_{ijk} \sim Bernoulli(\pi_{ij'})$

  $\text{logit}(\pi_{i1'}) = \beta'_0 + \beta'_1 + b'_{0i}$

  $\text{logit}(\pi_{i2'}) = \beta'_0 + b'_{0i}$

- priors: $\beta'_{0i} \sim Normal(0, 0.0001)$

  $\beta'_{1i} \sim Normal(0, 0.0001)$

  $b'_{0i'} \sim Normal(0, s)$

  $s \sim Gamma(0.001, 0.001)$

WinBUGS has been used to perform this Bayesian analysis. Due to the computational intensity, simulation studies applying the Bayesian method have not been done, but the results for the mammogram data are presented and compared in Table 6.4.

6.1.4 Simulations

Different exploratory approaches has been investigated using simulations as well as with the real data analysis. We found that only the identity link and the logit link with Adaptive Gauss-Hermite Quadrature using PROC GLIMMIX give satisfactory results. In this simulation study, the same data generation and true value setups were applied as introduced in Chapter 3. Good, moderate as well as poor agreement scenarios were considered. Identity link, Gaussian approximation and the logit link with Adaptive Gauss-Hermite Quadrature using PROC GLIMMIX were considered. Table 6.1 presents the result from the identity link. Sample sizes 50,100 and 200 were included. The biases were small, the two standard error estimations were similar and the coverage probability approached the nominal 0.95 when the sample size increase. Due to the limitation of using identity link on binary outcomes, the proportions of extreme π_i and λ_i in the simulation population were calculated. We define the extreme proportions as the percentages less than 0.1 or greater than 0.9. Under the setup of good agreement, the extreme proportions of π_i and λ_i were 14.1% and 13.4%, respectively. Similarly, under moderate and poor agreement, the extreme proportions of π where the same due the way we generated the data. The extreme proportions of λ were 11.6% under moderate agreement and 10.9% under poor agreement. Figure 6.1 and 6.2 show the histogram of true values of π and λ under moderate agreement, respectively. Table 6.2 presents the simulation results of cumulative gaussian approximation. Apparently the cumulative gaussian approximation dose not work very well, especially when the sample size is small. The coverage probabilities of cumulative gaussian approximation were much very lower than 95% when the agreement is poor. Simulation results for the logit link with Adaptive Gauss-Hermite Quadrature are shown in Table 6.3. Only sample size 100 was included due to the computational intensity.

Table 6.1: Simulation of Replicated CIEA using Identity Link

Sample size	**k**	**l**	μ_B	**True**	**Bias**	SE^1	SE^2	CP
50	1	2	1	0.942	-0.084	0.139	0.136	0.958
100	1	2	1	0.942	-0.062	0.117	0.103	0.949
200	1	2	1	0.942	-0.038	0.083	0.078	0.955
50	3	3	1	0.922	-0.017	0.078	0.070	0.932
100	3	3	1	0.922	-0.008	0.059	0.055	0.931
200	3	3	1	0.922	-0.004	0.046	0.043	0.923
50	4	2	1	0.908	-0.007	0.083	0.075	0.916
100	4	2	1	0.908	0.003	0.065	0.059	0.926
200	4	2	1	0.908	0.007	0.051	0.045	0.919
50	1	2	3	0.783	-0.049	0.157	0.138	0.926
100	1	2	3	0.783	-0.033	0.114	0.109	0.936
200	1	2	3	0.783	-0.016	0.086	0.083	0.955
50	3	3	3	0.766	-0.004	0.087	0.085	0.946
100	3	3	3	0.766	0.000	0.066	0.065	0.942
200	3	3	3	0.766	0.000	0.050	0.048	0.939
50	4	2	3	0.755	0.008	0.097	0.089	0.919
100	4	2	3	0.755	0.008	0.071	0.069	0.950
200	4	2	3	0.755	0.010	0.054	0.051	0.923
50	1	2	5	0.554	0.007	0.151	0.135	0.939
100	1	2	5	0.554	0.010	0.111	0.105	0.941
200	1	2	5	0.554	0.016	0.080	0.079	0.951
50	3	3	5	0.576	0.005	0.083	0.082	0.930
100	3	3	5	0.576	0.003	0.058	0.060	0.940
200	3	3	5	0.576	0.001	0.043	0.043	0.937
50	4	2	5	0.592	-0.007	0.086	0.086	0.947
100	4	2	5	0.592	-0.009	0.062	0.063	0.948
200	4	2	5	0.592	-0.010	0.045	0.045	0.931

