The Pennsylvania State University

The Graduate School

Department of Statistics

A GENERAL CLASS OF AGREEMENT COEFFICIENTS FOR CATEGORICAL AND CONTINUOUS RESPONSES

A Dissertation in

Statistics

by

Wei Zhang

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

August 2008

The dissertation of Wei Zhang was read and approved* by the following:

Vernon M. Chinchilli
Distinguished Professor of Public Health Sciences and Statistics
Dissertation Adviser
Chair of Committee

Donald St. P. Richards
Professor of Statistics

Steven F. Arnold
Professor of Statistics

Tonya S. King
Associate Professor of Public Health Sciences

Damla Senturk
Assistant Professor of Statistics

Peter C. M. Molenaar
Professor of Human Development and Family Studies

Bruce G. Lindsay
Willaman Professor of Statistics
Head of the Department of Statistics

* Signatures on file in the Graduate School.

Abstract

We propose a general class of agreement coefficients for categorical and continuous responses. An agreement coefficient is used to measure the interrater agreement. Motivated by the traditional Cohen's kappa, concordance correlation coefficient (CCC), and the recent random marginal agreement coefficient (RMAC), we formulate this task using a parameter a, which reflects the distance between marginal distributions. Our approach generalizes Cohen's kappa as the upper bound and RMAC as the lower bound for categorical data, and generalizes Lin's CCC as the upper bound and RMAC as the lower bound for continuous data, in a class of appropriate measurements of interrater agreement based on the discrepancy of marginal distributions. We study the large sample properties for the estimators of members of this class and conduct simulation studies to assess and compare the accuracy and precision of the estimators. Some real data examples are also discussed to demonstrate their use.

Table of Contents

List of Tables

List of Figures

Acknowledgments

I would like to gratefully and sincerely thank my dissertation adviser Professor Vernon M. Chinchilli, for his guidance, understanding, encouragement, and patience during my dissertation research. His mentorship was paramount in not only giving valuable feedback and suggestions to my research, but also providing me a well-rounded experience from which I benefit a lot to grow as a statistician.

I would also like to thank Professor Steven F. Arnold, Professor Donald St. P. Richards, Professor Damla Senturk, Professor Tonya S. King, and Professor Peter C. M. Molenaar, the committee members of my dissertation committee, for their valuable comments and suggestions. Special thanks go to Professor Donald St. P. Richards and Professor Damla Senturk, for devoting their precious time to discuss and solve theoretical problems with me.

I would like to thank the Department of Statistics at Penn State University, especially those professors who taught me and colleagues I worked with, for their input, valuable discussions and accessibility. In particular, I would like to thank Professor Naomi S. Altman, Professor Thomas P. Hettmansperger, Professor Arkady A. Tempelman, Jingyun Yang and Sasiprapa Hiriote for their expertise and hard work.

Finally, I would like to thank my wife Rong Liu. Her support, encouragement, quiet patience and unwavering love were undeniably the bedrock upon which the past five years of my life have been built. I thank my parents, Guohua Zhang and Ping Wang, for their faith in me and allowing me to be as ambitious as I wanted. I thank my daughter, Jasmine Zhang, for bringing so much happy laughter during my graduate study.

Dedicated to my parents, Guohua Zhang and Ping Wang,
my wife, Rong Liu and my daughter Jasmine Zhang.

Chapter 1

Literature Review

In medical and social science research, analysis of observer or interrater agreement data is often used in research. Researchers are interested in the agreement between two raters or two methods of measuring a response. Subjects are usually being classified more than once, by more than one rater or at different time points, and outcomes are recorded on a nominal, ordinal or continuous scale. This can be summarized into the following three cases: (1) two observers evaluate the experimental units in a study, and it is of interest to measure how well these observers agree. (2) in the biomedical sciences, it is of interest to assess the amount of agreement between two distinct methods of measuring the same response variable, or the same method measuring the variable at different time points. (3) when a new assay or instrument is developed, question arises as to whether the new assay or instrument can reproduce the results of a traditional gold-standard.

In this chapter, we review various approaches to the study of interrater agreement, for which the relevant data comprise nominal, ordinal or continuous scales. These approaches are summarized into three structures: kappa agreement coefficient, concordance correlation coefficient, and random marginal agreement coefficient.

1.1 Kappa Agreement Coefficient

1.1.1 Cohen's Kappa

Suppose that each of a sample of n subjects is rated independently into one of $C + 1$ mutually exclusive and exhaustive nominal or ordinal categories, by the same two raters. We will express these measures in terms of estimates of probabilities (i.e. sample proportions), as they were originally proposed, throughout this section. Let $\hat{p}_{ij}$ be the proportion of subjects that were classified into (i, j) cell, i.e. classified into the ith category by the first rater and into the jth category by the second rater, $i, j = 0, 1, ..., C$. Further, let $\hat{p}_{i.} = \sum_{j=0}^{C} \hat{p}_{ij}$ and $\hat{p}_{.j} = \sum_{i=0}^{C} \hat{p}_{ij}$.

The simplest and most primitive index of interrater agreement is the overall proportion of agreement, say

$$\hat{p}_o = \sum_{i=0}^{C} \hat{p}_{ii}$$

This approach is not adequate, since it does not account for the fact that a certain amount of agreement is to be expected by chance alone. Other early approaches to assess agreement were a chi-square statistic (χ^2) as a test statistic of the hypothesis of chance agreement and a contingency coefficient (C) as a measure of degree of agreement [25]. Again, these measures are improperly used for measuring agreement. When applied to a contingency table, chi-square and contingency coefficient (which is based on chi-square) measure the degree of association, not agreement. Therefore, these measures will be inflated quite impartially by any departure from chance association, either disagreement or agreement [9].

Considering nominal categories, Scott [35] proposed a chance-corrected agreement measure,

$$\pi = \frac{\hat{p}_o - \hat{p}_e}{1 - \hat{p}_e}$$

where $\hat{p}_e = \sum_{i=0}^{C} \hat{p}_{i.} \hat{p}_{.i}$, which represents the proportion of agreement expected by chance.

This measure corrects for chance by removing chance agreement from consideration. It is defined under the assumption that the distribution of proportions over the categories for the population is known and is equal for the two raters. The former assumption is reasonable in practice, but the latter assumption does not seem to be appropriate in some circumstances, since disagreement is partially due to the discrepancy in distributing raters' judgements over the categories.

Cohen [9] extended Scott's measure, and it is known as Cohen's kappa,

$$\hat{\kappa} = \frac{\hat{p}_o - \hat{p}_e}{1 - \hat{p}_e}$$

It has the same form as Scott's measure, except it does not assume the distribution of proportions over the categories for the population is equal for the two raters. Therefore, if the two raters are interchangeable, in the sense that the marginal distributions are the same, then Cohen's and Scott's measures are equivalent.

Cohen's kappa has desirable properties. If there is a complete agreement, then $\hat{\kappa} = +1$. If observed agreement is greater than or equal to chance agreement, $\hat{\kappa} \geq 0$, and if observed agreement is less than or equal to chance agreement, $-1 \leq \hat{\kappa} \leq 0$.

For testing the hypothesis that κ is equal to zero, i.e. all the observed agreement is due to chance, one could use the estimated large-sample standard error of kappa shown by Fleiss, Cohen and Everitt [17],

$$\hat{se}_0(\hat{\kappa}) = \frac{1}{(1-\hat{p}_e)\sqrt{n}}\sqrt{\hat{p}_e + \hat{p}_e^2 - \sum_{i=0}^{C}\hat{p}_{i.}\hat{p}_{.i}(\hat{p}_{i.} + \hat{p}_{.i})}$$

To test the hypothesis against the alternative that agreement is better than chance, one could refer the quantity

$$z = \frac{\hat{\kappa}}{\hat{se}(\hat{\kappa})}$$

to the standard normal distribution and reject the hypothesis if z is sufficiently large.

To set confidence limit of kappa, one needs to use the estimated large-sample standard error of kappa shown by Fleiss, Cohen and Everitt [17],

$$\hat{se}(\hat{\kappa}) = \frac{\sqrt{A+B-C}}{(1-\hat{p}_e)\sqrt{n}}$$

where

$$A = \sum_{i=1}^{k}\hat{p}_{ii}[1-(\hat{p}_{i.}+\hat{p}_{.i})(1-\hat{\kappa})]^2$$

$$B = (1-\hat{\kappa})^2\sum\sum_{i\neq j}\hat{p}_{ij}(\hat{p}_{.i}+\hat{p}_{j.})^2$$

$$C = [\hat{\kappa} - \hat{p}_e(1-\hat{\kappa})]^2$$

An approximate $100(1-\alpha)\%$ confidence interval for κ is

$$\hat{\kappa} - z_{\alpha/2}\hat{se}(\hat{\kappa}) \leq \kappa \leq \hat{\kappa} + z_{\alpha/2}\hat{se}(\hat{\kappa})$$

Landis and Koch [29] have characterized different ranges of values for kappa with respect to the degree of agreement they suggest. For most purposes, $\hat{\kappa} \geq 0.75$ may be taken to represent excellent agreement beyond chance; $0.4 < \hat{\kappa} < 0.75$ may be taken to represent fair to good agreement beyond chance; and $\hat{\kappa} \leq 0.4$ may be taken to represent poor agreement beyond chance.

However, many researchers have pointed out that Cohen's kappa yields non-intuitive results when the marginal distributions are very different. In particular, its dependence on the marginal distributions makes it difficult to use and interpret under certain circumstances.

For example, Feinstein and Cicchetti [15] identified two paradoxes associated with kappa:

- Paradox 1: if p_e is large, the chance correction process can convert a relatively high value of p_o into a relatively low value of κ.
- Paradox 2: Unbalanced marginal totals produce higher values of κ than more balanced totals.

Byrt et al. [7] discussed further issues about bias and prevalence that cause these paradoxes. "Bias" between the raters refers to the difference between two raters in their assessment of the frequency of occurrence of a condition in a study group. When this occurs the marginal distributions for the raters are unequal. "Prevalence" is defined as the proportions of cases of the various types in the population.

Researchers have pointed out the unsatisfactory features about kappa, which includes: (i) difficult interpretation and comparison of a single coefficient of agreement reported, (ii) dependence on the marginal distributions and (iii) misleading result in reporting kappa values alone when comparisons are made between agreement studies. These suggest that no single omnibus index of agreement can be satisfactory for all purposes.

In practice, as Zwick [39] recommended, rather than ignoring marginal disagreement or attempting to correct for it, researchers should be studying it to determine whether it reflects important rater differences or merely a random error. She proposed that one should begin the assessment of rater agreement with the investigation of marginal homogeneity, if there is no significant difference use Cohen's kappa; otherwise one can explain the degree of agreement between raters in terms of the discrepancies between marginal distributions.

1.1.2 Weighted Kappa

The motivation of using weighted kappa is that some disagreements are of greater importance than others. For example, as Cohen [10] provided in an assessment of the reliability of psychiatric diagnosis in the categories personality disorder, neurosis and psychosis, a clinician would likely consider a diagnostic disagreement between neurosis and psychosis to be more serious than between neurosis and personality disorder. Cohen's kappa does not make such distinction, as it treats all disagreements equally seriously.

Cohen [10] proposed a weighted kappa to measure the proportion of weighted agreement corrected by chance. Again, assume that each of a sample of n subjects is rated independently into one of $C+1$ mutually exclusive and exhaustive nominal or ordinal categories, with $\hat{p}_{ij}$ as the proportion of subjects that are classified into (i, j)

cell. Agreement weights, say $w_{ij}, i, j = 0, ..., C$, are assigned to the $(C+1)^2$ cells. The range of weights is $0 \leq w_{ij} \leq 1$ such that $w_{ij} = 1$ for $i = j$, $0 \leq w_{ij} < 1$ for $i \neq j$ and $w_{ij} = w_{ji}$. The observed weighted proportion of agreement is defined to be

$$\hat{p}_{o(w)} = \sum_{i=0}^{C} \sum_{j=0}^{C} w_{ij} \hat{p}_{ij}$$

and the chance-expected weighted proportion of agreement is defined to be

$$\hat{p}_{e(w)} = \sum_{i=0}^{C} \sum_{j=0}^{C} w_{ij} \hat{p}_{i.} \hat{p}_{.j}$$

Weighted kappa is then given by

$$\hat{\kappa}_w = \frac{\hat{p}_{o(w)} - \hat{p}_{e(w)}}{1 - \hat{p}_{e(w)}}$$

Cohen's kappa is a special case of weighted kappa, when $w_{ij} = 0$, for all $i \neq j$, as weighted kappa is identical to Cohen's kappa if all disagreements are assigned the same weight. The interpretation of the magnitude of weighted kappa is like that of unweighted kappa.

In particular, suppose the $(C+1)^2$ categories are ordered. Weighted kappa is capable of accounting for severity of discordance or size of the discrepancy. However, as Maclure and Willett [33] pointed out, since the magnitude of weighted kappa is greatly influenced by the relative magnitude of its weights, some standardization of weights should be used so that the index of agreement is interpretable.

Two commonly used weighting scheme are Fleiss-Cohen weights

$$w_{ij} = 1 - \frac{(i-j)^2}{C^2}$$

and Cicchetti-Allison weights

$$w_{ij} = 1 - \frac{|i-j|}{C}$$

The sampling distribution of weighted kappa was derived by Fleiss et al. [17] as

$$\hat{se}_0(\hat{\kappa}_w) = \frac{1}{(1-\hat{p}_{e(w)})\sqrt{n}} \sqrt{\sum_{i=0}^{C}\sum_{j=0}^{C} \hat{p}_{i.}\hat{p}_{.j}[w_{ij} - (\bar{w}_{i.} + \bar{w}_{.j})]^2 - \hat{p}^2_{e(w)}}$$

where $\bar{w}_{i.} = \sum_{j=0}^{C} \hat{p}_{.j} w_{ij}$ and $\bar{w}_{.j} = \sum_{i=0}^{C} \hat{p}_{i.} w_{ij}$.

1.1.3 Intraclass Kappa

Bloch and Kraemer [5] introduced intraclass kappa in the context of agreement. Context of agreement assumes that the responses of the m ($m \geq 2$) ratings on a subject are interchangeable (i.e. no rater bias), the raters are being asked the same question and the ratings are said to be in agreement if and only if they are equal.

The intraclass kappa was introduced in the special case of independent blinded dichotomous ratings (success and failure) on each subject by two fixed raters. Let X_{ij} denote the rating for the ith subject assigned by the jth rater, $i = 1, 2, ..., n$, $j = 1, 2$. Then for each subject i, over ratings, let $P(X_{ij} = 1) = p_i$ be the probability that the rating for the ith subject is a success, and $p'_i = 1 - p_i$. Over the population of subjects, let $E(p_i) = P$ and $Var(p_i) = \sigma_P^2$. Then the intraclass kappa is defined as

$$\kappa_I = \frac{\sigma_P^2}{PP'}$$

where $P' = 1 - P$.

Bloch and Kraemer [5] recommended the use of intraclass kappa for any situation in which subjects are independently rated using the same rating instrument in order to assess the reliability or reproducibility of the instrument or its users.

The theoretical model for this case is shown in Table 1.1.

Table 1.1 Theoretical model for 2 × 2 data: model for agreement

	Rater B		
Rater A	Success	Failure	Total
Success	$E(p_i^2)$	$E(p_i p_i')$	P
Failure	$E(p_i p_i')$	$E(p_i'^2)$	P'
Total	P	P'	1

The probability of agreement between two raters is $p_i^2 + p_i'^2$. Over the population of subjects, the expected agreement is

$$E(p_i^2 + p_i'^2) = 2\sigma_P^2 + P^2 + P'^2$$

Agreement is random if the probability of agreement is $P^2 + P'^2$. Then the kappa agreement coefficient defined as

$$\frac{\text{Expected agreement-Random agreement}}{\text{Maximum expected agreement-Random agreement}} = \frac{\sigma_P^2}{PP'}$$

is equivalent to the intraclass kappa defined above.

In the context of agreement, Bloch and Kreamer [5] derived the maximum likelihood estimator (MLE) of P and κ_I. Suppose that in a sample of n subjects, the observed frequencies of responses are shown in Table 1.2.

Table 1.2 Frequencies of responses for 2 × 2 data

	Rater B		
Rater A	Success	Failure	Total
Success	n_1	n_2	$n_1 + n_2$
Failure	n_3	n_4	$n_3 + n_4$
Total	$n_1 + n_3$	$n_2 + n_4$	n

And the model given by Table 1.1 is equivalent to Table 1.3:

Table 1.3 Expected probability of joint responses for 2 × 2 data

	Rater B		
Rater A	Success	Failure	Total
Success	$P^2 + \kappa_I PP'$	$PP'(1 - \kappa_I)$	P
Failure	$PP'(1 - \kappa_I)$	$P'^2 + \kappa_I PP'$	P'
Total	P	P'	1

Then the log-likelihood function is

$$\begin{aligned}&\ln L(P, \kappa_I | n_1, n_2, n_3, n_4)\\ &= n_1 \ln(P^2 + \kappa_I PP') + (n_2 + n_3) \ln[PP'(1 - \kappa_I)] + n_4 \ln(P'^2) + \kappa_I PP'\end{aligned}$$

Thus the MLEs of P and κ_I are

$$\hat{p} = \frac{2n_1 + n_2 + n_3}{2n}$$

$$\hat{\kappa}_I = \frac{4(n_1 n_4 - n_2 n_3) - (n_2 - n_3)^2}{(2n_1 + n_2 + n_3)(2n_4 + n_2 + n_3)} \tag{1.1}$$

If $p_0 = (n_1 + n_4)/n$, $p_c = \hat{p}^2 + \hat{p}'^2$ and $p_{\max} = 1$, then the right-hand side of equation (1.1) is of the form $(p_0 - p_c)/(p_{\max} - p_c)$. Therefore, the MLE of κ_I is algebraically equivalent to the index of agreement proposed by Scott [35] as a measure of agreement between two raters when their underlying marginal distributions are the same.

By using the result due to Fisher [16], the asymptotic variance for $\hat{\kappa}_I$ is given by ([5]):

$$\lim_{n\to\infty} n\mathrm{var}(\hat{\kappa}_I) = (1 - \kappa_I)[(1 - \kappa_I)(1 - 2\kappa_I) + \frac{\kappa_I(2 - \kappa_I)}{2PP'}]$$

A variance-stabilizing transformation for $\hat{\kappa}_I$, derived by Bloch and Kraemer [5], provides improved accuracy for confidence interval calculation and power of tests.

A jackknife estimator of κ_I, $\hat{\kappa}_J$, obtained by averaging the estimators $\hat{\kappa}_{-i}$, where $\hat{\kappa}_{-i}$ is the value of $\hat{\kappa}_I$ obtained over all subjects except the ith, was also proposed ([5]). It is verified that the asymptotic variances for $\hat{\kappa}_I$ and $\hat{\kappa}_J$ are equal.

However, as the authors pointed out, for small samples, the asymptotic normal approximation may yield inaccurate results when P is near 0 or 1, and when κ_I is near 0 or 1, because of the occurrence of degenerate ($\hat{\kappa}_I$ undefined) or extreme values of $\hat{\kappa}_I$. Nonetheless, based on the simulation study, the authors recommended the use of the jackknife estimator of κ_I for small samples for its smaller bias and more accurate sampling variance.

Donner and Eliasziw [12] proposed an approach for constructing a confidence interval that has more accurate coverage levels in samples of small sample size than the methods discussed above. This approach is based on a chi-square goodness-of-fit test as applied to a model frequently used for clustered binary data.

Suppose the model is shown in Table 1.2. A $100(1-\alpha)\%$ confidence interval about $\hat{\kappa}_I$ is $(\hat{\kappa}^2_L, \hat{\kappa}^2_U)$, where

$$\hat{\kappa}_L = (\frac{1}{9}y_3^2 - \frac{1}{3}y_2)^{\frac{1}{2}}[\cos(\frac{\theta + 2\pi}{3}) + \sqrt{3}\sin(\frac{\theta + 2\pi}{3})] - \frac{1}{3}y_3$$

$$\hat{\kappa}_U = 2(\frac{1}{9}y_3^2 - \frac{1}{3}y_2)^{\frac{1}{2}}[\cos(\frac{\theta + 5\pi}{3})] - \frac{1}{3}y_3$$

where $\theta = \arccos(\frac{V}{W})$; $V = \frac{1}{27}y_3^3 - \frac{1}{6}(y_2y_3 - 3y_1)$; $W = (\frac{1}{9}y_3^2 - \frac{1}{3}y_2)^{\frac{3}{2}}$ and

$$y_1 = \frac{[n_2 - 2n\hat{p}(1-\hat{p})]^2 + 4n^2\hat{p}^2(1-\hat{p})^2}{4n\hat{p}^2(1-\hat{p})^2(\chi^2_{1,1-\alpha} + n)} - 1$$

$$y_2 = \frac{n_2^2 - 4n\hat{p}(1-\hat{p})[1 - 4\hat{p}(1-\hat{p})]\chi^2_{1,1-\alpha}}{4n\hat{p}^2(1-\hat{p})^2(\chi^2_{1,1-\alpha} + n)} - 1$$

$$y_3 = \frac{n_2 + [1 - 2\hat{p}(1-\hat{p})]\chi^2_{1,1-\alpha}}{\hat{p}(1-\hat{p})(\chi^2_{1,1-\alpha} + n)} - 1$$

The simulation study showed that the goodness-of-fit procedure provides improved coverage levels across almost all values of κ_I and P for samples as few as 25 subjects. This procedure can also be used for hypothesis testing and sample size calculation.

1.1.4 Multiple Ratings per Subject with Different Raters

Fleiss [18] proposed a generalization of Cohen's kappa statistic for the case where each of a sample of subjects is rated on a nominal scale by the same number of raters, but where the raters rating one subject are not necessarily the same as those rating another.

Suppose that a sample of N subjects has been studied, with n being the number of ratings per subject and k being the number of categories into which assignments are made. Let n_{ij}, $i = 1, ..., N$, $j = 1, ..., k$, be the number of raters who assigned the ith subject to the jth category, and define

$$\hat{p}_j = \frac{1}{Nn}\sum_{i=1}^{N} n_{ij}$$

So $\hat{p}_j$ is the proportion of all assignments which were to the jth category. Also, $\sum_j n_{ij} = n$ and $\sum_j \hat{p}_j = 1$.

