The Pennsylvania State University

The Graduate School

Department of Statistics

MEASUREMENT OF AGREEMENT FOR CATEGORICAL DATA

A Thesis in

Statistics

by

Jingyun Yang

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2007

The thesis of Jingyun Yang was read and approved[1] by the following:

Vernon M. Chinchilli
Professor of Statistics
Chairman, Department of Public Health Sciences
Thesis Co-Adviser
Co-Chair of Committee

Runze Li
Associate Professor of Statistics
Thesis Co-Adviser
Co-Chair of Committee

Damla Senturk
Assistant Professor of Statistics

Tonya S. King
Associate Professor of Public Health Sciences

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

[1]Signatures on file in the Graduate School.

Abstract

Measurements of agreement are used to assess the reproducibility of a new assay or instrument, the acceptability of a new or generic process, methodology or method comparison. Examples include the agreement when two methods or two raters simultaneously assess a response or when one rater makes the same assessment at two times, the agreement of a newly developed method with a gold standard method, and the agreement of observed values with predicted values.

Traditionally, kappa and weighted kappa coefficients are used for measurements of agreement when the responses are categorical. The concordance correlation coefficient is used when the responses are continuous. Cohen's kappa and weighted kappa coefficients have received many criticisms since they were proposed and they may fail to work well under certain situations. As a result, researchers have suggested the investigation of alternative methods when measuring agreement.

In this paper, we investigate several different alternatives to Cohen's kappa and weighted kappa coefficients. Their properties and asymptotic distributions are presented. Simulation performances are provided to compare with the performances of Cohen's kappa and weighted kappa coefficients.

Table of Contents

List of Tables

List of Figures

Acknowledgments

I would first like to thank my advisor, Dr. Vernon M. Chinchilli, for his guidance, support, patience and expertise in the past years as we worked through my research. His valuable feedback and suggestions have been instrumental in designing and writing this study. Many thanks should also be given to my co-advisor, Dr. Runze Li, for instructing me in dealing with some of the most difficult parts of my research. I would also like to thank the committee members: Dr. Damla Senturk and Dr. Tonya S. King. Thank you for your questions and suggestions throughout my doctoral studies.

I also appreciate the help from the faculty and graduate students at the Penn State Statistics Department. Special thanks to Weixin Yao for his valuable discussions and Bing Han for helpful suggestions.

I am truly grateful to my mother, for her endless love and support and my wife, Wenshu Li, for her love and continuous encouragement, support and understanding.

Finally, I am greatly indebted to my father. He passed away 12 years ago. He instilled in me the importance of education and had always expected that I could receive an excellent higher education. His expectations have always been present in my mind, without which this paper was not possible. I miss him so much.

During my thesis research, I have been supported as a summer research assistant by NSF grant DMS 0348869 and NIH grant P50-DA10075.

Chapter 1

Literature Review

Measurements of agreement are used to assess the reproducibility of a new assay or instrument, the acceptability of a new or generic process, methodology or method comparison. Examples include the agreement when two methods or two raters simultaneously assess a response or when one rater makes the same assessment at two times, the agreement of a newly developed method with a gold standard method, and the agreement of observed values with predicted values. In recent years, the question of agreement has received considerable attention. In this chapter, we review briefly the various measures in evaluating agreement.

1.1 Cohen's Kappa and Weighted Kappa Coefficients

Suppose there is a bivariate response, (X, Y), where each of X and Y yields a categorical response. For convenience, the categories are denoted as $0, 1, 2, \ldots, k$. Suppose the following is the bivariate distribution table:

		y									
		0	1	2	$\cdots$	i	$\cdots$	j	$\cdots$	k	$P_X(x)$
	0	p_{00}	p_{01}	p_{02}	$\cdots$	p_{0i}	$\cdots$	p_{0j}	$\cdots$	p_{0k}	$p_{0\cdot}$
	1	p_{10}	p_{11}	p_{12}	$\cdots$	p_{1i}	$\cdots$	p_{1j}	$\cdots$	p_{1k}	$p_{1\cdot}$
	2	p_{20}	p_{21}	p_{22}	$\cdots$	p_{2i}	$\cdots$	p_{2j}	$\cdots$	p_{2k}	$p_{2\cdot}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x	i	p_{i0}	p_{i1}	p_{i2}	$\cdots$	p_{ii}	$\cdots$	p_{ij}	$\cdots$	p_{ik}	$p_{i\cdot}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	j	p_{j0}	p_{j1}	p_{j2}	$\cdots$	p_{ji}	$\cdots$	p_{jj}	$\cdots$	p_{jk}	$p_{j\cdot}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	k	p_{k0}	p_{k1}	p_{k2}	$\cdots$	p_{ki}	$\cdots$	p_{kj}	$\cdots$	p_{kk}	$p_{k\cdot}$
	$P_Y(y)$	$p_{\cdot 0}$	$p_{\cdot 1}$	$p_{\cdot 2}$	$\cdots$	$p_{\cdot i}$	$\cdots$	$p_{\cdot j}$	$\cdots$	$p_{\cdot k}$	1

Define p_o and p_c as

$$p_o = \sum_{i=0}^{k} p_{ii}$$

and

$$p_c = \sum_{i=0}^{k} p_{i\cdot} p_{\cdot i}.$$

Cohen (1960) proposed a coefficient of agreement, called the kappa coefficient, for nominal scales of response, which is defined by

$$\kappa = \frac{p_o - p_c}{1 - p_c}, \tag{1.1}$$

where p_o is the observed proportion of agreement and p_c is the proportion of agreement expected by chance.

If $q_o = 1 - p_o$ and $q_c = 1 - p_c$, then (1.1) can also be written as:

$$\kappa = \frac{(1 - q_o) - (1 - q_c)}{q_c} = \frac{q_c - q_o}{q_c} = 1 - \frac{q_o}{q_c}. \tag{1.2}$$

Cohen also derived the standard error of an observed $\hat{\kappa}$ for a large sample of size N:

$$\sigma_\kappa \cong \sqrt{\frac{p_o\,(1 - p_o)}{N\,(1 - p_c)^2}}.$$

For a large sample, κ is approximately normally distributed. This expression was later found to be incorrect. Fleiss and Cohen (1969) derived a corrected version of the large sample variance of the estimated kappa:

$$\begin{aligned} \hat{\mathrm{Var}}(\hat{\kappa}) &= \frac{1}{N(1-p_c)^4}\{\sum_{i=0}^{k+1} p_{ii}[(1-p_c) - (p_{\cdot i} + p_{i\cdot})(1-p_o)]^2 \\ &\quad +(1-p_o)^2 \sum_{i=0}^{k+1} \sum_{\substack{j=0 \\ j\neq i}}^{k+1} p_{ij}(p_{\cdot i} + p_{j\cdot})^2 - (p_o p_c - 2p_c + p_o)^2\}, \end{aligned} \tag{1.3}$$

where N is the sample size.

We can use (1.3) to find confidence limits as well as performing a hypothesis test.

The kappa coefficient provides a simple way to measure agreement. (1.1) gives negative values when the observed agreement is less than that expected by chance. It yields 0 when the observed agreement can be exactly accounted for by chance and it yields 1 when there is complete agreement.

Landis and Koch (1977) have characterized different ranges of values for kappa with respect to the degree of agreement they suggest. For most purposes, values greater than 0.75 or so may be taken to represent excellent agreement beyond chance, values below 0.40 or so may be taken to represent poor agreement beyond chance, and values between 0.40 and 0.75 may be taken to represent fair to good agreement beyond chance.

All previous articles in the literature, when referring to Cohen's kappa coefficient, say that the lower bound is -1. It seems that there is no formal proof for this statement. In the following, through a new definition which is equivalent to Cohen's kappa coefficient, we show that the universal lower bound of Cohen's kappa is -1. We state this as a theorem.

Theorem 1.1.1. *The universal lower bound of the kappa coefficient is* -1.

Proof. Let X and Y be categorical variables with possible outcomes $0, 1, \cdots, k$. Let $P_{XY}(x, y) = \Pr[X = x, Y = y]$, where $X, Y = 0, 1, \cdots, k$ denote joint probabilities. Let $P_X(x) = \Pr[X = x]$ and $P_Y(y) = \Pr[Y = y]$ denote marginal probabilities. We define a new statistic

$$e_d = 1 - \frac{\sum\sum_{x \neq y} P_{XY}(x, y) d(x, y)}{\sum\sum_{x \neq y} P_X(x) P_Y(y) d(x, y)},$$

where $d(x, y)$ is the distance function such that $d(x, y) \geq 0$.

For categorical responses, we can use the following distance function:

$$\text{d(x,y)}=\begin{cases} 1, & \text{if } x \neq y \\ 0, & \text{if } x = y \end{cases}.$$

We can use the following estimator which is quite intuitive:

$$\hat{e}_d = 1 - \frac{\sum\sum_{k \neq l} \frac{n_{kl}}{n} d(k,l)}{\sum\sum_{k \neq l} \frac{n_{k+}}{n} \cdot \frac{n_{+l}}{n} d(k,l)} = 1 - \frac{n \cdot \sum\sum_{k \neq l} n_{kl} d(k,l)}{\sum\sum_{k \neq l} n_{k+} \cdot n_{+l} d(k,l)},$$

where n_{kl} is the number of observations when $x = k$, $y = l$. n_{k+} represents the number of observations when $x = k$ and n_{+l} represents the number of observations when $y = l$.

Apply the above distance function, we can get

$$\hat{e}_d = 1 - \frac{\sum\sum_{k \neq l} \frac{n_{kl}}{n}}{\sum\sum_{k \neq l} \frac{n_{k+}}{n} \cdot \frac{n_{+l}}{n}} = 1 - \frac{n \cdot \sum\sum_{k \neq l} n_{kl}}{\sum\sum_{k \neq l} n_{k+} \cdot n_{+l}}.$$

Let $q_o = \sum\sum_{k \neq l} \frac{n_{kl}}{n}$ and $q_c = \sum\sum_{k \neq l} \frac{n_{k+}}{n} \cdot \frac{n_{+l}}{n}$, then

$$\hat{e}_d = 1 - \frac{q_o}{q_c}, \tag{1.4}$$

which turns out to be the estimator of the simple Kappa coefficient. So e_d is equivalent to Cohen's kappa. Therefore, we need only to show that $e_d \geq -1$.

Remark 1. *For ordinal responses, if we use the absolute distance function,* $d(x, y) = |a - b|$, *then* e_d *can be estimated by*

$$\hat{e}_d = 1 - \frac{\sum\sum_{k \neq l} \frac{n_{kl}}{n} |k-l|}{\sum\sum_{k \neq l} \frac{n_{k+}}{n} \cdot \frac{n_{+l}}{n} |k-l|} = 1 - \frac{n \cdot \sum\sum_{k \neq l} n_{kl}}{\sum\sum_{k \neq l} n_{k+} \cdot n_{+l}}.$$

Let $v_{kl} = |k-l|$, $p_{okl} = \frac{n_{kl}}{n}$, $p_{ckl} = \frac{n_{k+}}{n} \cdot \frac{n_{+l}}{n}$.

Then $q_o' = \sum\sum_{k \neq l} p_{okl} v_{kl}$ *and* $q_c' = \sum\sum_{k \neq l} p_{ckl} v_{kl}$.

Consequently, $\hat{e}_d$ *can be rewritten as*

$$\hat{e}_d = 1 - \frac{q_o'}{q_c'}.$$

And this turns out to be the estimator for weighted kappa with weights being the absolute distance function.

Note that to show

$$e_d = 1 - \frac{\sum\sum_{x \neq y} P_{XY}(x,y) d(x,y)}{\sum\sum_{x \neq y} P_X(x) P_Y(y) d(x,y)} \geq -1,$$

it is equivalent to show that

$$\frac{\sum\sum_{x \neq y} P_{XY}(x,y) d(x,y)}{\sum\sum_{x \neq y} P_X(x) P_Y(y) d(x,y)} \leq 2.$$

If we use the above distance function, we need to show that

$$\sum\sum_{x \neq y} P_{XY}(x,y) \leq 2 \sum\sum_{x \neq y} P_X(x) P_Y(y).$$

That is we need

$$
\begin{aligned}
1-[P(0,0)+P(1,1)+\ldots+P(k,k)] \quad \leq \quad & 2\{P_X(0)\,[P_Y(1)+P_Y(2)+\ldots+P_Y(k)] \\
& +P_X(1)\,[P_Y(0)+P_Y(2)+\ldots+P_Y(k)] \\
& +\ldots+P_X(k)\,[P_Y(0)+P_Y(1)+\ldots+P_Y(k-1)]\}.
\end{aligned}
$$

This is equivalent to showing that

$$
\begin{aligned}
1-[P(0,0)+P(1,1)+\ldots+P(k,k)] \quad \leq \quad & 2\{P_X(0)\,[1-P_Y(0)]+P_X(1)\,[1-P_Y(1)] \\
& +\ldots+2P_X(k)\,[1-P_Y(k)]\},
\end{aligned}
$$

which is to show that

$$
\begin{aligned}
1-[P(0,0)+P(1,1)+\ldots+P(k,k)] \quad \leq \quad & 2\,[P_X(0)+P_X(1)+\ldots+P_X(k)] \\
& -2\,[P_X(0)P_Y(0)+P_X(1)P_Y(1)+\ldots P_X(k)P_Y(k)]\,.
\end{aligned}
$$

Since $2\,[P_X(0)+P_X(1)+\ldots+P_X(k)]=2$, it is equivalent to show that

$$
2\,[P_X(0)P_Y(0)+P_X(1)P_Y(1)+\ldots P_X(k)P_Y(k)] \leq 1+[P(0,0)+P(1,1)+\ldots+P(k,k)]\,.
$$

- The simple case is if all $P_X(i)$ or $P_Y(i)$,where $i = 0, 1, 2 \dots k$, satisfy $P_X(i) \leq \frac{1}{2}$ or $P_Y(i) \leq \frac{1}{2}$. Suppose all $P_X(i)$, where $i = 0, 1, 2 \dots k$, satisfy $P_X(i) \leq \frac{1}{2}$. Then

$$\begin{aligned}
& 2[P_X(0)P_Y(0) + P_X(1)P_Y(1) + \dots P_X(k)P_Y(k)] \\
\leq\ & 2[1/2P_Y(0) + 1/2P_Y(1) + \dots 1/2P_X(k)] \\
=\ & P_Y(0) + P_Y(1) + \dots P_X(k) = 1 \leq 1 + [P(0,0) + P(1,1) + \dots + P(k,k)].
\end{aligned}$$

  So [(1.1)] is satisfied.

- Now suppose there is one $P_X(i)$ such that $P_X(i) > 1/2$. We know that there can be only one such $P_X(i)$. Without loss of generality, we can let $i = 0$. That is, suppose the marginal probability for the first category is greater than 1/2. Let this probability be $1/2 + a$, where $0 < a \leq 1/2$. We know that for all $P_X(i)$, where $i = 1, 2 \dots, k$, $P_X(i) \leq 1/2 - a$. Then

$$\begin{aligned}
& 2[P_X(0)P_Y(0) + P_X(1)P_Y(1) + \dots P_X(k)P_Y(k)] \\
=\ & 2[(1/2 + a)P_Y(0) + P_X(1)P_Y(1) + \dots P_X(k)P_Y(k)] \\
\leq\ & 2[(1/2 + a)P_Y(0) + (1/2 - a)P_Y(1) + (1/2 - a)P_Y(2) + \dots + (1/2 - a)P_Y(k)] \\
=\ & 2[1/2P_Y(0) + 1/2P_Y(1) + 1/2P_Y(2) + \dots + 1/2P_Y(k) \\
+\ & aP_Y(0) - aP_y(1) - aP_Y(2) - \dots - aP_Y(k)] \\
=\ & [P_Y(0) + P_Y(1) + P_Y(2) + \dots + P_Y(k)] + a[P_Y(0) - (P_Y(1) + P_Y(2) + \dots + P_Y(k))] \\
=\ & 1 + a[P_Y(0) - (1 - P_Y(0))] \\
=\ & 1 + a(2P_Y(0) - 1).
\end{aligned}$$

Now if $P_Y(0) \leq 1/2$, then $(2P_Y(0)-1) \leq 0$. Hence,

$$1 + a(2P_Y(0)-1) \leq 1 + [P(0,0) + P(0,1) + \ldots + P(k,k)] .$$

So [(1.1)] is satisfied.