SE^1Standard errors based on simulations of CIE

SE^2Mean of estimated standard errors calculated from SE estimator

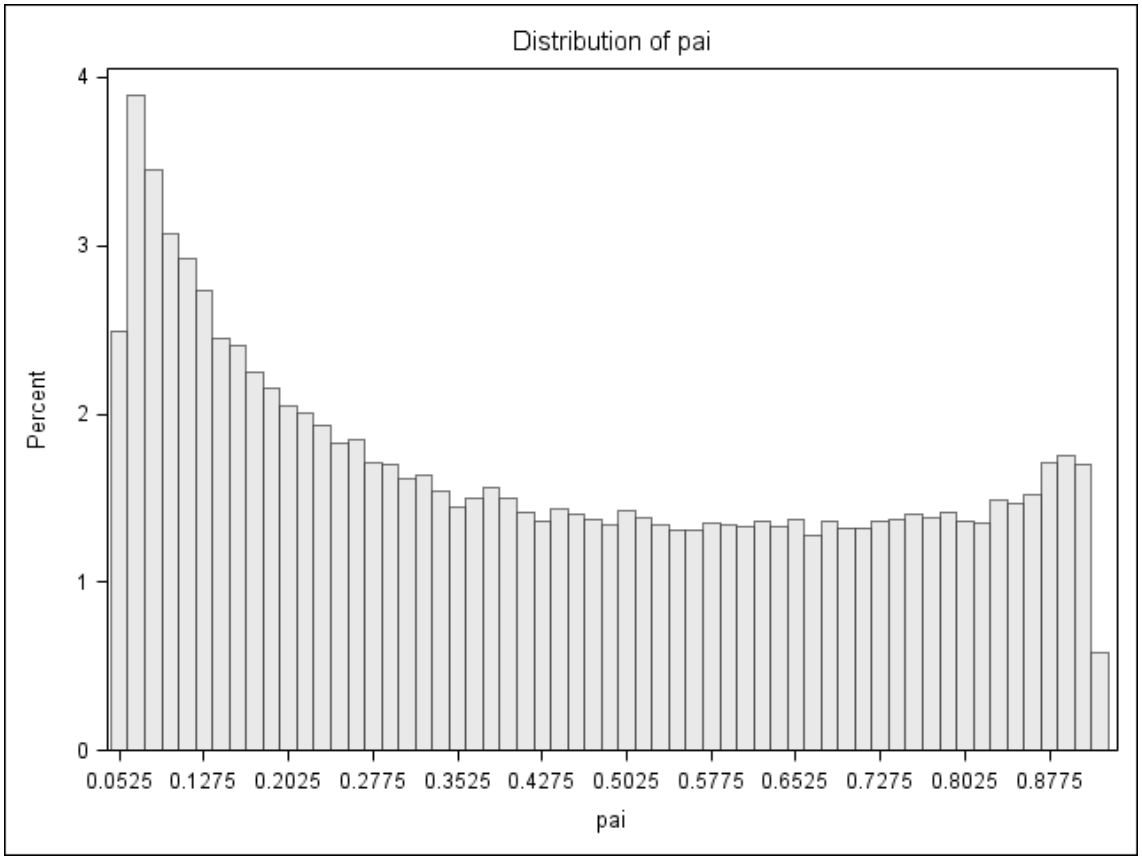

Figure 6.1: Histogram of π under Moderate Agreement

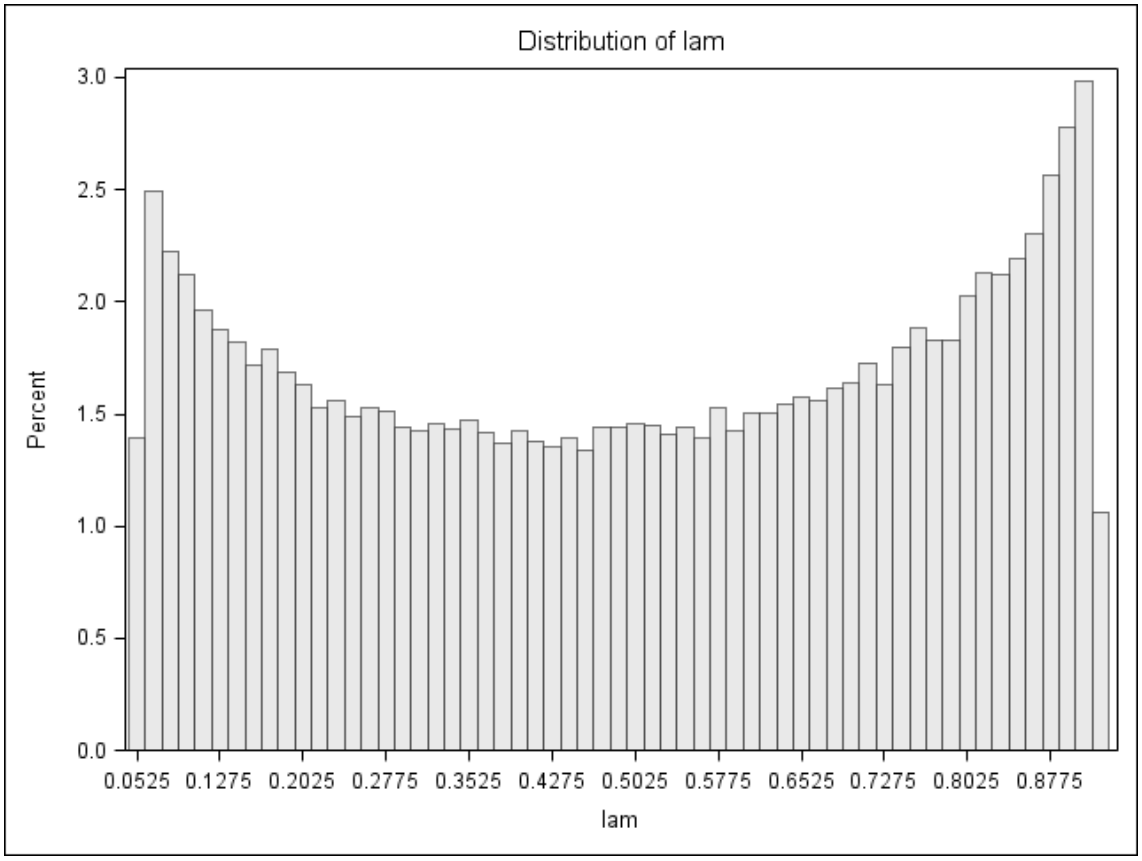

Figure 6.2: Histogram of λ under Moderate Agreement

Table 6.2: Simulation of Replicated CIEA with Logit Link using Variance Components with Cumulative Gaussian Approximation

Sample size	**k**	**l**	μ_B	**True**	**bias**	SE^1	SE^2	**CP**
100	1	2	1	0.942	-0.030	0.090	0.108	0.813
200	1	2	1	0.942	-0.015	0.065	0.080	0.900
400	1	2	1	0.942	-0.006	0.049	0.059	0.952
100	3	3	1	0.922	-0.002	0.086	0.074	0.792
200	3	3	1	0.922	0.010	0.054	0.056	0.900
400	3	3	1	0.922	0.017	0.038	0.040	0.890
100	4	2	1	0.908	-0.012	0.114	0.095	0.744
200	4	2	1	0.908	0.012	0.070	0.073	0.866
400	4	2	1	0.908	0.022	0.047	0.054	0.904
100	1	2	3	0.783	0.086	0.119	0.136	0.660
200	1	2	3	0.783	0.105	0.088	0.102	0.618
400	1	2	3	0.783	0.119	0.062	0.072	0.516
100	3	3	3	0.766	0.116	0.106	0.087	0.550
200	3	3	3	0.766	0.134	0.066	0.066	0.451
400	3	3	3	0.766	0.142	0.043	0.047	0.214
100	4	2	3	0.755	0.103	0.133	0.110	0.616
200	4	2	3	0.755	0.133	0.084	0.084	0.546
400	4	2	3	0.755	0.145	0.053	0.061	0.344
100	1	2	5	0.554	0.234	0.152	0.166	0.499
200	1	2	5	0.554	0.257	0.119	0.133	0.432
400	1	2	5	0.554	0.274	0.088	0.098	0.252
100	3	3	5	0.576	0.236	0.124	0.107	0.387
200	3	3	5	0.576	0.254	0.091	0.081	0.231
400	3	3	5	0.576	0.268	0.054	0.060	0.061
100	4	2	5	0.592	0.195	0.156	0.129	0.495
200	4	2	5	0.592	0.228	0.101	0.101	0.396
400	4	2	5	0.592	0.243	0.070	0.075	0.196

SE^1Standard errors based on simulations of CIE

SE^2Mean of estimated standard errors calculated from SE estimator

Table 6.3: Simulation of Replicated CIEA with Logit Link by Adaptive Gauss-Hermite Quadrature when Sample size=100

Sample size	**k**	**l**	μ_B	**True**	**Bias**	SE^1	SE^2	**CP**
100	1	2	1	0.942	-0.005	0.153	0.14	0.92
100	3	3	1	0.922	0.049	0.053	0.054	0.81
100	4	2	1	0.908	0.031	0.057	0.061	0.90
100	1	2	3	0.783	-0.01	0.134	0.135	0.91
100	3	3	3	0.766	0.035	0.063	0.064	0.91
100	4	2	3	0.755	0.021	0.068	0.071	0.96
100	1	2	5	0.554	-0.035	0.113	0.118	0.92
100	3	3	5	0.576	0.017	0.055	0.06	0.94
100	4	2	5	0.592	0.017	0.063	0.069	0.95

SE^1Standard errors based on simulations of CIE

SE^2Mean of estimated standard errors calculated from SE estimator

6.1.5 Mammogram Study

Pairwise agreement was assessed between radiologist A and each of the other 9 radiologists in the Mammogram study. Nonparametric estimation was introduced in Chapter 3. Here Table 6.4 presents the estimation results using identity link, logit link with the Adaptive Gauss-Hermite Quadrature as well as the Bayesian approach. All methods yield similar results compared to nonparametric estimation. The identity link seemed to be closet to nonparametric estimation in terms of the result in this Mammogram study. Standard error estimations of nonparametric approach, logic link and identity link were presented in Table 6.5. In most comparisons, the identity link yielded the smallest SE estimations among the three methods.

Table 6.4: Comparison of Estimations of CIEA in Mammogram Study

Comparison	Nonparametric	Logit Link	Identity Link	Bayesian
(A,B)	0.646	0.626	0.643	0.593
(A,C)	0.358	0.305	0.356	0.379
(A,D)	0.697	0.739	0.695	0.651
(A,E)	0.643	0.669	0.641	0.608
(A,F)	0.763	0.713	0.760	0.658
(A,G)	0.541	0.587	0.553	0.565
(A,H)	0.486	0.450	0.486	0.482
(A,I)	0.739	0.798	0.745	0.676
(A,J)	0.619	0.685	0.617	0.614

Table 6.5: Comparison of Standard Error Estimations of CIEA in Mammogram Study

Comparison	Nonparametric	Logit Link	Identity Link
(A,B)	0.141	0.130	0.134
(A,C)	0.093	0.098	0.082
(A,D)	0.126	0.213	0.105
(A,E)	0.141	0.253	0.115
(A,F)	0.147	0.108	0.128
(A,G)	0.111	0.185	0.083
(A,H)	0.108	0.143	0.103
(A,I)	0.111	0.293	0.103
(A,J)	0.108	0.235	0.104

6.2 CIEA for Repeated Binary Outcomes

So far, we estimated the CIE from data with unmatched replications which are measured under the "same" condition, i.e. when the true values remain fixed. This is the ideal scenario. More often, in practice, the number of readings made by each observer on each subject is fixed and these readings correspond to the levels of an additional factor whose levels can be considered as "conditions". In this case, the agreement study may be designed such that multiple matched observations with two (or more) observers are conducted on each subject under specific "conditions" where the subjects' true values may change across conditions. These observations are then considered matched repeated measurements. Here we distinguish "replicated" measurements and "repeated" measurements where for "replicated" measurements, the true values of subjects are assumed to be the same over replications, while for "repeated" measurements, the true values of the subject can vary under different conditions.

In this section, we extend the concepts and ideas of the CIE for assessing observer agreement from data consisting of matched repeated binary observations made with the same observer under different conditions. These conditions may correspond to different time points, laboratories, devices, treatments and so forth. Our approach allows the values of the measured variables and the magnitude of disagreement to vary across the conditions. First, a variance components linear mixed model with identity link will be used. Next, a variance components generalized linear mixed model, using adaptive Gaussian Hermite Quadrature is applied to estimate CIE. Simulation studies with $K = L = 2$ and $K = L = 3$ are conducted. The mammogram study are used again, treating the two readings of the same radiologist as repeated measurements, taken at two different time points.