The extent of agreement among the n raters for the ith subject is denoted by the proportion of agreeing pairs out of all $n(n-1)$ possible pairs of assignments, P_i,

$$\begin{aligned} P_i &= \frac{1}{n(n-1)} \sum_{j=1}^{k} n_{ij}(n_{ij} - 1) \\ &= \frac{1}{n(n-1)} (\sum_{j=1}^{k} n_{ij}^2 - n) \end{aligned}$$

The overall extent of agreement is measured by the mean of the P_is, $\bar{P}$,

$$\begin{aligned} \bar{P} &= \frac{1}{N} \sum_{i=1}^{N} P_i \\ &= \frac{1}{Nn(n-1)} (\sum_{i=1}^{N} \sum_{j=1}^{k} n_{ij}^2 - Nn) \end{aligned}$$

The extent of agreement expected by chance alone is

$$\bar{P}_e = \sum_{j=1}^{k} \hat{p}_j^2$$

Then the measure of overall agreement beyond chance is

$$\hat{\kappa} = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

Fleiss, Nee and Landis [19] derived the estimated large sample variance of κ, which is appropriate for testing the hypothesis that the underlying value is zero,

$$\begin{aligned}\mathrm{Var}(\hat{\kappa}) =& \frac{2}{Nn(n-1)(\sum \hat{p}_j\hat{q}_j)^2} \\ & \times [(\sum \hat{p}_j\hat{q}_j)^2 - \sum \hat{p}_j\hat{q}_j(\hat{q}_j - \hat{p}_j)]\end{aligned}$$

Fleiss and Cuzick [20] proposed a kappa statistic for unequal number of judges per subject. Suppose that a sample of n subjects has been studied, with m_i being the number of ratings on the ith subject. The raters rating one subject are not assumed to be the same as those rating another. Consider ratings consisting of classifications into one of two categories. Let x_i be the number of positive ratings on subject i. Define the overall proportion of positive ratings to be

$$\bar{p} = \frac{\sum_{i=1}^{n} x_i}{n\bar{m}}$$

where

$$\bar{m} = \frac{\sum_{i=1}^{n} m_i}{n}$$

the mean number of ratings per subject. If the number of subjects is greater than or equal to 20, then the mean square between subjects (BMS) is approximately equal to

$$\mathrm{BMS} = \frac{1}{n}\sum_{i=1}^{n} \frac{(x_i - m_i\bar{p})^2}{m_i}$$

and the mean square within subjects (WMS) is equal to

$$\text{WMS} = \frac{1}{n(\bar{m}-1)} \sum_{i=1}^{n} \frac{x_i(m_i - x_i)}{m_i}$$

Then the kappa statistic is

$$\begin{aligned}\hat{\kappa} &= \frac{\text{BMS} - \text{WMS}}{\text{BMS} + (\bar{m}-1)\text{WMS}} \\ &= 1 - \frac{\sum_{i=1}^{n} \frac{x_i(m_i - x_i)}{m_i}}{n(\bar{m}-1)\bar{p}\bar{q}}\end{aligned}$$

where $\bar{q} = 1 - \bar{p}$. Fleiss and Cuzick [20] derived the estimated variance of $\hat{\kappa}$

$$\begin{aligned}\text{Var}(\hat{\kappa}) = &\frac{1}{(\bar{m}-1)^2 n\bar{m}_H} \\ &\times [2(\bar{m}_H - 1) + \frac{(\bar{m} - \bar{m}_H)(1 - 4\bar{p}\bar{q})}{m\bar{p}\bar{q}}]\end{aligned}$$

where

$$\bar{m}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{m_i}}$$

1.2 Concordance Correlation Coefficient

1.2.1 Lin's Concordance Correlation Coefficient

Lin [31] proposed the concordance correlation coefficient (CCC) to evaluate the agreement between two readings by measuring the variation from the 45° line through the origin. It assumes that pairs of samples (Y_{i1}, Y_{i2}), $i = 1, 2, ..., n$, are independently selected from a bivariate population with means μ_1 and μ_2 and covariance matrix

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

Then the CCC, ρ_c, is defined as

$$\begin{aligned} \rho_c &= 1 - \frac{E[(Y_1 - Y_2)^2]}{E_{\text{independence}}[(Y_1 - Y_2)^2]} \\ &= 1 - \frac{E[(Y_1 - Y_2)^2]}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2} \\ &= \frac{2\sigma_{12}}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2} \\ &= \rho C_b \end{aligned}$$

where

$$\begin{aligned} &C_b = [(\upsilon + 1/\upsilon + u^2)/2]^{-1} \\ &\upsilon = \sigma_1/\sigma_2 = \text{scale shift} \\ &u = (\mu_1 - \mu_2)/\sqrt{\sigma_1\sigma_2} = \text{location shift relative to the scale} \end{aligned}$$

and ρ is Pearson correlation coefficient.

The CCC has the following properties ([31]):

(i) $-1 \le -|\rho| \le \rho_c \le |\rho| \le 1$.

(ii) $\rho_c = 0$ if and only if $\rho = 0$.

(iii) $\rho_c = \rho$ if and only if $\sigma_1 = \sigma_2$ and $\mu_1 = \mu_2$.

(iv) $\rho_c = \pm 1$ if and only if

(a) $(\mu_1 - \mu_2)^2 + (\sigma_1 - \sigma_2)^2 + 2\sigma_1\sigma_2(1 \mp \rho) = 0$, or equivalently,

(b) $\rho = \pm 1$, $\sigma_1 = \sigma_2$, and $\mu_1 = \mu_2$, or equivalently,

(c) each pair of readings is in perfect agreement (1) or in perfect reversed agreement (-1).

The CCC evaluates the degree to which pairs fall on the 45° line, where C_b measures how far the best-fit line deviates from the 45° line (accuracy) and ρ measures how far each observation deviates from the best-fit line (precision). If $C_b = 1$, then it indicates that there is no deviation from the 45° line. Otherwise, the further C_b is from 1, the greater the deviation is from the 45° line.

If we assume paired samples from a bivariate normal distribution, then we define the estimated CCC as

$$\hat{\rho}_c = \frac{2S_{12}}{S_1^2 + S_2^2 + (\bar{Y}_1 - \bar{Y}_2)^2}$$

where $\bar{Y}_j = \frac{1}{n}\sum_{i=1}^{n} Y_{ij}$, $S_j^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_{ij} - \bar{Y}_j)^2$, $j = 1, 2$; and

$$S_{12} = \frac{1}{n}\sum_{i=1}^{n}(Y_{i1} - \bar{Y}_1)(Y_{12} - \bar{Y}_2)$$

Lin [31] showed that $\hat{\rho}_c$ is a consistent estimator of ρ_c. And by using the inverse hyperbolic tangent transformation (or Z-transformation), the normal approximation of $\hat{\rho}_c$ can be improved

$$\hat{Z} = \tanh^{-1}(\hat{\rho}_c) = \frac{1}{2}\ln\frac{1 + \hat{\rho}_c}{1 - \hat{\rho}_c}$$

$$\sigma^2_{\hat{Z}} = \frac{1}{n-2}[\frac{(1-\rho^2)\rho^2_c}{(1-\rho^2_c)\rho^2} + \frac{2\rho^3_c(1-\rho_c)u^2}{\rho(1-\rho^2_c)^2} - \frac{\rho^4_c u^4}{2\rho^2(1-\rho^2_c)^2}]$$

Then ρ_c is asymptotically normal with mean ρ_c and variance

$$(1-\rho^2_c)^2\sigma^2_{\hat{Z}}$$

1.2.2 Generalized Concordance Correlation Coefficient

King and Chinchilli [27] proposed a generalization of the concordance correlation coefficient because Lin's [31] CCC is not robust when the underlying bivariate distribution is heavy-tailed.

Assume that observations (X_i, Y_i), $i = 1, 2, ..., n$ are independently selected from a bivariate population with cumulative distribution function (CDF) F_{XY}. Let F_X and F_Y be the marginal CDFs of X and Y, respectively. Further, let $g(.)$ be a convex function of distance defined on the real line and $g(X-Y)$ denote an integrable function with respect to F_{XY}. The generalized CCC is defined as

$$\rho_g = \frac{[E_{F_X F_Y} g(X-Y) - E_{F_X F_Y} g(X+Y)] - [E_{F_{XY}} g(X-Y) - E_{F_{XY}} g(X+Y)]}{E_{F_X F_Y} g(X-Y) - E_{F_X F_Y} g(X+Y) + \frac{1}{2} E_{F_{XY}}[g(2X) + g(2Y)]}$$

The estimator of ρ_g is

$$\hat{\rho}_g = \frac{\frac{1}{n}\sum_i \sum_j [g(X_i - Y_j) - g(X_i + Y_j)] - \sum_i [g(X_i - Y_i) - g(X_i + Y_i)]}{\frac{1}{n}\sum_i \sum_j [g(X_i - Y_j) - g(X_i + Y_j)] + \frac{1}{2}\sum_i [g(2X_i) + g(2Y_i)]}$$

King and Chinchilli [26] derived the asymptotic distribution of $\hat{\rho}_g$ by expressing this estimator in terms of U-statistics. They also showed that the normal approximation of the U-statistic estimator of CCC can be improved by using Fisher's Z-transformation

$$\hat{Z} = \tanh^{-1}(\hat{\rho}_g) = \frac{1}{2} \ln \frac{1+\hat{\rho}_g}{1-\hat{\rho}_g}$$

They demonstrated that the generalized CCC can be used to reproduce Cohen's kappa, weighted kappa and stratified CCC and can be extended to evaluate agreement among more than two raters or assays.

1.3 Random Marginal Agreement Coefficient

Fay [14] proposed the random marginal agreement coefficient (RMAC) to solve the problem that many standard agreement coefficients (e.g. kappa, weighted kappa and CCC) may yield larger agreement as the marginal distributions of the two raters become more different.

The RMAC adjusts for chance by modeling two independent readings both from the mixture distribution that averages the two marginal distributions. In other words, the RMAC models disagreement by chance by first randomly choosing an instrument and then randomly drawing from the marginal distribution of that instrument. Then the RMAC cannot be inflated by the difference between the marginal distributions. The RMAC can be used for both categorical responses and continuous responses.

For categorical responses, let (X,Y) denote a bivariate categorical response in which X and Y can take on the values $0, 1, ..., C$. Let $p_{ij} = P(X = i, Y = j)$, $i, j =$

$0, 1, ..., C$, denote the bivariate probabilities, and let $p_{i.} = P(X = i)$ and $p_{.j} = P(Y = j)$ denote the marginal probabilities. The RMAC is defined as

$$\kappa_{RMAC} = \frac{p_o - \sum_{i=0}^{C} (\frac{1}{2}p_{i.} + \frac{1}{2}p_{.i})^2}{1 - \sum_{i=0}^{C} (\frac{1}{2}p_{i.} + \frac{1}{2}p_{.i})^2}$$

where $p_o = \sum_{i=0}^{C} p_{ii}$, $p_{i.} = \sum_{j=0}^{C} p_{ij}$ and $p_{.i} = \sum_{j=0}^{C} p_{ji}$.

Fay [14] derived the asymptotic variance for κ_{RMAC}, which is

$$[1 - \kappa_{RMAC}^2]^2 V$$

where

$$V = \sum_{g=0}^{C} \sum_{h=0}^{C} p_{gh} (D_{gh})^2 - (\sum_{g=0}^{C} \sum_{h=0}^{C} p_{gh} D_{gh})^2$$

$$D_{gh} = \frac{1}{2(1 + \Pi_0 - 2\Pi az)} [I(g = h) - (p_{g.} + p_{.h}) - (p_{h.} + p_{.g})] + \frac{1}{2(1 - \Pi_0)} I(g = h)$$

and $\Pi_0 = \sum_{i=0}^{C} p_{ii}, \Pi_{az} = \sum_{i=0}^{C} [\frac{1}{2}p_{i.} + \frac{1}{2}p_{.i}]^2$.

For continuous responses, the RMAC assumes that pairs of samples (Y_{i1}, Y_{i2}), $i = 1, 2, ..., n$, are independently selected from a bivariate population with means μ_1 and μ_2 and covariance matrix

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

If $d(x, y) = (x - y)^2$, then the RMAC is

$$A_R(d) = \frac{2\rho - \frac{1}{2}u^2}{\upsilon + 1/\upsilon + \frac{1}{2}u^2}$$

where u, v and ρ are as defined for the CCC [31].

Fay [14] also conducted a simulation study to evaluate the performance of the RMAC, using the delta method and the bias-corrected and accelerated (BCa) method. Both the delta method intervals and the BCa intervals gave reasonably adequate coverage.

Chapter 2

A General Class of Agreement Coefficients

2.1 Motivation and General Form

We use a simple example to illustrate the use and limitations of Cohen's kappa and the RMAC. In Table 2.1, two observers classified 100 subjects into two categories, and the marginal distributions are identical. Cohen's kappa is $\hat{\kappa} = 0.167$, $\hat{SE}(\hat{\kappa}) = 0.1$, whereas the RMAC is $\hat{\kappa}_{RMAC} = 0.167$, $\hat{SE}(\hat{\kappa}_{RMAC}) = 0.1$. So in the presence of equal underlying marginal distributions, Cohen's kappa and the RMAC measure the same parameter, and therefore in the presence of equal observed marginal distributions, they yield the same estimates of agreement and standard error. Comparing to this table, Table 2.2 has identical diagonal values but different marginal distributions. Cohen's kappa becomes $\hat{\kappa} = 0.238$, $\hat{SE}(\hat{\kappa}) = 0.078$, whereas the RMAC is $\hat{\kappa}_{RMAC} = 0.167$, $\hat{SE}(\hat{\kappa}_{RMAC}) = 0.1$. Cohen's kappa shows better estimate of agreement for Table 2.2 over Table 2.1, and smaller SE than the one for RMAC for Table 2.2, despite the fact that Table 2.2 has different marginal distributions and identical diagonal values (exact matches) to Table 2.1.

From this example, we can see that, given fixed diagonal values, when the marginal distributions are equal or approximately so, both Cohen's kappa and the RMAC are equivalent in terms of estimates of agreement and standard error; when the marginal distributions are different, Cohen's kappa yields a non-intuitive result whereas RMAC is

Table 2.1 Original data

	Observer A		
Observer B	Yes	No	Total
Yes	40	20	60
No	20	20	40
Total	60	40	100

Table 2.2 Modified data

	Observer A		
Observer B	Yes	No	Total
Yes	40	35	75
No	5	20	25
Total	45	55	100

robust, but Cohen's kappa has a smaller standard error than the RMAC. We would like to use Cohen's kappa or the RMAC when the marginal distributions are similar, and a mixture of Cohen's kappa and RMAC when the marginal distributions are very different to balance between robustness and efficiency. The decision of choosing an appropriate measure should therefore depend on the difference between marginal distributions. We define parameter a, which measures the difference between marginal distributions. The value of a can be calculated using the Kolmogorov-Smirnov criterion, $a = \sup_z |F_X(z) - F_Y(z)|$, or the square root of the average squared distance, $a = \sqrt{\frac{1}{C+1}\sum_{j=0}^{C}[F_X(j) - F_Y(j)]^2}$.

Let (X,Y) denote a bivariate categorical or continuous response, with cumulative distribution function (CDF), F_{XY}, and marginal CDFs of X and Y, F_X and F_Y , respectively. Define the cost of disagreement as $c(x, y)$, which equals zero when $x = y$ and is non-negative otherwise, and $c(x, y) = c(y, x)$ for all x, y. Then the general form of the general class of agreement coefficients can be written as

$$A(c) = 1 - \frac{E_{F_{XY}}\{c(X,Y)\}}{E_{F_{U_1}} E_{F_{U_2}}\{c(X,Y)\}} \tag{2.1}$$

where U_1 and U_2 are independent random variables, having distributions $F_{U_1} = 0.5aF_X + (1-0.5a)F_Y$ and $F_{U_2} = (1-0.5a)F_X + 0.5aF_Y$.

The true value of parameter a can be obtained only if the true underlying bivariate distribution of X and Y is known. However, this distribution is never known to us. Alternatively, in practice, the value of a is determined according to the researcher's preference or estimated from sample proportions. In our scheme, a is viewed as either fixed and known or a parameter estimate. Performances of the proposed agreement coefficients are investigated under these two situations.

2.2 Nominal Data

2.2.1 Definition

For nominal data, we usually use nominal cost function, that is, $c(x,y) = 0$ if $x = y$ and 1 otherwise. Let (X,Y) denote a bivariate categorical response in which X and Y can take on the values 0, 1, ..., C. Let $p_{ij} = P(X = i, Y = j)$, i, j=0, 1, ..., C, denote the bivariate probabilities, and let $p_{i.} = P(X = i)$ and $p_{.j} = P(Y = j)$ denote the marginal probabilities. Equation (2.1) is reduced to the general class of agreement coefficients for nominal data as

$$\kappa(a) = \frac{\sum_{i=0}^{C} p_{ii} - \sum_{i=0}^{C}[0.5ap_{i.} + (1-0.5a)p_{.i}][(1-0.5a)p_{i.} + 0.5ap_{.i}]}{1 - \sum_{i=0}^{C}[0.5ap_{i.} + (1-0.5a)p_{.i}][(1-0.5a)p_{i.} + 0.5ap_{.i}]}$$

A value of $a = 0$ yields Cohen's kappa, a value of $a = 1$ yields the RMAC, and a value of a between 0 and 1 yields mixtures of Cohen's kappa and the RMAC.

2.2.2 Asymptotic Distribution

The multivariate delta method for a multinomial distribution with parameters n and $\boldsymbol{\theta} = (\theta_1, ..., \theta_K)^T$, where not all $\theta_i = 0$, states that

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, \Lambda(\boldsymbol{\theta}))$$

where $\Lambda(\boldsymbol{\theta}) = \mathbf{D}_{\boldsymbol{\theta}} - \boldsymbol{\theta}\boldsymbol{\theta}^{\mathbf{T}}$, $\mathbf{D}_{\boldsymbol{\theta}}$ denotes the diagonal matrix based on $\boldsymbol{\theta}$ such that $\Lambda_{ii}(\boldsymbol{\theta}) = \theta_i - \theta_i^2$ and $\Lambda_{ij}(\boldsymbol{\theta}) = -\theta_i\theta_j, i \neq j$.

Let f: $\Re^T \longrightarrow \Re$, then if $f(\boldsymbol{\theta}) \neq 0$,

$$\sqrt{n}(f(\hat{\boldsymbol{\theta}}) - f(\boldsymbol{\theta})) \xrightarrow{d} N(0, \Sigma_f)$$

where

$$\begin{aligned}
\Sigma_f &= \nabla f(\boldsymbol{\theta})^T \Lambda(\boldsymbol{\theta}) \nabla f(\boldsymbol{\theta}) \\
&= (\frac{\partial f}{\partial \theta_1} \cdots \frac{\partial f}{\partial \theta_T}) \Lambda(\boldsymbol{\theta}) (\frac{\partial f}{\partial \theta_1} \cdots \frac{\partial f}{\partial \theta_T})^T \\
&= \sum_{i=1}^{K} \sum_{j=1}^{K} \Lambda_{ii}(\boldsymbol{\theta}) (\frac{\partial f}{\partial \theta_i})(\frac{\partial f}{\partial \theta_j}) \\
&= \sum_{i=1}^{K} \theta_i (\frac{\partial f}{\partial \theta_i})^2 - \sum_{i=1}^{K} \theta_i^2 (\frac{\partial f}{\partial \theta_i})^2 - 2 \sum_{i<j} \theta_i \theta_j (\frac{\partial f}{\partial \theta_i})(\frac{\partial f}{\partial \theta_j}) \\
&= \sum_{i=1}^{K} \theta_i (\frac{\partial f}{\partial \theta_i})^2 - [\sum_{i=1}^{K} \theta_i (\frac{\partial f}{\partial \theta_i})]^2 \qquad (2.2)
\end{aligned}$$

For $\kappa(a)$ on categorical responses, let $K = (C+1)^2$, and let

$$\boldsymbol{\theta} = (p_{00}, p_{01}, ..., p_{CC})^T$$

denote the vector of probability parameters. We can work with equation (2.2) in the simpler form of replacing each θ_i with a value for p_{gh}

$$\sum_{g=0}^{C}\sum_{h=0}^{C} p_{gh}(\frac{\partial f}{\partial p_{gh}})^2 - [\sum_{g=0}^{C}\sum_{h=0}^{C} p_{gh}(\frac{\partial f}{\partial p_{gh}})]^2$$

2.2.2.1 Fixed and Known a

Let

$$\begin{aligned} f(\boldsymbol{\theta}) = \tanh^{-1}[\kappa(a)] & \\ &= \frac{1}{2}\log(\frac{1+\kappa(a)}{1-\kappa(a)}) \\ &= \frac{1}{2}\log(\frac{1+\sum_{i=0}^{C} p_{ii} - 2\sum_{i=0}^{C}[0.5ap_{i.} + (1-0.5a)p_{.i}][(1-0.5a)p_{i.} + 0.5ap_{.i}]}{1-\sum_{i=0}^{C} p_{ii}}) \end{aligned}$$

Then

$$\begin{aligned} f(\hat{\boldsymbol{\theta}}) = \tanh^{-1}[\hat{\kappa}(a)] & \\ &= \frac{1}{2}\log(\frac{1+\hat{\kappa}(a)}{1-\hat{\kappa}(a)}) \\ &= \frac{1}{2}\log(\frac{1+\sum_{i=0}^{C} \hat{p}_{ii} - 2\sum_{i=0}^{C}[0.5a\hat{p}_{i.} + (1-0.5a)\hat{p}_{.i}][(1-0.5a)\hat{p}_{i.} + 0.5a\hat{p}_{.i}]}{1-\sum_{i=0}^{C} \hat{p}_{ii}}) \end{aligned}$$

where $\hat{\boldsymbol{\theta}} = (\hat{p}_{00}, \hat{p}_{01}, ..., \hat{p}_{CC})^T$ denotes the vector of estimated probability parameters.

Let $\Pi_0 = \sum_{i=0}^{C} p_{ii}, \Pi_{az} = \sum_{i=0}^{C} [0.5ap_{i.} + (1-0.5a)p_{.i}][(1-0.5a)p_{i.} + 0.5ap_{.i}]$. Then

$$\begin{aligned} D_{gh} &= \frac{\partial f}{\partial p_{gh}} \\ &= \frac{1-\Pi_0}{2(1+\Pi_0-2\Pi_{az})}[\frac{\frac{\partial(\Pi_0-2\Pi_{az})}{\partial p_{gh}}(1-\Pi_0) - \frac{\partial(1-\Pi_0)}{\partial p_{gh}}(1+\Pi_0-2\Pi_{az})}{(1-\Pi_0)^2}] \\ &= \frac{1}{2(1+\Pi_0-2\Pi_{az})}(\frac{\partial \Pi_0}{\partial p_{gh}} - 2\frac{\partial \Pi_{az}}{\partial p_{gh}}) + \frac{1}{2(1-\Pi_0)}(\frac{\partial \Pi_0}{\partial p_{gh}}) \end{aligned}$$

Now

$$\frac{\partial \Pi_0}{\partial p_{gh}} = I(g=h)$$

where $I(g=h)$ is an indicator function such that $I(g=h)=1$ if $g=h$ and $I(g=h)=0$ if $g \neq h$.

$$\begin{aligned} \frac{\partial \Pi_{az}}{\partial p_{gh}} &= \sum_{i=0}^{C}\{[0.5ap_{i.} + (1-0.5a)p_{.i}]\frac{\partial}{\partial p_{gh}}[(1-0.5a)p_{i.} + 0.5ap_{.i}] \\ &+ [(1-0.5a)p_{i.} + 0.5ap_{.i}]\frac{\partial}{\partial p_{gh}}[0.5ap_{i.} + (1-0.5a)p_{.i}]\} \\ &= \sum_{i=0}^{C}\{[0.5ap_{i.} + (1-0.5a)p_{.i}][(1-0.5a)I(g=i) + 0.5aI(h=i)] \\ &+ [(1-0.5a)p_{i.} + 0.5ap_{.i}][0.5aI(g=i) + (1-0.5a)I(h=i)]\} \\ &= 0.5a(1-0.5a)p_{g.} + 0.25a^2 p_{h.} + (1-0.5a)^2 p_{.g} + 0.5a(1-0.5a)p_{.h} \\ &+ 0.5a(1-0.5a)p_{g.} + (1-0.5a)^2 p_{h.} + 0.25a^2 p_{.g} + 0.5a(1-0.5a)p_{.h} \\ &= a(1-0.5a)(p_{g.} + p_{.h}) + (1-a+0.5a^2)(p_{h.} + p_{.g}) \end{aligned}$$

Thus

$$\begin{aligned} D_{gh} =& \frac{1}{2(1+\Pi_0-2\Pi az)}[I(g=h) \\ & -2a(1-0.5a)(p_{g.}+p_{.h})-2(1-a+0.5a^2)(p_{h.}+p_{.g})]+\frac{1}{2(1-\Pi_0)}I(g=h) \end{aligned}$$

The asymptotic variance of $\sqrt{n}(\tanh^{-1}(\hat{\kappa}(a))-\tanh^{-1}(\kappa(a)))$ is

$$V=\sum_{g=0}^{C}\sum_{h=0}^{C}p_{gh}(D_{gh})^2-(\sum_{g=0}^{C}\sum_{h=0}^{C}p_{gh}D_{gh})^2$$