- Now if $P_Y(0) > 1/2$. Let $P_X(0) + P_Y(0) = b$, then $1 < b \leq 2$.

$$\begin{aligned}
& 2[P_X(0)P_Y(0) + P_X(1)P_Y(1) + \ldots P_X(k)P_Y(k)] \\
\leq\ & 2\left(P_X(0)P_Y(0) + (1-P_X(0))[P_Y(1) + P_Y(2) + \ldots + P_Y(k)]\right) \\
=\ & 2\left[P_X(0)P_Y(0) + (1-P_X(0))(1-P_Y(0))\right] \\
\leq\ & \left(P_X(0) + P_Y(0)\right)^2/2 + \left(2 - P_X(0) - P_Y(0)\right)^2/2 \\
=\ & b^2/2 + (2-b)^2/2 \\
=\ & \frac{2b^2 - 4b + 4}{2} = b^2 - 2b + 2,
\end{aligned}$$

where $b > 1$.

Furthermore, since we know that $P_X(0) + P_Y(0) - P(0,0) \leq 1$, so we can get $P(0,0) \geq P_X(0) + P_Y(0) - 1$. And because

$$\begin{aligned}
& 1 + [P(0,0) + P(0,1) + \ldots + P(k,k)] \\
\geq\ & 1 + P(0,0) \\
\geq\ & 1 + P_X(0) + P_Y(0) - 1 \\
=\ & P_X(0) + P_Y(0) \\
=\ & b,
\end{aligned}$$

so if we can prove that $b^2 - 3b + 2 \leq 0$ when $1 < b \leq 2$, then we can say that (1.1) is satisfied. A simple check shows that this is indeed true. And this completes the proof.

□

Through our equivalent definition of Cohen's kappa, it is straightforward to notice that κ can only reach +1 only when $\sum\sum_{x \neq y} P_{XY}(x, y) = 0$, which means that all off-diagonal cells are 0.

From the definition of the kappa coefficient, we can see that it makes no distinction between different disagreements, assuming equal weights to all off-diagonal cells. But in many situations, some off-diagonal cells are of greater gravity than others. Cohen (1968) later proposed the weighted kappa coefficient to deal with ordinal categorical responses.

Define

$$q_o^{'} = \frac{\sum v_{ij} p_{oij}}{v_{max}}$$

and

$$q_c^{'} = \frac{\sum v_{ij} p_{cij}}{v_{max}},$$

where p_{oij} is the proportion of joint judgement in the ij cell and p_{cij} is the proportion in the same cell expected by chance. In particular, $p_{cij} = p_{i.} p_{.j}$. And v_{ij} is the disagreement whose assignment is determined, in many instances, by consensus of a committee of substantive experts or even by the investigator's own judgement. v_{max} is the weight we give to the maximum disagreement.

Then the weighted kappa coefficient can be defined as:

$$\kappa_w = 1 - \frac{q_o'}{q_c'} = 1 - \frac{\sum v_{ij} p_{oij}}{\sum v_{ij} p_{cij}}. \tag{1.5}$$

Cohen presented the standard error of $\hat{\kappa}_w$ for a large sample of size N, which is:

$$\sigma_{\kappa_w} \cong \sqrt{\frac{\sum v_{ij}^2 p_{oij} - \left(\sum v_{ij} p_{oij}\right)^2}{N\left(\sum v_{ij}^2 p_{cij}\right)^2}}.$$

Again, Fleiss and Cohen (1969) presented the corrected version of the large sample variance of the estimated weighted kappa coefficient:

$$\begin{aligned} \hat{\text{Var}}(\hat{\kappa}_w) &= \frac{1}{N(1-p_c)^4}\{\sum_{i=0}^{k+1}\sum_{j=0}^{k+1} p_{ij}[w_{ij}(1-p_c) - (\bar{w}_{i\cdot} + \bar{w}_{\cdot j})(1-p_o)]^2 \\ &\quad -(p_o p_c - 2p_c + p_o)^2\}. \end{aligned} \tag{1.6}$$

Since kappa is a special case of weighted kappa with $w_{ij} = 1$ for $i = j$ and $w_{ij} = 0$ for $i \neq j$, it can be shown that (1.3) is a special case of (1.6).

The interpretation of the magnitude of weighted kappa is like that of unweighted kappa: $\hat{\kappa}_w \geq 0.75$ or so signifies excellent agreement, for most purposes, $\hat{\kappa}_w \leq 0.40$ or so signifies poor agreement and $0.4 < \hat{\kappa}_w < 0.75$ signifies fair to good agreement.

1.2 Extensions of Kappa and Weighted Kappa Coefficients

Since the time of Cohen's proposed kappa and weighted kappa coefficient, many researchers have worked on this topic to develop extensions.

Gonin et al. (2000) suggested modeling κ as a function of covariates and using generalized estimating equations to estimate the weighted κ coefficient.

Suppose in an agreement study, object i is rated by R_i raters. The rating from observer r $(r = 1, \ldots, R_i)$ on subject i is denoted by Y_{ir}. Each subject has a subject-specific covariate vector g_i of dimension $P_1 \times 1$ and also R_i rater-specific covariates g_{ir}, $r = 1, \cdots, R_i$ of dimension $P_2 \times 1$. Let $X_i = (x_{i1}, \ldots, x_{iR_i})$ represent the $R_i \times (P_1 + P_2)$ matrix of covariates for individual i. Denote the probability of rating $(Y_{ir} = j, Y_{is} = k)$ from observer r and s on subject i by

$$p_{irs,jk} = \Pr(Y_{ir} = j, Y_{is} = k | x_{ir}, x_{is}).$$

Then the weighted κ coefficient is defined as

$$\kappa_{irs} = \frac{\pi_{irs,\text{agree}} - \pi_{irs,\text{chance}}}{1 - \pi_{irs,\text{chance}}},$$

where $\pi_{irs,\text{agree}}$ is the weighted combination of joint probabilities between r and s:

$$\pi_{irs,\text{agree}} = \sum_{j=1}^{K} \sum_{k=1}^{K} w_{jk} p_{irs,jk},$$

and $\pi_{irs,\text{chance}}$ is the weighted combination of cell probabilities between r and s under an independence assumption:

$$\pi_{irs,\text{chance}} = \sum_{j=1}^{K} \sum_{k=1}^{K} w_{jk} p_{ir,j} p_{is,k}.$$

The regression model to estimate the weighted κ coefficient is

$$g(\kappa_{irs}) = \log\left(\frac{1+\kappa_{irs}}{1-\kappa_{irs}}\right) = Z'_{irs}\gamma, \tag{1.7}$$

where z_{irs} is some function of covariates x_{ir} and x_{is}.

Gonin et al. (2000) proposed the estimating equations for γ as well as the specification of the working covariances. Variance of $\hat{\gamma}$ can be obtained using a jackknife estimate. An advantage of the proposed method is that it applies to situations where there are two or more raters. It also works when the data are unbalanced.

Kraemer (1980) proposed an extension of the kappa coefficient which deals appropriately with multiple responses per observation and which permits multiple and not necessarily equal numbers of observations per subject. Let $r_i (i = 1, 2, \cdots, N)$ be the average Spearman rank correlation coefficient among $\frac{1}{2}m_i(m_i - 1)$ pairs of observations of subject i, and r_I be the average of $r_1, r_2, \cdots, r_N$. Let r_T be the average Spearman rank correlation coefficient among all pairs of observations. Then the extended kappa is defined as

$$\kappa_0 = \frac{r_I - r_T}{1 - r_T}.$$

The nonnull conditional distribution, when assuming a fixed marginal distribution, can be easily obtained and can be used to construct confidence intervals or performing test. For the unconditional case, jackknife procedure can be applied to build confidence intervals or perform hypothesis tests.

1.3 Models for Measurement of Agreement

There also have been many models to measure agreement.

Agresti (1988) proposed an agreement plus linear-by-linear association model which decomposes the overall agreement into three parts: chance agreement (agreement when the classifications are independent), agreement due to a baseline association between the ratings, and an increment that reflects the agreement beyond that from the chance agreement or from the baseline association. This decomposition can be summarized as

$$\log m_{ij} = \mu + \lambda_j^A + \lambda_j^B + \beta_{ij} + \delta(i,j),$$

where β_{ij} represents a given structural form which reflects the expected baseline association and

$$\delta(i,j) = \begin{cases} \delta & i = j \\ 0 & i \neq j \end{cases}.$$

Since this is a log-linear model, it can be fitted using any software that has log-linear options, such as SAS, GLIM, etc. If we want to test whether the ratings are independent or not, the null hypothesis is $H_o : \beta = \delta = 0$. To test whether there is extra agreement beyond the agreement due to baseline association, the null hypothesis is $H_o : \delta = 0$. To test whether there is extra association beyond that due to exact agreement, the null hypothesis is $H_o : \beta = 0$.

Some of the positive features of this model are:

1. Making no independence assumption given that the raters agree.

2. Easy interpretation of the baseline association and extra agreement.

3. Utilization of orderings of the response categories.

Laurent (1998) proposed a measure to evaluate the agreement between an approximate method and a gold standard. The model that is considered is

$$X_i = G_i + \varepsilon_i,$$

where X_i is the approximate measurement on the ith unit, $i = 1, \ldots, n$. G_i is a random variable with mean μ and variance σ_G^2, which represents the corresponding gold standard measurement. ε_i is measurement error for the approximate method with mean 0 and variance σ^2. It is also assumed that ε_i is independent of G_i.

Laurent proposed to use

$$\rho = \frac{\sigma_G^2}{\sigma_G^2 + \sigma^2}$$

as the measure for evaluating the agreement between the approximate and the gold standard. The preferred estimating method is

$$\frac{1}{1+(n-1)S_{DD}/(nS_{GG})},$$

where $S_{GG} = \sum(G_i - \bar{G})^2$ and $S_{DD} = \sum D_i^2$ with $\bar{G} = \sum G_i/n$ and $D_i = X_i - G_i$.

When there is more than one approximate method and one gold standard, the proposed model is

$$X_{ij} = G_i + \varepsilon_{ij},$$

where ε_{ij} is the measurement error on the ith unit by the jth approximate method.

We can use

$$\rho_j = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_{jj}}$$

as the measure for evaluating the agreement between the jth approximate method and the gold standard. Under certain assumptions, the maximum likelihood estimator of ρ_j is

$$r_{g(j)}^2 = \frac{1}{1 + S_{DD(j)}/S_{GG}},$$

where $S_{DD(j)} = \sum(X_{ij} - G_i)^2$ for $j = 1, \ldots, J$.

Carrasco and Jover (2005) proposed the following measurement model for count data:

$$Y_{ij}|\alpha_i \sim \text{Poisson}(\mu_{ij}),\ \alpha_i \sim N(0, \sigma_\alpha^2),\ \log(\mu_{ij}) = \beta_0 + \alpha_i + \beta_j,$$

where Y_{ij} is the measurement from the ith subject obtained by the jth observer, μ_{ij} is the mean of the measurement from the ith subject and jth observer, β_0 is the baseline, α_i is the subject-effect and β_j is the observer effect.

Using this model, the intraclass correlation coefficient (ICC) can be expressed as a function of between-subject variance, between-observer variability, and the marginal mean. Penalized quasi-likelihood (PQL) can be used to estimate the ICC using the Poisson-Normal model. Confidence interval of the ICC was calculated by the Fisher's Z-transformation.

The Poisson distribution assumption can be checked by comparing the estimated variance with the estimated mean. When the assumption of a Poisson distribution is not fulfilled, some modification measures, such as allowing within-subject variation, must be taken. Simulation study shows that penalized quasi-likelihood gives correct estimators of the variance components and the concordance correlation coefficient. In some cases, there is a certain bias which may be caused by the behavior of the penalized quasi-likelihood estimator. When the residual distribution is Binomial or Poisson with a small mean, it gives biased estimates.

1.4 Concordance Correlation Coefficient

1.4.1 Lin's Concordance Correlation Coefficient

When the responses are measured on a continuous scale, Lin (1989) proposed a new reproducibility index called the concordance correlation coefficient. Assume that we have n pairs of samples (Y_{i1}, Y_{i2}), $i = 1, 2, ..., n$, which are selected randomly from a bivariate population with means μ_1 , μ_2 respectively and Covariance matrix

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

The concordance correlation coefficient is defined as

$$\rho_c = 1 - \frac{E[(Y_1 - Y_2)^2]}{E_{\text{indep}}[(Y_1 - Y_2)^2]} \tag{1.8}$$

$$= 1 - \frac{E[(Y_1 - Y_2)^2]}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2}$$

$$= \frac{2\sigma_{12}}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2}. \tag{1.9}$$

The concordance correlation coefficient can be estimated by substituting the sample moments of an independent bivariate sample into (1.8). That is,

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

$\hat{\rho}_c$ is a consistent estimator of ρ_c and has an asymptotic normal distribution with mean ρ_c and variance

$$\sigma_{\hat{\rho}_c}^2 = \frac{1}{n-2}\left[(1-\rho^2)\rho_c^2(1-\rho_c^2)/\rho^2 + 4\rho_c^3(1-\rho_c)u^2/\rho - 2\rho_c^4 u^4/\rho^2\right].$$

The inverse hyperbolic tangent transformation (or Z-transformation) can be used to improve the normal approximation:

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

This lead to improved approximate normality with mean

$$Z = \frac{1}{2}\ln\frac{1+\rho_c}{1-\rho_c}$$

and variance

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

1.4.2 Extensions of Concordance Correlation Coefficient

King and Chinchilli (2001) pointed out that when there exist outliers, the coefficient proposed by Lin (1989) may not be robust and may fail to accurately assess the agreement that may exist in the majority of the data. Instead, they proposed robust versions of the concordance correlation coefficient which allow alternative distance functions and may produce robust versions of the concordance correlation coefficient.

Let $g(\cdot)$ be the distance function which satisfies the following properties:

1. $g(0) = 0$
2. $g(z)$ is an even function, i.e., $g(-z) = g(z)$ for all z
3. $g(z)$ is a non-decreasing function of z for all $z \geq 0$

The generalized concordance correlation coefficient is defined as:

$$\rho_g = \frac{[E_{F_X F_Y} g(X-Y) - E_{F_X F_Y} g(X+Y)] - [E_{F_{XY}} g(X-Y) - E_{F_{XY}} g(X-Y)]}{E_{F_X F_Y} g(X-Y) - E_{F_X F_Y} g(X+Y) + \frac{1}{2} E_{F_X} g(2X) + \frac{1}{2} E_{F_Y} g(2Y)}, \tag{1.10}$$

where F_{XY} is the cumulative distribution function (cdf) of (X, Y); F_X and F_Y are the marginal cdf's of X and Y, respectively.

It can be shown that when choosing appropriate distance functions, the generalized concordance correlation coefficient reduces to Cohen's kappa or weighted kappa coefficient. Let $q = 1, 2, \ldots, p-1$ and $r = 2, 3, \ldots, p$ index the pairwise combinations of the p responses. Then the extended concordance correlation coefficient can be expressed as

$$\bar{\rho}_g = \frac{\sum_{\substack{qr \\ q<r}} E_{F_{X_q} F_{X_r}} [g(X_q - X_r) - g(X_q + X_r)] - \sum_{\substack{qr \\ q<r}} E_{F_{X_q X_r}} [g(X_q - X_r) - g(X_q + X_r)]}{\sum_{\substack{qr \\ q<r}} E_{F_{X_q} F_{X_r}} [g(X_q - X_r) - g(X_q + X_r)] + \frac{1}{2} \sum_{\substack{qr \\ q<r}} E_{F_{X_q X_r}} [g(2X_q) + g(2X_r)]},$$

where $q < r$ so that all pairwise comparisons are incorporated.

Fay (2005) pointed out that standard agreement coefficients (called the fixed marginal agreement coefficients, or simply FMACs), such as the concordance correlation coefficient, may indicate increased agreement as the marginal distributions of the two instruments become more different. Instead, he proposed the random marginal agreement coefficients (RMACs) to deal with this problem.

All the agreement coefficients are in the form:

$$A = 1 - \frac{E_{F_{XY}} \{c(X, Y)\}}{E_{F_U} E_{F_V} \{c(U, V)\}},$$

where $c(x, y)$ is the cost of disagreement when $X = x$ and $Y = y$.

If we let U and V be independent responses from the same distribution $F_Z = 0.5F_X + 0.5F_Y$, this will lead to the proposed RMACs:

$$A_R(c) = 1 - \frac{E_{F_{XY}}\{c(X,Y)\}}{E_{F_{Z_1}}E_{F_{Z_2}}\{c(Z_1, Z_2)\}}.$$

If we have a $k \times k$ table for categorical responses, then X and Y both represent categorical responses with k possible responses. Let e_j be a $k \times 1$ vector of zeros except with a 1 in the jth row. Let $\pi_{ab} = \Pr\{X = e_a, Y = e_b\}$. Let $c_{ij} = (e_i, e_j)$, then

$$A_R(c) = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} c_{ij}\pi_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} c_{ij}(0.5\pi_{i\cdot} + 0.5\pi_{\cdot i})(0.5\pi_{j\cdot} + 0.5\pi_{\cdot j})}.$$

Define $\Pi_o = \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}\pi_{ij}$ and $\Pi_Z = \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}(0.5\pi_{i\cdot} + 0.5\pi_{\cdot i})(0.5\pi_{j\cdot} + 0.5\pi_{\cdot j})$, then

$$A_R(c) = \frac{\Pi_o - \Pi_Z}{1 - \Pi_Z}.$$

Thus it has the same form as the weighted kappa.