6.2.1 Repeated Outcomes with the Identity Link

Since the binary measurements considered here do not include replicated observations, Y_j and Y_j', made with the same method on the same subject under the same condition, we cannot apply the nonparametric approach proposed in chapter 2 to estimate $\text{MSD}_h(Y_j, Y_j')$ were used, where h denotes the repeated condition, i.e. time point ($h = 1, ..., K$). Instead, we propose estimating MSDs from linear mixed models for repeated binary outcomes. For simplicity, we start with identity link function.

Let Y_{ijh} be the observation made on the i^{th} subject with the j^{th} observer under the h^{th} condition. To estimate $\widehat{MSD}_{ih}(Y_1, Y_2)$, we consider *subject* as a random factor while *observer* and *condition* are fixed factors. We construct a mixed ANOVA model as

$$Y_{ijh} \quad = \quad \mu + \alpha_i + \beta_j + \gamma_h + (\alpha\beta)_{ij} + (\beta\gamma)_{jh} + \varepsilon_{ijh} \tag{6.9}$$

The α's are the subjects' random effects while the β's and γ's are the fixed effects of the observers and the conditions, respectively. We assume that the random main effects, interactions and errors are independent and normally distributed with mean 0 and $\text{Var}(\alpha_i) = \sigma_\alpha^2$, $\text{Var}[(\alpha\beta)_{ij}] = \sigma_{\alpha\beta}^2$, and $\text{Var}(\varepsilon_{ijh}) = \sigma_\varepsilon^2$. Regarding the fixed effects, we make the common assumption that the sum of the coefficients over every index is zero, i.e., $\sum_j \beta_j = \sum_h \gamma_h = \sum_j (\beta\gamma)_{jh} = \sum_h (\beta\gamma)_{jh} = 0$. This model is very similar to the one used to estimate CIA for repeated continuous outcomes (Haber et al. (2010)). Interaction between the subject and the condition was not included due to the concern of model convergence.

It is important to note that this model allows the measurements Y_{ijh} for the same subject-method combination (i, j) to vary across the h conditions. If we consider two (hypothetical) replicated observations, Y_{jh} and Y_{jh}', that could be made by method j

on the same subject under the same condition, then

$$\begin{aligned} \mathrm{MSD}_h(Y_j, Y_j') &= E(Y_{ijh} - Y_{ijh}')^2 \\ &= 2\sigma_\varepsilon^2 \end{aligned}$$

as we assume that $E(Y_j) = E(Y_j')$ and $\mathrm{Var}(Y_j) = \mathrm{Var}(Y_j') = \sigma_\varepsilon^2$, as we assume homogeneity of the error terms across observer. It leads to $\mathrm{MSD}(Y_1, Y_1') = \mathrm{MSD}(Y_2, Y_2')$.

From the above model, it is evident that the disagreement between the two observers may depend on the condition. The $\mathrm{MSD}_h(Y_1, Y_2)$ for the h^{th} condition can be obtained from the parameters of the model as follows

$$\begin{aligned} \mathrm{MSD}_h(Y_1, Y_2) &= E(Y_{i1h} - Y_{i2h})^2 \\ &= E\left\{[\mu + \alpha_i + \beta_1 + \gamma_h + (\alpha\beta)_{i1} + (\beta\gamma)_{1h} + \varepsilon_{i1h}]\right. \\ &\quad \left.-[\mu + \alpha_i + \beta_2 + \gamma_h + (\alpha\beta)_{i2} + (\beta\gamma)_{2h} + \varepsilon_{i2h}]\right\}^2 \\ &= E\left\{(\beta_1 - \beta_2) + [(\alpha\beta)_{i1} - (\alpha\beta)_{i2}]\right. \\ &\quad \left.+[(\beta\gamma)_{1h} - (\beta\gamma)_{2h}] + (\varepsilon_{i1h} - \varepsilon_{i2h})\right\}^2 \\ &= (\beta_1 - \beta_2)^2 + \left[(\beta\gamma)_{1h} - (\beta\gamma)_{2h}\right]^2 + 2\sigma_{\alpha\beta}^2 + 2\sigma_\varepsilon^2 \end{aligned}$$

Also, as derived before, with $K = L$

$$CIE = \frac{C_2^K G((X, X') + G(Y, Y'))}{C_2^{2K} G(X, Y)} + \frac{K^2}{C_2^{2K}}$$

As a result, the CIEs for repeated binary observations under the h^{th} condition as

$$CIE_h = \frac{C_2^K (MSD(Y_1, Y_1') + MSD(Y_2, Y_2'))}{C_2^{2K} MSD_h(Y_1, Y_2)} + \frac{K^2}{C_2^{2K}}$$

Therefore,

$$CIEA_h = \frac{CIE_h - CIE_{min}}{1 - CIE_{min}},$$

where $CIE_{min} = K/(2K-1)$. One can then estimate the CIEAs from the estimated parameters and variances from fitting the mixed model.

6.2.2 Repeated Outcomes with the Logit Link

In this section, we further apply the logit link with adaptive Gaussian-Hermite Quadrature to estimate the CIEAs for evaluating agreement between observers when the measured variables are binary and the data consist of matched repeated observations made by different observers under different conditions. The idea of using three random effects logistic regression models are similar to what we described in estimating CIE/CIEA for replicated binary outcomes. We consider the cases where each of N subjects is evaluated by multiple observers under the same K conditions, where the condition is a categorial factor.

Oftentimes, for data with matched repeated measurements, replicated observations under each condition are not available. Then we propose to estimate the individual disagreement probabilities from fitted generalized linear mixed models.

Let Y_{ijh} be the observation made on the i^{th} subject with the j^{th} observer under the h^{th} condition. To estimate $MSD_{ih}(Y_1, Y_2)$, we consider *subject* as a random factor while *observer* and *condition* are fixed factors. We construct a generalized linear mixed model

$$\eta_{ijh} \quad = \quad \mu + \alpha_i + \beta_j + \gamma_h \tag{6.10}$$

where η_{ijh} is the linear predictor under the logit link function, i.e. $\text{logit}(\pi_{ijh}) = \eta_{ijh}$ ($i = 1, \ldots, N$ – subject; $j = 1, 2$ – observer, $h = 1, \ldots, H$ – condition), and μ is a constant, β_j, γ_h are fixed effects and α_i is an independent normal random variable

with expectation zero and variance σ_α^2.

The individual between-observer disagreement for the h^{th} condition h can be written as

$$\begin{aligned} MSD_{ih}(Y_1, Y_2) &= \Pr(Y_{i1h} \neq Y_{i2h}|i,h) \\ &= \Pr(Y_{i1h}=1|i,h)\Pr(Y_{i2h}=0|i,h) + \Pr(Y_{i1h}=0|i,h)\Pr(Y_{i2h}=1|i,h) \\ &= \pi_{i1h} + \pi_{i2h} - 2\pi_{i1h}\pi_{i2h} \end{aligned} \tag{6.11}$$

as $\pi_{i1h} = \Pr(Y_{i1h} = 1|i,h)$ and $\pi_{i2h} = \Pr(Y_{i2h} = 1|i,h)$ for $(i = 1,\ldots,N;\ h = 1,\ldots,H)$.

For linear mixed models, the likelihood function has a closed form. Consequently, efficient computational algorithms have been proposed for maximum likelihood and restricted maximum likelihood estimations. However, in the case of generalized linear mixed models, the likelihood function usually cannot be expressed as in a closed form which causes problems in estimating parameters. Similar to Section 6.1.3, where we have replicated binary outcomes, adaptive Gauss-Hermite quadrature has been used to approximates the marginal log likelihood.