The estimator of D_{gh}, $\hat{D}_{gh}$, replaces all p_{ij} with $\hat{p}_{ij}$ (the sample proportions). So the asymptotic variance estimate of $\sqrt{n}(\tanh^{-1}(\hat{\kappa}(a))-\tanh^{-1}(\kappa(a)))$ is

$$\hat{V}=\sum_{g=0}^{C}\sum_{h=0}^{C}\hat{p}_{gh}(\hat{D}_{gh})^2-(\sum_{g=0}^{C}\sum_{h=0}^{C}\hat{p}_{gh}\hat{D}_{gh})^2$$

The asymptotic normality of $\hat{\kappa}(a)$ can be obtained by letting

$$\hat{\kappa}(a)=\tanh[f(\hat{\boldsymbol{p}})]=g(f(\hat{\boldsymbol{p}}))$$

Since $g^{'}(t) = \frac{\partial}{\partial t}\tanh(t) = \frac{4e^{2t}}{(1+e^{2t})^2}$ and $\sqrt{n}\{\tanh^{-1}[f(\hat{\boldsymbol{\theta}})] - \tanh^{-1}[f(\boldsymbol{\theta})]\} \xrightarrow{d} N(0, V)$,

$$\begin{aligned} g^{'}(\tanh^{-1}[f(\boldsymbol{\theta})]) &= \frac{4\exp(\log\frac{1+\kappa(a)}{1-\kappa(a)})}{[1+\exp(\log\frac{1+\kappa(a)}{1-\kappa(a)})]^2} \\ &= \frac{4[1+\kappa(a)]}{1-\kappa(a)}\frac{[1-\kappa(a)]^2}{4} \\ &= 1-\kappa^2(a) \end{aligned}$$

Hence the asymptotic variance estimate for $\sqrt{n}(\hat{\kappa}(a) - \kappa(a))$ is

$$[1-\hat{\kappa}^2(a)]^2\hat{V}$$

Now we show that when $a = 0$ and $a = 1$, the asymptotic variances of $\hat{\kappa}(a)$ are equivalent to the ones of Cohen's kappa and the RMAC, respectively. Recall that the asymptotic variance for $\sqrt{n}(\hat{\kappa}(a) - \kappa(a))$ is

$$\begin{aligned} &[1-\kappa^2(a)]^2V \\ &= [1-\kappa(a)]^2[1+\kappa(a)]^2V \\ &= (\frac{1+p_o-2p_e}{1-p_e})^2[1-\kappa(a)]^2[\sum_{g=0}^{C}\sum_{h=0}^{C}p_{gh}(D_{gh})^2 - (\sum_{g=0}^{C}\sum_{h=0}^{C}p_{gh}D_{gh})^2] \\ &= (\frac{1+p_o-2p_e}{1-p_e})^2[1-\kappa(a)]^2[\sum_{g=0}^{C}p_{gg}(D_{gg})^2 + \sum\sum_{g\neq h}p_{gh}(D_{gh})^2 - (\sum_{g=0}^{C}\sum_{h=0}^{C}p_{gh}D_{gh})^2] \end{aligned} \tag{2.3}$$

When $a = 0$,

$$
\begin{aligned}
&\sum_{g=0}^{C} p_{gg}(D_{gg})^2 \\
&= \sum_{g=0}^{C} p_{gg}\{\frac{1}{2(1+p_o-2p_e)}[1-2(p_{g.}+p_{.g})]+\frac{1}{2(1-p_o)}\}^2 \\
&= \sum_{g=0}^{C} p_{gg}[\frac{(1-p_e)-(1-p_o)(p_{g.}+p_{.g})}{(1+p_o-2p_e)(1-p_o)}]^2 \\
&= \sum_{g=0}^{C} p_{gg}[\frac{1-p_e-(1-p_e)(1-\kappa)(p_{g.}+p_{.g})}{(1+p_o-2p_e)(1-p_e)(1-\kappa)}]^2 \\
&= \sum_{g=0}^{C} p_{gg}[\frac{1-(1-\kappa)(p_{g.}+p_{.g})}{(1+p_o-2p_e)(1-\kappa)}]^2
\end{aligned}
$$

$$
\begin{aligned}
&\sum\sum_{g\neq h} p_{gh}(D_{gh})^2 \\
&= \sum\sum_{g\neq h} p_{gh}\{\frac{1}{2(1+p_o-2p_e)}[-2(p_{h.}+p_{.g})]\}^2 \\
&= \frac{1}{(1+p_o-2p_e)^2}\sum\sum_{g\neq h} p_{gh}(p_{g.}+p_{.h})^2
\end{aligned}
$$

$$
\begin{aligned}
&(\sum_{g=0}^{C}\sum_{h=0}^{C} p_{gh} D_{gh})^2 \\
&= \{\frac{1}{2(1+p_o-2p_e)}[p_o - 2\sum_{g=0}^{C}\sum_{h=0}^{C} p_{gh}(p_{h.}+p_{.g})] + \frac{p_o}{2(1-p_o)}\}^2 \\
&= [\frac{p_o(1-p_e) - (1-p_o)(\sum_{g=0}^{C} p_{.g}\sum_{h=0}^{C} p_{gh} + \sum_{h=0}^{C} p_{h.}\sum_{g=0}^{C} p_{gh})}{(1+p_o-2p_e)(1-p_o)}]^2 \\
&= [\frac{[1-(1-p_e)(1-\kappa)] - (1-\kappa)(2p_e)}{(1+p_o-2p_e)(1-\kappa)}]^2 \\
&= [\frac{\kappa - p_e(1-\kappa)}{(1+p_o-2p_e)(1-\kappa)}]^2
\end{aligned}
$$

Plug these terms in (2.3), we have

$$
\sum_{g=0}^{C}\{\frac{[1-(p_{g.}+p_{.g})(1-\kappa)]^2}{(1-p_e)^2}\} + \frac{(1-\kappa)^2\sum\sum_{g\neq h} p_{gh}(p_{g.}+p_{.h})^2}{(1-p_e)^2} - \frac{\kappa - p_e(1-\kappa)}{(1-p_e)^2}
$$

Divide the above expression by n and use estimated probabilities $\hat{p}$ instead of p, it yields the same result as the one derived by Fleiss, Cohen and Everitt [17].

When a=1, our expression of D_{gh} becomes

$$
D_{gh} = \frac{1}{2(1+\Pi_0 - 2\Pi az)}[I(g=h) - (p_{g.}+p_{.h}) - (p_{h.}+p_{.g})] + \frac{1}{2(1-\Pi_0)}I(g=h)
$$

and the expression of $\frac{\partial \Pi_z}{\partial \pi_{gh}}$ in the RMAC [14], with $w_{ij} = 1$ when $i = j$ and $w_{ij} = 0$ when $i \neq j$, is

$$\sum_{i=0}^{C} \frac{1}{4}[(\pi_{i.} + \pi_{.i})\{I(g = j) + I(h = j)\} + (\pi_{j.} + \pi_{.j})\{I(g = i) + I(h = i)\}]$$

$$= \frac{1}{2}(\pi_{g.} + \pi_{h.} + \pi_{.g} + \pi_{.h})$$

So the expression of D_{gh} in the RMAC is the same as our expression.

2.2.2.2 Estimated a

Let a be defined based on the square root of the average squared distance between the marginal distributions, i.e.

$$a = \sqrt{\frac{1}{C+1}\sum_{j=0}^{C}[F_X(j) - F_Y(j)]^2}$$
$$= \sqrt{\frac{1}{C+1}\sum_{j=0}^{C}[\sum_{i=0}^{j}(p_{i.} - p_{.i})]^2}$$

Here, a is a function of $\boldsymbol{\theta}$.

Again, let

$$\boldsymbol{\theta} = (p_{00}, p_{01}, ..., p_{CC})^T$$

denote the vector of probability parameters.

Let

$$\begin{aligned} f(\boldsymbol{\theta}) &= \tanh^{-1}[\kappa(a)] \\ &= \frac{1}{2}\log(\frac{1+\kappa(a)}{1-\kappa(a)}) \\ &= \frac{1}{2}\log(\frac{1+\sum_{i=0}^{C} p_{ii} - 2\sum_{i=0}^{C}[0.5ap_{i.} + (1-0.5a)p_{.i}][(1-0.5a)p_{i.} + 0.5ap_{.i}]}{1-\sum_{i=0}^{C} p_{ii}}) \end{aligned}$$

Then

$$\begin{aligned} f(\hat{\boldsymbol{\theta}}) &= \tanh^{-1}[\hat{\kappa}(\hat{a})] \\ &= \frac{1}{2}\log(\frac{1+\hat{\kappa}(\hat{a})}{1-\hat{\kappa}(\hat{a})}) \\ &= \frac{1}{2}\log(\frac{1+\sum_{i=0}^{C} \hat{p}_{ii} - 2\sum_{i=0}^{C}[0.5\hat{a}\hat{p}_{i.} + (1-0.5\hat{a})\hat{p}_{.i}][(1-0.5\hat{a})\hat{p}_{i.} + 0.5\hat{a}\hat{p}_{.i}]}{1-\sum_{i=0}^{C} \hat{p}_{ii}}) \end{aligned}$$

where $\hat{\boldsymbol{\theta}} = (\hat{p}_{00}, \hat{p}_{01}, ..., \hat{p}_{CC})^T$ denotes the vector of estimated probability parameters and

$$\hat{a} = \sqrt{\frac{1}{C+1}\sum_{j=0}^{C}[\sum_{i=0}^{j}(\hat{p}_{i.} - \hat{p}_{.i})]^2}$$

That is, a is estimated from the observed proportions $\hat{\boldsymbol{\theta}}$.

If $a \neq 0, 1$, we can work with equation (2.2) in the simpler form of replacing each θ_i with a value for p_{gh}

$$\sum_{g=0}^{C}\sum_{h=0}^{C} p_{gh}(\frac{\partial f}{\partial p_{gh}})^2 - [\sum_{g=0}^{C}\sum_{h=0}^{C} p_{gh}(\frac{\partial f}{\partial p_{gh}})]^2 \tag{2.4}$$

Let $\Pi_0 = \sum_{i=0}^{C} p_{ii}, \Pi_{az} = \sum_{i=0}^{C}[0.5ap_{i.} + (1-0.5a)p_{.i}][(1-0.5a)p_{i.} + 0.5ap_{.i}]$.

Then

$$\begin{aligned}
D_{gh} &= \frac{\partial f}{\partial p_{gh}} \\
&= \frac{1-\Pi_0}{2(1+\Pi_0-2\Pi_{az})}[\frac{\frac{\partial(\Pi_0-2\Pi_{az})}{\partial p_{gh}}(1-\Pi_0)-\frac{\partial(1-\Pi_0)}{\partial p_{gh}}(1+\Pi_0-2\Pi_{az})}{(1-\Pi_0)^2}] \\
&= \frac{1}{2(1+\Pi_0-2\Pi_{az})}(\frac{\partial\Pi_0}{\partial p_{gh}}-2\frac{\partial\Pi_{az}}{\partial p_{gh}})+\frac{1}{2(1-\Pi_0)}(\frac{\partial\Pi_0}{\partial p_{gh}}) \qquad (2.5)
\end{aligned}$$

Now

$$\frac{\partial\Pi_0}{\partial p_{gh}} = I(g=h)$$

$$\begin{aligned}
\frac{\partial\Pi_{az}}{\partial p_{gh}} &= \sum_{i=0}^{C}\{[0.5ap_{i.}+(1-0.5a)p_{.i}]\frac{\partial}{\partial p_{gh}}[(1-0.5a)p_{i.}+0.5ap_{.i}] \\
&+[(1-0.5a)p_{i.}+0.5ap_{.i}]\frac{\partial}{\partial p_{gh}}[0.5ap_{i.}+(1-0.5a)p_{.i}]\} \\
&= \sum_{i=0}^{C}\{[0.5ap_{i.}+(1-0.5a)p_{.i}][-0.5p_{i.}(\frac{\partial}{\partial p_{gh}}a)+(1-0.5a)I(g=i) \\
&+0.5p_{.i}(\frac{\partial}{\partial p_{gh}}a)+0.5aI(h=i)]+[(1-0.5a)p_{i.}+0.5ap_{.i}] \\
&[0.5p_{i.}(\frac{\partial}{\partial p_{gh}}a)+0.5aI(g=i)-0.5p_{.i}(\frac{\partial}{\partial p_{gh}}a)+(1-0.5a)I(h=i)]\}
\end{aligned}$$

and

$$\Sigma_{gh} = \frac{\partial}{\partial p_{gh}} a = \frac{\partial}{\partial p_{gh}} \sqrt{\frac{1}{C+1} \sum_{j=0}^{C} [\sum_{i=0}^{j} (p_{i.} - p_{.i})]^2}$$
$$= \frac{1}{2} \{ \frac{1}{C+1} \sum_{j=0}^{C} [\sum_{i=0}^{j} (p_{i.} - p_{.i})]^2 \}^{-\frac{1}{2}} (\frac{2}{C+1}) \sum_{j=0}^{C} \{ [\sum_{i=0}^{j} (p_{i.} - p_{.i})]$$
$$[\sum_{i=0}^{j} (I(g = i) - I(h = i))] \}$$

Calculate each of these terms and plug them into equation (2.5), then plug equation (2.5) into equation (2.4). The asymptotic variance of $\sqrt{n}(\tanh^{-1}(\hat{\kappa}(\hat{a})) - \tanh^{-1}(\kappa(a)))$ is

$$V = \sum_{g=0}^{C} \sum_{h=0}^{C} p_{gh} (D_{gh})^2 - (\sum_{g=0}^{C} \sum_{h=0}^{C} p_{gh} D_{gh})^2$$

If $a = 0$, $\kappa(a)$ is reduced to Cohen's kappa; if $a = 1$, $\kappa(a)$ is reduced to the RMAC. One can then use the asymptotic results of Cohen's kappa and the RMAC, respectively.

If $\hat{a} \neq 0, 1$, replacing all p_{ij} with $\hat{p}_{ij}$, yields the estimated asymptotic variance of $\sqrt{n}(\tanh^{-1}(\hat{\kappa}(\hat{a})) - \tanh^{-1}(\kappa(a)))$

$$\hat{V} = \sum_{g=0}^{C} \sum_{h=0}^{C} \hat{p}_{gh} (\hat{D}_{gh})^2 - (\sum_{g=0}^{C} \sum_{h=0}^{C} \hat{p}_{gh} \hat{D}_{gh})^2$$

Hence the asymptotic variance estimate for $\sqrt{n}(\hat{\kappa}(\hat{a}) - \kappa(a))$ is

$$[1 - \hat{\kappa}^2(\hat{a})]^2 \hat{V}$$

If $\hat{a} = 0$, $\hat{\kappa}(\hat{a})$ is reduced to the estimate of Cohen's kappa; if $\hat{a} = 1$, $\hat{\kappa}(\hat{a})$ is reduced to the estimate of the RMAC. One can then use the asymptotic variance estimate of the estimate of Cohen's kappa and the RMAC, respectively.

2.3 Ordinal Data

2.3.1 Definition

Since the idea of the general class of agreement coefficients was motivated by Cohen's kappa and the RMAC, and both of these measures have been extended to incorporate weights for ordinal data, our proposed weighted $\kappa(a)$ and $\kappa(\hat{a})$ should have a general form with weighted kappa and weighted RMAC being special cases.

Let (X,Y) denote a bivariate categorical response in which X and Y can take on the values $0, 1, ..., C$. Let $p_{ij} = P(X = i, Y = j)$, $i, j = 0, 1, ..., C$, denote the bivariate probabilities, and let $p_{i.} = P(X = i)$ and $p_{.j} = P(Y = j)$ denote the marginal probabilities. Agreement weights, say $w_{ij}, i, j = 0, ..., C$, are assigned to the $(C + 1)^2$ cells. The range of weights is $0 \leq w_{ij} \leq 1$ such that $w_{ij} = 1$ for $i = j$, $0 \leq w_{ij} < 1$ for $i \neq j$ and $w_{ij} = w_{ji}$. Assume arbitrary cost function, equation (2.1) is reduced to the general class of agreement coefficients for ordinal data as

$$\kappa_w(a) = \frac{\sum_{i=0}^{C} \sum_{j=0}^{C} w_{ij} p_{ij} - \sum_{i=0}^{C} \sum_{j=0}^{C} w_{ij} [(1 - 0.5a)p_{i.} + 0.5ap_{.i}][0.5ap_{j.} + (1 - 0.5a)p_{.j}]}{1 - \sum_{i=0}^{C} \sum_{j=0}^{C} w_{ij} [(1 - 0.5a)p_{i.} + 0.5ap_{.i}][0.5ap_{j.} + (1 - 0.5a)p_{.j}]}$$

A value of $a = 0$ yields weighted kappa, a value of $a = 1$ yields the RMAC, and a value of a between 0 and 1 yields mixtures of weighted kappa and the RMAC.

For ordinal data, better weights should be proportional to the distance (or its square) between the two points i and j on the ordinal scale, in order to account for severity of discordance or size of discrepancy. Two commonly used weights for ordinal data are Cicchetti-Allison Weights, defined as $w_{ij} = 1 - \frac{|i-j|}{C}$, and Fleiss-Cohen Weights, defined as $w_{ij} = 1 - \frac{(i-j)^2}{C^2}$.

2.3.2 Asymptotic Distribution

2.3.2.1 Fixed and Known a

For $\kappa(a)$ on categorical responses, let $K = (C+1)^2$, and let

$$\boldsymbol{\theta} = (p_{00}, p_{01}, ..., p_{CC})^T$$

denote the vector of probability parameters. We can work with equation (2.2) in the simpler form of replacing each θ_i with a value for p_{gh}

$$\sum_{g=0}^{C}\sum_{h=0}^{C} p_{gh}(\frac{\partial f}{\partial p_{gh}})^2 - [\sum_{g=0}^{C}\sum_{h=0}^{C} p_{gh}(\frac{\partial f}{\partial p_{gh}})]^2 \tag{2.6}$$

Let

$$\begin{aligned}
f(\boldsymbol{\theta}) &= \tanh^{-1}[\kappa_w(a)] \\
&= \frac{1}{2}\log(\frac{1+\kappa_w(a)}{1-\kappa_w(a)}) \\
&= \frac{1}{2}\log(\frac{1+\sum_{i=0}^{C}\sum_{j=0}^{C} w_{ij}p_{ij} - 2\sum_{i=0}^{C}\sum_{j=0}^{C} w_{ij}[0.5ap_{i.} + (1-0.5a)p_{.i}][(1-0.5a)p_{j.} + 0.5ap_{.j}]}{1-\sum_{i=0}^{C}\sum_{j=0}^{C} w_{ij}p_{ij}})
\end{aligned}$$

Then

$$
\begin{aligned}
f(\hat{\boldsymbol{\theta}}) &= \tanh^{-1}[\hat{\kappa}_w(a)] \\
&= \frac{1}{2}\log(\frac{1+\hat{\kappa}_w(a)}{1-\hat{\kappa}_w(a)}) \\
&= \frac{1}{2}\log(\frac{1+\sum_{i=0}^{C}\sum_{j=0}^{C} w_{ij}\hat{p}_{ij} - 2\sum_{i=0}^{C}\sum_{j=0}^{C} w_{ij}[0.5a\hat{p}_{i.} + (1-0.5a)\hat{p}_{.i}][(1-0.5a)\hat{p}_{j.} + 0.5a\hat{p}_{.j}]}{1-\sum_{i=0}^{C}\sum_{j=0}^{C} w_{ij}\hat{p}_{ij}})
\end{aligned}
$$

where $\hat{\boldsymbol{\theta}} = (\hat{p}_{00}, \hat{p}_{01}, ..., \hat{p}_{CC})^T$ denotes the vector of estimated probability parameters.

Let $\Pi_0 = \sum_{i=0}^{C}\sum_{j=0}^{C} w_{ij}p_{ij}$, $\Pi_{az} = \sum_{i=0}^{C}\sum_{j=0}^{C} w_{ij}[0.5ap_{i.} + (1-0.5a)p_{.i}][(1-0.5a)p_{j.} + 0.5ap_{.j}]$.

Then

$$
\begin{aligned}
D_{gh} &= \frac{\partial f}{\partial p_{gh}} \\
&= \frac{1-\Pi_0}{2(1+\Pi_0-2\Pi_{az})}[\frac{\frac{\partial(\Pi_0-2\Pi_{az})}{\partial p_{gh}}(1-\Pi_0) - \frac{\partial(1-\Pi_0)}{\partial p_{gh}}(1+\Pi_0-2\Pi_{az})}{(1-\Pi_0)^2}] \\
&= \frac{1}{2(1+\Pi_0-2\Pi_{az})}(\frac{\partial\Pi_0}{\partial p_{gh}} - 2\frac{\partial\Pi_{az}}{\partial p_{gh}}) + \frac{1}{2(1-\Pi_0)}(\frac{\partial\Pi_0}{\partial p_{gh}}) \qquad (2.7)
\end{aligned}
$$

Now

$$
\frac{\partial\Pi_0}{\partial p_{gh}} = w_{gh} \qquad (2.8)
$$

$$\begin{aligned}\frac{\partial \Pi_{az}}{\partial p_{gh}} &= \sum_{i=0}^{C}\sum_{j=0}^{C}\{w_{ij}[0.5ap_{i.} + (1-0.5a)p_{.i}]\frac{\partial}{\partial p_{gh}}[(1-0.5a)p_{j.} + 0.5ap_{.j}] \\ &+ w_{ij}[(1-0.5a)p_{j.} + 0.5ap_{.j}]\frac{\partial}{\partial p_{gh}}[0.5ap_{i.} + (1-0.5a)p_{.i}]\} \\ &= \sum_{i=0}^{C}\sum_{j=0}^{C}\{w_{ij}[0.5ap_{i.} + (1-0.5a)p_{.i}][(1-0.5a)I(g=j) + 0.5aI(h=j)] \\ &+ w_{ij}[(1-0.5a)p_{j.} + 0.5ap_{.j}][0.5aI(g=i) + (1-0.5a)I(h=i)]\} \\ &= \sum_{i=0}^{C}\{w_{ig}[0.5ap_{i.} + (1-0.5a)p_{.i}](1-0.5a) + w_{ih}[0.5ap_{i.} + (1-0.5a)p_{.i}](0.5a)\} \\ &+ \sum_{j=0}^{C}\{w_{gj}[(1-0.5a)p_{j.} + 0.5ap_{.j}](0.5a) + w_{hj}[(1-0.5a)p_{j.} + 0.5ap_{.j}](1-0.5a)\}\end{aligned} \tag{2.9}$$

Plug (2.8) and (2.9) into (2.7), D_{gh} can be calculated.

The asymptotic variance of $\sqrt{n}(\tanh^{-1}(\hat{\kappa}_w(a)) - \tanh^{-1}(\kappa_w(a)))$ is

$$V = \sum_{g=0}^{C}\sum_{h=0}^{C} p_{gh}(D_{gh})^2 - (\sum_{g=0}^{C}\sum_{h=0}^{C} p_{gh}D_{gh})^2$$

The estimator of D_{gh}, $\hat{D}_{gh}$ replaces all p_{ij} with $\hat{p}_{ij}$ (the sample proportions). So the asymptotic variance estimate of $\sqrt{n}(\tanh^{-1}(\hat{\kappa}_w(a)) - \tanh^{-1}(\kappa_w(a)))$ is

$$\hat{V} = \sum_{g=0}^{C}\sum_{h=0}^{C} \hat{p}_{gh}(\hat{D}_{gh})^2 - (\sum_{g=0}^{C}\sum_{h=0}^{C} \hat{p}_{gh}\hat{D}_{gh})^2$$

Finally, we use the delta method to transform back to get the asymptotic variance estimate for $\sqrt{n}(\hat{\kappa}_w(a) - \kappa_w(a))$, which is

$$[1 - \hat{\kappa}_w^2(a)]^2 \hat{V}$$

With tedious but straightforward algebra, we can show that when $a = 0$ and $a = 1$, the asymptotic variances of $\hat{\kappa}_w(a)$ are equivalent to the ones of weighted kappa and the RMAC, respectively.