Although the concordance correlation coefficient, generally speaking, performs well in measuring agreement when the responses are continuous, its agreement index is defined in the context of comparing two observers. Barnhart et.al (2002) proposed the overall concordance correlation coefficient for assessing agreement among multiple fixed observers.

Define $V = \sum_{j=1}^{J}(Y_j - Y_\bullet)^2/(J-1)$, where $Y_\bullet$ represents the arithmetic mean. The overall concordance correlation coefficient (OCCC) can then be defined as

$$\rho_o^c = 1 - \frac{E(V)}{E(V|Y_1, Y_2, \ldots, Y_J \text{ are uncorrelated})}.$$

It can be shown that the OCCC is a natural extension of the CCC. It can be interpreted as a weighted average of all pairwise CCCs, where higher weights are given to the pairs of observers whose readings have higher variances and larger mean differences.

Barnhart and Williamson (2001) proposed a generalized estimating equations (GEE) approach to model both the concordance correlation coefficient and the marginal distribution while adjusting for covariates.

Three sets of estimating equations were proposed. It turns out that when there are no covariates, the estimation process requires no iteration and the resulting estimate is equivalent to Lin's concordance correlation coefficient. When there are covariates, the parameter estimates are consistent if the three models are correctly specified, no matter whether the working correlation matrices are correctly specified or not. Confidence intervals for the resulting estimate can also be derived using Fisher's Z-transformation.

Carrasco and Jover (2003) showed that the concordance correlation coefficient is a special case of the intraclass correlation coefficient (ICC) when the observers are considered as a fixed effect. This implies that the concordance correlation coefficient, just as the ICC, can be estimated by variance components.

Estimating the CCC by estimating the variance, covariance and means of the observer (moment method) as suggested by Lin (1989) will lead to a biased estimator.

Instead, a less biased estimator is

$$\hat{\rho}_C = \frac{2\sum\limits_{i=1}^{k-1}\sum\limits_{j=i+1}^{k} S_{ij}}{(k-1)\sum\limits_{i=1}^{k-1} S_i^2 + \sum\limits_{i=1}^{k-1}\sum\limits_{j=i+1}^{k}(\bar{Y}_i - \bar{Y}_j)^2 - \frac{k(k-1)}{n}\sum\limits_{i=1}^{k-1}\sum\limits_{j=i+1}^{k}(S_i^2 + S_j^2 - 2S_{ij})}.$$

The variance component of a mixed effect model was employed with subjects being a random effect and observers being fixed effects. The inverse hyperbolic tangent transformation was used to improve asymptotic normality.

The moment estimation method provides a biased estimation of between-observers variability which produces a biased CCC, which is confirmed by simulation study. The variance component method, on the other hand, provides systematically closer estimates to the true value than the moment method.

Chapter 2

Pairwise Conditional Measures of Agreement Minus Disagreement

2.1 Definition of the Pairwise Conditional Measures

Suppose there is a bivariate response, (X, Y), where each of X and Y yields a categorical response. For convenience, the categories are denoted as $0, 1, 2, \ldots, k$. A coefficient that reflects agreement minus disagreement between categories i and j, where $i \neq j$, conditioned on only responses in categories i and j, is as follows:

$$\begin{aligned} \kappa_{ij} &= \Pr(X = Y | X, Y = i, j) - \Pr(X \neq Y | X, Y = i, j) \\ &= \frac{(p_{ii} + p_{jj}) - (p_{ij} + p_{ji})}{p_{ii} + p_{ij} + p_{ji} + p_{jj}} \\ &= 1 - \frac{2(p_{ij} + p_{ji})}{p_{ii} + p_{ij} + p_{ji} + p_{jj}}. \end{aligned} \tag{2.1}$$

For convenience, we express κ as the $(k+1) \times (k+1)$ symmetric matrix:

$$\kappa = \begin{bmatrix} 1 & \kappa_{01} & \cdots & \kappa_{0k} \\ \kappa_{01} & 1 & \cdots & \kappa_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ \kappa_{0k} & \kappa_{1k} & \ldots & 1 \end{bmatrix}.$$

We can also construct a $\frac{1}{2}k(k+1)$ vector whose elements are the unique pairwise kappa's as we just defined:

$$\text{Vech}(\kappa) = [\ \kappa_{01} \quad \kappa_{02} \quad \ldots \quad \kappa_{0k} \quad | \quad \kappa_{12} \quad \ldots \quad \kappa_{1k} \quad | \quad \ldots \quad | \quad \kappa_{k-1,k}\ ]^{\mathrm{T}}. \tag{2.2}$$

2.2 Properties of κ_{ij}

1. $\kappa_{ij} \leq 1$, for all $i, j = 0, 1, 2, \ldots, k$, where $i \neq j$.

2. Exclude the degenerate cases, $\kappa_{ij} = 1$ if and only if $p_{ij} + p_{ji} = 0$. This means that $p_{ij} = p_{ji} = 0$. Intuitively this makes sense in that this means that there is no disagreement conditioned on only responses in i and j. This can be considered as perfect conditional agreement.

3. Exclude the degenerate cases, $\kappa_{ij} = -1$ if and only if $p_{ii} = p_{jj} = 0$. Intuitively this makes sense in that this means that there is no agreement conditioned on only responses in i and j. This can be considered as complete conditional disagreement.

4. $\kappa_{ij} \geq -1$.

Proof. 1. Apparently this is true since

$$\frac{p_{ij} + p_{ji}}{p_{ii} + p_{ij} + p_{ji} + p_{jj}} \geq 0.$$

2. Obvious and no need to prove.

3. Obvious and no need to prove.

4. Apparently this is true since

$$\frac{(p_{ij} + p_{ji})}{p_{ii} + p_{ij} + p_{ji} + p_{jj}} \leq 1.$$

□

2.3 Asymptotic Distribution of the Estimators of the Pairwise Conditional Measures

κ_{ij} can be estimated in the following way:

$$\hat{\kappa}_{ij} = 1 - \frac{2(\hat{p}_{ij} + \hat{p}_{ji})}{\hat{p}_{ii} + \hat{p}_{ij} + \hat{p}_{ji} + \hat{p}_{jj}},$$

where $\hat{p}_{ij}$ denotes the observed proportions, $i, j = 0, 1, 2, \ldots, k$. When the sample size is large, the estimated probabilities should be close to their true values. And hence the estimated κ_{ij}'s should be close to their true values.

In most situations, the above estimator will work. However, when the number of categories is large, it is possible in simulation studies as well as in practical situations that $\hat{p}_{ii} = \hat{p}_{ij} = \hat{p}_{ji} = \hat{p}_{jj} = 0$. Consequently we get 0 for the denominator and thus $\hat{\kappa}_{ij}$ is undefined. One solution to this problem is to set the $\hat{\kappa}_{ij}$ to be 0. This is reasonable in that when $\hat{p}_{ii} = \hat{p}_{ij} = \hat{p}_{ji} = \hat{p}_{jj} = 0$, no information concerning agreement can be obtained through the four cell probabilities used for the calculation of κ_{ij}.

Similarly, Vech(κ) can be estimated by

$$\text{Vech}(\hat{\kappa}) = [\ \hat{\kappa}_{01} \quad \hat{\kappa}_{02} \quad \ldots \quad \hat{\kappa}_{0k} \quad | \quad \hat{\kappa}_{12} \quad \ldots \quad \hat{\kappa}_{1k} \quad | \quad \ldots \quad | \quad \hat{\kappa}_{k-1,k}\ ]^{\mathrm{T}}.$$

We want to derive the asymptotic distribution of $\hat{\kappa}_{ij}$ and $\mathrm{Vech}(\hat{\kappa})$. Suppose the sample size is n. First note that for a bivariate sample of size n,

- $\hat{p}_{ii} = \frac{1}{n}\sum\sum \mathrm{I}\{X = i, Y = i\}$
- $\hat{p}_{ij} = \frac{1}{n}\sum\sum \mathrm{I}\{X = i, Y = j\}$
- $\hat{p}_{ji} = \frac{1}{n}\sum\sum \mathrm{I}\{X = j, Y = i\}$
- $\hat{p}_{jj} = \frac{1}{n}\sum\sum \mathrm{I}\{X = j, Y = j\}$.

When n is large, we know that there is strong convergence, i.e.,

- $\hat{p}_{ii} \longrightarrow \mathrm{E}(\mathrm{I}\{X = i, Y = i\}) = p_{ii}$
- $\hat{p}_{ij} \longrightarrow \mathrm{E}(\mathrm{I}\{X = i, Y = j\}) = p_{ij}$
- $\hat{p}_{ji} \longrightarrow \mathrm{E}(\mathrm{I}\{X = j, Y = i\}) = p_{ji}$
- $\hat{p}_{jj} \longrightarrow \mathrm{E}(\mathrm{I}\{X = j, Y = j\}) = p_{jj}$.

Now let $v_1 = p_{ii}$, $v_2 = p_{ij}$, $v_3 = p_{ji}$, $v_4 = p_{jj}$ and $\hat{v}_1 = \hat{p}_{ii}$, $\hat{v}_2 = \hat{p}_{ij}$, $\hat{v}_3 = \hat{p}_{ji}$, $\hat{v}_4 = \hat{p}_{jj}$. Let $v = (v_1, v_2, v_3, v_4)$ and $\hat{v} = (\hat{v}_1, \hat{v}_2, \hat{v}_3, \hat{v}_4)$. Let $g(v)$ be a function of v such that

$$g(v) = 1 - \frac{2(v_2 + v_3)}{v_1 + v_2 + v_3 + v_4}.$$

The vector $\hat{v}$ has an asymptotic normal distribution with

$$\begin{aligned}\mathrm{E}(\hat{v}) &= (v_1, v_2, v_3, v_4) \\ &= (p_{ii}, p_{ij}, p_{ji}, p_{jj})\end{aligned}$$

and variance $n^{-1}\Sigma$, where

$$\Sigma = \{W_{ij}\}_{4\times 4}$$

$$W_{11} = v_1(1-v_1),\; W_{12} = W_{21} = -v_1v_2,\; W_{13} = W_{31} = -v_1v_3,\; W_{14} = W_{41} = -v_1v_4$$

$$W_{22} = v_2(1-v_2),\; W_{23} = W_{32} = -v_2v_3,\; W_{24} = W_{42} = -v_2v_4$$

$$W_{33} = v_3(1-v_3),\; W_{34} = W_{43} = -v_3v_4$$

$$W_{44} = v_4(1-v_4).$$

Using the theory on functions of asymptotically normal vectors, we know that $g(\hat{v})$ has asymptotically a normal distribution with mean $g(v)$ and variance $n^{-1}d\Sigma d'$, where

$$d = \left(\left.\frac{\partial g}{\partial v_1}\right|_{v=\mathrm{E}(v)}, \ldots, \left.\frac{\partial g}{\partial v_4}\right|_{v=\mathrm{E}(v)} \right).$$

The elements of d are

$$d_1 = \left.\tfrac{\partial g}{\partial v_1}\right|_{v=\mathrm{E}(v)} = \tfrac{2}{Q^2}(v_2+v_3),$$

$$d_2 = \left.\tfrac{\partial g}{\partial v_2}\right|_{v=\mathrm{E}(v)} = \tfrac{2}{Q^2}\,[-(v_1+v_4)],$$

$$d_3 = \left.\tfrac{\partial g}{\partial v_3}\right|_{v=\mathrm{E}(v)} = \tfrac{2}{Q^2}\,[-(v_1+v_4)],$$

$$d_4 = \left.\tfrac{\partial g}{\partial v_4}\right|_{v=\mathrm{E}(v)} = \tfrac{2}{Q^2}(v_2+v_3),$$

where $Q = v_1 + v_2 + v_3 + v_4$. After some algebraic calculations, it can be shown that the variance of $g(\hat{v})$ is

$$\begin{aligned} \sigma^2_{g(v)} &= n^{-1} d\Sigma d \\ &= \frac{4}{nQ^3}(v_1 + v_4)(v_2 + v_3) \\ &= \frac{4(v_1 + v_4)(v_2 + v_3)}{n(v_1 + v_2 + v_3 + v_4)^3}. \end{aligned}$$

This implies that

$$\sqrt{n}(\hat{\kappa}_{ij} - \kappa_{ij}) \xrightarrow[n\to\infty]{L} \mathrm{N}\left(0, \frac{4(p_{ii} + p_{jj})(p_{ij} + p_{ji})}{n(p_{ii} + p_{ij} + p_{ji} + p_{jj})^3}\right). \tag{2.3}$$

Now we want to find the asymptotic distribution of $\mathrm{Vech}(\hat{\kappa})$. From the above derivation, we can get the asymptotic variance for $\hat{\kappa}_{ij}$, where $i, j = 0, 1, 2, \ldots, k$, $i \neq j$. We also need to determine the asymptotic covariances. First, we derive the asymptotic covariance between $\hat{\kappa}_{ij}$ and $\hat{\kappa}_{ik}$, where $j \neq k$. Note that $\begin{pmatrix} \hat{\kappa}_{ij} \\ \hat{\kappa}_{ik} \end{pmatrix}$ can be written as

$$\begin{pmatrix} \hat{\kappa}_{ij} \\ \hat{\kappa}_{ik} \end{pmatrix} = g\left(\underline{\hat{p}}_{ijk}\right),$$

where $\underline{\hat{p}}_{ijk}$ is

$$\begin{pmatrix} \hat{p}_{ii} & \hat{p}_{ij} & \hat{p}_{ji} & \hat{p}_{jj} & \hat{p}_{ik} & \hat{p}_{ki} & \hat{p}_{kk} \end{pmatrix}^{\mathrm{T}}.$$

As before, we can let $v_1 = p_{ii}$, $v_2 = p_{ij}$, $v_3 = p_{ji}$, $v_4 = p_{jj}$, $v_5 = p_{ik}$, $v_6 = p_{ki}$, $v_7 = p_{kk}$ and $\hat{v}_1 = \hat{p}_{ii}$, $\hat{v}_2 = \hat{p}_{ij}$, $\hat{v}_3 = \hat{p}_{ji}$, $\hat{v}_4 = \hat{p}_{jj}$, $\hat{v}_5 = \hat{p}_{ik}$, $\hat{v}_6 = \hat{p}_{ki}$, $\hat{v}_7 = \hat{p}_{kk}$. We can let

$$v = \begin{pmatrix} v_1 & v_2 & v_3 & v_4 & v_5 & v_6 & v_7 \end{pmatrix}^{\mathrm{T}}.$$

Then

$$\sqrt{n}\left(\underline{\hat{p}}_{ijk} - \underline{p}_{ijk}\right) \xrightarrow[n\to\infty]{\mathrm{L}} \mathrm{N}\left(\underline{0}, \Sigma_{ijk}\right),$$

where

$\Sigma_{ijk} = \{W_{ijk}\}_{7\times 7}$,

$W_{11} = v_1(1-v_1)$, $W_{12} = W_{21} = -v_1v_2$,$W_{13} = W_{31} = -v_1v_3$, $W_{14} = W_{41} = -v_1v_4$, $W_{15} = W_{51} = -v_1v_5$, $W_{16} = W_{61} = -v_1v_6$, $W_{17} = W_{71} = -v_1v_7$,

$W_{22} = v_2(1-v_2)$, $W_{23} = W_{32} = -v_2v_3$, $W_{24} = W_{42} = -v_2v_4$, $W_{25} = W_{52} = -v_2v_5$, $W_{26} = W_{62} = -v_2v_6$, $W_{27} = W_{72} = -v_2v_7$,

$W_{33} = v_3(1-v_3)$, $W_{34} = W_{43} = -v_3v_4$, $W_{35} = W_{53} = -v_3v_5$, $W_{36} = W_{63} = -v_3v_6$, $W_{37} = W_{73} = -v_3v_7$,

$W_{44} = v_4(1-v_4)$, $W_{45} = W_{54} = -v_4v_5$, $W_{46} = W_{64} = -v_4v_6$,, $W_{47} = W_{74} = -v_4v_7$,

$W_{55} = v_5(1-v_5)$, $W_{56} = W_{65} = -v_5v_6$, $W_{57} = W_{75} = -v_5v_7$,

$W_{66} = v_6(1-v_6)$, $W_{67} = W_{76} = -v_6v_7$,

$W_{77} = v_7(1-v_7)$.