The MSD in (6.11) can be estimated as

$$\begin{aligned} \widehat{MSD}_{ih}(Y_1, Y_2) &= \hat{\pi}_{i1h} + \hat{\pi}_{i2h} - 2\hat{\pi}_{i1h}\hat{\pi}_{i2h} \\ &= \frac{\exp(\hat{\mu}+\hat{\alpha}_i+\hat{\beta}_1+\hat{\gamma}_h)}{1+\exp(\hat{\mu}+\hat{\alpha}_i+\hat{\beta}_1+\hat{\gamma}_h)} + \frac{\exp(\hat{\mu}+\hat{\alpha}_i+\hat{\beta}_2+\hat{\gamma}_h)}{1+\exp(\hat{\mu}+\hat{\alpha}_i+\hat{\beta}_2+\hat{\gamma}_h)} \\ &\quad -2\frac{\exp(\hat{\mu}+\hat{\alpha}_i+\hat{\beta}_1+\hat{\gamma}_h)}{1+\exp(\hat{\mu}+\hat{\alpha}_i+\hat{\beta}_1+\hat{\gamma}_h)} \times \frac{\exp(\hat{\mu}+\hat{\alpha}_i+\hat{\beta}_2+\hat{\gamma}_h)}{1+\exp(\hat{\mu}+\hat{\alpha}_i+\hat{\beta}_2+\hat{\gamma}_h)} \end{aligned}$$

To estimate $\text{MSD}_h(Y_1, Y_1')$ and $\text{MSD}_h(Y_2, Y_2')$, we fit two additional models sepa-

rately

$$\eta_{i1h} = \mu_1 + \alpha_{1i} + \gamma_{1h} \tag{6.12}$$

$$\eta_{i2h} = \mu_2 + \alpha_{2i} + \gamma_{2h} \tag{6.13}$$

where $\text{logit}(\pi_{i1h}) = \eta_{i1h}$ and $\text{logit}(\pi_{i2h}) = \eta_{i2h}$, *subject* is a random factor and *condition* is a fixed factor. Moreover, α_{1i} and α_{2i} are independent normal with zero expectations respectively.

Consequently, CIEAs under different conditions can be estimated based on the three MSD functions.

$$\begin{aligned} MSD_{ih}(Y_j, Y_j') &= E[(Y_{ijh} - Y_{ijh}')^2 | i, j, h] \\ &= \Pr(Y_{ijh} \neq Y_{ijh}' | i, j, h) \\ &= \Pr(Y_{ijh} = 1 | i, j, h) \Pr(Y_{ijh} = 0 | i, j, h) \\ &\quad + \Pr(Y_{ijh} = 0 | i, j, h) \Pr(Y_{ijh} = 1 | i, j, h) \\ &= 2\pi_{ijh}(1 - \pi_{ijh}) \end{aligned} \tag{6.14}$$

where $\pi_{ijh} = \Pr(Y_{ijh} = 1 | i, j, h)$ for $(i = 1, \ldots, N;\ j = 1, 2;\ h = 1, \ldots, K)$, i.e. the probability of the outcome being one for subject i under condition h for a specific observer j.

Consequently, CIEAs under different conditions can be estimated based on the

three MSD functions.

$$\begin{aligned} MSD_{ih}(Y_j, Y_j') &= E[(Y_{ijh} - Y_{ijh}')^2 | i, j, h] \\ &= \Pr(Y_{ijh} \neq Y_{ijh}' | i, j, h) \\ &= \Pr(Y_{ijh} = 1 | i, j, h) \Pr(Y_{ijh} = 0 | i, j, h) \\ &\quad + \Pr(Y_{ijh} = 0 | i, j, h) \Pr(Y_{ijh} = 1 | i, j, h) \\ &= 2\pi_{ijh}(1 - \pi_{ijh}) \end{aligned} \tag{6.15}$$

where $\pi_{ijh} = \Pr(Y_{ijh} = 1 | i, j, h)$ for $(i = 1, \ldots, N;\ j = 1, 2;\ h = 1, \ldots, K)$, i.e. the probability of the outcome being one for subject i under condition h for a specific observer j.

We do not initially include interaction terms in the models with logit link due to the concern of common convergence issue in generalized linear mixed models with limited number of subjects. As a result, the within MSD's can be estimated as

$$\widehat{MSD}_{ih}(Y_1, Y_1') = 2\frac{\exp(\hat{\mu}_1 + \hat{\alpha}_{1i} + \hat{\gamma}_{1h})}{[1 + \exp(\hat{\mu}_1 + \hat{\alpha}_{1i} + \hat{\gamma}_{1h})]^2} \tag{6.16}$$

$$\widehat{MSD}_{ih}(Y_2, Y_2') = 2\frac{\exp(\hat{\mu}_2 + \hat{\alpha}_{2i} + \hat{\gamma}_{2h})}{[2 + \exp(\hat{\mu}_2 + \hat{\alpha}_{2i} + \hat{\gamma}_{2h})]^2} \tag{6.17}$$

Denote

$$\overline{MSD}_h(Y_j, Y_j') = \frac{1}{N}\sum_{i=1}^{N} MSD_{ih}(Y_j, Y_j')\ (j = 1, 2) \tag{6.18}$$

$$\overline{MSD}_h(Y_1, Y_2) = \frac{1}{N}\sum_{i=1}^{N} MSD_{ih}(Y_1, Y_2) \tag{6.19}$$

Then, the CIEs for binary observations under the h^{th} condition as

$$CIE_h = \frac{C_2^K(MSD(Y_1, Y_1') + MSD(Y_2, Y_2'))}{C_2^{2K} MSD_h(Y_1, Y_2)} + \frac{K^2}{C_2^{2K}}$$

Therefore,

$$CIEA_h = \frac{CIE_h - CIE_{min}}{1 - CIE_{min}},$$

where $CIE_{min} = K/(2K-1)$. Confidence intervals for the estimated coefficients can be computed using the nonparametric bootstrap approach.

6.2.3 Simulation Studies

6.2.3.1 Data Generation

We assume that observers' binary readings are based on the value of a 'true' unobserved variable T, where $T \sim N(\mu_T, \sigma_T^2)$ and σ_T^2 is the between subjects variability. Suppose we're interested in looking at the agreement of repeated binary outcomes under two conditions. Let A_{1i}, B_{1i} and A_{2i}, B_{2i} be the biases for observers X and Y under conditions 1 and 2, respectively. We assume that $A_{1i} \sim N(\mu_A, \sigma_A^2), B_{1i} \sim N(\mu_B, \sigma_B^2)$. Furthermore, $A_{2i} \sim N(A_{1i}, \sigma_{A_2}), B_{2i} \sim N(B_{1i}, \sigma_{B_2})$. And T, A_1 and B_1 as well as T, A_2 and B_2 are mutually independent.

For subject i, let $T_i = t_i$, $A_{1i} = a_{1i}$, $B_{1i} = b_{1i}$, $A_{2i} = a_{2i}$, $B_{2i} = b_{2i}$. We use U_{i1} and V_{i1} to denote the observers' readings on the subject's true value for condition 1, $U_{i1} \sim N(t_i + a_{1i}, \sigma_U^2), V_{i1} \sim N(t_i + b_{1i}, \sigma_V^2)$ where σ_U^2 and σ_V^2 are the within-observer error variances for observers X and Y, respectively. Similarly we can define $U_{i2} \sim N(t_i + a_{1i}, \sigma_U^2), V_{i2} \sim N(t_i + b_{1i}, \sigma_V^2)$ for condition 2.

Finally, for a fixed common threshold C, we generate binary readings X_{i1}, X_{i2} and Y_{i1}, Y_{i2} by the criterion that if $U_{ik} > C$ then $X_{ik} = 1$, else then $X_{ik} = 0$, and if $V_{il} > C$ then $Y_{il} = 1$, else then $Y_{il} = 0$. Note that $X's$ and $Y's$ are conditionally independent given the subject.

For given t_i, a_{1i}, b_{1i}, a_{2i}, b_{2i} the values of $P(X_{i1} = 1)$ and $P(Y_{i1} = 1)$ are π_{1i} and

λ_{1i}, $P(X_{i2}=1)$ and $P(Y_{i2}=1)$ are π_{2i} and λ_{2i}as follows:

$$\pi_{1i} = P(X_{i1}=1) = P(U_{i1} > C) = 1 - \Phi\left(\frac{C-(t_i+a_{1i})}{\sigma_U}\right)$$

$$\lambda_{1i} = P(Y_{i1}=1) = P(V_{i1} > C) = 1 - \Phi\left(\frac{C-(t_i+b_{1i})}{\sigma_V}\right),$$

$$\pi_{2i} = P(X_{i2}=1) = P(U_{i2} > C) = 1 - \Phi\left(\frac{C-(t_i+a_{2i})}{\sigma_U}\right)$$

$$\lambda_{2i} = P(Y_{i2}=1) = P(V_{i2} > C) = 1 - \Phi\left(\frac{C-(t_i+b_{2i})}{\sigma_V}\right),$$

where Φ is the CDF of a standard normal.

Similarly, this can be extended to scenarios with three or more repeated conditions.