2.3.2.2 Estimated a

Let a be defined based on the square root of the average squared distance between the marginal distributions

$$\begin{aligned} a &\equiv \sqrt{\frac{1}{C+1}\sum_{j=0}^{C}[F_X(j) - F_Y(j)]^2} \\ &= \sqrt{\frac{1}{C+1}\sum_{j=0}^{C}[\sum_{i=0}^{j}(p_{i.} - p_{.i})]^2} \end{aligned}$$

Here, a is a function of $\boldsymbol{\theta}$.

Again, let

$$\boldsymbol{\theta} = (p_{00}, p_{01}, ..., p_{CC})^T$$

denote the vector of probability parameters.

Let

$$
\begin{aligned}
f(\boldsymbol{\theta}) &= \tanh^{-1}[\kappa_w(a)] \\
&= \frac{1}{2}\log(\frac{1+\kappa_w(a)}{1-\kappa_w(a)}) \\
&= \frac{1}{2}\log(\frac{1+\sum_{i=0}^{C}\sum_{j=0}^{C} w_{ij}p_{ij} - 2\sum_{i=0}^{C}\sum_{j=0}^{C} w_{ij}[0.5ap_{i.}+(1-0.5a)p_{.i}][(1-0.5a)p_{j.}+0.5ap_{.j}]}{1-\sum_{i=0}^{C}\sum_{j=0}^{C} w_{ij}p_{ij}})
\end{aligned}
$$

Then

$$
\begin{aligned}
f(\hat{\boldsymbol{\theta}}) &= \tanh^{-1}[\hat{\kappa}(\hat{a})] \\
&= \frac{1}{2}\log(\frac{1+\hat{\kappa}_w(\hat{a})}{1-\hat{\kappa}_w(\hat{a})}) \\
&= \frac{1}{2}\log(\frac{1+\sum_{i=0}^{C}\sum_{j=0}^{C} w_{ij}\hat{p}_{ij} - 2\sum_{i=0}^{C}\sum_{j=0}^{C} w_{ij}[0.5\hat{a}\hat{p}_{i.}+(1-0.5\hat{a})\hat{p}_{.i}][(1-0.5\hat{a})\hat{p}_{j.}+0.5\hat{a}\hat{p}_{.j}]}{1-\sum_{i=0}^{C}\sum_{j=0}^{C} w_{ij}\hat{p}_{ij}})
\end{aligned}
$$

where $\hat{\boldsymbol{\theta}} = (\hat{p}_{00}, \hat{p}_{01}, ..., \hat{p}_{CC})^T$ denotes the vector of estimated probability parameters and

$$
\hat{a} = \sqrt{\frac{1}{C+1}\sum_{j=0}^{C}[\sum_{i=0}^{j}(\hat{p}_{i.}-\hat{p}_{.i})]^2}
$$

That is, a is estimated from the observed proportions $\hat{\boldsymbol{\theta}}$.

If $a \neq 0, 1$, we can work with equation (2.6).

Let $\Pi_0 = \sum_{i=0}^{C}\sum_{j=0}^{C} w_{ij}p_{ij}$, $\Pi_{az} = \sum_{i=0}^{C}\sum_{j=0}^{C} w_{ij}[0.5ap_{i.} + (1-0.5a)p_{.i}][(1-0.5a)p_{j.} + 0.5ap_{.j}]$.

Then we have equation (2.7), and equation (2.8) holds as well.

$$
\begin{aligned}
\frac{\partial \Pi_{az}}{\partial p_{gh}} &= \sum_{i=0}^{C}\sum_{j=0}^{C}\{w_{ij}[0.5ap_{i.} + (1-0.5a)p_{.i}]\frac{\partial}{\partial p_{gh}}[(1-0.5a)p_{j.} + 0.5ap_{.j}] \\
&\quad + w_{ij}[(1-0.5a)p_{j.} + 0.5ap_{.j}]\frac{\partial}{\partial p_{gh}}[0.5ap_{i.} + (1-0.5a)p_{.i}]\} \\
&= \sum_{i=0}^{C}\sum_{j=0}^{C}\{w_{ij}[0.5ap_{i.} + (1-0.5a)p_{.i}][-0.5p_{j.}(\frac{\partial}{\partial p_{gh}}a) + (1-0.5a)I(g=j) \\
&\quad + 0.5p_{.j}(\frac{\partial}{\partial p_{gh}}a) + 0.5aI(h=j)] + w_{ij}[(1-0.5a)p_{j.} + 0.5ap_{.j}][0.5p_{i.}(\frac{\partial}{\partial p_{gh}}a) \\
&\quad + 0.5aI(g=i) - 0.5p_{.i}(\frac{\partial}{\partial p_{gh}}a) + (1-0.5a)I(h=i)]\} \qquad (2.10)
\end{aligned}
$$

where

$$
\begin{aligned}
\Sigma_{gh} = \frac{\partial}{\partial p_{gh}}a &= \frac{\partial}{\partial p_{gh}}\sqrt{\frac{1}{C+1}\sum_{j=0}^{C}[\sum_{i=0}^{j}(p_{i.} - p_{.i})]^2} \\
&= 0.5\{\frac{1}{C+1}\sum_{j=0}^{C}[\sum_{i=0}^{j}(p_{i.} - p_{.i})]^2\}^{-0.5}(\frac{2}{C+1})\sum_{j=0}^{C}\{[\sum_{i=0}^{j}(p_{i.} - p_{.i})] \\
&\quad [\sum_{i=0}^{j}(I(g=i) - I(h=i))]\}
\end{aligned}
$$

Plug equations (2.8) and (2.10) into equation (2.7), then plug equation (2.7) into equation (2.6). The asymptotic variance of $\sqrt{n}(\tanh^{-1}(\hat{\kappa}_w(\hat{a})) - \tanh^{-1}(\kappa_w(a)))$ is

$$
V = \sum_{g=0}^{C}\sum_{h=0}^{C}p_{gh}(D_{gh})^2 - (\sum_{g=0}^{C}\sum_{h=0}^{C}p_{gh}D_{gh})^2
$$

If $a = 0$, $\kappa_w(a)$ is reduced to Cohen's weighted kappa; if $a = 1$, $\kappa_w(a)$ is reduced to the weighted RMAC. One can then use the asymptotic results of Cohen's weighted kappa and the weighted RMAC, respectively.

If $\hat{a} \neq 0, 1$, replacing all p_{ij} with $\hat{p}_{ij}$, yields the estimated asymptotic variance of $\sqrt{n}(\tanh^{-1}(\hat{\kappa}_w(\hat{a})) - \tanh^{-1}(\kappa_w(a)))$

$$\hat{V} = \sum_{g=0}^{C}\sum_{h=0}^{C} \hat{p}_{gh}(\hat{D}_{gh})^2 - (\sum_{g=0}^{C}\sum_{h=0}^{C} \hat{p}_{gh}\hat{D}_{gh})^2$$

Hence the asymptotic variance estimate for $\sqrt{n}(\hat{\kappa}_w(\hat{a}) - \kappa_w(a))$ is

$$[1 - \hat{\kappa}_w^2(\hat{a})]^2 \hat{V}$$

If $\hat{a} = 0$, $\hat{\kappa}_w(\hat{a})$ is reduced to the estimate of Cohen's weighted kappa; if $\hat{a} = 1$, $\hat{\kappa}_w(\hat{a})$ is reduced to the estimate of the weighted RMAC. One can then use the asymptotic variance estimate of the estimate of Cohen's weighted kappa and the weighted RMAC, respectively.

2.4 Continuous Data

2.4.1 Definition

Let (X,Y) denote a bivariate continuous response with means μ_X and μ_Y and covariance matrix

$$\begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}$$

with cumulative distribution function (CDF) F_{XY}. In addition, let F_X and F_Y denote the marginal CDFs of X and Y, respectively.

From equation (2.1) with squared difference cost function, $c(x,y) = (x-y)^2$, it follows that

$$
\begin{aligned}
&E_{F_{U_1}} E_{F_{U_2}} (U_1 - U_2)^2 \\
&= E_{F_{U_1}} E_{F_{U_2}} (U_1^2 - 2U_1U_2 + U_2^2) \\
&= E_{U_1}(U_1^2) - 2\mu_{U_1}\mu_{U_2} + E_{U_2}(U_2^2) \\
&= 0.5aE_X(X^2) + (1-0.5a)E_Y(Y^2) - 2[0.5a\mu_X + (1-0.5a)\mu_Y][(1-0.5a)\mu_X + \\
&0.5a\mu_Y] + (1-0.5a)E_X(X^2) + 0.5aE_Y(Y^2) \\
&= \sigma_X^2 + \sigma_Y^2 + (0.5a^2 - a + 1)(\mu_X - \mu_Y)^2
\end{aligned}
$$

Then the general class of agreement coefficients for continuous responses is

$$
\begin{aligned}
\rho(a) &= 1 - \frac{\sigma_X^2 + \sigma_Y^2 + (\mu_X - \mu_Y)^2 - 2\rho\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2 + (0.5a^2 - a + 1)(\mu_X - \mu_Y)^2} \\
&= \frac{2\rho\sigma_X\sigma_Y + a(0.5a-1)(\mu_X - \mu_Y)^2}{\sigma_X^2 + \sigma_Y^2 + (0.5a^2 - a + 1)(\mu_X - \mu_Y)^2} \\
&= \frac{2\sigma_{XY} + a(0.5a-1)(\mu_X - \mu_Y)^2}{\sigma_X^2 + \sigma_Y^2 + (0.5a^2 - a + 1)(\mu_X - \mu_Y)^2}
\end{aligned}
$$

where ρ is Pearson correlation coefficient; and $\rho(a)$ in terms of u, υ and ρ is

$$
\rho(a) = \frac{2\rho + a(0.5a-1)u^2}{\upsilon + 1/\upsilon + (0.5a^2 - a + 1)u^2}
$$

where

$$\upsilon = \sigma_X/\sigma_Y = \text{scale shift}$$

$$u = (\mu_X - \mu_Y)/\sqrt{\sigma_X \sigma_Y} = \text{location shift relative to the scale}$$

A value of $a = 0$ yields the CCC, a value of $a = 1$ yields the RMAC, and a value of a between 0 and 1 yields mixtures of the CCC and the RMAC.

The true value of a is calculated as

$$a = \sqrt{\int_{-\infty}^{\infty} \{F_{n,X}(x) - F_{n,Y}(x)\}^2 dH(x)} \tag{2.11}$$

where $H(x) = 0.5F_X(x)+0.5F_Y(x)$, $F_{n,X}(t) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq t)$, $F_{n,Y}(t) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \leq t)$, and $I(.)$ is an indicator function. This is analogous to the two-sample Cramer-von Mises criterion [2]. In Anderson's situation, $H(x)$ is defined as $H(x) = \lambda F_X(x) + (1 - \lambda)F_Y(x)$, where

$$\lambda = \lim \frac{n_X}{n_X + n_Y}$$

here n_X and n_Y are the sample sizes for X and Y, respectively. Because we have a bivariate sampling case, $\lambda = 0.5$.

The Lebesgue-Stieltjes integral (2.11) is approximately equivalent to

$$\hat{a} = \sqrt{0.5\{\frac{1}{n}\sum_{i=1}^{n}[F_{n,X}(x_i) - F_{n,Y}(x_i)]^2 + \frac{1}{n}\sum_{i=1}^{n}[F_{n,X}(y_i) - F_{n,Y}(y_i)]^2\}} \tag{2.12}$$

which means $\hat{a}$ can be used to estimate a.

2.4.2 Asymptotic Distribution

2.4.2.1 Fixed and Known a

Let (X_1, Y_1), ..., (X_n, Y_n) be independent observations from a bivariate distribution such that the fourth-order moments exist, i.e. $EX_i^4 < \infty$ and $EY_i^4 < \infty$. The estimate of $\rho(a)$ is

$$\hat{\rho}(a) = \frac{2S_{XY} + a(0.5a - 1)(\bar{X} - \bar{Y})^2}{S_X^2 + S_Y^2 + (0.5a^2 - a + 1)(\bar{X} - \bar{Y})^2}$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$, $S_X^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$, $S_Y^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})^2$ and $S_{XY} = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$.

Let

$$\boldsymbol{\theta} = (E(X_1), E(Y_1), E(X_1^2), E(Y_1^2), E(X_1Y_1))^T$$

and

$$\hat{\boldsymbol{\theta}} = (\frac{1}{n}\sum_{i=1}^n X_i, \frac{1}{n}\sum_{i=1}^n Y_i, \frac{1}{n}\sum_{i=1}^n X_i^2, \frac{1}{n}\sum_{i=1}^n Y_i^2, \frac{1}{n}\sum_{i=1}^n X_iY_i)^T$$

By the Central Limit Theorem,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, \Sigma)$$

where

$$\Sigma = \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, Y_1) & \mathrm{Cov}(X_1, X_1^2) & \mathrm{Cov}(X_1, Y_1^2) & \mathrm{Cov}(X_1, X_1Y_1) \\ \mathrm{Cov}(X_1, Y_1) & \mathrm{Var}(Y_1) & \mathrm{Cov}(Y_1, X_1^2) & \mathrm{Cov}(Y_1, Y_1^2) & \mathrm{Cov}(Y_1, X_1Y_1) \\ \mathrm{Cov}(X_1, X_1^2) & \mathrm{Cov}(Y_1, X_1^2) & \mathrm{Var}(X_1^2) & \mathrm{Cov}(X_1^2, Y_1^2) & \mathrm{Cov}(X_1^2, X_1Y_1) \\ \mathrm{Cov}(X_1, Y_1^2) & \mathrm{Cov}(Y_1, Y_1^2) & \mathrm{Cov}(X_1^2, Y_1^2) & \mathrm{Var}(Y_1^2) & \mathrm{Cov}(Y_1^2, X_1Y_1) \\ \mathrm{Cov}(X_1, X_1Y_1) & \mathrm{Cov}(Y_1, X_1Y_1) & \mathrm{Cov}(X_1^2, X_1Y_1) & \mathrm{Cov}(Y_1^2, X_1Y_1) & \mathrm{Var}(X_1Y_1) \end{pmatrix}$$

is the variance-covariance matrix for $\boldsymbol{\theta}$.

Let $f : \Re^5 \longrightarrow \Re$ be the Fisher's Z-transformation on $\rho(a)$

$$\begin{aligned} f(\boldsymbol{\theta}) &= \tanh^{-1}[(\rho(a)] \\ &= \frac{1}{2}\log(\frac{1+\rho(a)}{1-\rho(a)}) \\ &= \frac{1}{2}\log(\frac{\sigma^2_{X_1} + \sigma^2_{Y_1} + 2\sigma_{X_1Y_1} + (a-1)^2(\mu_{X_1} - \mu_{Y_1})^2}{\sigma^2_{X_1} + \sigma^2_{Y_1} - 2\sigma_{X_1Y_1} + (\mu_{X_1} - \mu_{Y_1})^2}) \end{aligned}$$

and hence

$$\begin{aligned} f(\hat{\boldsymbol{\theta}}) &= \tanh^{-1}[\hat{\rho}(a)] \\ &= \frac{1}{2}\log(\frac{1+\hat{\rho}(a)}{1-\hat{\rho}(a)}) \\ &= \frac{1}{2}\log(\frac{S^2_X + S^2_Y + 2S_{XY} + (a-1)^2(\bar{X} - \bar{Y})^2}{S^2_X + S^2_Y - 2S_{XY} + (\bar{X} - \bar{Y})^2}) \end{aligned}$$

Then let $\theta_1, ..., \theta_5$ denote the elements of $\boldsymbol{\theta}$. Note that $\sigma^2_{X_1} = E(X_1^2) - [E(X_1)]^2 = \theta_3 - \theta_1^2$, $\sigma^2_{Y_1} = E(Y_1^2) - [E(Y_1)]^2 = \theta_4 - \theta_2^2$, $\sigma_{X_1Y_1} = E(X_1Y_1) - E(X_1)E(Y_1) = \theta_5 - \theta_1\theta_2$, $\mu_{X_1} = \theta_1$

and $\mu_{Y_1} = \theta_2$. Therefore,

$$f(\boldsymbol{\theta}) = \frac{1}{2}\log(\frac{\theta_3 - \theta_1^2 + \theta_4 - \theta_2^2 + 2(\theta_5 - \theta_1\theta_2) + (a-1)^2(\theta_1 - \theta_2)^2}{\theta_3 - \theta_1^2 + \theta_4 - \theta_2^2 - 2(\theta_5 - \theta_1\theta_2) + (\theta_1 - \theta_2)^2})$$

For simplicity, let

$$\Pi_1 = \theta_3 - \theta_1^2 + \theta_4 - \theta_2^2 + 2(\theta_5 - \theta_1\theta_2) + (a-1)^2(\theta_1 - \theta_2)^2$$

and

$$\Pi_2 = \theta_3 - \theta_1^2 + \theta_4 - \theta_2^2 - 2(\theta_5 - \theta_1\theta_2) + (\theta_1 - \theta_2)^2$$

Then we have

$$\frac{\partial \Pi_1}{\partial \theta_1} = -2\theta_1 - 2\theta_2 + 2(a-1)^2(\theta_1 - \theta_2)$$

$$\frac{\partial \Pi_2}{\partial \theta_1} = -2\theta_1 + 2\theta_2 + 2(\theta_1 - \theta_2) = 0$$

$$\frac{\partial \Pi_1}{\partial \theta_2} = -2\theta_2 - 2\theta_1 - 2(a-1)^2(\theta_1 - \theta_2)$$

$$\frac{\partial \Pi_2}{\partial \theta_2} = -2\theta_2 + 2\theta_1 - 2\theta_1 + 2\theta_2 = 0$$

$$\frac{\partial \Pi_1}{\partial \theta_3} = \frac{\partial \Pi_2}{\partial \theta_3} = 1$$

$$\frac{\partial \Pi_1}{\partial \theta_4} = \frac{\partial \Pi_2}{\partial \theta_4} = 1$$

$$\frac{\partial \Pi_1}{\partial \theta_5} = 2$$

and

$$\frac{\partial \Pi_2}{\partial \theta_5} = -2$$

So the gradient matrix of f, evaluated at $\boldsymbol{\theta}$ is

$$\nabla f(\boldsymbol{\theta}) = \frac{1}{2}(\frac{\Pi_2}{\Pi_1}) \begin{pmatrix} \frac{\frac{\partial \Pi_1}{\partial \theta_1}\Pi_2 - \frac{\partial \Pi_2}{\partial \theta_1}\Pi_1}{\Pi_2^2} \\ \frac{\frac{\partial \Pi_1}{\partial \theta_2}\Pi_2 - \frac{\partial \Pi_2}{\partial \theta_2}\Pi_1}{\Pi_2^2} \\ \frac{\frac{\partial \Pi_1}{\partial \theta_3}\Pi_2 - \frac{\partial \Pi_2}{\partial \theta_3}\Pi_1}{\Pi_2^2} \\ \frac{\frac{\partial \Pi_1}{\partial \theta_4}\Pi_2 - \frac{\partial \Pi_2}{\partial \theta_4}\Pi_1}{\Pi_2^2} \\ \frac{\frac{\partial \Pi_1}{\partial \theta_5}\Pi_2 - \frac{\partial \Pi_2}{\partial \theta_5}\Pi_1}{\Pi_2^2} \end{pmatrix}$$

Therefore, if we let

$$\Sigma^* = (\nabla f)^T \Sigma (\nabla f)$$

then by the delta method,

$$\sqrt{n}(f(\hat{\boldsymbol{\theta}}) - f(\boldsymbol{\theta})) \xrightarrow{d} N(0, \Sigma^*)$$

To find the asymptotic normality of $\hat{\rho}(a)$, we let

$$\rho(a) = \tanh(f(\boldsymbol{\theta}))$$

By using the delta method one more time, we get

$$\sqrt{n}(\hat{\rho}(a) - \rho(a)) \xrightarrow{d} N(0, (1 - \rho^2(a))^2 \Sigma^*)$$

The estimated asymptotic standard error of $\hat{\rho}(a)$, obtained by substituting sample counterparts that are consistent estimates, is $\sqrt{\frac{1}{n}(1-\hat{\rho}^2(a))^2\hat{\Sigma}^*}$. By the Slutsky theorem, the $100 \times (1-\alpha)\%$ asymptotic confidence interval for $\rho(a)$ is

$$(\hat{\rho}(a) - z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{n}(1-\hat{\rho}^2(a))^2\hat{\Sigma}^*}, \hat{\rho}(a) + z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{n}(1-\hat{\rho}^2(a))^2\hat{\Sigma}^*})$$

With tedious but straightforward algebra, we can show that when $a = 0$ and $a = 1$, the asymptotic variances of $\hat{\rho}(a)$ are equivalent to the ones of Lin's CCC and the RMAC, respectively.