Therefore,

$$\sqrt{n}\left(g(\hat{\underline{p}}_{ijk}) - g(\underline{p}_{ijk})\right) \xrightarrow[n\to\infty]{\mathrm{L}} \mathrm{N}\left(\underline{0}, \dot{g}(\underline{p}_{ijk})\Sigma_{ijk}\dot{g}(\underline{p}_{ijk})^{\mathrm{T}}\right),$$

where $\dot{g}(\underline{p}_{ijk})$ is

$$\begin{pmatrix} \frac{2(v_2+v_3)}{Q^2} & \frac{-2(v_1+v_4)}{Q^2} & \frac{-2(v_1+v_4)}{Q^2} & \frac{2(v_2+v_3)}{Q^2} & 0 & 0 & 0 \\ \frac{2(v_5+v_6)}{R^2} & 0 & 0 & 0 & \frac{-2(v_1+v_7)}{R^2} & \frac{-2(v_1+v_7)}{R^2} & \frac{2(v_5+v_6)}{R^2} \end{pmatrix}.$$

After tedious algebraic calculations, it can be shown that $\dot{g}(\underline{p}_{ijk})\Sigma_{ijk}\dot{g}(\underline{p}_{ijk})^{\mathrm{T}}$ is

$$\begin{pmatrix} \frac{4(v_1+v_4)(v_2+v_3)}{Q^3} & \frac{4v_1(v_2+v_3)(v_5+v_6)}{Q^2R^2} \\ \frac{4v_1(v_2+v_3)(v_5+v_6)}{Q^2R^2} & \frac{4(v_1+v_7)(v_5+v_6)}{R^3} \end{pmatrix}.$$

Therefore,

$$\mathrm{Cov}(\hat{\kappa}_{ij}, \hat{\kappa}_{ik}) = \frac{4v_1(v_2+v_3)(v_5+v_6)}{nQ^2R^2} \quad \text{for } j \neq k,$$

where $Q = v_1 + v_2 + v_3 + v_4$ and $R = v_1 + v_5 + v_6 + v_7$. That is,

$$\mathrm{Cov}(\hat{\kappa}_{ij}, \hat{\kappa}_{ik}) = \frac{4p_{ii}(p_{ij}+p_{ji})(p_{ik}+p_{ki})}{n(p_{ii}+p_{ij}+p_{ji}+p_{jj})^2(p_{ii}+p_{ik}+p_{ki}+p_{kk})^2} \quad \text{for } j \neq k. \tag{2.4}$$

Similarly, we can also obtain the asymptotic covariance between $\hat{\kappa}_{is}$ and $\hat{\kappa}_{js}$, where $i \neq j$:

$$\mathrm{Cov}(\hat{\kappa}_{is}, \hat{\kappa}_{js}) = \frac{4p_{ss}(p_{is}+p_{si})(p_{js}+p_{sj})}{n(p_{ii}+p_{is}+p_{si}+p_{ss})^2(p_{ss}+p_{js}+p_{sj}+p_{jj})^2} \quad \text{for } j \neq k. \tag{2.5}$$

Now we need to find the asymptotic covariance between $\hat{\kappa}_{ik}$ and $\hat{\kappa}_{jl}$, where $i \neq j$ and $k \neq l$. As before, we can let $v_1 = p_{ii}$, $v_2 = p_{ik}$, $v_3 = p_{ki}$, $v_4 = p_{kk}$, $v_5 = p_{jj}$, $v_6 = p_{jl}$,

$v_7 = p_{lj}$, $v_8 = p_{ll}$ and $\hat{v}_1 = \hat{p}_{ii}$, $\hat{v}_2 = \hat{p}_{ik}$, $\hat{v}_3 = \hat{p}_{ki}$, $\hat{v}_4 = \hat{p}_{kk}$, $\hat{v}_5 = \hat{p}_{jj}$, $\hat{v}_6 = \hat{p}_{jl}$, $\hat{v}_7 = \hat{p}_{lj}$, $\hat{v}_8 = \hat{p}_{ll}$. We can let

$$v = \begin{pmatrix} v_1 & v_2 & v_3 & v_4 & v_5 & v_6 & v_7 & v_8 \end{pmatrix}^{\mathrm{T}}.$$

We know that

$$\sqrt{n}(\hat{v} - v) \xrightarrow[n\to\infty]{\mathrm{L}} \mathrm{N}\left(\underline{0}, \frac{1}{n}\Sigma_{ijkl}\right),$$

where

$\Sigma_{ijkl} = \{W_{ijkl}\}_{8\times 8}$,

$W_{11} = v_1(1 - v_1)$, $W_{12} = W_{21} = -v_1v_2$, $W_{13} = W_{31} = -v_1v_3$, $W_{14} = W_{41} = -v_1v_4$, $W_{15} = W_{51} = -v_1v_5$, $W_{16} = W_{61} = -v_1v_6$, $W_{17} = W_{71} = -v_1v_7$, $W_{18} = W_{81} = -v_1v_8$,

$W_{22} = v_2(1 - v_2)$, $W_{23} = W_{32} = -v_2v_3$, $W_{24} = W_{42} = -v_2v_4$, $W_{25} = W_{52} = -v_2v_5$, $W_{26} = W_{62} = -v_2v_6$, $W_{27} = W_{72} = -v_2v_7$, $W_{28} = W_{82} = -v_2v_8$,

$W_{33} = v_3(1 - v_3)$, $W_{34} = W_{43} = -v_3v_4$, $W_{35} = W_{53} = -v_3v_5$, $W_{36} = W_{63} = -v_3v_6$, $W_{37} = W_{73} = -v_3v_7$, $W_{38} = W_{83} = -v_3v_8$,

$W_{44} = v_4(1 - v_4)$, $W_{45} = W_{54} = -v_4v_5$, $W_{46} = W_{64} = -v_4v_6$,, $W_{47} = W_{74} = -v_4v_7$, $W_{48} = W_{84} = -v_4v_8$,

$W_{55} = v_5(1 - v_5)$, $W_{56} = W_{65} = -v_5v_6$, $W_{57} = W_{75} = -v_5v_7$, $W_{58} = W_{85} = -v_5v_8$,

$W_{66} = v_6(1 - v_6)$, $W_{67} = W_{76} = -v_6v_7$, $W_{68} = W_{86} = -v_6v_8$,

$W_{77} = v_7(1 - v_7)$, $W_{78} = W_{87} = -v_7 v_8$,

$W_{88} = v_8(1 - v_8)$.

Therefore,

$$\sqrt{n}\left(g(\hat{v} - g(v)\right) \xrightarrow[n\to\infty]{\text{L}} \text{N}\left(\underline{0}, \dot{g}(v)\Sigma_{ijkl}\dot{g}(v)^{\text{T}}\right),$$

where $\dot{g}(v)$ is

$$\begin{pmatrix} \frac{2(v_2+v_3)}{Q^2} & \frac{-2(v_1+v_4)}{Q^2} & \frac{-2(v_1+v_4)}{Q^2} & \frac{2(v_2+v_3)}{Q^2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{2(v_6+v_7)}{R^2} & \frac{-2(v_5+v_8)}{R^2} & \frac{-2(v_5+v_8)}{R^2} & \frac{2(v_6+v_7)}{R^2} \end{pmatrix}$$

and where $Q = v_1 + v_2 + v_3 + v_4$ and $R = v_5 + v_6 + v_7 + v_8$. Again, after long and tedious algebraic calculations, it can be shown that $\dot{g}(v)\Sigma_{ijkl}\dot{g}(v)^{\text{T}}$ is

$$\begin{pmatrix} \frac{4(v_1+v_4)(v_2+v_3)}{Q^3} & 0 \\ 0 & \frac{4(v_5+v_8)(v_6+v_7)}{R^3} \end{pmatrix}.$$

So the asymptotic variance between $\hat{\kappa}_{ik}$ and $\hat{\kappa}_{jl}$, where $i \neq j$ and $k \neq l$, is 0. And now we can get the joint asymptotic distribution of Vech($\hat{\kappa}$). Since Vech($\hat{\kappa}$) can also be considered as a function of $\hat{p}_{ij}$'s, for $i, j = 0, 1, 2, \ldots, k$, its joint distribution is multivariate normal:

$$\sqrt{n}\left[\begin{pmatrix} \hat{\kappa}_{00} \\ \hat{\kappa}_{01} \\ \vdots \\ \hat{\kappa}_{k-1,k} \end{pmatrix} - \begin{pmatrix} \kappa_{00} \\ \kappa_{01} \\ \vdots \\ \kappa_{k-1,k} \end{pmatrix}\right] \underset{n\to\infty}{\longrightarrow} \mathrm{N}\left(\underline{0}, \Sigma_{\mathrm{v}}\right), \tag{2.6}$$

where Σ_{v} is a $\frac{k(k+1)}{2} \times \frac{k(k+1)}{2}$ matrix and can be obtained using the previously derived variances and covariances.

2.4 An Overall Measure of Agreement Based on the Pairwise Conditional Measures

2.4.1 Definition

κ_{ij}, where $i, j = 0, 1, \ldots, k$, reflects the pairwise agreement minus disagreement conditioned on only responses in categories i and j. We would like to define an overall agreement minus disagreement coefficient which utilizes the information contained in the pairwise agreements. The overall measure can be constructed as the linear combination:

$$\kappa_w = w^{\mathrm{T}}\mathrm{Vech}(\kappa), \tag{2.7}$$

where

$$w = [\begin{array}{cccc|ccc|c|c} w_{01} & w_{02} & \ldots & w_{0k} & w_{12} & \ldots & w_{1k} & \ldots & w_{k-1,k} \end{array}]^{\mathrm{T}}$$

is a $\frac{1}{2}k(k+1)$ weight vector with each $w_{ij} \geq 0$ and $\sum w_{ij} = 1$.

2.4.2 Properties of κ_w

1. $-1 \leq \kappa_w \leq 1$.

2. $\kappa_w = 1$ if and only if $\kappa_{ij} = 1$ for all $i, j = 0, 1, 2, \ldots, k$, where $i \neq j$. This means that we have complete agreement for each κ_{ij}. As we have already shown, this means that that is no disagreement for each κ_{ij}. Consequently, there is no observation for all off-diagonal cells, which implies that we get perfect agreement.

3. $\kappa_w = -1$ if and only if $\kappa_{ij} = -1$ for all $i, j = 0, 1, 2, \ldots, k$, where $i \neq j$. This means that we have complete disagreement for each κ_{ij}. As we have already shown, this means that that is no agreement for each κ_{ij}. Consequently, there is no observation for all diagonal cells, which implies that we get perfect disagreement.

4. $\kappa_{min} \leq \kappa_w \leq \kappa_{max}$, where $\kappa_{min} = \min(\kappa_{01}, \kappa_{02}, \ldots, \kappa_{k-1,k})$ and $\kappa_{max} = \max(\kappa_{01}, \kappa_{02}, \ldots, \kappa_{k-1,k})$. This means that the final agreement coefficient is between the minimum pairwise conditional agreement and the maximum pairwise conditional agreement.

Proof. 1. We know that

$$\begin{aligned} \kappa_w &= w^{\mathrm{T}} \mathrm{Vech}(\kappa) \\ &= \sum_{i<j} w_{ij} \kappa_{ij}. \end{aligned} \tag{2.8}$$

We have already shown that $-1 \leq \kappa_{ij} \leq 1$. Therefore, from (2.8), we know that

$$-1 = -\sum_{i<j} w_{ij} \leq \kappa_w = \sum_{i<j} w_{ij} \kappa_{ij} \leq \sum_{i<j} w_{ij} = 1.$$

That is,

$$-1 \leq \kappa_w \leq 1.$$

2. Obvious and no need to prove.

3. Obvious and no need to prove.

4. This is because

$$\kappa_w = \sum_{i<j} w_{ij}\kappa_{ij} \leq \sum_{i<j} w_{ij}\kappa_{max} = \kappa_{max}\sum_{i<j} w_{ij} = \kappa_{max}.$$

Similarly,

$$\kappa_w = \sum_{i<j} w_{ij}\kappa_{ij} \geq \sum_{i<j} w_{ij}\kappa_{min} = \kappa_{min}\sum_{i<j} w_{ij} = \kappa_{min}.$$

Therefore, $\kappa_{min} \leq \kappa_w \leq \kappa_{max}$.

□

2.4.3 Asymptotic Distribution of $\hat{\kappa}_w$

In equation(2.6), we derived the joint asymptotic distribution for pairwise κ's. It remains straightforward to derive the asymptotic distribution for κ_w:

$$\kappa_w = w^{\mathrm{T}}\mathrm{Vech}(\kappa),$$

where

$$w = [\ w_{01} \quad w_{02} \quad \ldots \quad w_{0k} \quad | \quad w_{12} \quad \ldots \quad w_{1k} \quad | \quad \ldots \quad | \quad w_{k-1,k}\ ]^{\mathrm{T}} \tag{2.9}$$

is a $\frac{1}{2}k(k+1)$ weight vector with each $w_{ij} \geq 0$ and $\sum w_{ij} = 1$. If w does not depend on the data, then we know that

$$\sqrt{n}\left[w^{\mathrm{T}}\begin{pmatrix}\hat{\kappa}_{00}\\ \hat{\kappa}_{01}\\ \vdots\\ \hat{\kappa}_{k-1,k}\end{pmatrix} - w^{\mathrm{T}}\begin{pmatrix}\kappa_{00}\\ \kappa_{01}\\ \vdots\\ \kappa_{k-1,k}\end{pmatrix}\right] \xrightarrow[n\to\infty]{\mathrm{L}} \mathrm{N}\left(\underline{0}, w^{\mathrm{T}}\Sigma_{\mathrm{v}}w\right). \tag{2.10}$$

That is,

$$\sqrt{n}\left(\hat{\kappa}_w - \kappa_w\right) \xrightarrow[n\to\infty]{\mathrm{L}} \mathrm{N}\left(\underline{0}, w^{\mathrm{T}}\Sigma_{\mathrm{v}}w\right), \tag{2.11}$$

where Σ_{v} is the same as defined before.

2.4.4 Choices of Weight for Nominal Responses

If X and Y are nominal variables, then a reasonable choice for w is to set each w_{ij} to be $\left[\frac{1}{2}k(k+1)\right]^{-1}$, where $i, j = 0, 1, \ldots, k$. This means that we assign equal weight to each pairwise measure of agreement. This makes sense when both X and Y are nominal.

However, in some cases this choice of weight may cause trouble. Suppose we get a high agreement κ_{rs} conditioned on responses in r and s. If $p_{rr} + p_{rs} + p_{sr} + p_{ss}$ is very small, which means that the two raters are unlikely to give such ratings, then we may exaggerate the overall agreement if we give equal weight to κ_{rs}. In other words, high pairwise agreement with small rating probabilities may lead to the overestimation of κ_w. Similarly, a small pairwise agreement with small probabilities may lead to the underestimation of κ_w.

To account for the probabilities associated with the calculation of the pairwise conditional measure of agreement, we can use another weight for nominal responses which we call equal weight adjusted for probabilities.

- In calculating κ_{01}, we need p_{00}, p_{01}, p_{10}, p_{11}. So the sum of the four cell probabilities for the calculation of κ_{01} is $P_{01} = p_{00} + p_{01} + p_{10} + p_{11}$.

- The sum of the four cell probabilities for the calculation of κ_{02} is $P_{02} = p_{00} + p_{02} + p_{20} + p_{02}$.

  ......

- The sum of the four cell probabilities for the calculation of $\kappa_{k-1,k}$ is $P_{k-1,k} = p_{k-1,k-1} + p_{k-1,k} + p_{k,k-1} + p_{k,k}$.