6.2.3.2 True Values

Suppose T stands for the systolic blood pressure and we consider a hypothetical population of obese type II diabetes, i.e., persons with $\mu_T = 138mmHg$ and a between-subjects $\sigma_T = 5mmHg$. The usual cut-off point for hypertension is 140mmHg for systolic blood pressure, which is the threshold C in our model. We set the within-observers standard errors to 3 ($\sigma_U = \sigma_V = 3$). Furthermore, we set the average bias of observer X, μ_A, to 0 and gradually increase the average bias of observer Y, μ_B, from 1 to 3 to 5 in order to accommodate good, moderate and poor agreement, respectively. The closer μ_A and μ_B, the better the agreement when other conditions remain unchanged. Standard deviation for biases in observers X and Y were set to 1 ($\sigma_A = \sigma_B = 1$). And also $\sigma_{A2} = \sigma_{B2} = 1$. For example, under good agreement with a random sample (n=100), the mean of the readings from the first observer under condition 1, $\bar{Y}_{11} = 0.377$. Furthermore, $\bar{Y}_{12} = 0.375$, denoting the mean of measurements from the first observer under condition 2. Similarly, $\bar{Y}_{21} = 0.438$ and $\bar{Y}_{22} = 0.420$.

Tables 6.6 and 6.7 present the simulation results for CIEA estimation with re-

peated binary data using the identity link under 2 conditions (K=L=2) and three conditions (K=L=3). Generally, satisfactory simulation results were obtained. Tables 6.8 and 6.9 present the results for CIEA with repeated binary data under logit link when K=L=2 and K=L=3 for sample sizes 100 and 200, respectively. Minimal biases were obtained and the similarity between two standard error estimates indicates the robustness of standard error estimations. With the increase of sample size, we tend to have satisfactory coverage probabilities.

Table 6.6: Simulation Results For CIEA Estimation with Repeated Binary Data under Identity Link, K=L=2

Sample size	**K**	**L**	**Condition**	μ_B	**True**	**Bias**	SE^1	SE^2	**CP**
50	2	2	1	1	0.922	-0.077	0.115	0.116	0.924
50	2	2	2	1	0.878	-0.005	0.112	0.113	0.932
100	2	2	1	1	0.922	-0.037	0.082	0.084	0.960
100	2	2	2	1	0.878	0.022	0.080	0.079	0.943
200	2	2	1	1	0.922	-0.018	0.066	0.061	0.940
200	2	2	2	1	0.878	0.034	0.065	0.062	0.942
50	2	2	1	3	0.766	-0.029	0.119	0.115	0.928
50	2	2	2	3	0.740	0.018	0.122	0.119	0.941
100	2	2	1	3	0.766	-0.010	0.096	0.088	0.930
100	2	2	2	3	0.740	0.028	0.097	0.095	0.933
200	2	2	1	3	0.766	-0.002	0.073	0.067	0.928
200	2	2	2	3	0.740	0.030	0.074	0.068	0.931
50	2	2	1	5	0.575	-0.004	0.112	0.107	0.913
50	2	2	2	5	0.565	0.020	0.116	0.113	0.927
100	2	2	1	5	0.575	0.007	0.086	0.083	0.917
100	2	2	2	5	0.565	0.025	0.088	0.086	0.920
200	2	2	1	5	0.575	0.009	0.063	0.061	0.922
200	2	2	2	5	0.565	0.023	0.064	0.063	0.923

SE^1 Standard errors based on simulations of CIE

SE^2 Mean of estimated standard errors calculated from SE estimator

Table 6.7: Simulation Results for CIEA Estimation with Repeated Binary Data under Identity Link, K=L=3

Sample size	**K**	**L**	**Condition**	μ_B	**True**	**Bias**	SE^1	SE^2	**CP**
50	3	3	1	1	0.921	-0.057	0.092	0.116	0.955
50	3	3	2	1	0.910	-0.045	0.094	0.103	0.934
50	3	3	3	1	0.899	-0.006	0.086	0.103	0.933
100	3	3	1	1	0.921	-0.031	0.068	0.084	0.977
100	3	3	2	1	0.910	-0.019	0.068	0.078	0.946
100	3	3	3	1	0.899	0.007	0.065	0.078	0.940
200	3	3	1	1	0.921	-0.018	0.052	0.062	0.963
200	3	3	2	1	0.910	-0.005	0.052	0.059	0.933
200	3	3	3	1	0.899	0.013	0.051	0.059	0.925
50	3	3	1	3	0.766	-0.027	0.109	0.115	0.953
50	3	3	2	3	0.759	-0.019	0.109	0.115	0.944
50	3	3	3	3	0.752	0.010	0.111	0.115	0.946
100	3	3	1	3	0.766	-0.010	0.078	0.089	0.964
100	3	3	2	3	0.759	-0.002	0.078	0.089	0.959
100	3	3	3	3	0.752	0.015	0.079	0.089	0.959
200	3	3	1	3	0.766	-0.004	0.058	0.067	0.968
200	3	3	2	3	0.759	0.003	0.058	0.068	0.965
200	3	3	3	3	0.752	0.016	0.058	0.068	0.958
50	3	3	1	5	0.576	-0.004	0.101	0.107	0.948
50	3	3	2	5	0.573	0.000	0.101	0.113	0.964
50	3	3	3	5	0.571	0.016	0.106	0.113	0.963
100	3	3	1	5	0.576	-0.001	0.073	0.082	0.966
100	3	3	2	5	0.573	0.002	0.073	0.085	0.972
100	3	3	3	5	0.571	0.011	0.074	0.085	0.973
200	3	3	1	5	0.576	0.003	0.052	0.061	0.974
200	3	3	2	5	0.573	0.006	0.052	0.063	0.976
200	3	3	3	5	0.571	0.012	0.053	0.063	0.976

SE^1 Standard errors based on simulations of CIE

SE^2 Mean of estimated standard errors calculated from SE estimator

Table 6.8: Simulation Results for CIEA Estimation with Repeated Binary Data under Logit Link, K=L=2

Sample size	**k**	**l**	**Condition**	μ_B	**True**	**Bias**	SE^1	SE^2	**CP**
100	2	2	1	1	0.922	0.044	0.090	0.092	0.85
100	2	2	2	1	0.878	0.053	0.090	0.090	0.87
200	2	2	1	1	0.924	0.056	0.059	0.062	0.86
200	2	2	2	1	0.887	0.063	0.060	0.062	0.88
100	2	2	1	3	0.766	0.035	0.096	0.094	0.93
100	2	2	2	3	0.740	0.040	0.093	0.094	0.90
200	2	2	1	3	0.766	0.041	0.062	0.066	0.92
200	2	2	2	3	0.740	0.028	0.060	0.067	0.93
100	2	2	1	5	0.575	0.019	0.093	0.084	0.91
100	2	2	2	5	0.565	0.033	0.091	0.084	0.91
200	2	2	1	5	0.575	0.014	0.064	0.061	0.93
200	2	2	2	5	0.565	0.029	0.063	0.062	0.94

SE^1 Standard errors based on simulations of CIE

SE^2 Mean of estimated standard errors calculated from SE estimator

Table 6.9: Simulation Results for CIEA Estimation with Repeated Binary Data under Logit Link, K=L=3

Sample size	**k**	**l**	**Condition**	μ_B	**True**	**Bias**	SE^1	SE^2	**CP**
100	3	3	1	1	0.958	0.040	0.061	0.062	0.84
100	3	3	2	1	0.956	0.049	0.062	0.062	0.84
100	3	3	3	1	0.957	0.054	0.063	0.062	0.80
200	3	3	1	1	0.963	0.045	0.052	0.051	0.87
200	3	3	2	1	0.963	0.046	0.053	0.051	0.87
200	3	3	3	1	0.963	0.051	0.053	0.051	0.88
100	3	3	1	3	0.800	0.035	0.075	0.069	0.87
100	3	3	2	3	0.794	0.038	0.075	0.069	0.89
100	3	3	3	3	0.795	0.041	0.075	0.069	0.89
200	3	3	1	3	0.793	0.028	0.045	0.048	0.91
200	3	3	2	3	0.790	0.034	0.046	0.048	0.88
200	3	3	3	3	0.792	0.037	0.048	0.048	0.88
100	3	3	1	5	0.586	0.008	0.074	0.065	0.92
100	3	3	2	5	0.582	0.010	0.069	0.066	0.95
100	3	3	3	5	0.586	0.015	0.070	0.065	0.93
200	3	3	1	5	0.589	0.011	0.040	0.045	0.96
200	3	3	2	5	0.587	0.016	0.041	0.045	0.94
200	3	3	3	5	0.590	0.018	0.043	0.045	0.97

SE^1 Standard errors based on simulations of CIE

SE^2 Mean of estimated standard errors calculated from SE estimator

6.2.4 Mammogram Study

Tables 6.10 and 6.11 present the estimates of CIEA from the Mammogram study with two time points as conditions. Also, results from the nonparametric replicated binary method are shown for comparison. The p-values testing for homogeneity of CIEA under both conditions are also presented. Wald type test was used and bootstrap SE was applied to obtain those p-values. From both estimation approaches, homogeneity of CIEA's was established for all comparisons.