2.4.2.2 Estimated a

Let (X_1, Y_1), ..., (X_n, Y_n) be independent observations from a bivariate distribution such that the fourth-order moments exist, i.e. $EX_i^4 < \infty$ and $EY_i^4 < \infty$. The estimate of $\rho(a)$ is

$$\hat{\rho}(\hat{a}) = \frac{2S_{XY} + \hat{a}(0.5\hat{a}-1)(\bar{X}-\bar{Y})^2}{S_X^2 + S_Y^2 + (0.5\hat{a}^2 - \hat{a} + 1)(\bar{X}-\bar{Y})^2}$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$, $S_X^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$, $S_Y^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})^2$, $S_{XY} = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$, and

$$\hat{a} = \sqrt{0.5\{\frac{1}{n}\sum_{i=1}^n [F_{n,X}(x_i) - F_{n,Y}(x_i)]^2 + \frac{1}{n}\sum_{i=1}^n [F_{n,X}(y_i) - F_{n,Y}(y_i)]^2\}}$$

Define

$$U = [F_{n,X}(X_1) - F_{n,Y}(X_1)]^2$$

$$u = \frac{1}{n}\sum_{i=1}^{n}[F_{n,X}(X_i) - F_{n,Y}(X_i)]^2$$

$$V = [F_{n,X}(Y_1) - F_{n,Y}(Y_1)]^2$$

$$v = \frac{1}{n}\sum_{i=1}^{n}[F_{n,X}(Y_i) - F_{n,Y}(Y_i)]^2$$

Let

$$\boldsymbol{\delta} = (E(X_1), E(Y_1), E(X_1^2), E(Y_1^2), E(X_1Y_1), E(U), E(V))^T$$

and

$$\hat{\boldsymbol{\delta}} = (\frac{1}{n}\sum_{i=1}^{n}X_i, \frac{1}{n}\sum_{i=1}^{n}Y_i, \frac{1}{n}\sum_{i=1}^{n}X_i^2, \frac{1}{n}\sum_{i=1}^{n}Y_i^2, \frac{1}{n}\sum_{i=1}^{n}X_iY_i, u, v)^T$$

By the Central Limit Theorem,

$$\sqrt{n}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}) \xrightarrow{d} N(\mathbf{0}, \Omega)$$

Here

$$\Omega = \begin{pmatrix} \Sigma & \Psi_1 \\ \Psi_1^T & \Psi_2 \end{pmatrix}$$

is the variance-covariance matrix for $\boldsymbol{\delta}$, where Σ was defined in section 2.4.2.1 and

$$\Psi_1 = \begin{pmatrix} \mathrm{Cov}(X_1, U) & \mathrm{Cov}(X_1, V) \\ \mathrm{Cov}(Y_1, U) & \mathrm{Cov}(Y_1, V) \\ \mathrm{Cov}(X_1^2, U) & \mathrm{Cov}(X_1^2, V) \\ \mathrm{Cov}(Y_1^2, U) & \mathrm{Cov}(Y_1^2, V) \\ \mathrm{Cov}(X_1Y_1, U) & \mathrm{Cov}(X_1Y_1, V) \end{pmatrix}$$

$$\Psi_2 = \begin{pmatrix} \mathrm{Var}(U) & \mathrm{Cov}(U, V) \\ \mathrm{Cov}(U, V) & \mathrm{Var}(V) \end{pmatrix}$$

Since

$$\begin{aligned} \sqrt{0.5[E(U) + E(V)]} &= \sqrt{0.5 \int_{-\infty}^{\infty} [F_{n,X}(x) - F_{n,Y}(x)]^2 f_X(x) + [F_{n,X}(x) - F_{n,Y}(x)]^2 f_Y(x)} \\ &= \sqrt{0.5 \int_{-\infty}^{\infty} \{F_{n,X}(x) - F_{n,Y}(x)\}^2 d(F_X(x) + F_Y(x))} \\ &= \sqrt{\int_{-\infty}^{\infty} \{F_{n,X}(x) - F_{n,Y}(x)\}^2 dH(x)} \\ &= a \end{aligned}$$

let $g : \Re^7 \longrightarrow \Re$ be the Fisher's Z-transformation on $\rho(a)$

$$\begin{aligned} g(\boldsymbol{\delta}) &= \tanh^{-1}[(\rho(a)] \\ &= \frac{1}{2}\log(\frac{1+\rho(a)}{1-\rho(a)}) \\ &= \frac{1}{2}\log(\frac{\sigma^2_{X_1} + \sigma^2_{Y_1} + 2\sigma_{X_1Y_1} + (a-1)^2(\mu_{X_1} - \mu_{Y_1})^2}{\sigma^2_{X_1} + \sigma^2_{Y_1} - 2\sigma_{X_1Y_1} + (\mu_{X_1} - \mu_{Y_1})^2}) \end{aligned}$$

and hence

$$\begin{aligned} g(\hat{\boldsymbol{\delta}}) &= \tanh^{-1}[\hat{\rho}(\hat{a})] \\ &= \frac{1}{2}\log(\frac{1+\hat{\rho}(\hat{a})}{1-\hat{\rho}(\hat{a})}) \\ &= \frac{1}{2}\log(\frac{S^2_X + S^2_Y + 2S_{XY} + (\hat{a}-1)^2(\bar{X} - \bar{Y})^2}{S^2_X + S^2_Y - 2S_{XY} + (\bar{X} - \bar{Y})^2}) \end{aligned}$$

Then let $\delta_1, ..., \delta_7$ denote the elements of $\boldsymbol{\delta}$. Note that $\sigma^2_{X_1} = E(X^2_1) - [E(X_1)]^2 = \theta_3 - \theta^2_1$, $\sigma^2_{Y_1} = E(Y^2_1) - [E(Y_1)]^2 = \theta_4 - \theta^2_2$, $\sigma_{X_1Y_1} = E(X_1Y_1) - E(X_1)E(Y_1) = \theta_5 - \theta_1\theta_2$, $\mu_{X_1} = \theta_1$ and $\mu_{Y_1} = \theta_2$. Therefore,

$$g(\boldsymbol{\delta}) = \frac{1}{2}\log(\frac{\delta_3 - \delta^2_1 + \delta_4 - \delta^2_2 + 2(\delta_5 - \delta_1\delta_2) + (\sqrt{0.5(\delta_6 + \delta_7)} - 1)^2(\delta_1 - \delta_2)^2}{\delta_3 - \delta^2_1 + \delta_4 - \delta^2_2 - 2(\delta_5 - \delta_1\delta_2) + (\delta_1 - \delta_2)^2})$$

For simplicity, let

$$\Lambda_1 = \delta_3 - \delta^2_1 + \delta_4 - \delta^2_2 + 2(\delta_5 - \delta_1\delta_2) + (\sqrt{0.5(\delta_6 + \delta_7)} - 1)^2(\delta_1 - \delta_2)^2$$

$$\Lambda_2 = \delta_3 - \delta_1^2 + \delta_4 - \delta_2^2 - 2(\delta_5 - \delta_1\delta_2) + (\delta_1 - \delta_2)^2$$

Then we have

$$\frac{\partial\Lambda_1}{\partial\delta_1} = -2\delta_1 - 2\delta_2 + 2(\sqrt{0.5(\delta_6 + \delta_7)} - 1)^2(\delta_1 - \delta_2)$$

$$\frac{\partial\Lambda_2}{\partial\delta_1} = -2\delta_1 + 2\delta_2 + 2(\delta_1 - \delta_2) = 0$$

$$\frac{\partial\Lambda_1}{\partial\delta_2} = -2\delta_2 - 2\delta_1 - 2(\sqrt{0.5(\delta_6 + \delta_7)} - 1)^2(\delta_1 - \delta_2)$$

$$\frac{\partial\Lambda_2}{\partial\delta_2} = -2\delta_2 + 2\delta_1 - 2\delta_1 + 2\delta_2 = 0$$

$$\frac{\partial\Lambda_1}{\partial\delta_3} = \frac{\partial\Lambda_2}{\partial\delta_3} = 1$$

$$\frac{\partial\Lambda_1}{\partial\delta_4} = \frac{\partial\Lambda_2}{\partial\delta_4} = 1$$

$$\frac{\partial\Lambda_1}{\partial\delta_5} = 2$$

$$\frac{\partial\Lambda_2}{\partial\delta_5} = -2$$

$$\frac{\partial\Lambda_1}{\partial\delta_6} = \frac{\partial\Lambda_1}{\partial\delta_7} = 0.5(\sqrt{0.5(\delta_6 + \delta_7)} - 1)(\delta_1 - \delta_2)^2[0.5(\delta_6 + \delta_7)]^{-\frac{1}{2}}$$

$$\frac{\partial\Lambda_2}{\partial\delta_6} = \frac{\partial\Lambda_2}{\partial\delta_7} = 0$$

So the gradient matrix of g, evaluated at $\boldsymbol{\delta}$ is

$$\nabla g(\boldsymbol{\delta}) = \frac{1}{2}(\frac{\Lambda_2}{\Lambda_1}) \begin{pmatrix} \frac{\frac{\partial \Lambda_1}{\partial \delta_1}\Lambda_2 - \frac{\partial \Lambda_2}{\partial \delta_1}\Lambda_1}{\Lambda_2^2} \\ \frac{\frac{\partial \Lambda_1}{\partial \delta_2}\Pi_2 - \frac{\partial \Lambda_2}{\partial \delta_2}\Lambda_1}{\Lambda_2^2} \\ \frac{\frac{\partial \Lambda_1}{\partial \delta_3}\Lambda_2 - \frac{\partial \Lambda_2}{\partial \delta_3}\Lambda_1}{\Lambda_2^2} \\ \frac{\frac{\partial \Lambda_1}{\partial \delta_4}\Lambda_2 - \frac{\partial \Lambda_2}{\partial \delta_4}\Lambda_1}{\Lambda_2^2} \\ \frac{\frac{\partial \Lambda_1}{\partial \delta_5}\Lambda_2 - \frac{\partial \Lambda_2}{\partial \delta_5}\Lambda_1}{\Lambda_2^2} \\ \frac{\frac{\partial \Lambda_1}{\partial \delta_6}\Lambda_2 - \frac{\partial \Lambda_2}{\partial \delta_6}\Lambda_1}{\Lambda_2^2} \\ \frac{\frac{\partial \Lambda_1}{\partial \delta_7}\Lambda_2 - \frac{\partial \Lambda_2}{\partial \delta_7}\Lambda_1}{\Lambda_2^2} \end{pmatrix}$$

Therefore, if we let

$$\Omega^* = (\nabla g)^T \Omega (\nabla g)$$

then by the delta method,

$$\sqrt{n}(g(\hat{\boldsymbol{\delta}}) - g(\boldsymbol{\delta})) \xrightarrow{d} N(0, \Omega^*)$$

To find the asymptotic normality of $\hat{\rho}(\hat{a})$, we let

$$\rho(a) = \tanh(g(\boldsymbol{\delta}))$$

By using the delta method one more time, we get

$$\sqrt{n}(\hat{\rho}(\hat{a}) - \rho(a)) \xrightarrow{d} N(0, (1 - \rho^2(a))^2 \Omega^*)$$

The estimated asymptotic standard error of $\hat{\rho}(\hat{a})$, obtained by substituting sample counterparts that are consistent estimates, is $\sqrt{\frac{1}{n}(1-\hat{\rho}^2(\hat{a}))^2\hat{\Omega}^*}$. By the Slutsky theorem, the $100 \times (1-\alpha)\%$ asymptotic confidence interval for $\rho(a)$ is

$$(\hat{\rho}(\hat{a}) - z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{n}(1-\hat{\rho}^2(\hat{a}))^2\hat{\Omega}^*}, \hat{\rho}(\hat{a}) + z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{n}(1-\hat{\rho}^2(\hat{a}))^2\hat{\Omega}^*})$$

Chapter 3

Simulations

3.1 Simulation Study for Categorical Responses

We conducted analysis using simulated data to assess the accuracy and precision of $\hat{\kappa}(a)$ and $\hat{\kappa}(\hat{a})$. Two cases were considered: a was fixed and known and a was estimated from the sample proportions. For each case, we considered three distributions with $C = 1$ and denoted by the vector $\boldsymbol{\theta} = (p_{11}, p_{21}, p_{12}, p_{22})$; these distributions are (i) (0.6,0.1,0.1,0.2); (ii) (0.58,0.22,0.02,0.18) and (iii) (0.5,0.4,0,0.1). Further, we considered the following distributions, (iv) the distribution with $C = 2$ formed by taking the probabilities from Table 3.1; (v) the distribution with $C = 3$ formed by taking the sample proportions from data from Westlund and Kurkland [37](Table 3.2) and (vi) the distribution with $C = 4$ formed by taking the sample proportions from data from Brostoff, et al [6](Table 3.3). For each distribution, we simulated 100000 data sets of size n=20, 50 and 100.

Table 3.1 The fourth distribution

	Y			
X	0	1	2	Total
0	0.25	0.01	0.24	0.5
1	0.02	0.28	0.1	0.4
2	0.03	0.01	0.06	0.1
Total	0.3	0.3	0.4	1

Case 1. Fixed and Known a:

Table 3.2 The fifth distribution

	Neurologist 2				
Neurologist 1	1	2	3	4	Total
1	38	5	0	1	44
2	33	11	3	0	47
3	10	14	5	6	35
4	3	7	3	10	23
Total	84	37	11	17	149

Table 3.3 The sixth distribution

	RAST					
MAST	Negative	Weak	Moderate	High	Very High	Total
Negative	86	3	14	0	2	105
Weak	26	0	10	4	0	40
Moderate	20	2	22	4	1	49
High	11	1	37	16	14	79
Very High	3	0	15	24	48	90
Total	146	6	98	48	65	363

For each distribution, we used the square root of the average squared distance to calculate a. Then we computed $\kappa(a)$ using one or more weighting schemes and treated it as the true parameter value for that weighting scheme. For each simulated data set of a distribution, we used a=0,0.2,0.4,0.6,0.8 and 1 respectively as fixed and known values of a, to compute $\hat{\kappa}(a)$, $\hat{SE}(\hat{\kappa}(a))$ and confidence interval. The series of values of a reflected the uncertainty of selecting a a priori in practice and enabled us to compare the performance of the estimators with different values of a. In particular, since a=0 yields Cohen's kappa and a=1 yields the RMAC, we were able to assess the performance of these two important measures as well.

The "a" column contains the true values of a for the distributions. The "$\kappa(a)$" column contains the true values of $\kappa(a)$. The "Mean(std)" column contains the mean values of $\hat{\kappa}(a)$, which were the averages of valid $\hat{\kappa}(a)$ computed from the simulated data sets of the distribution, and the standard deviations of valid $\hat{\kappa}(a)$ computed from the simulated data sets of the distribution. The "Relative Bias(%)" column contains the relative bias of the agreement estimate, which is calculated by [(mean $\hat{\kappa}(a)$-true $\kappa(a)$)/mean $\hat{\kappa}(a)$]× 100. The "SE(std)" column contains the mean values of $\hat{SE}(\hat{\kappa}(a))$, which were the averages of valid $\hat{SE}(\hat{\kappa}(a))$ computed from the simulated data sets of the distribution, and the standard deviations of valid $\hat{SE}(\hat{\kappa}(a))$ computed from the simulated data sets of the distribution. Each confidence interval computed from the generated data set was evaluated to determine whether the true parameter value was contained within the confidence interval, and the coverage is contained in the "Coverage of CI" column.

Case 2. Estimated a:

The simulation was conducted differently from Case 1 only in that, instead of using a series of fixed and known a, we estimated a from each generated data set using the square root of the average squared distance. As in Case 1, the true value of a was calculated from the underlying distribution using the square root of the average squared distance, so both cases share the same true value of a and true value of $\kappa(a)$ for each distribution. Same quantities were calculated as in Case 1.

Results:

The simulation results are listed in Table 3.4 - Table 3.10. For each of the distributions with noticeable difference between marginal distributions and each sample size, the selected value of a=0 yields the largest agreement and the selected value of a=1 yields the smallest agreement, that is, the Cohen's kappa and the RMAC form the upper and lower bounds of the class of agreement. The larger the difference between two marginal distributions (i.e. the larger the value in "a" column), the larger the difference between the bounds, suggesting the impact of different marginal distributions on agreement.

For agreement estimation, if the selected a is less than true a, then $\hat{\kappa}(a)$ tends to overestimate $\kappa(a)$; if the selected a is greater than true a, then $\hat{\kappa}(a)$ tends to underestimate $\kappa(a)$. An accurate estimate is obtained only if the selected a is very close to true a. On the other hand, $\hat{\kappa}(\hat{a})$ always resembles the performance of the estimator with selected a close to true a.

For standard error estimation, the estimates become smaller as sample size increases. The estimated SE of $\hat{\kappa}(a)$ increases as the value of the selected a increases, and the difference between the estimated SEs of $\hat{\kappa}(0)$ and $\hat{\kappa}(1)$ gets larger as the true a gets larger. The estimated SE of $\hat{\kappa}(\hat{a})$ is slightly larger than the one of $\hat{\kappa}(a)$, which is due to the fact that $\hat{\kappa}(a)$ tends to underestimate the variability because the variability associated with choosing a is not taken into account, whereas $\hat{\kappa}(\hat{a})$ accounts for the estimation of a and yields the correct level of precision. The difference between the average of estimated SEs and the standard deviations of $\hat{\kappa}(a)$ and the difference between the average of estimated SEs and the standard deviations of $\hat{\kappa}(\hat{a})$, which can be regarded as the true values of the standard deviations of $\hat{\kappa}(a)$ and $\hat{\kappa}(\hat{a})$, are all less than one-half standard deviation of the estimated SEs. This implies that the SE formulas proposed are fairly accurate even for small sample sizes. We also notice that the SE is always less than the true standard deviation. This may imply that the SE formulas proposed slightly underestimate the true SE.

For coverage of confidence interval, the coverage improves as sample size increases. For selected a, quite a few coverage probabilities are between 0.93 and 0.94, close to the true probability 0.95, except for some selected values of a (e.g. selected a=0, 0.8 and 1); whereas for estimated a, most coverage probabilities are between 0.93 and 0.94. This implies that the proposed confidence interval formula for $\hat{\kappa}(\hat{a})$ works better than the one for $\hat{\kappa}(a)$ when a is selected a priori, even for small sample sizes.

3.2 Simulation Study for Continuous Responses

We conducted analysis using simulated data to assess the accuracy and precision of $\hat{\rho}(a)$ and $\hat{\rho}(\hat{a})$ for continuous responses. Two cases were considered: a was fixed and known and a was estimated. For each case, we considered the following three bivariate normal distributions, which were selected from those in Lin's simulation study [31]:

Case 1. Mean $(0, 0)$ and covariance matrix

$$\begin{pmatrix} 1 & 0.95 \\ 0.95 & 1 \end{pmatrix}$$

with no difference in location and scale parameters, and strong positive correlation.

Case 2. Mean $(-\frac{\sqrt{0.1}}{2}, \frac{\sqrt{0.1}}{2})$ and covariance matrix

$$\begin{pmatrix} 1.1^2 & 0.95 \times 1.1 \times 0.9 \\ 0.95 \times 1.1 \times 0.9 & 0.9^2 \end{pmatrix}$$

with slight difference in both location and scale parameters, and strong positive correlation.

Table 3.4 Simulation results for $\hat{\kappa}(0)$

Distribution	a	w	$\kappa(a)$	n	Mean(std)	Relative Bias(%)	SE(std)	Coverage of CI
(i)	0	0-1	0.524	20	0.505(0.218)	-3.8	0.198(0.038)	0.907
				50	0.517(0.134)	-1.4	0.129(0.015)	0.933
				100	0.521(0.094)	-0.6	0.092(0.007)	0.939
(ii)	0.2	0-1	0.445	20	0.442(0.198)	-0.7	0.18(0.036)	0.885
				50	0.450(0.122)	1.1	0.119(0.011)	0.924
				100	0.453(0.086)	1.8	0.085(0.005)	0.937
(iii)	0.4	0-1	0.109	20	0.198(0.138)	44.9	0.118(0.054)	0.836
				50	0.199(0.084)	45.2	0.081(0.018)	0.883
				100	0.2(0.059)	45.5	0.058(0.009)	0.718
(iv)	0.216	0-1	0.394	20	0.397(0.14)	0.8	0.133(0.017)	0.929
				50	0.402(0.088)	2	0.086(0.006)	0.942
				100	0.404(0.062)	2.5	0.061(0.003)	0.945
		C-A	0.24	20	0.26(0.159)	7.7	0.149(0.028)	0.924
				50	0.26(0.1)	7.7	0.098(0.012)	0.941
				100	0.261(0.071)	8	0.07(0.006)	0.939
		F-C	0.084	20	0.119(0.2)	29.4	0.183(0.044)	0.924
				50	0.117(0.125)	28.2	0.121(0.018)	0.938
				100	0.116(0.088)	27.6	0.087(0.009)	0.933
(v)	0.161	0-1	0.199	20	0.201(0.137)	1	0.128(0.019)	0.918
				50	0.205(0.087)	2.9	0.085(0.007)	0.938
				100	0.207(0.062)	3.9	0.061(0.004)	0.943
		C-A	0.371	20	0.364(0.142)	-1.9	0.13(0.022)	0.893
				50	0.374(0.089)	0.8	0.087(0.008)	0.93
				100	0.376(0.063)	1.3	0.062(0.004)	0.939
		F-C	0.517	20	0.503(0.167)	-2.8	0.146(0.04)	0.86
				50	0.517(0.105)	0	0.1(0.017)	0.913
				100	0.52(0.074)	0.6	0.072(0.008)	0.93
(vi)	0.102	0-1	0.316	20	0.31(0.129)	-1.9	0.123(0.012)	0.922
				50	0.315(0.082)	-0.3	0.08(0.004)	0.939
				100	0.317(0.058)	0.3	0.057(0.002)	0.946
		C-A	0.558	20	0.544(0.124)	-2.6	0.117(0.021)	0.916
				50	0.553(0.077)	-0.9	0.076(0.008)	0.938
				100	0.556(0.055)	-0.4	0.054(0.004)	0.942
		F-C	0.711	20	0.695(0.128)	-2.3	0.114(0.042)	0.872
				50	0.706(0.079)	-0.7	0.075(0.018)	0.913
				100	0.709(0.055)	-0.3	0.054(0.009)	0.930

Table 3.5 Simulation results for $\hat{\kappa}(0.2)$

Distri- bution	a	w	$\kappa(a)$	n	Mean(std)	Relative Bias(%)	SE(std)	Coverage of CI
(i)	0	0-1	0.524	20	0.503(0.219)	-4.2	0.2(0.038)	0.908
				50	0.516(0.134)	-1.6	0.13(0.015)	0.935
				100	0.52(0.094)	-0.8	0.093(0.007)	0.941
(ii)	0.2	0-1	0.445	20	0.428(0.208)	-4	0.19(0.035)	0.887
				50	0.438(0.128)	-1.6	0.125(0.011)	0.932
				100	0.442(0.091)	-0.7	0.0892(0.005)	0.938
(iii)	0.4	0-1	0.109	20	0.142(0.162)	23.2	0.14(0.054)	0.84
				50	0.148(0.101)	26.4	0.098(0.017)	0.925
				100	0.15(0.072)	27.3	0.07(0.008)	0.927
(iv)	0.216	0-1	0.394	20	0.383(0.148)	-2.9	0.141(0.014)	0.925
				50	0.39(0.092)	-1	0.091(0.005)	0.942
				100	0.392(0.065)	-0.5	0.065(0.002)	0.946
		C-A	0.24	20	0.235(0.17)	-2.1	0.16(0.023)	0.914
				50	0.239(0.108)	-0.4	0.105(0.009)	0.936
				100	0.24(0.076)	0	0.075(0.004)	0.943
		F-C	0.084	20	0.083(0.215)	-1.1	0.198(0.038)	0.906
				50	0.085(0.134)	0.8	0.131(0.015)	0.936
				100	0.086(0.095)	2	0.094(0.008)	0.942
(v)	0.161	0-1	0.199	20	0.185(0.143)	-7.6	0.135(0.018)	0.911
				50	0.192(0.09)	-3.6	0.089(0.007)	0.935
				100	0.195(0.064)	-2.1	0.063(0.003)	0.943
		C-A	0.371	20	0.348(0.151)	-6.6	0.139(0.021)	0.893
				50	0.36(0.095)	-3.1	0.092(0.008)	0.93
				100	0.365(0.067)	-1.6	0.066(0.004)	0.939
		F-C	0.517	20	0.488(0.177)	-5.9	0.155(0.041)	0.865
				50	0.505(0.11)	-2.4	0.105(0.017)	0.919
				100	0.51(0.077)	-1.4	0.076(0.009)	0.936
(vi)	0.102	0-1	0.316	20	0.301(0.133)	-5	0.127(0.012)	0.92
				50	0.309(0.084)	-2.3	0.082(0.004)	0.938
				100	0.312(0.059)	-1.3	0.059(0.002)	0.943
		C-A	0.558	20	0.54(0.127)	-3.3	0.12(0.021)	0.918
				50	0.551(0.079)	-1.3	0.077(0.008)	0.938
				100	0.554(0.055)	-0.7	0.055(0.004)	0.945
		F-C	0.711	20	0.693(0.13)	-2.6	0.116(0.043)	0.877
				50	0.705(0.08)	-0.9	0.076(0.019)	0.915
				100	0.707(0.056)	-0.6	0.055(0.009)	0.93

Table 3.6 Simulation results for $\hat{\kappa}(0.4)$

Distri- bution	a	w	$\kappa(a)$	n	Mean(std)	Relative Bias(%)	SE(std)	Coverage of CI
(i)	0	0-1	0.524	20	0.5(0.221)	-4.8	0.202(0.039)	0.913
				50	0.516(0.134)	-1.6	0.131(0.015)	0.935
				100	0.519(0.094)	-1	0.093(0.007)	0.942
(ii)	0.2	0-1	0.445	20	0.418(0.218)	-6.5	0.199(0.035)	0.886
				50	0.43(0.134)	-3.5	0.13(0.011)	0.932
				100	0.434(0.094)	-2.5	0.093(0.005)	0.942
(iii)	0.4	0-1	0.109	20	0.095(0.187)	-14.7	0.165(0.051)	0.833
				50	0.103(0.118)	-5.8	0.114(0.015)	0.917
				100	0.106(0.083)	-2.8	0.082(0.007)	0.936
(iv)	0.216	0-1	0.394	20	0.37(0.156)	-6.5	0.149(0.012)	0.924
				50	0.38(0.097)	-3.7	0.095(0.004)	0.941
				100	0.383(0.068)	-2.9	0.068(0.002)	0.945
		C-A	0.24	20	0.215(0.181)	-11.6	0.172(0.018)	0.911
				50	0.222(0.115)	-8.1	0.112(0.007)	0.93
				100	0.224(0.081)	-7.1	0.08(0.003)	0.938
		F-C	0.084	20	0.052(0.231)	-61.5	0.213(0.033)	0.895
				50	0.058(0.145)	-44.8	0.141(0.012)	0.926
				100	0.06(0.102)	-40	0.101(0.006)	0.933
(v)	0.161	0-1	0.199	20	0.17(0.15)	-17.1	0.141(0.017)	0.905
				50	0.182(0.094)	-9.3	0.092(0.006)	0.931
				100	0.186(0.066)	-7	0.066(0.003)	0.937
		C-A	0.371	20	0.335(0.158)	-10.7	0.146(0.022)	0.895
				50	0.351(0.098)	-5.7	0.096(0.008)	0.93
				100	0.355(0.069)	-4.5	0.068(0.004)	0.938
		F-C	0.517	20	0.478(0.186)	-8.2	0.162(0.042)	0.868
				50	0.496(0.115)	-4.2	0.109(0.018)	0.923
				100	0.501(0.08)	-3.2	0.079(0.009)	0.939
(vi)	0.102	0-1	0.316	20	0.295(0.136)	-7.1	0.13(0.011)	0.92
				50	0.304(0.085)	-3.9	0.084(0.004)	0.937
				100	0.308(0.06)	-2.6	0.06(0.002)	0.943
		C-A	0.558	20	0.537(0.129)	-3.9	0.122(0.022)	0.919
				50	0.548(0.079)	-1.8	0.078(0.009)	0.939
				100	0.552(0.056)	-1.1	0.055(0.004)	0.945
		F-C	0.711	20	0.691(0.132)	-2.9	0.118(0.044)	0.878
				50	0.703(0.081)	-1.1	0.077(0.019)	0.917
				100	0.707(0.057)	-0.6	0.055(0.01)	0.933