Therefore,

$$
\begin{aligned}
P &= P_{01} + P_{02} + \ldots + P_{k-1,k} \\
&= [p_{00} + p_{01} + p_{10} + p_{11}] + [p_{00} + p_{02} + p_{20} + p_{02}] \\
&\quad + \ldots + \left[p_{k-1,k-1} + p_{k-1,k} + p_{k,k-1} + p_{k,k}\right] \\
&= 1 + (k-1)\left(p_{11} + p_{22} + \ldots + p_{k,k}\right) \\
&= 1 + (k-1)\sum_{i=0}^{k} p_{ii}.
\end{aligned}
$$

Thereafter, we can set the weight w_{rs} for κ_{rs}, where

$$
w_{rs} = \frac{p_{rr} + p_{rs} + p_{sr} + p_{ss}}{1 + (k-1)\sum_{i=0}^{k} p_{ii}}.
$$

By using the equal weight adjusted for probabilities, we can avoid the disadvantage of using equal weight, i.e. overestimation (high agreement for small cell probabilities) or underestimation (poor agreement for small cell probabilities). Another advantage of this type of weighting scheme occurs when $p_{rr} + p_{rs} + p_{sr} + p_{ss} = 0$. In such a situation, κ_{rs} is undefined but receives a weight of zero. The disadvantage of this weight scheme is that

$$\begin{aligned}
\kappa_w &= \sum_{i<j}\sum w_{ij}\kappa_{ij} \\
&= \sum_{i<j}\sum \frac{(p_{ii}+p_{jj})-(p_{ij}+p_{ji})}{p_{ii}+p_{ij}+p_{ji}+p_{jj}} \times \frac{p_{ii}+p_{ij}+p_{ji}+p_{jj}}{1+(k-1)\sum_{i=0}^{k} p_{ii}} \\
&= \sum_{i<j}\sum \frac{(p_{ii}+p_{jj})-(p_{ij}+p_{ji})}{1+(k-1)\sum\limits_{i=0}^{k} p_{ii}} \\
&= \frac{k\sum\limits_{i=0}^{k} p_{ii} - \sum\limits_{i\neq j} p_{ii}}{1+(k-1)\sum\limits_{i=0}^{k} p_{ii}} \\
&= \frac{k\sum\limits_{i=0}^{k} p_{ii} - (1-\sum\limits_{i=0}^{k} p_{ii})}{1+(k-1)\sum\limits_{i=0}^{k} p_{ii}} \\
&= \frac{(k+1)\sum\limits_{i=0}^{k} p_{ii} - 1}{1+(k-1)\sum\limits_{i=0}^{k} p_{ii}}. \qquad (2.12)
\end{aligned}$$

In other words, it results in a statistic that ignores the pairwise conditional measures.

There certainly exist some other weighting schemes:

$$w_{rs} = \frac{max\left(p_{rr}, p_{rs}, p_{sr}, p_{ss}\right)}{\sum\sum_{i<j} max\left(p_{ii}, p_{ij}, p_{ji}, p_{jj}\right)}$$

or

$$\begin{aligned} w_{rs} &= \frac{\sqrt{\frac{1}{4}\left(p_{rr}^2 + p_{rs}^2 + p_{sr}^2 + p_{ss}^2\right)}}{\sum\sum_{i<j}\sqrt{\frac{1}{4}\left(p_{ii}^2 + p_{ij}^2 + p_{ji}^2 + p_{jj}^2\right)}} \\ &= \frac{\sqrt{\left(p_{rr}^2 + p_{rs}^2 + p_{sr}^2 + p_{ss}^2\right)}}{\sum\sum_{i<j}\sqrt{\left(p_{ii}^2 + p_{ij}^2 + p_{ji}^2 + p_{jj}^2\right)}}. \end{aligned}$$

These two schemes will maintain the advantage of a zero weight when κ_{rs} is undefined, but do not have the disadvantage noted above. Also, because these weights must be estimated by plugging in the estimation of the probabilities, which are the observed proportions, we have to find the asymptotic distribution of $\hat{\kappa}_w$.

To derive the asymptotic distribution for the first weighting scheme, first note that

- $\hat{p}_{00} \xrightarrow{P} \mathrm{E}(\mathrm{I}\{X = 0, Y = 0\}) = p_{00}$
- $\hat{p}_{01} \xrightarrow{P} \mathrm{E}(\mathrm{I}\{X = 0, Y = 1\}) = p_{01}$
- $\hat{p}_{02} \xrightarrow{P} \mathrm{E}(\mathrm{I}\{X = 0, Y = 2\}) = p_{02}$

  $\vdots$

- $\hat{p}_{kk} \xrightarrow{P} \mathrm{E}(\mathrm{I}\{X = k, Y = k\}) = p_{kk}.$

Then we can get

$$\begin{pmatrix} \hat{p}_{00} \\ \hat{p}_{01} \\ \vdots \\ \hat{p}_{k,k} \end{pmatrix} \xrightarrow[n\to\infty]{\text{P}} \begin{pmatrix} p_{00} \\ p_{01} \\ \vdots \\ p_{k,k} \end{pmatrix}.$$

Let $g : \mathbb{R}^{(k+1)^2\times 1} \mapsto \mathbb{R}^{\frac{1}{2}k(k+1)\times 1}$ be a continuous function such that

$$g\begin{pmatrix} p_{00} \\ p_{01} \\ \vdots \\ p_{k,k} \end{pmatrix} = \begin{pmatrix} \frac{max(p_{00},p_{01},p_{10},p_{11})}{\sum\sum_{i<j} max\left(p_{ii},p_{ij},p_{ji},p_{jj}\right)} \\ \frac{max(p_{00},p_{02},p_{20},p_{22})}{\sum\sum_{i<j} max\left(p_{ii},p_{ij},p_{ji},p_{jj}\right)} \\ \vdots \\ \frac{max\left(p_{k-1,k-1},p_{k-1,k},p_{k,k-1},p_{k,k}\right)}{\sum\sum_{i<j} max\left(p_{ii},p_{ij},p_{ji},p_{jj}\right)} \end{pmatrix}.$$

Then we know that

$$g\begin{pmatrix} \hat{p}_{00} \\ \hat{p}_{01} \\ \vdots \\ \hat{p}_{k,k} \end{pmatrix} \xrightarrow[n\to\infty]{\text{P}} g\begin{pmatrix} p_{00} \\ p_{01} \\ \vdots \\ p_{k,k} \end{pmatrix}.$$

Let $\hat{W}$ and W be a $\frac{1}{2}k(k+1)$ vector such that

$$\hat{W} = g \begin{pmatrix} \hat{p}_{00} \\ \hat{p}_{01} \\ \vdots \\ \hat{p}_{k,k} \end{pmatrix}$$

and

$$W = g \begin{pmatrix} p_{00} \\ p_{01} \\ \vdots \\ p_{k,k} \end{pmatrix}.$$

That is,

$$\hat{W} \xrightarrow[n \to \infty]{\mathrm{P}} W.$$

For a given distribution, W is a constant vector. From 2.6, we know that the asymptotic distribution of $\mathrm{Vech}(\hat{\kappa})$ is

$$\sqrt{n} \left[\begin{pmatrix} \hat{\kappa}_{00} \\ \hat{\kappa}_{01} \\ \vdots \\ \hat{\kappa}_{k-1,k} \end{pmatrix} - \begin{pmatrix} \kappa_{00} \\ \kappa_{01} \\ \vdots \\ \kappa_{k-1,k} \end{pmatrix} \right] \xrightarrow[n \to \infty]{\mathrm{L}} \mathrm{N}\left(\underline{0}, \Sigma_{\mathrm{v}}\right).$$

That is,

$$\sqrt{n}\left[\mathrm{Vech}(\hat{\kappa}) - \mathrm{Vech}(\kappa)\right] \xrightarrow[n\to\infty]{\mathrm{L}} \mathrm{N}\left(\underline{0}, \Sigma_{\mathrm{v}}\right).$$

Let $f : \mathbb{R}^{k(k+1)} \mapsto \mathbb{R}^1$ be a continuous function such that

$$f\begin{pmatrix} \mathrm{Vech}(\kappa) \\ W^{\mathrm{T}} \end{pmatrix} = W^{\mathrm{T}}\mathrm{Vech}(\kappa).$$

Then we can get

$$\sqrt{n}\left(\hat{W}^{\mathrm{T}}\mathrm{Vech}(\hat{\kappa}) - W^{\mathrm{T}}\mathrm{Vech}(\kappa)\right) \xrightarrow[n\to\infty]{\mathrm{L}} \mathrm{N}\left(\underline{0}, W^{\mathrm{T}}\Sigma_{\mathrm{v}}W\right). \tag{2.13}$$

That is,

$$\sqrt{n}\left(\hat{\kappa}_w - \kappa_w\right) \xrightarrow[n\to\infty]{\mathrm{L}} \mathrm{N}\left(\underline{0}, W^{\mathrm{T}}\Sigma_{\mathrm{v}}W\right). \tag{2.14}$$

This asymptotic distribution also applies when we use other weighting schemes. We can see that this is exactly the same form as the previous asymptotic distribution we have obtained earlier. So in general, we do not need to use different asymptotic distributions for different weighting schemes.

However, there is also a disadvantage in using the max weighting scheme and the square weighting scheme. Suppose we have a joint distribution where all off-diagonal elements are small while some of the diagonal elements have relatively large probabilities. Hence we will have at least several pairwise conditional measures of high values. However, at the same time, we will also allocate high weights for all of these high pairwise conditional measures since all the off-diagonal elements are small. This will obviously overestimate the agreement level. In other words, we give high weights to high pairwise conditional measures too many times because of some high agreement probabilities.

2.4.5 Choices of Weight for Ordinal Responses

If X and Y are ordinal variables, then w should be chosen such that $w_{ij} \geq w_{i,j-1}$ and $w_{ij} \geq w_{i-1,j}$ for $0 \leq i < j \leq k$. The rationale for this is that the further apart the indices i and j, the more weight should be given to each κ_{ij} because the pairwise agreement (or disagreement) should be stronger. A simple way is to set the weight proportional to the difference between i and j. To meet the requirement that $\sum w_{ij} = 1$, we can construct this simple weight in the following manner:

For $i = 0, j = 1, 2, \ldots, k$, $|i - j| = 1, 2, \ldots, k$. And we know that $\sum |i - j| = \frac{k(k+1)}{2}$.

For $i = 1, j = 2, 3, \ldots, k$, $|i-j| = 1, 2, \ldots, (k-1)$. And we know that $\sum |i-j| = \frac{k(k-1)}{2}$.

For $i = 2, j = 3, \ldots, k$, $|i-j| = 1, 2, \ldots, (k-2)$. And we know that $\sum |i-j| = \frac{(k-1)(k-2)}{2}$.

$\ldots$

For $i = (k-1), j = k$, $|i - j| = 1$.

And it can be shown that

$$\frac{k(k+1)}{2} + \frac{k(k-1)}{2} + \frac{(k-1)(k-2)}{2} + \cdots + 1 = \frac{k(k+1)(2k+4)}{12}.$$

Therefore, we can set the weight for κ_{ij} to be

$$w_{ij} = \frac{|i-j|}{\frac{k(k+1)(2k+4)}{12}} = \frac{12|i-j|}{k(k+1)(2k+4)}.$$

For example, suppose $k = 2$, that is, we have 3 categories. Then

$$w_{ij} = \frac{|i-j|}{4}.$$

Therefore, the weight for κ_{01} is $\frac{1}{4}$, the weight for κ_{02} is $\frac{2}{4}$ and the weight for κ_{12} is $\frac{1}{4}$. We can see that the sum of weights is equal to 1 and the weights are proportional to the differences between index i and j.

We know that Cohen's weighted kappa does not change if we multiply all the weights by a certain constant. Therefore, in the calculation of Cohen's weighted kappa we can use similar weights that are proportional to the differences between index i and j. That is, let $w_{ij} = |i-j|$.

In general, we can set the disagreement weight for w_{ij} using a general distance function $g(|i-j|)$, where $g(\cdot)$ is a positive-valued distance function.

For $i = 0, j = 1, 2, \ldots, k$, we have $g(1), g(2), \ldots, g(k)$.

For $i = 1, j = 2, 3, \ldots, k$, we have $g(1), g(2), \ldots, g(k-1)$.

For $i = 2, j = 3, \ldots, k$,we have $g(1), g(2), \ldots, g(k-2)$.

$\ldots$

For $i = (k-1), j = k$, we have $g(1)$.

And it can be shown that

$$
\begin{aligned}
& g(1)+g(2)+\ldots+g(k)+g(1)+g(2)\ldots+g(k-1)+\cdots+g(1) \\
= \;& kg(1)+(k-1)g(2)+\ldots+g(k) \\
= \;& \sum_{t=1}^{k}(k-t+1)\,g(t).
\end{aligned}
$$

Therefore, the weight for w_{ij} is

$$
w_{ij}=\frac{g(|i-j|)}{\sum\limits_{t=1}^{k}(k-t+1)\,g(t)}.
$$

For example, suppose we use the squared distance function, that is, let $g(|i-j|)=(i-j)^2$, for $i,j=0,1,2\ldots,k$. Then

$$
\begin{aligned}
& \sum_{t=1}^{k}(k-t+1)\,g(t) \\
= \;& \sum_{t=1}^{k}(k-t+1)\,t^2 \\
= \;& \frac{k(k+1)^2(k+2)}{12}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
w_{ij} &= \frac{|i-j|^2}{\frac{k(k+1)^2(k+2)}{12}} \\
&= \frac{12|i-j|^2}{k(k+1)^2(k+2)}.
\end{aligned}
$$

For example, suppose $k = 2$, that is, we have 3 categories. Then

$$w_{ij} = \frac{(i-j)^2}{6}.$$

Therefore, the weight for κ_{01} is $\frac{1}{6}$, the weight for κ_{02} is $\frac{4}{6}$ and the weight for κ_{12} is $\frac{1}{6}$. We can see that the sum of weights is equal to 1.

We know that Cohen's weighted kappa does not change if we multiply all the weights by a certain constant. Therefore, in the calculation of Cohen's weighted kappa we can use similar weights. That is, let $w_{ij} = (i-j)^2$.

On the other hand, suppose we want to use exponential distance function. That is, let $g(|i-j|) = \exp(|i-j|) - 1$, for $i, j = 0, 1, 2 \ldots, k$. Then

$$\begin{aligned} & \sum_{t=1}^{k} (k-t+1)\, g(t) \\ = & \sum_{t=1}^{k} (k-t+1)\, (\exp(t) - 1) \\ = & \frac{e}{1-e}\left[\frac{e(e^k-1)}{1-e} + k\right] - \frac{k(k+1)}{2}. \end{aligned}$$

Therefore,

$$w_{ij} = \frac{e^{|i-j|} - 1}{\frac{e}{1-e}\left[\frac{e(e^k-1)}{1-e} + k\right] - \frac{k(k+1)}{2}}.$$

For example, suppose $k = 2$, that is, we have 3 categories. Then

$$\frac{e}{1-e}\left[\frac{e(e^k-1)}{1-e} + k\right] - \frac{k(k+1)}{2} = e^2 + 2e - 3.$$

Hence,

$$w_{ij} = \frac{e^{|i-j|} - 1}{e^2 + 2e - 3}.$$

Therefore, the weight for κ_{01} is $\frac{e-1}{e^2+2e-3}$, the weight for κ_{02} is $\frac{e^2-1}{e^2+2e-3}$ and the weight for κ_{12} is $\frac{e-1}{e^2+2e-3}$. We can see that the sum of weights is equal to 1.

We know that Cohen's weighted kappa does not change if we multiply all the weights by a certain constant. Therefore, in the calculation of Cohen's weighted kappa we can use similar weights. That is, let $w_{ij} = e^{|i-j|} - 1$.

Now suppose $k = 3$, that is , we have 4 categories. Then

$$\frac{e}{1-e}\left[\frac{e(e^k-1)}{1-e} + k\right] - \frac{k(k+1)}{2} = e^3 + 2e^2 + 3e - 6.$$

Hence,

$$w_{ij} = \frac{e^{|i-j|} - 1}{e^3 + 2e^2 + 3e - 6}.$$

Therefore, the weight for κ_{01} is $\frac{e-1}{e^3+2e^2+3e-6}$, the weight for κ_{02} is $\frac{e^2-1}{e^3+2e^2+3e-6}$, the weight for κ_{03} is $\frac{e^3-1}{e^3+2e^2+3e-6}$, the weight for κ_{12} is $\frac{e-1}{e^3+2e^2+3e-6}$, the weight for κ_{13} is $\frac{e^2-1}{e^3+2e^2+3e-6}$ and the weight for κ_{23} is $\frac{e-1}{e^3+2e^2+3e-6}$. We can see that the sum of weights is equal to 1. And Cohen's kappa remains the same as the previous case.

The weight assignment for κ_w, the proposed weighted sum of pairwise agreement measures, is different from that for Cohen's weighted kappa in that here we require that the sum of individual weights must be 1 while for Cohen's weighted kappa, the weight can be any positive numbers since multiplication does not change its value.

2.4.6 Advantages and Disadvantages of κ_w

As can be seen from the definition of κ_w, it utilizes all conditional pairwise information concerning agreements which both Cohen's kappa and weighted kappa do not include. Furthermore, it is easy to assign different weights to individual κ_{ij}'s to represent the researcher's perception of the different levels of gravity of individual pairwise agreements. We hope that the κ_w defined in this way can have good performance, especially when there are more than two categories. Performances of the proposed κ_w will be investigated through simulation studies, where Cohen's kappa and weighted kappa are also presented to compare their performances with the proposed method.