Table 6.10: Results of Repeated Binary at Two Time Conditions in Mammogram Study with Identity Link

Comparison	$CIEA_1$	$CIEA_2$	SE_1	SE_2	Replicated Nonparametric	P-value
(A,B)	0.624	0.628	0.133	0.136	0.646	0.63
(A,C)	0.347	0.348	0.086	0.087	0.358	0.64
(A,D)	0.640	0.671	0.094	0.100	0.697	0.93
(A,E)	0.642	0.642	0.124	0.125	0.643	0.50
(A,F)	0.732	0.740	0.125	0.126	0.763	0.70
(A,G)	0.547	0.556	0.098	0.099	0.541	0.76
(A,H)	0.490	0.501	0.095	0.108	0.486	0.71
(A,I)	0.693	0.713	0.093	0.094	0.739	0.77
(A,J)	0.611	0.614	0.109	0.109	0.619	0.60

$CIEA_1$ CIEA Estimation under Time Condition 1

$CIEA_2$ CIEA Estimation under Time Condition 2

SE_1 Standard Errors Estimation under Time Condition 1

SE_2 Standard Errors Estimation under Time Condition 2

Table 6.11: Results of Repeated Binary at Two Time Conditions in Mammogram Study with Logit Link

Comparison	$CIEA_1$	$CIEA_2$	SE_1	SE_2	Replicated Nonparametric	P-value
(A,B)	0.584	0.619	0.211	0.207	0.646	0.69
(A,C)	0.269	0.301	0.096	0.112	0.358	0.78
(A,D)	0.690	0.669	0.219	0.238	0.697	0.35
(A,E)	0.619	0.707	0.203	0.207	0.643	0.89
(A,F)	0.624	0.711	0.185	0.216	0.763	0.88
(A,G)	0.563	0.569	0.185	0.187	0.541	0.56
(A,H)	0.400	0.439	0.152	0.150	0.486	0.82
(A,I)	0.761	0.787	0.202	0.232	0.739	0.71
(A,J)	0.661	0.697	0.196	0.203	0.619	0.86

$CIEA_1$ CIEA Estimation under Time Condition 1

$CIEA_2$ CIEA Estimation under Time Condition 2

SE_1 Standard Errors Estimation under Time Condition 1

SE_2 Standard Errors Estimation under Time Condition 2

6.3 Discussion

This chapter introduced model-based estimation of CIE with replicated and repeated binary outcomes. Variance components linear mixed models as well as variance components generalized linear mixed models were applied with the identity and logit links, respectively. The cumulative Gaussian approximation was used to approximate the marginal likelihood in order to estimate CIE when using logit link. However, it cannot achieve satisfactory results . Adaptive Gaussian-Hermite Quadrature was also used within SAS PROC GLIMMIX, using three simpler logistic random effects models. Model-based estimation requires more model assumptions compared to the nonparametric method. However, as shown in Table 6.12, compared to the nonparametric approach, estimates of model-based CIE with identity link have smaller root mean squared errors (RMSE) when we have a medium sample size of 100. On the other hand, RMSEs of model-based CIE estimates with logit link seem to be similar to nonparametric method. Model-based methods require more assumptions than nonparametric methods so that we expect better RMSEs from model-based approach. In the numerical approximation for CIE estimates with logit link, we fitted models without the interaction term. This may decrease the precision and result in a larger RMSEs as observed in our estimation.

It is also interesting to further investigate the scenario when fitting the generalized linear mixed model with adaptive gaussian hermite quadrature and include the interaction term. The simulation results of CIEA are shown as follows in Tables 6.13 with sample size=100 only. As compared to Table 6.2, where the Cumulative Gaussian Approximation algorithm has been applied in approximating the parameters from the generalized linear mixed model with logit link, we found that both approaches did not gave us overall good coverage probabilities. The biases were bigger with the AGHQ approximation approach. Standard error estimations were smaller with Cumulative Gaussian approximation.

Table 6.12: Comparison of Estimations with Replicated Binary Outcomes

Sample size	k	l	μ_B	True	$RMSE_1$	$RMSE_2$	$RMSE_3$
100	1	2	1	0.942	0.178	0.120	0.140
100	3	3	1	0.942	0.070	0.056	0.073
100	4	2	1	0.942	0.083	0.059	0.068
100	1	2	3	0.922	0.151	0.114	0.135
100	3	3	3	0.922	0.069	0.065	0.073
100	4	2	3	0.922	0.079	0.069	0.074
100	1	2	5	0.908	0.119	0.105	0.123
100	3	3	5	0.908	0.061	0.060	0.062
100	4	2	5	0.908	0.070	0.064	0.071

$RMSE_1$ Root Mean Squared Error of Nonparametric Estimation
$RMSE_2$ Root Mean Squared Error of Model-Based Estimation with Identity Link
$RMSE_3$ Root Mean Squared Error of Model-Based Estimation with Logit Link

Table 6.13: Simulation Results of CIEA using Logit Link (AGHQ), one model with interaction

Sample size	K	L	μ_B	TRUE	Bias	SE^1	SE^2	CP
100	1	2	1	0.942	-0.129	0.105	0.111	0.66
100	3	3	1	0.922	-0.217	0.108	0.117	0.59
100	4	2	1	0.908	-0.359	0.166	0.149	0.38
100	1	2	3	0.783	-0.197	0.129	0.148	0.76
100	3	3	3	0.766	-0.127	0.159	0.153	0.56
100	4	2	3	0.755	-0.137	0.115	0.113	0.48
100	1	2	5	0.554	-0.03	0.061	0.057	0.851
100	3	3	5	0.576	-0.076	0.093	0.089	0.807
100	4	2	5	0.592	-0.109	0.102	0.105	0.77

SE^1 Standard errors based on simulations of CIEA

SE^2 Mean of estimated standard Errors

When we have repeated measurements where the true values can change across different conditions, it is much easier to estimate CIE using model-based estimations with either the identity link or the logit link. We can estimate CIE under each conditions and test the homogeneity across different conditions.

Chapter 7

Summary and Future Research

In this dissertation, we first introduced the concept of agreement and the motivation for agreement studies. Then we reviewed existing methods for assessing agreement with both qualitative and quantitative measurements, as well as agreement for repeated outcomes. For categorical outcomes, we reviewed Kappa and weighted Kappa. Although we admit that Kappa is the most popular statistic for agreement studies with categorical data, the Kappa coefficient has been criticized for a very long time as it is very likely to attain unreasonable values. Feinstein and Cicchetti (1990) summarized two paradoxes for Kappa statistic: (i) the marginal distributions of the two observers are highly asymmetric unbalanced, and (ii) there exists a large discrepancy between the marginal distributions. Furthermore, because of the heavy dependence of Kappa on the prevalence of a condition being studied, a high value of Kappa is nearly unachievable for a rare disease with a low prevalence.

For continuous measurements, we reviewed both unscaled (LOA) and scaled indices of agreement (ICC, CCC and CIA). Unscaled indices have the advantage of interpretation based on the original units of measurement, but it may prove difficult to ascertain the limit for acceptable agreement without sufficient knowledge of the measured variable and measurement unit. Scaled indices have the advantage of

judging the degree of agreement based on a standardized value, but the agreement values may not be compared across very different populations, and sometime artificially high or low agreement values may be obtained due to the dependence of these indices (except the CIA) on between-subject variability.

In detail, the scaled agreement indices ICC, CCC, CIA are all standardized to have values between -1 and 1 (for CIA, it's 0 to 1). The ICC, CCC are related and depend on between-subject variability, hence they may produce high values for heterogeneous populations (Atkinson and Nevill (1997)). The inflation of CCC when we increase the between subjects variability was demonstrated in Chapter 5. The ICC can accommodate multiple fixed and random observers, but is not suited for cases with a reference observer or for data with repeated measures, without additional assumptions. The CCC is mainly used for fixed observers and reduces to the ICC for fixed observers under additional assumptions. The CCC formulation may be used for the case with random observers (Chen and Barnhart 2008), but additional care is needed for inference. The CIA is fundamentally different from the ICC, CCC because it uses within-subject variability, rather than between-subject variability, as the scaling factor. It is possible to have high ICC/CCC value and low CIA value (and vice versa) from the same data set.