Table 3.7 Simulation results for $\hat{\kappa}(0.6)$

Distri- bution	a	w	$\kappa(a)$	n	Mean(std)	Relative Bias(%)	SE(std)	Coverage of CI
(i)	0	0-1	0.524	20	0.499(0.221)	-5	0.203(0.039)	0.913
				50	0.515(0.134)	-1.7	0.131(0.0145)	0.936
				100	0.52(0.094)	-0.8	0.093(0.007)	0.942
(ii)	0.2	0-1	0.445	20	0.407(0.225)	-9.3	0.206(0.035)	0.885
				50	0.424(0.138)	-5	0.135(0.011)	0.932
				100	0.429(0.097)	-3.7	0.096(0.005)	0.942
(iii)	0.4	0-1	0.109	20	0.056(0.209)	-94.6	0.186(0.048)	0.831
				50	0.068(0.132)	-60.3	0.127(0.013)	0.902
				100	0.072(0.093)	-51.4	0.092(0.005)	0.911
(iv)	0.216	0-1	0.394	20	0.361(0.161)	-9.1	0.155(0.013)	0.926
				50	0.372(0.1)	-5.9	0.099(0.004)	0.94
				100	0.376(0.071)	-4.8	0.07(0.002)	0.940
		C-A	0.24	20	0.2(0.191)	-20	0.181(0.015)	0.908
				50	0.208(0.12)	-15.4	0.118(0.005)	0.928
				100	0.211(0.085)	-13.7	0.084(0.002)	0.929
		F-C	0.084	20	0.029(0.244)	-190	0.226(0.03)	0.89
				50	0.037(0.153)	-127	0.149(0.011)	0.918
				100	0.039(0.108)	-115.4	0.106(0.005)	0.919
(v)	0.161	0-1	0.199	20	0.162(0.153)	-22.8	0.146(0.017)	0.906
				50	0.174(0.096)	-14.4	0.095(0.006)	0.928
				100	0.179(0.068)	-11.2	0.067(0.003)	0.930
		C-A	0.371	20	0.326(0.164)	-13.8	0.152(0.023)	0.897
				50	0.343(0.102)	-8.2	0.099(0.009)	0.928
				100	0.349(0.071)	-6.3	0.071(0.004)	0.937
		F-C	0.517	20	0.469(0.194)	-10.2	0.168(0.045)	0.869
				50	0.489(0.119)	-5.7	0.113(0.019)	0.924
				100	0.496(0.083)	-4.2	0.081(0.009)	0.939
(vi)	0.102	0-1	0.316	20	0.289(0.138)	-9.3	0.133(0.01)	0.917
				50	0.301(0.086)	-5	0.085(0.003)	0.935
				100	0.305(0.061)	-3.6	0.06(0.002)	0.941
		C-A	0.558	20	0.534(0.131)	-4.5	0.124(0.022)	0.921
				50	0.546(0.08)	-2.2	0.079(0.009)	0.941
				100	0.551(0.056)	-1.3	0.056(0.004)	0.945
		F-C	0.711	20	0.69(0.133)	-3	0.12(0.045)	0.88
				50	0.702(0.082)	-1.3	0.078(0.019)	0.917
				100	0.705(0.057)	-0.9	0.056(0.01)	0.935

Table 3.8 Simulation results for $\hat{\kappa}(0.8)$

Distri- bution	a	w	$\kappa(a)$	n	Mean(std)	Relative Bias(%)	SE(std)	Coverage of CI
(i)	0	0-1	0.524	20	0.498(0.222)	-5.2	0.204(0.04)	0.914
				50	0.514(0.135)	-1.9	0.131(0.015)	0.935
				100	0.519(0.094)	-1	0.093(0.007)	0.942
(ii)	0.2	0-1	0.445	20	0.403(0.23)	-10.4	0.211(0.036)	0.885
				50	0.419(0.141)	-6.2	0.137(0.012)	0.929
				100	0.425(0.099)	-4.7	0.097(0.006)	0.941
(iii)	0.4	0-1	0.109	20	0.029(0.222)	-275.9	0.2(0.046)	0.832
				50	0.045(0.141)	-142.2	0.136(0.011)	0.886
				100	0.05(0.1)	-118	0.098(0.004)	0.891
(iv)	0.216	0-1	0.394	20	0.355(0.166)	-11	0.159(0.014)	0.927
				50	0.368(0.102)	-7.1	0.101(0.005)	0.938
				100	0.371(0.072)	-6.2	0.071(0.002)	0.937
		C-A	0.24	20	0.191(0.197)	-25.7	0.187(0.014)	0.909
				50	0.2(0.124)	-20	0.121(0.005)	0.926
				100	0.204(0.087)	-17.6	0.086(0.002)	0.924
		F-C	0.084	20	0.013(0.253)	-546.2	0.235(0.029)	0.890
				50	0.026(0.158)	-223.1	0.154(0.01)	0.915
				100	0.028(0.111)	-200	0.11(0.005)	0.912
(v)	0.161	0-1	0.199	20	0.155(0.156)	-28.4	0.149(0.017)	0.906
				50	0.17(0.098)	-17.1	0.096(0.006)	0.925
				100	0.175(0.069)	-13.7	0.068(0.003)	0.928
		C-A	0.371	20	0.319(0.169)	-16.3	0.156(0.023)	0.897
				50	0.338(0.104)	-9.8	0.101(0.009)	0.929
				100	0.344(0.073)	-7.8	0.072(0.004)	0.932
		F-C	0.517	20	0.463(0.197)	-11.7	0.173(0.046)	0.872
				50	0.485(0.121)	-6.6	0.115(0.02)	0.925
				100	0.491(0.085)	-5.3	0.083(0.01)	0.939
(vi)	0.102	0-1	0.316	20	0.286(0.139)	-10.5	0.134(0.01)	0.917
				50	0.299(0.087)	-5.7	0.086(0.003)	0.934
				100	0.303(0.061)	-4.3	0.061(0.002)	0.939
		C-A	0.558	20	0.532(0.132)	-4.9	0.125(0.023)	0.92
				50	0.546(0.081)	-2.2	0.079(0.009)	0.941
				100	0.55(0.057)	-1.5	0.056(0.004)	0.945
		F-C	0.711	20	0.688(0.134)	-3.3	0.121(0.046)	0.879
				50	0.701(0.082)	-1.4	0.078(0.019)	0.918
				100	0.705(0.057)	-0.9	0.056(0.01)	0.934

Table 3.9 Simulation results for $\hat{\kappa}(1)$

Distri- bution	a	w	$\kappa(a)$	n	Mean(std)	Relative Bias(%)	SE(std)	Coverage of CI
(i)	0	0-1	0.524	20	0.497(0.223)	-5.4	0.205(0.04)	0.914
				50	0.515(0.135)	-1.7	0.131(0.015)	0.934
				100	0.519(0.094)	-1	0.093(0.007)	0.941
(ii)	0.2	0-1	0.445	20	0.4(0.232)	-11.3	0.212(0.036)	0.885
				50	0.418(0.142)	-6.5	0.138(0.012)	0.931
				100	0.424(0.099)	-5	0.098(0.006)	0.940
(iii)	0.4	0-1	0.109	20	0.02(0.227)	-445	0.206(0.045)	0.835
				50	0.037(0.144)	-194.6	0.14(0.011)	0.883
				100	0.042(0.102)	-159.5	0.1(0.004)	0.873
(iv)	0.216	0-1	0.394	20	0.353(0.166)	-11.6	0.16(0.014)	0.928
				50	0.366(0.103)	-7.7	0.101(0.005)	0.940
				100	0.37(0.072)	-6.5	0.072(0.002)	0.938
		C-A	0.24	20	0.187(0.199)	-28.3	0.19(0.014)	0.910
				50	0.198(0.124)	-21.2	0.122(0.005)	0.926
				100	0.202(0.088)	-18.8	0.087(0.002)	0.922
		F-C	0.084	20	0.009(0.256)	-833.3	0.238(0.028)	0.892
				50	0.019(0.161)	-342.1	0.156(0.01)	0.913
				100	0.024(0.113)	-250	0.112(0.005)	0.906
(v)	0.161	0-1	0.199	20	0.151(0.158)	-31.8	0.15(0.017)	0.905
				50	0.168(0.099)	-18.5	0.097(0.006)	0.923
				100	0.173(0.069)	-15	0.069(0.003)	0.926
		C-A	0.371	20	0.317(0.169)	-17	0.157(0.024)	0.899
				50	0.336(0.105)	-10.4	0.102(0.009)	0.927
				100	0.342(0.073)	-8.5	0.072(0.004)	0.932
		F-C	0.517	20	0.462(0.199)	-11.9	0.174(0.047)	0.874
				50	0.483(0.122)	-7	0.116(0.02)	0.926
				100	0.49(0.085)	-5.5	0.083(0.01)	0.939
(vi)	0.102	0-1	0.316	20	0.285(0.14)	-10.9	0.135(0.01)	0.918
				50	0.299(0.087)	-5.7	0.086(0.003)	0.935
				100	0.302(0.061)	-4.6	0.061(0.002)	0.938
		C-A	0.558	20	0.532(0.132)	-4.9	0.125(0.023)	0.923
				50	0.545(0.081)	-2.4	0.079(0.009)	0.941
				100	0.549(0.057)	-1.6	0.056(0.004)	0.946
		F-C	0.711	20	0.688(0.135)	-3.3	0.121(0.046)	0.88
				50	0.701(0.082)	-1.4	0.079(0.019)	0.919
				100	0.705(0.057)	-0.9	0.056(0.01)	0.935

Table 3.10 Simulation results for $\hat{\kappa}(\hat{a})$

Distribution	a	w	$\kappa(a)$	n	Mean(std)	Relative Bias(%)	SE(std)	Coverage of CI
(i)	0	0-1	0.524	20	0.502(0.22)	-4.4	0.2(0.038)	0.908
				50	0.516(0.134)	-1.6	0.13(0.015)	0.934
				100	0.52(0.094)	-0.8	0.093(0.007)	0.94
(ii)	0.2	0-1	0.445	20	0.424(0.216)	-5	0.197(0.033)	0.893
				50	0.437(0.132)	-1.8	0.128(0.01)	0.931
				100	0.441(0.092)	-0.9	0.091(0.005)	0.942
(iii)	0.4	0-1	0.109	20	0.083(0.208)	-31.3	0.188(0.039)	0.84
				50	0.099(0.131)	-10.1	0.126(0.009)	0.928
				100	0.104(0.091)	-4.8	0.09(0.003)	0.938
(iv)	0.216	0-1	0.394	20	0.376(0.154)	-4.8	0.148(0.012)	0.931
				50	0.387(0.095)	-1.8	0.094(0.004)	0.943
				100	0.39(0.067)	-1	0.066(0.002)	0.946
		C-A	0.24	20	0.227(0.18)	-5.7	0.171(0.015)	0.921
				50	0.234(0.112)	-2.6	0.11(0.005)	0.939
				100	0.237(0.079)	-1.3	0.078(0.003)	0.946
		F-C	0.084	20	0.068(0.225)	-23.5	0.21(0.028)	0.909
				50	0.078(0.14)	-7.7	0.137(0.011)	0.935
				100	0.081(0.099)	-3.7	0.097(0.005)	0.941
(v)	0.161	0-1	0.199	20	0.183(0.146)	-8.7	0.138(0.016)	0.914
				50	0.193(0.091)	-3.1	0.09(0.006)	0.937
				100	0.196(0.064)	-1.5	0.064(0.003)	0.944
		C-A	0.371	20	0.346(0.153)	-7.2	0.142(0.021)	0.898
				50	0.362(0.094)	-2.5	0.093(0.008)	0.934
				100	0.366(0.067)	-1.4	0.066(0.004)	0.942
		F-C	0.517	20	0.488(0.179)	-5.9	0.157(0.0412)	0.868
				50	0.506(0.11)	-2.2	0.105(0.017)	0.922
				100	0.511(0.078)	-1.2	0.076(0.009)	0.935
(vi)	0.102	0-1	0.316	20	0.302(0.133)	-4.6	0.127(0.011)	0.922
				50	0.311(0.084)	-1.6	0.082(0.004)	0.938
				100	0.314(0.059)	-0.6	0.058(0.002)	0.945
		C-A	0.558	20	0.54(0.126)	-3.3	0.12(0.021)	0.919
				50	0.552(0.078)	-1.1	0.077(0.008)	0.939
				100	0.554(0.055)	-0.7	0.055(0.004)	0.944
		F-C	0.711	20	0.694(0.13)	-2.4	0.116(0.044)	0.876
				50	0.705(0.08)	-0.9	0.076(0.019)	0.913
				100	0.708(0.056)	-0.4	0.055(0.009)	0.931

Case 3. Mean $(-\frac{\sqrt{0.25}}{2}, \frac{\sqrt{0.25}}{2})$ and covariance matrix

$$\begin{pmatrix} (\frac{4}{3})^2 & 0.5 \times \frac{4}{3} \times \frac{2}{3} \\ 0.5 \times \frac{4}{3} \times \frac{2}{3} & (\frac{2}{3})^2 \end{pmatrix}$$

with large difference in both location and scale parameters, and weaker positive correlation, with correlation coefficient 0.5.

The small sample properties of $\hat{\rho}(a)$ and $\hat{\rho}(\hat{a})$ for data from underlying normal distributions with contaminations were also examined. Additional simulations were performed with 90% of the data distributed as in Cases 1-3, and 10% of the data following the same normal distribution but with three times the standard deviation of the data. Data were also simulated from a log-normal distribution to evaluate the small sample properties of the proposed measure when applied to non-normal data. Data were generated as described in Cases 1-3.

Case 1. Fixed and Known a:

For each distribution, we simulated 100000 data sets of size n=20, 50 and 100. For each distribution, we also calculated $\rho(a)$ and treated it as the true parameter value. For each simulated data set of a distribution, we used a=0,0.2,0.4,0.6,0.8 and 1 respectively, as fixed and known values of a, to compute agreement coefficients, SEs and confidence intervals. As we stated in the simulation study for categorical responses, the series of values of a reflected the uncertainty of selecting a a priori in practice and enabled us to compare the performance of the estimators with different values of a. In particular, since a=0 yields Lin's CCC and a=1 yields the RMAC, we were able to assess the performance of these two important measures as well.

The "a" column contains the true values of a for the distributions. Due to the complexity of the formula involved in the evaluation of the integral (2.11), we approximate a by $\hat{a}$ calculated from a data set with sample size 10000 generated from underlying distribution. The "$\rho(a)$" column contains the true values of $\rho(a)$. The "Mean(std)" column contains the mean values of $\hat{\rho}(a)$, which were the averages of valid $\hat{\rho}(a)$ computed from the simulated data sets of the distribution, and the standard deviations of valid $\hat{\rho}(a)$ computed from the simulated data sets of the distribution. The "Relative Bias(%)" column contains the relative bias of the agreement estimate, which is calculated by [(mean $\hat{\rho}(a)$-true $\rho(a)$)/mean $\hat{\rho}(a)$]× 100. The "SE(std)" column contains the mean values of $\hat{SE}(\hat{\rho}(a))$, which were the averages of valid $\hat{SE}(\hat{\rho}(a))$ computed from the simulated data sets of the distribution, and the standard deviations of valid $\hat{SE}(\hat{\rho}(a))$ computed from the simulated data sets of the distribution. Each confidence interval computed from the generated data set was evaluated to determine whether the true parameter value was contained within the confidence interval, and the coverage is contained in the "Coverage of CI" column.

Case 2. Estimated a:

The simulation was conducted differently from previous case only in that, instead of using a series of fixed and known a, we estimated a from each generated data set analogously to the Cramer-von Mises criterion, using (2.12). As in previous case, the true value of a was calculated from the underlying distribution analogously to the Cramer-von Mises criterion, so both cases share the same true value of a and true value of $\rho(a)$ for each distribution. Same quantities were calculated as in previous case.

Contaminated Normal Distribution:

To calculate the true value of $\rho(a)$ for the underlying contaminated normal distribution, we need to calculate the mean vector and the variance-covariance matrix of the random vector with the distribution. Let

$$\boldsymbol{U} = \begin{pmatrix} X \\ Y \end{pmatrix} \sim N(\boldsymbol{\theta}, \Sigma)$$

where $\boldsymbol{\theta} = (\mu_X, \mu_Y)^T$ and

$$\Sigma = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}$$

be the random vector with one of the distributions in Cases 1-3. Since the contaminated normal random vector follows bivariate normal distribution 90% of the time and follows a bivariate normal distribution with 9 times the variance 10% of the time, in addition let

$$\boldsymbol{V} = \begin{pmatrix} X^* \\ Y^* \end{pmatrix} \sim N(\boldsymbol{\theta}, \Sigma^*)$$

$$\Sigma^* = \begin{pmatrix} 9\sigma_X^2 & 9\sigma_{XY} \\ 9\sigma_{XY} & 9\sigma_Y^2 \end{pmatrix}$$

and let $I_{0.9} = 1$ with probability 0.9 and $I_{0.9} = 0$ with probability 0.1, and assume that $\boldsymbol{U}$ and $I_{0.9}$ are independent, $\boldsymbol{V}$ and $I_{0.9}$ are independent. Then

$$\boldsymbol{W} = \boldsymbol{U} I_{0.9} + \boldsymbol{V}\ (1 - I_{0.9})$$

follows the desired distribution.

By independence,

$$\begin{aligned} E(\boldsymbol{W}) &= E(\boldsymbol{U})E(I_{0.9}) + E(\boldsymbol{V})E(1 - I_{0.9}) \\ &= 0.9\boldsymbol{\theta} + 0.1\boldsymbol{\theta} \\ &= \boldsymbol{\theta} \end{aligned}$$

$$\begin{aligned} Var(\boldsymbol{W}) &= Var[\boldsymbol{U}I_{0.9} + \boldsymbol{V}(1 - I_{0.9})] \\ &= Var(\boldsymbol{U}I_{0.9}) + Var[\boldsymbol{V}(1 - I_{0.9})] + 2\text{Cov}[\boldsymbol{U}I_{0.9}, \boldsymbol{V}(1 - I_{0.9})] \\ &= E(I_{0.9}^2\boldsymbol{U}\boldsymbol{U}^T) - [E(\boldsymbol{U}I_{0.9})][E(\boldsymbol{U}I_{0.9})]^T + E[(1 - I_{0.9})^2\boldsymbol{V}\boldsymbol{V}^T] \\ &- \{E[\boldsymbol{V}(1 - I_{0.9})]\}\{E[\boldsymbol{V}(1 - I_{0.9})]\}^T + 2\{E[I_{0.9}(1 - I_{0.9})\boldsymbol{U}\boldsymbol{V}^T] \\ &- E(\boldsymbol{U}I_{0.9})\{E[\boldsymbol{V}(1 - I_{0.9})]\}^T\} \end{aligned}$$

where

$$\begin{aligned} E(I_{0.9}^2\boldsymbol{U}\boldsymbol{U}^T) &= E(I_{0.9}^2)E(\boldsymbol{U}\boldsymbol{U}^T) \\ &= 0.9\{\Sigma + E(\boldsymbol{U})[E(\boldsymbol{U})]^T\} \\ &= 0.9\begin{pmatrix} \sigma_X^2 + \mu_X^2 & \sigma_{XY} + \mu_X\mu_Y \\ \sigma_{XY} + \mu_X\mu_Y & \sigma_Y^2 + \mu_Y^2 \end{pmatrix} \end{aligned}$$

$$\begin{aligned}[E(\boldsymbol{U}I_{0.9})][E(\boldsymbol{U}I_{0.9})]^T &= [E(I_{0.9})]^2 E(\boldsymbol{U})[E(\boldsymbol{U})]^T \\ &= 0.81 \begin{pmatrix} \mu_X^2 & \mu_X\mu_Y \\ \mu_X\mu_Y & \mu_Y^2 \end{pmatrix}\end{aligned}$$

$$\begin{aligned}E[(1-I_{0.9})^2 \boldsymbol{V}\boldsymbol{V}^T] &= E[(1-I_{0.9})^2]E(\boldsymbol{V}\boldsymbol{V}^T) \\ &= 0.1\{\Sigma^* + E(\boldsymbol{V})[E(\boldsymbol{V})]^T\} \\ &= 0.1 \begin{pmatrix} 9\sigma_X^2 + \mu_X^2 & 9\sigma_{XY} + \mu_X\mu_Y \\ 9\sigma_{XY} + \mu_X\mu_Y & 9\sigma_Y^2 + \mu_Y^2 \end{pmatrix}\end{aligned}$$

$$\begin{aligned}\{E[\boldsymbol{V}(1-I_{0.9})]\}\{E[\boldsymbol{V}(1-I_{0.9})]\}^T &= [E(1-I_{0.9})]^2 E(\boldsymbol{V})[E(\boldsymbol{V})]^T \\ &= 0.01 \begin{pmatrix} \mu_X^2 & \mu_X\mu_Y \\ \mu_X\mu_Y & \mu_Y^2 \end{pmatrix}\end{aligned}$$

$$\begin{aligned}E[I_{0.9}(1-I_{0.9})\boldsymbol{U}\boldsymbol{V}^T] &= E[I_{0.9}(1-I_{0.9})]E(\boldsymbol{U}\boldsymbol{V}^T) \\ &= \boldsymbol{0}\end{aligned}$$

and

$$E(\boldsymbol{U}I_{0.9})\{E[\boldsymbol{V}(1-I_{0.9})]\}^T = 0.09 \begin{pmatrix} \mu_X^2 & \mu_X\mu_Y \\ \mu_X\mu_Y & \mu_Y^2 \end{pmatrix}$$

Therefore,

$$Var(\boldsymbol{W}) = 1.8 \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}$$

Log-normal Distribution:

To calculate the true value of $\rho(a)$ for the underlying log-normal distribution, we need to calculate the mean vector and the variance-covariance matrix of the random vector with the distribution. By definition, if

$$\boldsymbol{U} = \begin{pmatrix} X \\ Y \end{pmatrix} \sim N(\boldsymbol{\theta}, \Sigma)$$

where $\boldsymbol{\theta} = (\mu_X, \mu_Y)^T$ and

$$\Sigma = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}$$

is the random vector with one of the distributions in Cases 1-3, then

$$\boldsymbol{U}^* = \begin{pmatrix} e^X \\ e^Y \end{pmatrix}$$

follows log-normal distribution, where mean vector is $\boldsymbol{\theta}^* = (exp(\mu_X + \sigma_X^2/2), exp(\mu_Y + \sigma_Y^2/2))^T$ and variance-covariance matrix is

$$\Sigma^* = \begin{pmatrix} exp(2\mu_X + \sigma_X^2)[exp(\sigma_X^2) - 1] & exp[\mu_X + \mu_Y + 0.5(\sigma_X^2 + \sigma_Y^2)][exp(\sigma_{XY}) - 1] \\ exp[\mu_X + \mu_Y + 0.5(\sigma_X^2 + \sigma_Y^2)][exp(\sigma_{XY}) - 1] & exp(2\mu_Y + \sigma_Y^2)[exp(\sigma_Y^2) - 1] \end{pmatrix}$$

Results:

The simulation results are listed in Table 3.11 - Table 3.17. For each of the distributions with noticeable difference between marginal distributions and each sample size, the selected value of $a = 0$ yields the largest agreement and the selected value of $a = 1$ yields the smallest agreement, that is, the Lin's CCC and the RMAC form the upper and lower bounds of the class of agreement coefficients. The larger the difference between two marginal distributions, the larger the difference between the bounds, suggesting the impact of different marginal distributions on agreement.