Another advantage is to avoid the so-called kappa's paradox. Kappa will be higher with an asymmetrical rather than symmetrical imbalance in marginal probabilities. Let's first consider one joint distribution:

		y = 0	1	2	$P_X(x)$
x	0	0.1	0.2	0.05	0.35
	1	0.2	0.1	0.1	0.4
	2	0.05	0.1	0.1	0.25
	$P_Y(y)$	0.35	0.4	0.25	1

In this example, the observed agreement, p_o, is 0.3. We can also see that the marginal probabilities are symmetric. Cohen's kappa gives -0.0687. The proposed method using equal weights gives 0. If we use max weights and square weights, we get -0.0833 and -0.0781, respectively, which indicates that there is extremely poor agreement.

Let us then consider another joint distribution:

		y			
		0	1	2	$P_X(x)$
	0	0.1	0.3	0.05	0.45
x	1	0.05	0.1	0	0.15
	2	0	0.3	0.1	0.4
	$P_Y(y)$	0.15	0.7	0.15	1

In this example, the observed agreement, p_o, is also 0.3. But there is asymmetrical imbalance of the marginal probabilities. Cohen's kappa gives 0.0879, which is higher than the previous kappa we got. This does not make sense since in this example, we should have poorer agreement. The proposed method using equal weights gives 0.0424. However, if we use max weights and square weights, we can get -0.1169 and -0.0830, respectively, which indicates that there is extremely poor agreement. This is in conformity with our distribution setting.

One disadvantage of the pairwise agreement minus disagreement is that some times it overestimates the overall level of agreement. When some diagonal cells have relatively large agreement probabilities, this will result in a series of large pairwise agreements, leading to an overall larger agreement coefficient. Another disadvantage is that the large sample formula is relatively too complicated. For example, if we have six categories, then Σ_v is a 15×15 matrix, making computations very difficult. In such cases, we may consider using the bootstrap to estimate the variances.

2.5 κ_w, Corrected for Chance?

We know that both Cohen's κ and weighted κ are chance corrected, which means that they are measuring agreement with chance agreement excluded. We had tried to construct a weighted sum of kappa statistics that is also corrected for chance. Originally we want to construct the following measure:

$$\begin{aligned} k_{ij} &= \frac{\Pr(X = Y|X, Y = i, j) - \Pr_{\text{indep}}(X = Y|X, Y = i, j)}{1 - \Pr_{\text{indep}}(X = Y|X, Y = i, j)} \\ &= \frac{\frac{p_{ii}+p_{jj}}{p_{ii}+p_{ij}+p_{jj}+p_{ji}} - \frac{P_X(i)P_Y(i)+P_X(j)P_Y(j)}{P_X(i)P_Y(i)+P_X(j)P_Y(j)+P_X(i)P_Y(j)+P_X(j)P_Y(i)}}{1 - \frac{P_X(i)P_Y(i)+P_X(j)P_Y(j)}{P_X(i)P_Y(i)+P_X(j)P_Y(j)+P_X(i)P_Y(j)+P_X(j)P_Y(i)}}, \end{aligned}$$

where $\Pr_{\text{indep}}$ is the probability when assuming that X and Y are independent. However, we encountered problems in the lower bound of κ_{ij}. The following process explains this situation:

To show that $\kappa_{ij} \geq -1$, we need to show that

$$\frac{2[P_X(i)P_Y(i) + P_X(j)P_Y(j)]}{P_X(i)P_Y(i) + P_X(j)P_Y(j) + P_X(i)P_Y(j) + P_X(j)P_Y(i)} \leq 1 + \frac{p_{ii} + p_{jj}}{p_{ii} + p_{ij} + p_{jj} + p_{ji}}.$$

1. If

$$\frac{P_X(i)P_Y(i) + P_X(j)P_Y(j)}{P_X(i)P_Y(i) + P_X(j)P_Y(j) + P_X(i)P_Y(j) + P_X(j)P_Y(i)} \leq \frac{1}{2},$$

then it is obvious that the inequality holds.

2. If, however,

$$\frac{P_X(i)P_Y(i) + P_X(j)P_Y(j)}{P_X(i)P_Y(i) + P_X(j)P_Y(j) + P_X(i)P_Y(j) + P_X(j)P_Y(i)} > \frac{1}{2}.$$

This implies that

$$2[P_X(i)P_Y(i) + P_X(j)P_Y(j)] > P_X(i)P_Y(i) + P_X(j)P_Y(j) + P_X(i)P_Y(j) + P_X(j)P_Y(i)$$

$$\Rightarrow \quad P_X(i)P_Y(i) + P_X(j)P_Y(j) > P_X(i)P_Y(j) + P_X(j)P_Y(i)$$

$$\Rightarrow \quad [P_X(i) - P_X(j)][P_Y(i) - P_Y(j)] > 0$$

$$\Rightarrow \quad P_X(i) > P_X(j) \text{ AND } P_Y(i) > P_Y(j) \text{ or } P_X(i) < P_X(j) \text{ AND } P_Y(i) < P_Y(j).$$

Now if $P_X(i) > P_X(j)$ AND $P_Y(i) > P_Y(j)$, then we can set $P_X(j) = m$, $P_Y(j) = n$, $P_X(i) = m + a$ and $P_Y(i) = n + b$, where $m \geq 0, n \geq 0, a > 0, b > 0$ and $m + a \leq 1, n + b \leq 1$. So we want to show that

$$\frac{2[(m+a)(n+b) + mn]}{(m+a)(n+b) + mn + (m+a)n + m(n+b)} < 1 + \frac{p_{ii} + p_{jj}}{p_{ii} + p_{ij} + p_{jj} + p_{ji}}$$

$$\Rightarrow \quad \frac{4mn + 2mb + 2an + 2ab}{4mn + 2mb + 2an + ab} < 1 + \frac{p_{ii} + p_{jj}}{p_{ii} + p_{ij} + p_{jj} + p_{ji}}$$

$$\Rightarrow \quad \frac{ab}{4mn + 2mb + 2an + ab} < \frac{p_{ii} + p_{jj}}{p_{ii} + p_{ij} + p_{jj} + p_{ji}}.$$

Apparently the inequality does not hold in general. One counterexample is when $p_{ii} + p_{jj} = 0$. So in general, we can not conclude that the lower bound of κ_{ij} is -1. Consider the following example:

		y = 0	1	2	$P_X(x)$
	0	0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{3}{8}$
x	1	$\frac{1}{8}$	0	0	$\frac{1}{8}$
	2	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{4}{8}$
	$P_Y(y)$	$\frac{3}{8}$	$\frac{2}{8}$	$\frac{3}{8}$	1

We get that $\kappa_{01} = -\frac{11}{9}$.

The uncertainty of the lower boundary of κ_{ij} is because the calculation of the conditional probability under the independence assumption involves probabilities that are not involved with the conditional probability under the no independence assumption.

We then tried to define the conditioning event more explicitly in the hope that it would yield a different result than we had previously. Let Z be the event that $(X, Y) = (i, i), (i, j), (j, i)$ or (j, j). Define κ_{ij} as

$$\kappa_{ij} = \frac{\Pr(X = Y|Z) - \Pr_{\text{indep}}(X = Y|Z)}{\Pr(Z) - \Pr_{\text{indep}}(X = Y|Z)}.$$

However, it turns out that this definition is exactly the same as the previous one. As long as we are dealing with the marginal probabilities in constructing the pairwise kappa, we may encounter the problem of the uncertainty of the lower bound. Therefore, we have to abandon this approach for a weighted sum of agreement measures.

2.6 Simulation for Nominal Responses

2.6.1 Simulation Design

We used Matlab to generate samples following multimoninal distributions. For each distribution, we calculated Cohen's simple kappa, the pairwise agreement minus disagreement using equal weights, weights corresponding to individual probabilities, maximum cell probabilities, and the square root weights as we previously introduced. Since in practice, we often do not know the actual joint distribution of X and Y, we can use the sample as the estimate of the joint distribution. Their variances are also given as well as the variances calculated using the sample counterparts calculated from the asymptotic distribution formula. To avoid the situation when there are no observations for cells with small probabilities, we use a relatively large sample, say, $n = 200$, 400, and 800. And each simulation is run 1000 times. As aforementioned, in practice, agreement measures greater than 0.75 or so may be taken to represent excellent agreement beyond chance, values between 0.40 and 0.75 may be taken to represent fair to good agreement beyond chance and values below 0.40 or so may be taken to represent poor agreement.

Case 1 The distribution is

		y			
		0	1	2	$P_X(x)$
	0	0.2	0.05	0.05	0.3
x	1	0.03	0.3	0.07	0.4
	2	0.05	0.07	0.18	0.3
	$P_Y(y)$	0.28	0.42	0.3	1

This case represents medium agreement. We can see that the diagonal probabilities add up to 0.68 and the marginal probabilities of X and Y are very close to each other.

Case 2 The distribution is

		y			
		0	1	2	$P_X(x)$
	0	0.05	0.1	0.05	0.2
x	1	0.22	0.05	0.03	0.3
	2	0.13	0.35	0.02	0.5
	$P_Y(y)$	0.4	0.5	0.1	1

This case represents poor agreement. It can be seen that there is more disagreement between marginal probabilities as well as high probabilities for off-diagonal probabilities.

Case 3 The distribution is

		y			
		0	1	2	$P_X(x)$
	0	0.21	0.01	0.02	0.24
x	1	0.02	0.4	0.02	0.44
	2	0.01	0.01	0.3	0.32
	$P_Y(y)$	0.24	0.42	0.34	1

This case represents high agreement. The diagonal probabilities add up to 0.91 and the marginal probabilities are very close to each other.

Case 4 The distribution is

		y			
		0	1	2	$P_X(x)$
	0	0.2	0	0	0.2
x	1	0	0.3	0	0.3
	2	0	0	0.5	0.5
	$P_Y(y)$	0.2	0.3	0.5	1

This case represents complete agreement. There is no disagreement.

Case 5 The distribution is

		y			
		0	1	2	$P_X(x)$
	0	0.125	0.125	0	0.25
x	1	0	0.25	0	0.25
	2	0	0.25	0.25	0.5
	$P_Y(y)$	0.125	0.625	0.25	1

Again this case represent medium agreement.

Case 6 The distribution is

		y				
		0	1	2	3	$P_X(x)$
	0	0.0455	0.1136	0.1364	0.0182	0.3137
x	1	0.1364	0.2273	0.0818	0.0136	0.4591
	2	0.0455	0.0545	0	0.0227	0.1227
	3	0.0045	0.0227	0.0318	0.0455	0.1045
	$P_Y(y)$	0.2319	0.4181	0.25	0.1	1

This case represents poor agreement. The diagonal probabilities add up to 0.3183. But the marginal probabilities are not very different from each other.

Case 7 The distribution is

		y				
		0	1	2	3	$P_X(x)$
	0	0.1	0.1	0.0875	0.0375	0.325
x	1	0	0.175	0	0	0.175
	2	0	0.1125	0.1125	0	0.225
	3	0.0875	0	0	0.1875	0.275
	$P_Y(y)$	0.1875	0.3875	0.2	0.225	1

This case represents medium agreement. The diagonal elements add up to 0.5750 and we can see that the off-diagonal elements are not quite different from each other.

Case 8 The distribution is

		y				
		0	1	2	3	$P_X(x)$
	0	0	0.05	0.1	0.2	0.35
x	1	0.05	0	0.2	0.15	0.4
	2	0.06	0.02	0.05	0	0.13
	3	0.05	0	0.02	0.05	0.12
	$P_Y(y)$	0.16	0.07	0.37	0.4	1

This case represents extremely poor agreement. In this case, we have complete disagreement given the first two categories. And we want to see if this significantly affect the agreement coefficient.

Case 9 The distribution is

		y = 0	1	2	3	$P_X(x)$
	0	0	0	0.1	0.05	0.15
x	1	0	0	0.08	0.02	0.1
	2	0.05	0.03	0.3	0.01	0.39
	3	0.06	0.02	0.03	0.25	0.36
	$P_Y(y)$	0.11	0.05	0.51	0.33	1

In this case, we can see that $p(1,1) = p(1,2) = p(2,1) = p(2,2) = 0$. And as we have suggested, $\kappa(0,1)$ should be set to 0. As a result, the covariance matrix should also be modified correspondingly. And we will see the performance of the proposed statistic κ_w through simulation.

Case 10 The distribution is

		y				
		0	1	2	3	$P_X(x)$
	0	0	0.01	0	0.01	0.02
x	1	0.02	0.05	0.05	0.3	0.42
	2	0	0.25	0	0.12	0.37
	3	0.08	0.03	0.05	0.03	0.19
	$P_Y(y)$	0.1	0.34	0.1	0.46	1

In this case, we still have four categories but poor agreement. The diagonal elements add up to 0.08 and the marginal probabilities are very different from each other.

Case 11 The distribution is

		y				
		0	1	2	3	$P_X(x)$
	0	0.2	0	0	0	0.2
x	1	0	0.1	0	0	0.1
	2	0	0	0.2	0	0.2
	3	0	0	0	0.5	0.5
	$P_Y(y)$	0.2	0.1	0.2	0.5	1

In this case, we have four categories but complete agreement. We would like to see whether the proposed method can correctly detect this or not.

Case 12 The distribution is

	y	0	1	2	3	4	$P_X(x)$
	0	0.2	0.01	0	0.01	0	0.22
	1	0	0.1	0	0.04	0.02	0.16
x	2	0.08	0	0.15	0	0.01	0.24
	3	0.01	0.04	0	0.2	0	0.25
	4	0	0.06	0	0.02	0.05	0.13
	$P_Y(y)$	0.29	0.21	0.15	0.27	0.08	1

In this case, we have five categories, and we have relatively large agreement. The sum of diagonal elements is 0.7 and all off-diagonal probabilities are small. We can also see that the marginal probabilities for X and Y are very close to each other.

Case 13 The distribution is

		y					
		0	1	2	3	4	$P_X(x)$
	0	0.01	0.05	0.2	0.06	0.03	0.35
	1	0	0.03	0.25	0.1	0.07	0.45
x	2	0	0.002	0.008	0.03	0.04	0.08
	3	0	0.004	0.02	0.03	0.003	0.057
	4	0.04	0.02	0.003	0	0	0.063
	$P_Y(y)$	0.05	0.106	0.481	0.22	0.143	1

We have five categories but very poor agreement.

Case 14 The distribution is

		y					
		0	1	2	3	4	$P_X(x)$
	0	0.1	0.01	0	0	0	0.11
	1	0	0.2	0.05	0	0	0.25
x	2	0.002	0.01	0	0	0.003	0.015
	3	0.003	0.002	0.01	0.3	0	0.315
	4	0.05	0.01	0	0	0.25	0.31
	$P_Y(y)$	0.155	0.232	0.06	0.3	0.253	1

We have five categories but high agreement. The diagonal elements add up to 0.85 . The marginal probabilities of X and Y are close to each other.

Case 15 The distribution is

		y 0	1	2	3	4	$P_X(x)$
	0	0.1	0	0	0	0	0.1
	1	0	0.05	0	0	0	0.05
x	2	0	0	0.25	0	0	0.25
	3	0	0	0	0.3	0	0.3
	4	0	0	0	0	0.3	0.3
	$P_Y(y)$	0.1	0.05	0.25	0.3	0.3	1

We have five categories and complete agreement. There is no disagreement.

2.6.2 Simulation Results

Table 2.1 gives the simulation results for case 1 to case 9 and table 2.2 gives the simulation results for case 10 to case 15.

Note: κ means Cohen's kappa, ke, km, ks are the proposed statistic using equal weights, max weights and square weights, respectively. v_κ is the observed variance of Cohen's kappa, v_{ke}, v_{km}, v_{ks} are the corresponding observed variances. v_1, v_2, v_3 are the corresponding variances calculated using the asymptotic formula derived previously.