This dissertation focuses on a new coefficient, the coefficient of individual equivalence (CIE), to assess agreement for binary and continuous outcomes. Both CIA and CIE are fundamentally different from ICC and CCC because CIA uses the within-subject, rather than between-subject, variability as the scaling factor. CIE, on the other hand, uses the expected disagreement under individual equivalence as the scaling factor, while sharing the same denominator as the CIA. Under individual equivalence, the conditional distribution of two observers are equal for each subject. The approach underlying the CIA's has two limitations: (a) it requires that the within-observer disagreement, which is used as a baseline for "good agreement", is accept-

able, and (b) the CIA comparing an observer to a reference or a gold standard, i.e. ψ^R, cannot be estimated when replicated readings on the reference observer are unavailable (as is the case when the reference value of each subject can be determined without an error). The CIE does not share these limitations. Similar to CIA, CIE can also be extended to repeated outcomes. In this dissertation, CIE has been extended to assess agreement on repeated binary outcomes.

Possible future research based on my dissertation are as follows: (1) The current approach for defining and estimating CIE is based on permutations. One may consider more gerneral approaches based on the requirement that the conditional distributions of the observers will be equal. (2) Currently, we only consider the fixed observer case through the 2-way linear mixed model. It is also possible to consider the case when the observer effect is treated as random. (3)It is useful to extend this index when there are multiple observers. (4) It is also useful to extend CIE to nominal and ordinal outcomes. (5) For continuous outcomes, the possibility of ranking the outcomes and performing the estimation using the ranks instead of the real measurements, can be explored. This method may provides more robust results. (6) Consider agreement studies with missing data. (7) Indices for continuous data with repeated measurements, censoring, outliers, and covariates. (8) Sample size calculation for design of agreement study. (9) How to better implement Bayesian approach to observer agreement studies with replicated and repeated measurements. (10) Whether it is possible to estimate CIE using one replication per rater based on random effects models, and (11) compare the estimated CIEA and CIA with respect to their root mean square error when the number of replications is not the same for different observers.

Bibliography

Agresti, A. (1989): An agreement model with kappa as parameter. *Statistics and Probability Letters* **7**:271–273.

Ahmed, M. and Shoukri, M. (2010): A Bayesian Estimator of the Intracluster Correlation Coefficient from Correlated Binary Responses. *Journal of Data Science* **8**:127–137.

Anderson, D. and Aitkin, M. (1985): Variance Components Models with Binary Response: Interviewer Variability. *Journal of the Royal Statistical Society. Series B (Methodological)* **47**:203–210.

Anderson, S. and Hauck, W. W. (1990): Considerations of individual bioequivalence. *Statistics in Medicine* **18**:259–273.

Atkinson, G. and Nevill, A. (1997): Comment on the use of concordance correlation coefficient to assess the agreement between two variables. *Biometrics* **53**:775–777.

Banerjee, M. (1999): Beyong Kappa: A review of interrater agreement measures. *The Canadian Journal of Statistics* **27**:3–23.

Barnhart, H. X., Haber, M. and Kosinski, A. S. (2007a): Assessing individual agreement. *Journal of Biopharmaceutical Statistics* **17**(4):697–719.

Barnhart, H. X., Haber, M. and Lin, L. I. (2007b): An overview on assessing

agreement with continuous measurements. *Journal of Biopharmaceutical Statistics* **17**(4):529–569.

Barnhart, H. X., Haber, M., Lokhnygina, Y. and Kosinski, A. S. (2007c): Comparison of concordance correlation coefficient and coefficient of individual agreement in assessing agreement. *Journal of Biopharmaceutical Statistics* **17**(4):721–738.

Barnhart, H. X., Haber, M. and Song, J. (2002): Overall concordance correlation coefficient for evaluating agreement among multiple observers. *Biometrics* **58**:1020–1027.

Barnhart, H. X., Song, J. and Haber, M. (2005): Assessing intra, inter, and total agreement with replicated measurements. *Statistics in Medicine* **24**:1371–1384.

Barnhart, H. X. and Williamson, J. M. (2002): Weighted least-squares approach for comparing correlated kappa. *Biometrics* **58**:1012–1019.

Bartko, J. J. (1966): The intraclass correlation coefficient as a measure of reliability. *Psychological Reports* **19**:3–11.

Bartko, J. J. (1974): Corrective note to the intraclass correlation coefficient as a measure of reliability. *Psychological Reports* **34**:418.

Basu, S., Banerjee, M. and Sen, A. (2000): Bayesian Inference for Kappa from single and multiple studies. *Biometrics* **56**:577–582.

Bland, J. M. and Altman, D. G. (1999): Measuring agreement in method comparison studies. *Statistical Methods in Medical Research* **8**:135–160.

Bland, J. M. and Altman, D. G. (2007): Agreement between methods of measurement with multiple observers per individual. *Journal of Biopharmaceutical Statistics* **17**:571–582.

Branscum, A. J., Gardner, I. A., Wagner, B. A., McInturff, P. and Salman, M. (2005): Effect of diagnostic testing error on intracluster correlation coefficient estimation. *Preventive Veterinary Medicine* **69**:63–75.

Broemeling, L. D. (2009): *Bayesian Methods for Measures of Agreement*. CRC Press.

Carrasco, J. and Jover, L. (2003): Estimation of the Generalized Concordance Correlation Coefficient through Variance Components. *Biometrics* **59**:849–858.

Carrasco, J., King, T. and Chinchilli, V. (2009): The Concordance Correlation Coefficient for Repeated Measures Estimated by Variance Components. *Journal of Biopharmaceutical Statistics* **19**:90–105.

Chen, C. and Barnhart, H. X. (2008): Comprison of ICC and CCC for assessing agreement for data without and with replications. *Computational Statistics and Data Analysis* **53**:554–564.

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

Cohen, J. (1968): Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* **70**:213–220.

Coughlin, S., Pickle, L., Goodman, M. and Wilkens, L. (1992): The Logistic Modeling of Interobserver Agreement. *Journal of Clinical Epidemiology* **45**:1237–1241.

Dawid, A. P. and Skene, A. M. (1979): Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied. Statist* **28**:20–28.

Donner, A., Shoukri, M. M., Klar, N. and Bartfay, E. (2000): Testing the equality of two dependent kappa statistics. *Statistics in Medicine* **19**:373–387.

Dunn, G. (2004): *Statistical Evaluation of Measurement Errors: Design and Analysis of Reliability Studies, second edition*. Oxford University Press Inc.

Eliasziw, M., Young, S. L., Woodbury, M. G. and Fryday-Field, K. (1994): Statistical methodology for the concurrent assessment of interrater and intrarater reliability: Using goniometric measurements as an example. *Physical Therapy* **74**:777–788.

Elmore, J. G., Wells, C. K., Lee, C. H., Howard, D. H. and Feinstein, A. (1994): Variability in radiologists' interpretation of mammograms. *New England Journal of Medicine* **331**:1493–1499.

Everitt, B. S. (1968): Moments of the statistics kappa and weighted kappa. *British Journal of Mathematical and Statistical Psychology* **21**:97–103.

Feinstein, A. R. and Cicchetti, D. V. (1990): High agreement but low kappa. *Journal of Clinical Epidemiology* **43**:543–558.

Fleiss, J. (1981): *Statistical methods for rates and proportions (2nd ed.)*. New York: Wiley.

Fleiss, J. L. (1971): Measuring nominal scale agreemetn among many raters. *Psychological Bulletin* **76**:378–382.

Fleiss, J. L. and Cohen, J. (1973): The equivalence of the weighted kappa and the intraclass correlation coefficient as a measure of reliability. *Educational and Psychological Measurements* **33**:613–619.

Fleiss, J. L., Cohen, J. and Everitt, B. S. (1969): Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin* **72**:323–327.

Fleiss, J. L., Nee, J. C. M. and Landies, J. R. (1979): Large sample variance of kappa in the case of different sets of raters. *Psychological Bulletin* **86**:974–977.

Gajewski, B. J., Hart, S., Bergquist-Beringer, S. and Dunton, N. (2007): Inter-rater reliability of pressure ulcer staging: Ordinal probit Bayesian hierarchical model that allows for uncertain rater response. *Statistics in Medicine* **26**:4602–4618.

Gao, J. J. (2010): Assessing Observer Agreement for Categorical Observations. *Ph.D. Dissertation* .

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004): *Bayesian Data Analysis*. CRC Press.

Guo, Y. (2004): Assessing Agreement for Survival Outcomes. *Ph.D. Dissertation* .

Guo, Y. and Manatunga, A. (2005): Modeling the Agreement of Discrete Bivariate Survival Times Using Kappa Coefficient. *Lifetime Data Analysis* **11**:309–332.

Guo, Y. and Manatunga, A. (2009): Measuring Agreement of Multivariate Discrete Survival Times Using a Modified Weighted Kappa Coefficient. *Biometrics* **65**:125–134.