For agreement estimation, the estimates become closer to the true parameter value as sample size increases for most cases. If the selected a is greater than the true a, then $\hat{\rho}(a)$ tends to underestimate $\rho(a)$ most of the time. If the selected a is less than the true a, then $\hat{\rho}(a)$ tends to overestimate $\rho(a)$ only for Case 3 of normal distributions. On the other hand, $\hat{\rho}(\hat{a})$ always resembles the performance of the estimator with selected a close to true a. The special case is Case 3 of log-normal distribution, where the agreement estimates all overestimate the true parameter value, regardless of the selected value of a.

For standard error estimation, the estimates become smaller as sample size increases. If marginal distributions are different, then estimated SE becomes larger as the selected a gets larger, indicating that the RMAC is less efficient in the presence of different marginal distributions. On the other hand, $\hat{\rho}(\hat{a})$ accounts for the estimation of a and yields a correct level of precision. The difference between the average of the estimated SEs and the standard deviation of $\hat{\rho}(\hat{a})$ and $\hat{\rho}(a)$, which can be regarded as the true value of the standard deviation of $\hat{\rho}(\hat{a})$ and $\hat{\rho}(a)$, is less than one-half standard deviation of the estimated SEs for all normal cases and contaminated normal cases with $n = 50$ and 100. But the difference is larger for log-normal cases. This implies that the SE formulas proposed are fairly accurate for normal distributions with small to large sample sizes and contaminated normal distributions with moderate to large sample sizes. We also notice that the SE is always less than the true standard deviation. This may imply that the SE formulas proposed slightly underestimate the true SE.

For coverage of confidence interval, the coverage improves as sample size increases. It shows that for both selected a and estimated a, quite a few coverage probabilities are between 0.93 and 0.94, close to the true probability 0.95 for normal cases, but the coverage probabilities are only between 0.85 and 0.9 for contaminated normal cases and between 0.7 and 0.8 for log-normal cases. This implies that the proposed confidence interval formula works better for normal underlying distributions.

Table 3.11 Simulation results for $\hat{\rho}(0)$

Distribution	Case	a	$\rho(a)$	n	Mean(std)	Relative Bias(%)	SE(std)	Coverage of CI
Normal	1	0	0.95	20	0.942(0.027)	-0.8	0.024(0.012)	0.926
				50	0.947(0.015)	-0.3	0.014(0.004)	0.938
				100	0.949(0.01)	-0.1	0.01(0.002)	0.944
	2	0.102	0.887	20	0.874(0.047)	-1.5	0.042(0.016)	0.931
				50	0.882(0.027)	-0.5	0.026(0.006)	0.941
				100	0.885(0.018)	-0.2	0.018(0.003)	0.944
	3	0.179	0.349	20	0.341(0.138)	-2.4	0.127(0.033)	0.913
				50	0.353(0.087)	1.1	0.083(0.014)	0.931
				100	0.356(0.061)	2	0.056(0.007)	0.937
Contaminated Normal	1	0	0.95	20	0.937(0.043)	-1.4	0.03(0.025)	0.856
				50	0.944(0.025)	-0.6	0.02(0.012)	0.882
				100	0.947(0.017)	-0.3	0.015(0.006)	0.904
	2	0.093	0.906	20	0.885(0.058)	-2.4	0.042(0.025)	0.884
				50	0.897(0.034)	-1	0.028(0.012)	0.9
				100	0.902(0.024)	-0.5	0.021(0.006)	0.914
	3	0.169	0.371	20	0.345(0.193)	-7.3	0.143(0.057)	0.841
				50	0.363(0.134)	-2.1	0.111(0.037)	0.879
				100	0.369(0.098)	-0.4	0.087(0.024)	0.904
Log-normal	1	0.002	0.923	20	0.901(0.065)	-2.4	0.037(0.025)	0.718
				50	0.912(0.046)	-1.2	0.028(0.016)	0.755
				100	0.916(0.037)	-0.7	0.024(0.012)	0.796
	2	0.1	0.889	20	0.878(0.074)	-1.3	0.042(0.026)	0.701
				50	0.891(0.054)	0.2	0.032(0.017)	0.717
				100	0.894(0.046)	0.5	0.027(0.014)	0.733
	3	0.179	0.177	20	0.282(0.193)	37.2	0.117(0.061)	0.69
				50	0.265(0.142)	33.4	0.088(0.043)	0.692
				100	0.248(0.112)	28.7	0.072(0.032)	0.708

Table 3.12 Simulation results for $\hat{\rho}(0.2)$

Distri- bution	Case	a	$\rho(a)$	n	Mean(std)	Relative Bias(%)	SE(std)	Coverage of CI
Normal	1	0	0.95	20	0.942(0.027)	-0.8	0.024(0.012)	0.925
				50	0.947(0.015)	-0.3	0.014(0.004)	0.939
				100	0.949(0.01)	-0.2	0.01(0.002)	0.943
	2	0.102	0.887	20	0.872(0.049)	-1.7	0.043(0.017)	0.932
				50	0.881(0.028)	-0.7	0.026(0.007)	0.942
				100	0.884(0.019)	-0.3	0.018(0.003)	0.948
	3	0.179	0.349	20	0.324(0.146)	-7.5	0.135(0.033)	0.913
				50	0.338(0.092)	-3.1	0.088(0.014)	0.931
				100	0.343(0.064)	-1.6	0.063(0.007)	0.942
Conta- minated Normal	1	0	0.95	20	0.936(0.044)	-1.5	0.03(0.026)	0.857
				50	0.944(0.025)	-0.6	0.02(0.012)	0.884
				100	0.947(0.017)	-0.3	0.015(0.006)	0.907
	2	0.1	0.906	20	0.884(0.059)	-2.5	0.043(0.026)	0.885
				50	0.897(0.035)	-1.1	0.029(0.012)	0.9
				100	0.901(0.024)	-0.5	0.021(0.007)	0.914
	3	0.169	0.371	20	0.334(0.2)	-11.1	0.148(0.057)	0.841
				50	0.354(0.137)	-4.9	0.114(0.038)	0.879
				100	0.361(0.101)	-2.7	0.09(0.024)	0.905
Log- normal	1	0.002	0.923	20	0.901(0.065)	-2.4	0.037(0.026)	0.719
				50	0.912(0.046)	-1.2	0.028(0.016)	0.755
				100	0.916(0.037)	-0.7	0.024(0.012)	0.794
	2	0.1	0.889	20	0.878(0.075)	-1.3	0.043(0.027)	0.705
				50	0.891(0.055)	0.1	0.032(0.017)	0.718
				100	0.894(0.046)	0.5	0.028(0.014)	0.734
	3	0.179	0.177	20	0.275(0.195)	35.6	0.12(0.062)	0.7
				50	0.263(0.143)	32.7	0.089(0.043)	0.701
				100	0.246(0.113)	28.2	0.073(0.032)	0.714

Table 3.13 Simulation results for $\hat{\rho}(0.4)$

Distribution	Case	a	$\rho(a)$	n	Mean(std)	Relative Bias(%)	SE(std)	Coverage of CI
Normal	1	0	0.95	20	0.942(0.027)	-0.8	0.024(0.012)	0.923
				50	0.947(0.015)	-0.3	0.014(0.004)	0.938
				100	0.949(0.01)	-0.2	0.01(0.002)	0.944
	2	0.102	0.887	20	0.871(0.05)	-1.8	0.044(0.018)	0.935
				50	0.88(0.028)	-0.7	0.027(0.007)	0.944
				100	0.883(0.019)	-0.4	0.019(0.003)	0.949
	3	0.179	0.349	20	0.311(0.155)	-12.2	0.142(0.035)	0.910
				50	0.327(0.096)	-6.6	0.092(0.015)	0.933
				100	0.333(0.067)	-4.8	0.066(0.008)	0.940
Contaminated Normal	1	0	0.95	20	0.936(0.044)	-1.5	0.03(0.026)	0.857
				50	0.944(0.025)	-0.6	0.02(0.011)	0.884
				100	0.947(0.017)	-0.3	0.015(0.006)	0.906
	2	0.1	0.906	20	0.882(0.06)	-2.7	0.044(0.026)	0.887
				50	0.896(0.036)	-1.1	0.029(0.012)	0.901
				100	0.9(0.025)	-0.6	0.022(0.007)	0.916
	3	0.169	0.371	20	0.324(0.206)	-14.6	0.154(0.058)	0.842
				50	0.347(0.141)	-6.8	0.117(0.038)	0.88
				100	0.355(0.102)	-4.5	0.092(0.024)	0.906
Log-normal	1	0.002	0.923	20	0.901(0.065)	-2.5	0.037(0.026)	0.718
				50	0.912(0.046)	-1.2	0.029(0.016)	0.757
				100	0.916(0.037)	-0.8	0.024(0.012)	0.796
	2	0.1	0.889	20	0.877(0.076)	-1.4	0.043(0.028)	0.704
				50	0.891(0.055)	0.1	0.032(0.017)	0.717
				100	0.894(0.046)	0.5	0.028(0.014)	0.734
	3	0.179	0.177	20	0.271(0.199)	34.7	0.123(0.062)	0.71
				50	0.26(0.143)	32	0.091(0.042)	0.711
				100	0.245(0.114)	27.9	0.073(0.032)	0.716

Table 3.14 Simulation results for $\hat{\rho}(0.6)$

Distribution	Case	a	$\rho(a)$	n	Mean(std)	Relative Bias(%)	SE(std)	Coverage of CI
Normal	1	0	0.95	20	0.942(0.027)	-0.9	0.024(0.012)	0.926
				50	0.947(0.015)	-0.3	0.014(0.004)	0.938
				100	0.949(0.01)	-0.2	0.01(0.002)	0.945
	2	0.102	0.887	20	0.87(0.051)	-1.9	0.045(0.019)	0.937
				50	0.879(0.029)	-0.9	0.027(0.007)	0.945
				100	0.882(0.019)	-0.5	0.019(0.003)	0.949
	3	0.179	0.349	20	0.3(0.161)	-16.3	0.148(0.037)	0.915
				50	0.319(0.1)	-9.3	0.096(0.016)	0.932
				100	0.325(0.07)	-7.3	0.068(0.008)	0.936
Contaminated Normal	1	0	0.95	20	0.936(0.044)	-1.5	0.03(0.026)	0.856
				50	0.944(0.025)	-0.6	0.02(0.012)	0.882
				100	0.947(0.017)	-0.3	0.015(0.006)	0.905
	2	0.1	0.906	20	0.882(0.061)	-2.7	0.044(0.027)	0.887
				50	0.896(0.036)	-1.2	0.029(0.012)	0.902
				100	0.9(0.025)	-0.6	0.022(0.007)	0.916
	3	0.169	0.371	20	0.314(0.213)	-17.9	0.159(0.06)	0.843
				50	0.341(0.143)	-8.8	0.119(0.039)	0.88
				100	0.35(0.105)	-5.9	0.093(0.025)	0.903
Lognormal	1	0.002	0.923	20	0.901(0.065)	-2.4	0.037(0.026)	0.72
				50	0.912(0.046)	-1.2	0.029(0.016)	0.757
				100	0.916(0.037)	-0.7	0.024(0.012)	0.796
	2	0.1	0.889	20	0.877(0.076)	-1.4	0.044(0.028)	0.708
				50	0.89(0.055)	0.1	0.032(0.017)	0.719
				100	0.894(0.047)	0.5	0.028(0.014)	0.736
	3	0.179	0.177	20	0.267(0.2)	33.8	0.126(0.062)	0.716
				50	0.259(0.143)	31.7	0.091(0.042)	0.713
				100	0.244(0.114)	27.6	0.073(0.031)	0.719

Table 3.15 Simulation results for $\hat{\rho}(0.8)$

Distribution	Case	a	$\rho(a)$	n	Mean(std)	Relative Bias(%)	SE(std)	Coverage of CI
Normal	1	0	0.95	20	0.942(0.028)	-0.9	0.024(0.012)	0.925
				50	0.947(0.015)	-0.3	0.014(0.004)	0.938
				100	0.949(0.01)	-0.1	0.001(0.002)	0.942
	2	0.102	0.887	20	0.869(0.051)	-2	0.046(0.019)	0.938
				50	0.879(0.029)	-0.9	0.027(0.007)	0.946
				100	0.882(0.019)	-0.5	0.019(0.003)	0.949
	3	0.179	0.349	20	0.294(0.165)	-18.7	0.153(0.039)	0.919
				50	0.314(0.101)	-11.2	0.098(0.017)	0.934
				100	0.321(0.071)	-8.8	0.07(0.008)	0.935
Contaminated Normal	1	0	0.95	20	0.936(0.043)	-1.4	0.03(0.025)	0.857
				50	0.944(0.025)	-0.6	0.02(0.012)	0.885
				100	0.947(0.017)	-0.3	0.015(0.006)	0.905
	2	0.1	0.906	20	0.882(0.061)	-2.8	0.045(0.027)	0.888
				50	0.895(0.036)	-1.2	0.029(0.012)	0.904
				100	0.9(0.025)	-0.6	0.022(0.007)	0.917
	3	0.169	0.371	20	0.312(0.214)	-19	0.162(0.061)	0.848
				50	0.337(0.144)	-9.9	0.121(0.039)	0.881
				100	0.348(0.105)	-6.7	0.094(0.025)	0.905
Log-normal	1	0.002	0.923	20	0.901(0.066)	-2.5	0.037(0.026)	0.722
				50	0.912(0.046)	-1.2	0.029(0.016)	0.759
				100	0.916(0.037)	-0.7	0.024(0.012)	0.796
	2	0.1	0.889	20	0.877(0.076)	-1.5	0.044(0.028)	0.711
				50	0.89(0.055)	0.1	0.032(0.017)	0.717
				100	0.894(0.046)	0.5	0.028(0.014)	0.734
	3	0.179	0.177	20	0.264(0.201)	33.1	0.127(0.063)	0.723
				50	0.258(0.144)	31.6	0.092(0.042)	0.714
				100	0.244(0.114)	27.7	0.074(0.031)	0.72

Table 3.16 Simulation results for $\hat{\rho}(1)$

Distribution	Case	a	$\rho(a)$	n	Mean(std)	Relative Bias(%)	SE(std)	Coverage of CI
Normal	1	0	0.95	20	0.942(0.028)	-0.9	0.024(0.012)	0.926
				50	0.947(0.015)	-0.3	0.014(0.004)	0.937
				100	0.949(0.01)	-0.2	0.01(0.002)	0.943
	2	0.102	0.887	20	0.869(0.051)	-2	0.046(0.019)	0.940
				50	0.879(0.029)	-0.9	0.027(0.007)	0.946
				100	0.882(0.02)	-0.6	0.019(0.003)	0.951
	3	0.179	0.349	20	0.292(0.166)	-19.7	0.154(0.039)	0.919
				50	0.312(0.102)	-11.8	0.099(0.017)	0.931
				100	0.319(0.071)	-9.5	0.07(0.009)	0.934
Contaminated Normal	1	0	0.95	20	0.936(0.044)	-1.5	0.03(0.026)	0.858
				50	0.944(0.025)	-0.6	0.02(0.012)	0.882
				100	0.947(0.017)	-0.3	0.015(0.006)	0.904
	2	0.1	0.906	20	0.882(0.061)	-2.8	0.045(0.027)	0.889
				50	0.895(0.036)	-1.2	0.029(0.012)	0.905
				100	0.9(0.025)	-0.7	0.022(0.007)	0.916
	3	0.169	0.371	20	0.309(0.216)	-20	0.163(0.061)	0.849
				50	0.336(0.145)	-10.3	0.121(0.039)	0.88
				100	0.367(0.105)	-6.9	0.094(0.025)	0.906
Log-normal	1	0.002	0.923	20	0.901(0.066)	-2.4	0.037(0.026)	0.719
				50	0.912(0.046)	-1.2	0.029(0.016)	0.758
				100	0.916(0.037)	-0.7	0.024(0.012)	0.796
	2	0.1	0.889	20	0.876(0.077)	-1.5	0.044(0.028)	0.71
				50	0.89(0.055)	0.1	0.032(0.017)	0.72
				100	0.894(0.046)	0.5	0.028(0.014)	0.737
	3	0.179	0.177	20	0.265(0.2)	33.2	0.128(0.063)	0.724
				50	0.259(0.144)	31.7	0.092(0.042)	0.713
				100	0.244(0.114)	27.5	0.074(0.031)	0.721

Table 3.17 Simulation results for $\hat{\rho}(\hat{a})$

Distribution	Case	a	$\rho(a)$	n	Mean(std)	Relative Bias(%)	SE(std)	Coverage of CI
Normal	1	0	0.95	20	0.942(0.028)	-0.9	0.024(0.012)	0.924
				50	0.947(0.015)	-0.3	0.014(0.004)	0.937
				100	0.949(0.01)	-0.2	0.01(0.002)	0.944
	2	0.102	0.887	20	0.872(0.049)	-1.7	0.043(0.017)	0.932
				50	0.881(0.027)	-0.6	0.026(0.006)	0.941
				100	0.884(0.019)	-0.3	0.018(0.003)	0.946
	3	0.179	0.349	20	0.321(0.15)	-8.5	0.136(0.034)	0.911
				50	0.338(0.092)	-3	0.088(0.014)	0.932
				100	0.343(0.065)	-1.5	0.063(0.007)	0.939
Contaminated Normal	1	0	0.95	20	0.936(0.044)	-1.5	0.03(0.026)	0.856
				50	0.944(0.025)	-0.6	0.02(0.011)	0.885
				100	0.947(0.017)	-0.3	0.015(0.006)	0.905
	2	0.1	0.906	20	0.884(0.059)	-2.5	0.043(0.026)	0.885
				50	0.897(0.035)	-1.1	0.029(0.012)	0.901
				100	0.901(0.024)	-0.5	0.021(0.007)	0.912
	3	0.169	0.371	20	0.331(0.203)	-12	0.149(0.057)	0.837
				50	0.354(0.138)	-4.7	0.114(0.038)	0.877
				100	0.362(0.1)	-2.5	0.089(0.024)	0.904
Lognormal	1	0.002	0.923	20	0.901(0.065)	-2.4	0.037(0.025)	0.719
				50	0.912(0.046)	-1.2	0.029(0.016)	0.759
				100	0.916(0.037)	-0.7	0.024(0.012)	0.795
	2	0.1	0.889	20	0.878(0.075)	-1.3	0.043(0.027)	0.706
				50	0.891(0.055)	0.2	0.032(0.017)	0.717
				100	0.894(0.046)	0.5	0.027(0.014)	0.73
	3	0.179	0.177	20	0.275(0.197)	35.7	0.12(0.062)	0.7
				50	0.259(0.144)	31.7	0.092(0.042)	0.713
				100	0.247(0.113)	28.5	0.073(0.032)	0.711

Chapter 4

Applications to Real Data

Six examples with categorical or continuous responses are described here. In practice, the value of fixed and known a is usually dependent on the researcher's preference. Some researchers may use a specific value of a in every circumstance, others may use different ones in different circumstances. We use a=0, 0.2, 0.4, 0.6, 0.8 and 1 in each of these six examples to illustrate the results of different choices of a. The estimate of a, $\hat{a}$, is calculated from the sample proportions using the Cramer-von Mises criterion.

4.1 Coffee Data

This example is based on the data of subjects' purchase choice of instant decaffeinated coffee at two times. The complete data set contains 4657 households that made two purchases of one or more of the 11 brands during the 12-month period and was discussed by Grover and Srinivasan [24]. For illustration purpose, we take a subsample that contains 541 observations and five categories. Note that the responses were recorded on a nominal scale.

The data are displayed in Table 4.1. These frequencies demonstrate moderate agreement between the purchase choices at two times, which is also reflected in the estimates of agreement coefficient. For the case where a is fixed and known, the results are shown in Table 4.2. For the case where a is estimated, $\hat{a} = 0.038$, $\hat{\kappa}(\hat{a}) = 0.476$, $\hat{SE} = 0.028$, and 95% asymptotic CI $= (0.421, 0.531)$.

Table 4.1 541 subjects' purchase choice of instant decaffeinated coffee at two times

	Second Purchase					
First Purchase	High Point	Taster's Choice	Sanka	Nescafe	Brim	Total
High Point	93	17	44	7	10	171
Taster's Choice	9	46	11	0	9	75
Sanka	17	11	155	9	12	204
Nescafe	6	4	9	15	2	36
Brim	10	4	12	2	27	55
Total	135	82	231	33	60	541

Table 4.2 Results of fixed and known a for example 1

a	$\hat{\kappa}(a)$	$\hat{SE}$	CI
0	0.476	0.028	(0.421, 0.531)
0.2	0.476	0.028	(0.421, 0.531)
0.4	0.476	0.028	(0.42, 0.531)
0.6	0.476	0.028	(0.42, 0.531)
0.8	0.476	0.028	(0.42, 0.531)
1	0.475	0.028	(0.42, 0.531)

Since the marginal proportions are similar, proposed agreement coefficients, Cohen's kappa and the RMAC produce very close estimates of agreement, estimates of asymptotic standard error and confidence intervals.

4.2 Carotid Artery Locations Data

We consider another example in which researchers interpreted Magnetic Resonance Images (MRIs) of 90 carotid artery locations and compared their findings with histopathologic examination. The researchers of study were interested in evaluating the MRI technique and improving its accuracy, specificity and sensitivity. The data are from Yuan et al. [38], and are displayed in Table 4.3. Note that the responses are binary.

The table shows that the MRIs and results from the histologic examination agreed that 56 of 90 were positive and 22 were negative. For the case where a is fixed and known, the results are displayed in Table 4.4. For the case where a is estimated, $\hat{a} = 0.089$, $\hat{\kappa}(\hat{a}) = 0.691$, $\hat{SE} = 0.081$, and 95% asymptotic CI=(0.532, 0.85).

Table 4.3 Comparison between MRI and histology findings

	Histology		
MRI	Positive	Negative	Total
Positive	56	2	58
Negative	10	22	32
Total	66	24	90

Table 4.4 Results of fixed and known a for example 2

a	$\hat{\kappa}(a)$	$\hat{SE}$	CI
0	0.692	0.081	(0.534, 0.85)
0.2	0.691	0.081	(0.531, 0.85)
0.4	0.69	0.082	(0.529, 0.851)
0.6	0.689	0.083	(0.528, 0.851)
0.8	0.689	0.083	(0.527, 0.851)
1	0.689	0.083	(0.526, 0.851)

Proposed agreement coefficients, Cohen's kappa and the RMAC produce very close estimates of agreement because the marginal proportions are similar. The agreement is quite respectable. The RMAC has the largest estimate of SE, whereas Cohen's kappa has the smallest estimate of SE.

4.3 Allergy Data

In this example, a radioallergosorbent (RAST) test and a multi-RAST (MAST) test on sera for specific IgE as a test of allergy in subjects for whom prick tests cannot be used were compared. The MAST was a new, simpler and cheaper method. The data are from Brostoff, et al. [6] and are displayed in Table 4.5. Note that the responses were recorded on an ordinal scale.