Table 2.1. Simulation results for pairwise agreement minus disagreement for categorical responses

case	n	κ	ke	km	ks	v_κ	v_{ke}	v_{km}	v_{ks}	v_1	v_2	v_3
1	200	0.5123	0.6183	0.6237	0.6221	0.0024	0.0021	0.0021	0.0021	0.0017	0.0017	0.0017
1	400	0.5113	0.6172	0.6219	0.6205	0.0012	0.0011	0.0011	0.0011	0.0008	0.0009	0.0009
1	800	0.5129	0.6181	0.6227	0.6214	0.0006	0.0006	0.0006	0.0006	0.0004	0.0004	0.0004
1	1000	0.5126	0.6179	0.6225	0.6212	0.0006	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004
2	200	-0.2208	-0.5489	-0.5925	-0.5851	0.0012	0.0056	0.0044	0.0047	0.0145	0.0070	0.0074
2	400	-0.2203	-0.5493	-0.5908	-0.5842	0.0006	0.0028	0.0024	0.0025	0.0071	0.0034	0.0036
2	800	-0.2220	-0.5503	-0.5910	-0.5847	0.0003	0.0013	0.0011	0.0012	0.0034	0.0017	0.0018
2	1000	-0.2216	-0.5500	-0.5903	-0.5840	0.0002	0.0011	0.0009	0.0009	0.0027	0.0013	0.0014
3	200	0.8595	0.9034	0.9051	0.9051	0.0009	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004
3	400	0.8605	0.9040	0.9055	0.9056	0.0005	0.0003	0.0003	0.0003	0.0002	0.0002	0.0002
3	800	0.8618	0.9048	0.9063	0.9064	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
3	1000	0.8616	0.9046	0.9061	0.9062	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
4	200	1.0000	1.0000	1.0000	1.0000	0	0.0000	0	0	0.0000	0.0000	0.0000
4	400	1.0000	1.0000	1.0000	1.0000	0	0.0000	0	0	0.0000	0.0000	0.0000
4	800	1.0000	1.0000	1.0000	1.0000	0	0.0000	0	0	0.0000	0.0000	0.0000
4	1000	1.0000	1.0000	1.0000	1.0000	0	0.0000	0	0	0.0000	0.0000	0.0000
5	200	0.4542	0.6124	0.6017	0.5665	0.0019	0.0018	0.0023	0.0025	0.0015	0.0015	0.0017
5	400	0.4548	0.6119	0.6040	0.5666	0.0009	0.0009	0.0012	0.0012	0.0007	0.0008	0.0009
5	800	0.4537	0.6106	0.6052	0.5656	0.0005	0.0005	0.0006	0.0006	0.0004	0.0004	0.0004
5	1000	0.4536	0.6103	0.6058	0.5655	0.0004	0.0003	0.0004	0.0005	0.0003	0.0003	0.0003
6	200	0.0174	0.1627	0.2023	0.1845	0.0018	0.0050	0.0061	0.0059	0.0051	0.0041	0.0040
6	400	0.0175	0.1609	0.2021	0.1863	0.0010	0.0026	0.0031	0.0031	0.0026	0.0021	0.0020
6	800	0.0183	0.1625	0.2044	0.1895	0.0005	0.0012	0.0016	0.0015	0.0013	0.0010	0.0010
6	1000	0.0175	0.1612	0.2000	0.1858	0.0004	0.0010	0.0013	0.0012	0.0010	0.0008	0.0008
7	200	0.4406	0.6176	0.6396	0.6276	0.0019	0.0017	0.0016	0.0018	0.0013	0.0012	0.0013
7	400	0.4440	0.6197	0.6405	0.6309	0.0009	0.0008	0.0007	0.0009	0.0007	0.0006	0.0006
7	800	0.4454	0.6208	0.6409	0.6324	0.0005	0.0004	0.0004	0.0004	0.0003	0.0003	0.0003
7	1000	0.4434	0.6188	0.6388	0.6305	0.0004	0.0003	0.0003	0.0004	0.0003	0.0002	0.0003
8	200	-0.0983	-0.4453	-0.5375	-0.5229	0.0008	0.0049	0.0083	0.0084	0.0045	0.0051	0.0051
8	400	-0.0976	-0.4422	-0.5347	-0.5187	0.0004	0.0023	0.0042	0.0043	0.0023	0.0026	0.0026
8	800	-0.0983	-0.4445	-0.5378	-0.5212	0.0002	0.0011	0.0020	0.0020	0.0011	0.0013	0.0013
8	1000	-0.0975	-0.4424	-0.5353	-0.5183	0.0001	0.0009	0.0017	0.0017	0.0009	0.0011	0.0011

Table 2.2. Simulation results for pairwise agreement minus disagreement for categorical responses (*continued*)

9	200	0.3163	0.4605	0.5566	0.5679	0.0019	0.0015	0.0020	0.0021	0.0015	0.0020	0.0019
9	400	0.3183	0.4616	0.5555	0.5684	0.0010	0.0008	0.0011	0.0011	0.0007	0.0010	0.0009
9	800	0.3185	0.4623	0.5553	0.5686	0.0005	0.0004	0.0006	0.0006	0.0004	0.0005	0.0005
9	1000	0.3187	0.4624	0.5552	0.5686	0.0004	0.0003	0.0004	0.0004	0.0003	0.0004	0.0004
10	200	-0.2581	-0.3811	-0.5966	-0.5936	0.0007	0.0062	0.0083	0.0080	0.0052	0.0059	0.0059
10	400	-0.2585	-0.3808	-0.5954	-0.5926	0.0003	0.0030	0.0040	0.0038	0.0026	0.0030	0.0030
10	800	-0.2586	-0.3791	-0.5922	-0.5896	0.0002	0.0016	0.0021	0.0019	0.0013	0.0015	0.0015
10	1000	-0.2583	-0.3794	-0.5922	-0.5897	0.0001	0.0013	0.0017	0.0016	0.0011	0.0012	0.0012
11	200	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
11	400	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
11	800	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
11	1000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
12	200	0.6177	0.8106	0.8324	0.8251	0.0016	0.0006	0.0005	0.0006	0.0005	0.0004	0.0004
12	400	0.6175	0.8096	0.8317	0.8249	0.0008	0.0003	0.0002	0.0003	0.0003	0.0002	0.0002
12	800	0.6185	0.8104	0.8316	0.8252	0.0004	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
12	1000	0.6195	0.8109	0.8319	0.8258	0.0003	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
13	200	-0.0532	-0.3392	-0.5432	-0.5266	0.0005	0.0089	0.0085	0.0093	0.0064	0.0046	0.0048
13	400	-0.0538	-0.3408	-0.5413	-0.5260	0.0002	0.0046	0.0042	0.0047	0.0033	0.0024	0.0025
13	800	-0.0540	-0.3404	-0.5388	-0.5243	0.0001	0.0022	0.0021	0.0024	0.0017	0.0012	0.0013
13	1000	-0.0536	-0.3387	-0.5371	-0.5228	0.0001	0.0017	0.0017	0.0019	0.0013	0.0010	0.0010
14	200	0.8010	0.9036	0.9143	0.9178	0.0010	0.0003	0.0002	0.0002	0.0003	0.0002	0.0002
14	400	0.8001	0.9031	0.9130	0.9169	0.0005	0.0002	0.0001	0.0001	0.0002	0.0001	0.0001
14	800	0.8002	0.9031	0.9126	0.9167	0.0003	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
14	1000	0.8006	0.9032	0.9128	0.9169	0.0002	0.0001	0.0000	0.0000	0.0001	0.0000	0.0000
15	200	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
15	400	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
15	800	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
15	1000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

2.6.3 Discussion

Generally speaking, the proposed method is good as an indicator of measurement of agreement. In most cases, the equal weighting scheme, the max weighting scheme and the square weighting scheme have similar performance and work well. Under some cases, such as when the diagonal probabilities predominate (for example, case 7, case 12, case 14), it overestimates the agreement level. The proposed method is very good at detecting poor or good agreement. We therefore recommend that a researcher look at the data before he actually does the analysis and if possible, calculate Cohen's kappa and the proposed statistic using different weights. If there is a significant difference, a decision can be made based on the distribution of the original data.

The measurement of agreement is itself a complicated issue. If there are more categories available, then the two raters have more options in giving their evaluations. And the difference between two categories will become more subtle, making it harder for two raters to completely agree with each other. For example, if there are only two categories, good and bad, then it is very easy for two skillful raters to agree on an individual rating. However, if there are more categories, say, extremely poor, poor, fair, good and excellent (here we only consider them as categorical), then it is natural that the two raters will disagree on some ratings. Suppose fair, good and excellent can be categorized as being good and extremely poor and poor can be categorized as being poor. Consider the following example:

		$y=0$	1	2	3	4	$P_X(x)$
	0	0	0.2	0	0	0	0.2
	1	0.3	0	0	0	0	0.3
x	2	0	0	0	0.1	0.1	0.2
	3	0	0	0.1	0	0.1	0.2
	4	0	0	0.1	0	0	0.1
	$P_Y(y)$	0.3	0.2	0.2	0.1	0.2	1

So Cohen's kappa gives -0.25 and the proposed method with max and square weight both give -0.4. However, if we have only two categories, then the distribution is

		$y=0$	1	$P_X(x)$
	0	0.5	0	0.5
x	2	0	0.5	0.5
	$P_Y(y)$	0.5	0.5	1

Cohen's kappa is one and the proposed method also gives one. This example shows that the degree of agreement really depends on the evaluation methods. Given such context, it makes sense intuitively that some agreement coefficients may be higher than the sum of diagonal elements. And in such cases, the proposed method makes sense.

2.7 Simulation for Ordinal Responses

2.7.1 Simulation Design

We used Matlab to generate samples following multinominal distributions. For each distribution, we calculated Cohen's weighted kappa using simple proportional weights, square weights and exponential weights. We also calculated the proposed statistic κ_w using corresponding weights. Their variances are also given as well as the variances calculated using the sample counterparts calculated using the asymptotic distribution formula. Each simulation is run 1000 times. Keep in mind that in practice, the interpretation of the magnitude of weighted sum of pairwise agreement measures is similar as that of simple kappa. Agreement measures with values greater than 0.75 or so may be taken to represent excellent agreement, values between 0.40 and 0.75 may be taken to represent fair to good agreement and values below 0.40 or so may be taken to represent poor agreement.

2.7.2 Simulation Results

Note: wk1, wk2, wk3 are Cohen's weighted kappa using proportional weights, square distance weights and exponential distance weights, respectively. mwk1, mwk2, mwk3 are the proposed statistic's κ_w *using proportional weights, square weights and exponential weights, respectively.* x_1, y_1 *are the variances of wk1 and mwk1, respectively.* z_1 *is the mean asymptotic variance for the proposed statistic* κ_w *using proportional weights.* x_2, y_2 *are the variances of wk2 and mwk2, respectively.* z_2 *is the mean asymptotic variance for the proposed statistic* κ_w *using square weights.* x_3, y_3 *are the variance of wk3*

Table 2.3. Simulation results for pairwise agreement minus disagreement for ordinal responses

case	n	wk1	wk2	wk3	mwk1	mwk2	mwk3	x_1	y_1	z_1	x_2	y_2	z_2	x_3	y
1	20	0.4772	0.4535	0.4560	0.6055	0.5964	0.5973	0.0310	0.0277	0.0198	0.0486	0.0377	0.0254	0.0461	0.036
1	40	0.4873	0.4654	0.4678	0.6081	0.5997	0.6006	0.0150	0.0131	0.0091	0.0242	0.0178	0.0111	0.0229	0.017
1	80	0.4924	0.4703	0.4726	0.6100	0.6006	0.6015	0.0077	0.0066	0.0043	0.0124	0.0090	0.0052	0.0117	0.008
1	100	0.4921	0.4708	0.4730	0.6093	0.6005	0.6014	0.0063	0.0055	0.0035	0.0102	0.0076	0.0041	0.0097	0.007
1	200	0.4947	0.4740	0.4762	0.6109	0.6022	0.6031	0.0032	0.0028	0.0017	0.0049	0.0037	0.0020	0.0046	0.003
2	200	-0.1268	-0.0265	-0.0366	-0.5227	-0.4936	-0.4964	0.0015	0.0068	0.0286	0.0029	0.0091	0.0481	0.0027	0.008
2	400	-0.1271	-0.0284	-0.0384	-0.5224	-0.4950	-0.4977	0.0008	0.0037	0.0135	0.0015	0.0048	0.0227	0.0014	0.004
2	800	-0.1276	-0.0290	-0.0389	-0.5230	-0.4953	-0.4981	0.0004	0.0018	0.0066	0.0008	0.0024	0.0110	0.0007	0.002
2	1000	-0.1277	-0.0293	-0.0392	-0.5230	-0.4951	-0.4978	0.0003	0.0015	0.0052	0.0006	0.0019	0.0087	0.0005	0.001
3	200	0.8509	0.8391	0.8403	0.9007	0.8967	0.8971	0.0013	0.0007	0.0003	0.0020	0.0010	0.0003	0.0019	0.000
3	400	0.8522	0.8410	0.8422	0.9015	0.8979	0.8982	0.0006	0.0003	0.0002	0.0010	0.0005	0.0002	0.0009	0.000
3	800	0.8515	0.8400	0.8412	0.9007	0.8969	0.8973	0.0003	0.0002	0.0001	0.0005	0.0003	0.0001	0.0005	0.000
3	1000	0.8506	0.8388	0.8400	0.9000	0.8960	0.8964	0.0003	0.0001	0.0001	0.0004	0.0002	0.0001	0.0004	0.000
4	200	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.000
4	400	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.000
4	800	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.000
4	1000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.000
5	200	0.5364	0.6441	0.6320	0.7084	0.8056	0.7960	0.0017	0.0009	0.0008	0.0015	0.0004	0.0004	0.0015	0.000
5	400	0.5386	0.6465	0.6344	0.7093	0.8062	0.7966	0.0009	0.0005	0.0004	0.0007	0.0002	0.0002	0.0007	0.000
5	800	0.5377	0.6462	0.6341	0.7081	0.8054	0.7958	0.0004	0.0002	0.0002	0.0004	0.0001	0.0001	0.0004	0.000
5	1000	0.5382	0.6466	0.6345	0.7085	0.8057	0.7961	0.0004	0.0002	0.0002	0.0003	0.0001	0.0001	0.0003	0.000
6	200	0.0565	0.1205	0.1577	0.2347	0.3146	0.3485	0.0031	0.0065	0.0058	0.0065	0.0094	0.0080	0.0070	0.011
6	400	0.0562	0.1204	0.1577	0.2320	0.3109	0.3446	0.0016	0.0031	0.0030	0.0033	0.0045	0.0041	0.0035	0.005
6	800	0.0561	0.1214	0.1591	0.2341	0.3141	0.3482	0.0008	0.0015	0.0015	0.0017	0.0021	0.0020	0.0018	0.002
6	1000	0.0567	0.1226	0.1605	0.2335	0.3139	0.3481	0.0007	0.0013	0.0012	0.0014	0.0018	0.0016	0.0015	0.002
7	200	0.3961	0.3289	0.2881	0.5907	0.5547	0.5350	0.0030	0.0019	0.0017	0.0058	0.0028	0.0027	0.0074	0.003
7	400	0.3982	0.3307	0.2898	0.5925	0.5565	0.5367	0.0014	0.0009	0.0009	0.0027	0.0013	0.0014	0.0035	0.001
7	800	0.3978	0.3302	0.2892	0.5917	0.5555	0.5356	0.0008	0.0005	0.0004	0.0015	0.0007	0.0007	0.0019	0.000
7	1000	0.3966	0.3294	0.2887	0.5905	0.5545	0.5348	0.0006	0.0004	0.0003	0.0013	0.0006	0.0005	0.0016	0.000
8	200	-0.1756	-0.2580	-0.2967	-0.4999	-0.5520	-0.5640	0.0016	0.0047	0.0060	0.0034	0.0055	0.0076	0.0044	0.005
8	400	-0.1727	-0.2551	-0.2949	-0.4995	-0.5517	-0.5641	0.0008	0.0025	0.0030	0.0016	0.0030	0.0038	0.0020	0.003
8	800	-0.1732	-0.2556	-0.2956	-0.5008	-0.5528	-0.5651	0.0004	0.0012	0.0015	0.0008	0.0014	0.0019	0.0010	0.001
8	1000	-0.1733	-0.2553	-0.2950	-0.5018	-0.5533	-0.5654	0.0003	0.0010	0.0012	0.0007	0.0011	0.0015	0.0008	0.001