Haber, M. and Barnhart, H. X. (2006): Coefficients of agreement for fixed observers. *Statistical Methods in Medical Research* **15**:1–17.

Haber, M. and Barnhart, H. X. (2008): A general approach to evaluating agreement between two observers or methods of measurement. *Statistical Methods in Medical Rsearch* **17**:151–169.

Haber, M., Barnhart, H. X., Song, J. and Gruden, J. (2005): Observer variability: a new approach in evaluating interobserver agreement. *Journal of Data Science* **3**:69–83.

Haber, M., Gao, J. and Barnhart, H. X. (2007): Assessing observer agreement in studies involving replicated binary observations. *Journal of Biopharmaceutical Statistics* **17**(4):757–766.

Haber, M., Gao, J. and Barnhart, H. X. (2010): Evaluation of Agreement between Measurement Methods from Data with Matched Repeated Measurements via the Coefficient of Individual Agreement. *Journal of Data Science* **8**:457–469.

Haggard, A. E. (1958): *Intraclass correlation and the analysis of variance*. New York: The Dryden Press, INC.

Hawkins, D. (2002): Diagnostics for conformity of paired quantitative measurements. *Statistics in Medicine* **21**:1913–1935.

Hutchison, T. (1993): Kappa muddlees together two sources of disagreement: Tetrachoric correlation is preferable. *Res. Nursing and Health* **16**:313–315.

Jason, H. and Olsson, U. (2001): A measure of agreement for interval or nominal multivariate observations. *Educational and Psychological Measurement* **61**:277–289.

Jason, H. and Olsson, U. (2004): A measure of agreement for interval or nominal multivariate observations by different sets of judges. *Educational and Psychological Measurement* **64**:62–70.

Johnson, N. L. and Kotz, S. (1970): *Distributions in statistics. Continuous univariate distributions*. New York: Wiley.

King, T. S. and Chinchilli, V. M. (2001a): A generalized concordance correlation coefficient for continuous and categorical data. *Statistics in Medicine* **20**:2131–2174.

King, T. S. and Chinchilli, V. M. (2001b): Robust estimators of the concordance correlation coefficients. *Journal of Biopharmaceutical Statistics* **11**:83–105.

King, T. S., Chinchilli, V. M. and Carrasco, J. L. (2007a): A repeated measures concordance correlation coefficient. *Statistics in Medicine* **16**:3096–3113.

King, T. S., Chinchilli, V. M., Carrasco, J. L. and Wang, K. (2007b): A class of repeated measures concordance correlation coefficients. *Journal of Biopharmaceutical Statistics* **17**(4):653–672.

Klar, N., R., L. S. and Ibrahim, J. G. (2000): An Estimating Equations Approach for Modeling Kappa. *Biometrical Journal* **45**:45–48.

Konishi, S., Khatri, C. G. and Rao, C. R. (1991): Inferences on multivariate measures of interclass and intraclass correlations in familiar data. *Journal of the Royal Statistical Society Series B* **53**:649–659.

Kraemer, H. C. (1979): Ramifications of a population model for kappa as a coefficient of reliability. *Psychometrika* **44**:461–472.

Landis, J. R. and Koch, G. G. (1977): The Measurement of Observer Agreement for Categorical Data. *Biometrics* **33**:159–174.

Liao, J. Z. and Lewis, W. J. (2000): A note on concordance correlation coefficient. *PDA Journal of Pharmaceutical Science and Technology* **54**:23–26.

Lin, L. I. (1989): A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **45**:255–268.

Lin, L. I. (1992): Assay validation using the concordance correlation coefficient. *Biometrics* **48**:599–604.

Lin, L. I. (2000a): A note on the concordance correlation coefficients. *Biometrics* **56**:324–325.

Lin, L. I., Hedayat, A. S., Sinha, B. and Yang, M. (2002): Statistical methods in assessing agreement: models, issues and tools. *Journal of American Statistical Association* **97**:257–270.

Lin, L. I., Hedayat, A. S. and Wenting, W. (2007): A unified approach for assessing agreement for continuous and categorical data. *Journal of Biopharmaceutical Statistics* **17**(4):629–652.

Lin, X. and Breslow, N. (1996): Bias Correction in Generalized Linear Mixed Models with Multiple Components of Dispersion. *Journal of American Statistical Association* **91**:1007–1016.

Lipsitz, S., Parzen, M., Fitzmaurice, G. and Klar, N. (2003): A Two-Stage Logistic Regression Model for Analyzing Inter-Rater Agreement. *Psychometrika* **68**:289–298.

Ma, Y., Tang, W., Feng, C. and Tu, X. (2008): Inference for Kappas for Longitudinal Study Data: Applications to Sexual Health Research. *Biometrics* **64**:781–789.

McCulloch, C. E. (1997): Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association* **92**:162–170.

McGraw, K. O. and Wong, S. P. (1996): Forming inferences about some intraclass correlation coefficients. *Psychological Methods* **1**:30–46.

Molenberghs, G., Vangeneugden, T. and Laenen, A. (2007): Estimating reliability and generalizability from hierarchical biomedical data. *Journal of Biopharmaceutical Statistics* **17**(4):595–627.

Muller, R. and Petra, B. (1994): A critical discussion of intraclass correlation coefficient. *Statistics in Medicine* **13**:2465–2476.

Nelson, K. and Edwards, D. (2007): A Model and Measure of Agreement for Population-based Studies **64**:781–789.

Pan, Y., Gao, J. J., Haber, M. and Barnhart, H. X. (2010): Estimation of Coefficients of Individual Agreement (CIAs) for Quantitative and Binary Data using SAS and R. *Computer Methods and Programs in Biomedicine* **98**:214–219.

Pan, Y., Haber, M., Barnhart, H. and Gao, J. (2011): A new permutation-based

method for assessing agreement between two observers making replicated binary readings. *Statistics in Medicine* **30**.

Quiroz, J. (2005): Assessment of equivalence using a concordance correlation coefficient in a repeated measurement design. *Journal of Biopharmaceutical Statistics* **15**:913–928.

Ridout, M., Demetrio, C. and Firth, D. (1999): Estimating Intraclass Correlation for Binary Data. *Biometrics* **55**:137–148.

Schall, R. and Luus, H. G. (1993): On population and individual bioequivalence. *Statistics in Medicine* **12**:1109–1124.

Scott, W. (1955): Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quart.* **19**:321–325.

Shao, J. and Zhong, B. (2004): Assessing the agreement between two quantitative assays with repeated measurements. *Journal of Biopharmaceutical Statistics* **14**:201–212.

Shoukri, M. M. (2004): *Measures of interobserver agreement*. Chapman and Hall.

Shrout, P. E. and Fleiss, J. L. (1979): Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin* **86**:420–428.

Sim, J. and Wright, C. C. (2005): The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical Therapy* **85**:257–268.

Song, J. L. (2003): Assessing Agreement/Association for Continuous Measurement Scales. *Ph.D. Dissertation* .

Thompson, W. D. and Walter, S. D. (1988): A reappraisal of the kappa coefficient. *Journal of Clinical Epidemiology* **41**:949–958.

Turner, B., Omar, R. Z. and G., T. S. (2001): Bayesian methods of analysis for cluster randomized trials with binary outcome data. *Statistics in Medicine* **20**:453–472.

Turner, B. M., Omar, R. Z. and G., T. S. (2006): Constructing intervals for the intracluster correlation coefficient using Bayesian modelling, and application in cluster randomized trials. *Statistics in Medicine* **25**:1443–1456.

Vangeneugden, T., Laenen, A., Geys, H., Renard, D. and Molenberghs, G. (2004): Applying linear mixed model to estimate reliability in clinical trials with repeated measurements. *Controlled Clinical Trials* **25**:13–30.

Vangeneugden, T., Laenen, A., Geys, H., Renard, D. and Molenberghs, G. (2005): Applying the concepts of generalizability therory on clinical trial data to investigate sources of variation and their impact on reliability. *Biometrics* **61**:295–304.

Von Eye, A. and Young Mun, E. (2005): *Analyzing Rater Agreement Manifest Variable Methods*. Lawrence Erlbaum Associates.

Wiener, J. (2009): Evaluating Agreement Among Observers or Methods of Measurements for Quantitative Data. *Ph.D. Dissertation* .

Williamson, J. M., Manatunga, A. and Lipsitz, S. (2000): Modeling kappa for measuring dependent categorical agreement data. *Biostatistics* **1**:191–202.

Zeger, S., Liang, K. Y. and Albert, P. (1988): Models for Longitudinal Data: A Generalized Estimating Equation Approach. *Biometrics* **44**:1049–1060.