Visual inspection shows there is considerable disagreement between the methods. Since the categories are ordered, some weighting schemes are needed to account for severity of discordance or size of the discrepancy. For the case where a is fixed and known, results are shown in Table 4.6 and Table 4.7, with Cicchetti-Allison weights and Fleiss-Cohen weights, respectively. For the case where a is estimated, $\hat{a} = 0.102$. When the Cicchetti-Allison weights are used, $\hat{\kappa}_w(\hat{a}) = 0.558$, $\hat{SE} = 0.029$, and 95% asymptotic CI $= (0.501, 0.614)$; when the Fleiss-Cohen weights are used, $\hat{\kappa}_w(\hat{a}) = 0.711$, $\hat{SE} = 0.029$, and 95% asymptotic CI $= (0.655, 0.768)$.

For each weighting scheme, since the marginal proportions are slightly different, the estimates of agreement, estimates of asymptotic standard error and confidence intervals for the agreement coefficients are slightly different. It is clear that weighted kappa

Table 4.5 Comparison of RAST and MAST methods of testing serum for allergies

	RAST					
MAST	Negative	Weak	Moderate	High	Very High	Total
Negative	86	3	14	0	2	105
Weak	26	0	10	4	0	40
Moderate	20	2	22	4	1	49
High	11	1	37	16	14	79
Very High	3	0	15	24	48	90
Total	146	6	98	48	65	363

Table 4.6 Results of fixed and known a for example 3, with Cicchetti-Allison weights

a	$\hat{\kappa}_w(a)$	$\hat{SE}$	CI
0	0.559	0.029	(0.503, 0.615)
0.2	0.557	0.029	(0.5, 0.614)
0.4	0.556	0.029	(0.5, 0.613)
0.6	0.555	0.029	(0.497, 0.612)
0.8	0.554	0.029	(0.496, 0.612)
1	0.554	0.029	(0.496, 0.611)

Table 4.7 Results of fixed and known a for example 3, with Fleiss-Cohen weights

a	$\hat{\kappa}_w(a)$	$\hat{SE}$	CI
0	0.712	0.029	(0.656, 0.769)
0.2	0.711	0.029	(0.654, 0.768)
0.4	0.71	0.029	(0.652, 0.767)
0.6	0.709	0.03	(0.651, 0.767)
0.8	0.709	0.03	(0.65, 0.767)
1	0.708	0.03	(0.65, 0.767)

and RMAC form the upper and lower bounds, whereas the proposed agreement coefficients have values in between them. The estimates of agreement with Fleiss-Cohen weights are much larger than the ones with Cicchetti-Allison weights, because Fleiss-Cohen weights tend to weight disagreements more highly than Cicchetti-Allison weights.

4.4 Disease Diagnosis Data

In Table 4.8 we present a subset of the data from Westlund and Kurkland [37]. The investigators were interested in comparing patient groups to study possible differences in the geographical distributions of the disease. For illustration purpose, we only consider the 149 patients in Winnipeg, Manitoba that were selected and were examined by a neurologist in Winnipeg and a neurologist in New Orleans, Louisiana. Two neurologists independently classified patients into four ordinal categories, 1=certain multiple sclerosis (MS), 2=probable MS, 3=possible MS (50:50 odds), and 4=doubtful, unlikely, or definitely not MS. For the case where a is fixed and known, results are displayed in Table 4.9 and Table 4.10, with Cicchetti-Allison weights and Fleiss-Cohen weights, respectively. For the case where a is estimated, $\hat{a}$, is 0.161. When the Cicchetti-Allison weights are used, $\hat{\kappa}_w(\hat{a}) = 0.371$, $\hat{SE} = 0.055$, and 95% asymptotic CI $= (0.263, 0.477)$; when the Fleiss-Cohen weights are used, $\hat{\kappa}_w(\hat{a}) = 0.517$, $\hat{SE} = 0.062$, and 95% asymptotic CI $=$ $(0.394, 0.639)$.

Comparing with Example 3, the difference between marginal proportions in this example is even larger. As a result, the estimate of weighted kappa gives the largest agreement, whereas the estimate of RMAC gives the smallest agreement, and they are significantly different. Weighted kappa possesses the smallest estimate of SE, whereas

Table 4.8 Multiple sclerosis diagnosis

	Neurologist 2				
Neurologist 1	1	2	3	4	Total
1	38	5	0	1	44
2	33	11	3	0	47
3	10	14	5	6	35
4	3	7	3	10	23
Total	84	37	11	17	149

Table 4.9 Results of fixed and known a for example 4, with Cicchetti-Allison weights

a	$\hat{\kappa}_w(a)$	$\hat{SE}$	CI
0	0.38	0.052	(0.278, 0.481)
0.2	0.369	0.054	(0.262, 0.475)
0.4	0.36	0.056	(0.249, 0.471)
0.6	0.354	0.058	(0.24, 0.468)
0.8	0.35	0.059	(0.234, 0.466)
1	0.348	0.06	(0.232, 0.465)

Table 4.10 Results of fixed and known a for example 4, with Fleiss-Cohen weights

a	$\hat{\kappa}_w(a)$	$\hat{SE}$	CI
0	0.525	0.06	(0.407, 0.642)
0.2	0.515	0.063	(0.392, 0.638)
0.4	0.507	0.065	(0.379, 0.635)
0.6	0.502	0.067	(0.37, 0.633)
0.8	0.498	0.068	(0.364, 0.632)
1	0.497	0.069	(0.362, 0.632)

the RMAC has the largest one. In contrast, $\hat{\kappa}(\hat{a})$ has moderate estimate of agreement and estimate of SE.

4.5 Blood draw data

The data for this example came from a study conducted by the Asthma Clinical Research Network (ACRN) entitled Dose of Inhaled Corticosteroids with Equisystemic Effects (DICE) [34]. One hundred and fifty-six corticosteroid-naive patients with asthma were recruited at six ACRN centers. Among them, one hundred and twenty-one patients completed the trial. The major objective of this study was to investigate dose-response relationships for six inhaled corticosteroids. After one week of treatment to evaluate drug adherence, the subjects were required to stay at the center overnight, and hourly blood sampling for cortisol was conducted between 8:00P.M. and 8:00A.M.. Plasma cortisol area under the curve (AUC) was calculated from the trapezoidal rule over the 12-hour period of the hourly blood draws. The actual time points of plasma sampling, rather than the nominal hourly time points, were used for the calculation and standardized to a 12-hour period. The secondary objective of this study was to assess the agreement between plasma cortisol AUC calculated from measurements taken every hour and measurements taken every 2 hours. The study involved repeated measures because plasma AUC data were from five visits on each of the subjects. Here, I consider only the first time point for each subject.

The scatter plot (Figure 4.1) shows the distribution of the 1- and 2-hour blood draw data at the first visit of the DICE trial. For the case where a is fixed and known, we use six values of a, 0, 0.2, 0.4, 0.6, 0.8 and 1, and the results are shown in Table 4.11.

For the case where a is estimated, $\hat{a} = 0.018$, $\hat{\rho}(\hat{a}) = 0.95$, $\hat{SE} = 0.011$, and 95% CI is (0.929,0.972). In this example, the agreement is fairly respectable. Since the observed marginal distributions are nearly equal, both cases yield identical results.

Table 4.11 Results for blood draw data, where a is fixed and known

a	$\hat{\rho}(a)$	$\hat{SE}$	95% CI
0	0.95	0.011	(0.929,0.972)
0.2	0.95	0.011	(0.929,0.972)
0.4	0.95	0.011	(0.929,0.972)
0.6	0.95	0.011	(0.929,0.972)
0.8	0.95	0.011	(0.929,0.972)
1	0.95	0.011	(0.929,0.972)

4.6 Body fat data

The data are taken from the Penn State Young Women's Health Study [32]. One hundred and twelve adolescent girls were enrolled in 1990, and results were based on measurements made on the eighty-two participants who remained in the study in 1996. The objective of the study was to obtain simultaneous and longitudinal measures of several growth parameters in the participants. In this example, among these growth parameters, we focus on only percentage of body fat. Percentage of body fat was estimated from skinfold calipers and dual-energy x-ray absorptiometry (DEXA). The secondary objective of the study is to find the agreement between the skinfold caliper and DEXA measurements of percentage of body fat. Since measurements were taken every 6 months, I consider only the first time point for each subject.

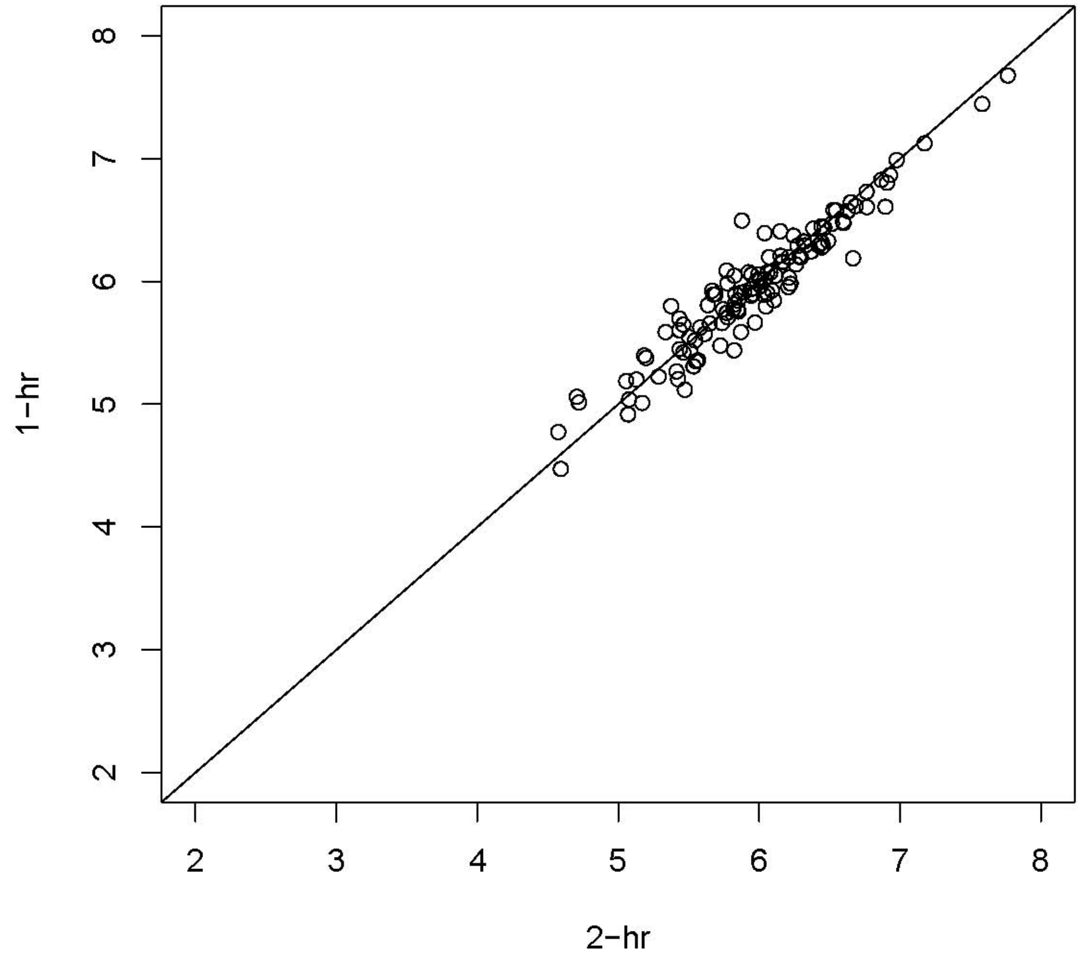

Fig. 4.1 Scatter plot of blood draw data

The scatter plot (Figure 4.2) shows the distribution of the DEXA and caliper body fat data at the first visit. For the case where a is fixed and known, we use six values of a, 0, 0.2, 0.4, 0.6, 0.8 and 1, and the results are shown in Table 4.12. For the case where a is estimated, $\hat{a} = 0.169$, $\hat{\rho}(\hat{a}) = 0.659$, $\hat{SE} = 0.053$, and 95% CI is (0.553,0.765). In this example, the agreement is moderate. Since the marginal distributions are slightly different, different values of fixed a yield slightly different estimates of agreement, standard error and confidence intervals, with decreasing agreement, increasing standard error, and confidence interval shifting to the left as a increases. On the other hand, the case where a is estimated provides estimates of agreement, standard error and confidence interval based on the difference between the marginal distributions, so the estimated agreement is not as inflated as CCC, and the estimated standard error is not as large as the one for RMAC.

Table 4.12 Results for body fat data, where a fixed and known

a	$\hat{\rho}(a)$	$\hat{SE}$	95% CI
0	0.667	0.051	(0.566,0.767)
0.2	0.658	0.054	(0.552,0.764)
0.4	0.651	0.056	(0.539,0.762)
0.6	0.646	0.058	(0.530,0.761)
0.8	0.643	0.059	(0.525,0.76)
1	0.641	0.06	(0.523,0.76)

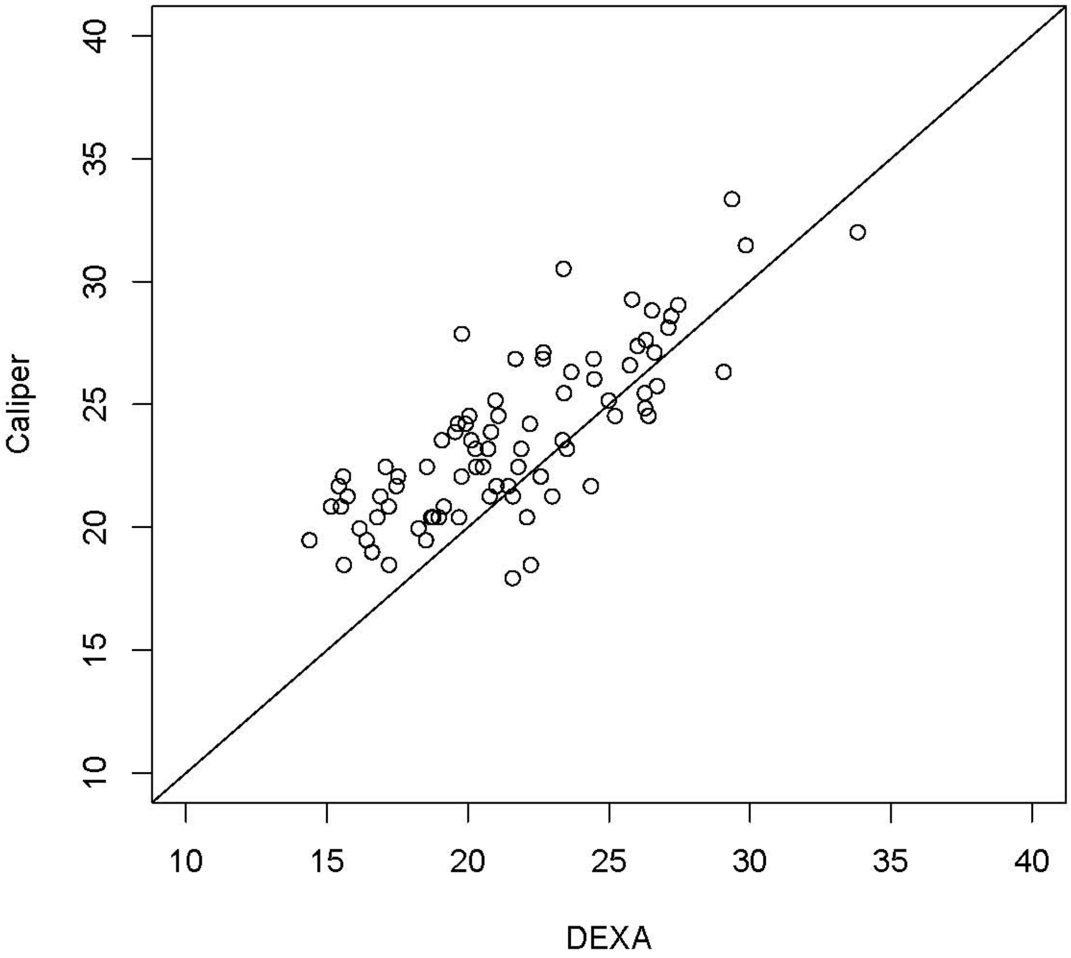

Fig. 4.2 Scatter plot of body fat data

Chapter 5

Conclusions and Future Work

5.1 Conclusions

In summary, we have proposed a general class of agreement coefficients for categorical and continuous responses based on the difference between marginal distributions. The difference, defined as a, is measured using Cramer-von Mises criterion. In practice, we consider case of fixed and known a and case of estimated a. If we call Cohen's kappa, weighted kappa and Lin's CCC the fixed marginal agreement coefficients (FMACs), then when the observed marginal proportions are similar, the proposed agreement coefficient produces similar estimates of agreement and asymptotic standard error, independent of a; when the observed marginal proportions are different, the proposed agreement coefficient with estimated a produces estimate of agreement and estimated asymptotic standard error in between FMAC's and RMAC's. We demonstrate that estimating a is preferred for most underlying distribution cases, for its better accuracy and precision comparing to fixed and known a. In conclusion, the general class of agreement coefficients is appropriate for categorical and continuous data and provides a good balance between efficiency and robustness.

In practice, the value of fixed and known a is chosen by the researcher. However, selecting a a priori has the (1) risk of yielding an inaccurate agreement, e.g., the true value of a might be very different from the value of a chosen by researcher; and (2) risk of underestimating the variability, because the variability associated with choosing a is not taken into account. On the other hand, estimating a will account for the estimation of a and yield the correct level of precision. Therefore, the case of estimated a is more appealing and should be considered for use.

5.2 Future Work

The general class of agreement coefficients has some desired properties and the potential to develop further. Future work would include:

(1) Extend the proposed method to handle multiple raters;

(2) Improve the performance of proposed agreement coefficients on non-normal data (e.g. log-normal distribution).

Bibliography

[1] Agresti, A. (1988). A model for agreement between ratings on an ordinal scale. *Biometrics.* **44** 539-548.

[2] Anderson, T. W. (1962). On the distribution of the two-sample Cramer-Von Mises criterion. *The Annals of Mathematical Statistics.* **33** 1148-1159.

[3] Banerjee, M., Capozzoli, M., McSweeney, L., and Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics.* **27** 3-23.

[4] Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and practice.* MIT Press, Cambridge, MA.

[5] Bloch, D.A. and Kraemer, H.C. (1989). 2 × 2 kappa coefficients: Measures of agreement or associatioin. *Biometrics.* **45**. 269-287.

[6] Brostoff, J., Pack, S. and Merrett, T. (1984). A new multiple specific IgE assay-Mast. *Lancet.* **i,** 748-749.

[7] Byrt, T., Bishop, J., and Carlin, J.B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology.* **46** 423-429.

[8] Cicchetti, D.V. and Allison, T. (1973). Assessing the reliability of scoring EEG sleep records: an improved method. *Proceedings and Journal of the Electro-physiological Technologist's Association.* **20**. 92-102.

[9] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurements.* **20** 37-46.

[10] Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin.* **70** 213-220.

[11] Darroch, J.N., and McCloud, P.I. (1986). Category distinguishability and observer agreement. *Australian Journal of Statistics.* **28** 371-388.

[12] Donner, A., and Eliasziw, M. (1992). A goodness-of-fit approach to inference procedures for the kappa statistic: Confidence interval construction, significance-testing and sample size estimation. *Statistics in Medicine.* **11** 1151-1519.

[13] Efron, B., and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap.* Chapman & Hall, New York.

[14] Fay, M.P. (2005). Random marginal agreement coefficients: Rethinking the adjustment for chance when measuring agreement. *Biostatistics.* **6** 171-180.

[15] Feinstein, A.R., and Cicchetti, D.V. (1990). High agreement but low kappa I: The problems of two paradoxes. *Journal of Clinical Epidemiology.* **43** 543-548.

[16] Fisher, R.A. (1958). *Statistical Methods for Research Workers.* 13th edition. Hafner, New York.

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

[18] Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin.* **76** 378-382.

[19] Fleiss, J.L., Nee, J.C.M., and Landis, J.R. (1979). Large sample variance of kappa in the case of different sets of raters. *Psychological Bulletin.* **86** 974-977.

[20] Fleiss, J.L., and Cuzick, J. (1979). The reliability of dichotomous judgements: unequal numbers of judges per subject. *Applied Psychological Measurement.* **3** 537-542.

[21] Fleiss, J.L. (2003). *Statistical Methods for Rates and Proportions.* 3rd edition. Wiley, New York.

[22] Goodman, L.A. (1979). Simple models for the analysis of association in cross-classification having ordered categories. *Journal of the American Statistical Association.* **74** 537-552.

[23] Gross. P. (1971). *A study of supervisor reliability.* Mimeograph, Laboratory of Human Development, Harvard Graduate School of Education, Cambridge, Mass.

[24] Grover, R. and Srinivasan. V. (1987). A simultaneous approach to market segmentation and market structuring. *Journal of Marketing Research.* **24,** 139-153.

[25] Guilford, J.P. (1950). *Fundamental Statistics in Psychology and Education.* 2nd edition. McGraw-Hill, New York.

[26] King, T.S., and Chinchilli, V.M. (1999). Developing robust estimators of the concordance correlation coefficient. *Institute of StatisticsMimeo Series No. 2197T.* The Univeristy of North Carolina: Chapel Hill, North Carolina.

[27] King, T.S., and Chinchili, V.M. (2001). Robust estimators for the concordance correlation coefficient. *Journal of Biopharmaceutical Statistics.* **11** 83-105.

[28] King, T.S., and Chinchilli, V.M. (2001). A generalized concordance correlation coefficient for continous and categorical data. *Statisics in Medicine.* **20** 2131-2147.

[29] Landis, R.J., and Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics.* **33** 159-174.

[30] Li, R., and Chow, M. (2001). Evaluation of reproducibility when the data are curves. *Technical Report*, **01-07**, Department of Statistics, The Pennsylvania State University.

[31] Lin, L. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, **45** 255-268.

[32] Lloyd, T., Chinchilli, V. M., Eggli, D. F., Rollings, N., and Kulin, K. E. (1998). Body composition development of adolescent white females. *Archives of Pediatric Adolescence Medicine*, **152** 998-1002.

[33] Maclure, M., and Willett, W.C. (1987). Misinterpretation and misuse of the kappa statistic. *American Journal of Epidemiology.* **126** 161-169.

[34] Martin, R. J., Szefler, S. J., Chinchilli, V. M. et al. (2002). Systemic effect comparisons of six inhaled corticosteroid preparations. *American Journal of Respiratory and Critical Care Medicine*, **165** 1377-1383.

[35] Scott, W.A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly.* **19** 321-325.

[36] Tanner, M.A., and Young, M.A. (1985). Modeling agreement among raters. *Journal of American Statistical Association.* **80** 175-180.

[37] Westlund, K. B. and Kurkland, L. T. (1953). Studies in multiple sclerosis in Winnipeg, Manitoba and New Orleans, Louisiana. *American Journal of Hygiene.* **57,** 380-396.

[38] Yuan, C., et al. (2001). In Vivo Accuracy of Multispectral Magnetic Resonance Imaging for Identifying Lipid-Rich Necrotic Cores and Intraplaque Hemorrhage in Advanced Human Carotid Plaques. *Circulation.* **104,** 2051-2056.

[39] Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin.* **103** 374-378.

Vita
WEI ZHANG

Education

The Pennsylvania State University University Park, Pennsylvania 2004–2008

Ph.D. in Statistics

University of Toledo Toledo, Ohio 2002–2004

M.S. in Statistics

Xi'an Institute of Finance and Economics Xi'an, Shaanxi, P. R. China1998–2002

Bachelor of Economics in Management Information Systems

Professional Experience

Research Assistant The Pennsylvania State University 2007–2008

Biostatistics Intern ProSanos Corporation 2007

Teaching Assistant The Pennsylvania State University 2004–2007

Teaching Assistant University of Toledo 2002–2004

Professional Affiliations

American Statistical Association

International Biometric Society-Eastern North American Region (ENAR)

Institute of Mathematical Statistics