Table 2.4. Simulation results for pairwise agreement minus disagreement for ordinal responses (*continued*)

case	n	wk1	wk2	wk3	mwk1	mwk2	mwk3	x_1	y_1	z_1	x_2	y_2	z_2	x_3	y
9	200	0.1474	-0.0075	-0.0263	0.4598	0.4517	0.4434	0.0022	0.0023	0.0025	0.0037	0.0034	0.0040	0.0052	0.003
9	400	0.1466	-0.0102	-0.0293	0.4593	0.4504	0.4419	0.0012	0.0012	0.0012	0.0021	0.0019	0.0020	0.0029	0.002
9	800	0.1492	-0.0066	-0.0250	0.4621	0.4542	0.4460	0.0006	0.0006	0.0006	0.0010	0.0009	0.0010	0.0014	0.001
9	1000	0.1484	-0.0076	-0.0260	0.4612	0.4531	0.4449	0.0005	0.0005	0.0005	0.0008	0.0007	0.0008	0.0012	0.000
10	200	-0.3166	-0.4207	-0.4922	-0.3886	-0.4060	-0.4182	0.0014	0.0071	0.0061	0.0036	0.0102	0.0089	0.0052	0.012
10	400	-0.3191	-0.4230	-0.4938	-0.3873	-0.4034	-0.4149	0.0007	0.0033	0.0031	0.0018	0.0046	0.0046	0.0025	0.005
10	800	-0.3181	-0.4224	-0.4941	-0.3888	-0.4054	-0.4173	0.0004	0.0017	0.0016	0.0010	0.0024	0.0023	0.0013	0.002
10	1000	-0.3193	-0.4242	-0.4959	-0.3878	-0.4044	-0.4161	0.0003	0.0014	0.0013	0.0007	0.0019	0.0018	0.0011	0.002
11	200	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.000
11	400	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.000
11	800	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.000
11	1000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.000
12	200	0.5630	0.5640	0.6317	0.7819	0.7823	0.8205	0.0024	0.0008	0.0008	0.0031	0.0009	0.0009	0.0027	0.000
12	400	0.5620	0.5633	0.6316	0.7816	0.7821	0.8204	0.0013	0.0004	0.0004	0.0016	0.0005	0.0004	0.0013	0.000
12	800	0.5637	0.5646	0.6326	0.7825	0.7828	0.8210	0.0006	0.0002	0.0002	0.0008	0.0002	0.0002	0.0007	0.000
12	1000	0.5625	0.5642	0.6327	0.7812	0.7816	0.8200	0.0005	0.0002	0.0002	0.0007	0.0002	0.0002	0.0005	0.000
13	200	-0.0736	-0.0973	-0.1288	-0.4378	-0.5073	-0.5488	0.0011	0.0092	0.0074	0.0037	0.0108	0.0093	0.0095	0.012
13	400	-0.0734	-0.0959	-0.1250	-0.4440	-0.5144	-0.5563	0.0005	0.0044	0.0037	0.0017	0.0049	0.0047	0.0046	0.005
13	800	-0.0744	-0.0980	-0.1286	-0.4429	-0.5132	-0.5551	0.0002	0.0022	0.0019	0.0008	0.0025	0.0024	0.0022	0.002
13	1000	-0.0743	-0.0966	-0.1260	-0.4448	-0.5143	-0.5554	0.0002	0.0018	0.0015	0.0007	0.0020	0.0019	0.0018	0.002
14	200	0.7883	0.7518	0.6821	0.9035	0.8908	0.8632	0.0017	0.0004	0.0004	0.0036	0.0007	0.0007	0.0068	0.001
14	400	0.7904	0.7529	0.6824	0.9046	0.8917	0.8639	0.0009	0.0002	0.0002	0.0018	0.0003	0.0003	0.0034	0.000
14	800	0.7915	0.7548	0.6850	0.9048	0.8923	0.8648	0.0004	0.0001	0.0001	0.0009	0.0002	0.0002	0.0017	0.000
14	1000	0.7906	0.7536	0.6834	0.9045	0.8920	0.8643	0.0003	0.0001	0.0001	0.0007	0.0001	0.0001	0.0012	0.000
15	200	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.000
15	400	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.000
15	800	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.000
15	1000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.000

and mwk3, respectively. z_3 *is the mean asymptotic variance for the proposed statistic* κ_w *using exponential weights.*

2.7.3 Discussion

It is more difficult and less intuitive to interpret weighted kappa since the degree of agreement is not as straightforward. Generally speaking, the proposed method is good as an indicator of measurement of agreement. In most cases, the equal weighting scheme, the max weighting scheme and the square weighting scheme have similar performances and work well. Under some cases, such as when the diagonal probabilities predominate (for example, case 7, case 12, case 14), it overestimates the agreement level. The proposed method is very good at detecting poor agreement. We therefore recommend that a researcher examine the data before he actually does the analysis and if possible, calculate Cohen's kappa and the proposed statistic using different weights. If there is significant difference, a decision can be made based on the distribution of the original data.

Chapter 3

Development, Properties and Estimators of the Re-defined Conditional κ

3.1 Definition of the Conditional κ

Another alternative that we have investigated is to construct the following measurement of agreement:

$$\kappa = \frac{\sum_{i=1}^{k} w_i(p_{o,i} - p_{e,i})}{\sum_{i=1}^{k} w_i(1 - p_{e,i})}, \tag{3.1}$$

where $p_{o,i}$ is defined as

$$p_{o,i} = \frac{p_{ii}}{p_{ii} + \sum_{\substack{j=0 \\ j\neq i}}^{k} (p_{ij} + p_{ji})}.$$

And $p_{e,i}$ is defined as

$$p_{e,i} = \frac{p_{i\cdot}p_{\cdot i}}{p_{i\cdot}p_{\cdot i} + \sum_{\substack{j=0 \\ j\neq i}}^{k} (p_{i\cdot}p_{\cdot j} + p_{j\cdot}p_{\cdot i})}.$$

It can be shown that

$$
\begin{aligned}
p_{o,i} &= \frac{p_{ii}}{p_{ii} + \sum\limits_{\substack{j=0 \\ j \neq i}}^{k} (p_{ij} + p_{ji})} \\
&= \frac{p_{ii}}{p_{i\cdot} + p_{\cdot i} - p_{ii}}.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
p_{e,i} &= \frac{p_{i\cdot}p_{\cdot i}}{p_{i\cdot}p_{\cdot i} + \sum\limits_{\substack{j=0 \\ j \neq i}}^{k} (p_{i\cdot}p_{\cdot j} + p_{j\cdot}p_{\cdot i})} \\
&= \frac{p_{i\cdot}p_{\cdot i}}{\sum\limits_{j=0}^{k} (p_{i\cdot}p_{\cdot j} + p_{j\cdot}p_{\cdot i}) - p_{i\cdot}p_{\cdot i}} \\
&= \frac{p_{i\cdot}p_{\cdot i}}{p_{i\cdot} \sum\limits_{j=0}^{k} p_{\cdot j} + p_{\cdot i} \sum\limits_{j=0}^{k} p_{j\cdot} - p_{i\cdot}p_{\cdot i}} \\
&= \frac{p_{i\cdot}p_{\cdot i}}{p_{i\cdot} + p_{\cdot i} - p_{i\cdot}p_{\cdot i}}.
\end{aligned}
$$

Therefore, the conditional κ can also be expressed as

$$
\kappa = \frac{\sum\limits_{i=0}^{k} w_i \left(\frac{p_{ii}}{p_{i\cdot} + p_{\cdot i} - p_{ii}} - \frac{p_{i\cdot}p_{\cdot i}}{p_{i\cdot} + p_{\cdot i} - p_{i\cdot}p_{\cdot i}} \right)}{\sum\limits_{i=0}^{k} w_i \left(1 - \frac{p_{i\cdot}p_{\cdot i}}{p_{i\cdot} + p_{\cdot i} - p_{i\cdot}p_{\cdot i}} \right)}. \tag{3.2}
$$

3.2 Properties of the Conditional κ

1. $-1 \leq \kappa \leq 1$.

2. $\kappa = 1$ if $p_{ij} = 0$, for $i, j = 0, 1, 2, \cdots, k$, $i \neq j$.

This property is nice in that when we have complete agreement, we will get 1 for the proposed method.

3. $\kappa = 0$ if $p_{ii} = p_{i\cdot}p_{\cdot i}$, $i = 0, 1, 2, \cdots, k$.

   This property means that when the observed agreement can be accounted exactly by chance, then the agreement coefficient is 0.

4. $\kappa = -1$ if $p_{i\cdot} + p_{\cdot i} = 3p_{i\cdot}p_{\cdot i}$ and $p_{ii} = 0$ for each $i = 0, 1, 2, \cdots, k$.

Proof. 1. From the definition of $p_{o,i}$ and $p_{e,i}$, it can be easily seen that $0 \leq p_{o,i} \leq 1$ and $0 \leq p_{e,i} \leq 1$. Hence it is straightforward to see that

$$\frac{p_{ii}}{p_{i\cdot} + p_{\cdot i} - p_{ii}} - \frac{p_{i\cdot}p_{\cdot i}}{p_{i\cdot} + p_{\cdot i} - p_{i\cdot}p_{\cdot i}} \leq 1 - \frac{p_{i\cdot}p_{\cdot i}}{p_{i\cdot} + p_{\cdot i} - p_{i\cdot}p_{\cdot i}},$$

for each $i = 0, 1, 2, \cdots, k$.

Therefore,

$$\begin{aligned} \kappa &= \frac{\sum_{i=0}^{k} w_i(\frac{p_{ii}}{p_{i\cdot}+p_{\cdot i}-p_{ii}} - \frac{p_{i\cdot}p_{\cdot i}}{p_{i\cdot}+p_{\cdot i}-p_{i\cdot}p_{\cdot i}})}{\sum_{i=0}^{k} w_i(1 - \frac{p_{i\cdot}p_{\cdot i}}{p_{i\cdot}+p_{\cdot i}-p_{i\cdot}p_{\cdot i}})} \\ &\leq 1. \end{aligned}$$

To show that

$$\kappa = \frac{\sum_{i=0}^{k} w_i(\frac{p_{ii}}{p_{i\cdot}+p_{\cdot i}-p_{ii}} - \frac{p_{i\cdot}p_{\cdot i}}{p_{i\cdot}+p_{\cdot i}-p_{i\cdot}p_{\cdot i}})}{\sum_{i=0}^{k} w_i(1 - \frac{p_{i\cdot}p_{\cdot i}}{p_{i\cdot}+p_{\cdot i}-p_{i\cdot}p_{\cdot i}})} \geq -1,$$

we need to show that

$$\sum_{i=0}^{k} w_i(1 + \frac{p_{ii}}{p_{i\cdot} + p_{\cdot i} - p_{ii}}) \geq \sum_{i=0}^{k} w_i(\frac{2p_{i\cdot}p_{\cdot i}}{p_{i\cdot} + p_{\cdot i} - p_{i\cdot}p_{\cdot i}}).$$

We want to show that

$$1 + \frac{p_{ii}}{p_{i\cdot} + p_{\cdot i} - p_{ii}} \geq \frac{2p_{i\cdot}p_{\cdot i}}{p_{i\cdot} + p_{\cdot i} - p_{i\cdot}p_{\cdot i}},$$

for $i = 0, 1, 2, \cdots, k$. Now if

$$\frac{p_{i\cdot}p_{\cdot i}}{p_{i\cdot} + p_{\cdot i} - p_{i\cdot}p_{\cdot i}} \leq \frac{1}{2},$$

then it is obvious that the inequality holds.

Suppose, on the other hand, that for some i,

$$\frac{p_{i\cdot}p_{\cdot i}}{p_{i\cdot} + p_{\cdot i} - p_{i\cdot}p_{\cdot i}} \geq \frac{1}{2}.$$

That is, $p_{i\cdot} + p_{\cdot i} \leq 3p_{i\cdot}p_{\cdot i}$. To avoid confusion, let $v_1 = p_{ii}$, $v_2 = p_{i\cdot}$ and $v_3 = p_{\cdot i}$. So in this case, $v_2 + v_3 \leq 3v_2v_3$. And we want to show that

$$\frac{v_2 + v_3}{v_2 + v_3 - v_1} \geq \frac{2v_2v_3}{v_2 + v_3 - v_2v_3}.$$

Note that $v_2 + v_3 \leq 3v_2v_3$ can also be written as $v_2(1-v_3) + v_3(1-2v_2) \leq 0$. Since $0 \leq v_2 \leq 1$, $0 \leq v_3 \leq 1$, for the left side to be smaller than 0, $v_2 \geq \frac{1}{2}$. Similarly, we can also get that $v_3 \geq \frac{1}{2}$. We know that there can be only one such i that satisfies these properties.

We can let $v_2 = \frac{1}{2} + a$ and $v_3 = \frac{1}{2} + b$, where $0 \leq a \leq \frac{1}{2}$ and $0 \leq b \leq \frac{1}{2}$. Then $v_2 + v_3 = 1 + a + b$.

Because we know that $v_2 + v_3 - v_1 \leq 1$. So we can get $v_1 \geq a + b$.

Therefore, we get

$$\frac{v_2 + v_3}{v_2 + v_3 - v_1} \geq \frac{v_2 + v_3}{v_2 + v_3 - (a + b)}.$$

We want to show that

$$\frac{v_2 + v_3}{v_2 + v_3 - (a + b)} \geq \frac{2v_2v_3}{v_2 + v_3 - v_2v_3}.$$

If we plug in the values for v_2 and v_3, respectively, the inequality becomes

$$1 + a + b \geq \frac{2(\frac{1}{2} + a)(\frac{1}{2} + b)}{1 + a + b - (\frac{1}{2} + a)(\frac{1}{2} + b)}.$$

That is, we need to show that

$$(1 + a + b)^2 - (\frac{1}{2} + a)(\frac{1}{2} + b)(1 + a + b) \geq 2(\frac{1}{2} + a)(\frac{1}{2} + b).$$

This means we need to prove that

$$(1 + a + b)^2 \geq (\frac{1}{2} + a)(\frac{1}{2} + b)(3 + a + b).$$

Tedious but straightforward algebraic calculation shows that we need to prove

$$\frac{1}{2}(a^2+b^2)+\frac{1}{4}(a+b)+\frac{1}{4} \geq 2ab+ab(a+b).$$

Since we know that $0 \leq a \leq \frac{1}{2}$ and $0 \leq b \leq \frac{1}{2}$, so $0 \leq ab \leq \frac{1}{4}$. So $\frac{1}{4}(a+b) \geq ab(a+b)$. And since $\frac{1}{2}(a^2+b^2) \geq ab$. Hence we can see that the inequality does hold. This finishes the proof.

2. If $p_{ij}=0$, for $i,j=0,1,2,\cdots,k$, $i \neq j$, then we can get

$$p_{i\cdot}+p_{\cdot i}=2p_{ii}.$$

Hence

$$\frac{p_{ii}}{p_{i\cdot}+p_{\cdot i}-p_{ii}}=1,$$

for $i=0,1,2,\cdots,k$

Therefore,

$$\begin{aligned}
&\sum_{i=0}^{k} w_i\left(\frac{p_{ii}}{p_{i\cdot}+p_{\cdot i}-p_{ii}}-1\right)=0\\
\Rightarrow\quad &\sum_{i=0}^{k} w_i\left(\frac{p_{ii}}{p_{i\cdot}+p_{\cdot i}-p_{ii}}-\frac{p_{i\cdot}p_{\cdot i}}{p_{i\cdot}+p_{\cdot i}-p_{i\cdot}p_{\cdot i}}\right)=\sum_{i=0}^{k} w_i\left(1-\frac{p_{i\cdot}p_{\cdot i}}{p_{i\cdot}+p_{\cdot i}-p_{i\cdot}p_{\cdot i}}\right)\\
\Rightarrow\quad &\kappa=1.
\end{aligned}$$

3. If $p_{ii}=p_{i\cdot}p_{\cdot i}$, $i=0,1,2,\cdots,k$, then

$$\frac{p_{ii}}{p_{i\cdot} + p_{\cdot i} - p_{ii}} = \frac{p_{i\cdot}p_{\cdot i}}{p_{i\cdot} + p_{\cdot i} - p_{i\cdot}p_{\cdot i}}.$$

Thus each term in the numerator becomes zero and hence $\kappa = 0$.

4. From the previous proof, it can be easily verified that if $p_{i\cdot} + p_{\cdot i} = 3p_{i\cdot}p_{\cdot i}$ and $p_{ii} = 0$ for each $i = 0, 1, 2, \cdots, k$, then

$$\frac{p_{ii}}{p_{i\cdot} + p_{\cdot i} - p_{ii}} - \frac{p_{i\cdot}p_{\cdot i}}{p_{i\cdot} + p_{\cdot i} - p_{i\cdot}p_{\cdot i}} = -1(1 - \frac{p_{i\cdot}p_{\cdot i}}{p_{i\cdot} + p_{\cdot i} - p_{i\cdot}p_{\cdot i}}),$$

for each $i = 0, 1, 2, \cdots, k$.

Hence

$$w_i(\frac{p_{ii}}{p_{i\cdot} + p_{\cdot i} - p_{ii}} - \frac{p_{i\cdot}p_{\cdot i}}{p_{i\cdot} + p_{\cdot i} - p_{i\cdot}p_{\cdot i}}) = -w_i(1 - \frac{p_{i\cdot}p_{\cdot i}}{p_{i\cdot} + p_{\cdot i} - p_{i\cdot}p_{\cdot i}}),$$

for each $i = 0, 1, 2, \cdots, k$.

Therefore,

$$\sum_{i=0}^{k} w_i(\frac{p_{ii}}{p_{i\cdot} + p_{\cdot i} - p_{ii}} - \frac{p_{i\cdot}p_{\cdot i}}{p_{i\cdot} + p_{\cdot i} - p_{i\cdot}p_{\cdot i}}) = -1 \times \sum_{i=0}^{k} w_i(1 - \frac{p_{i\cdot}p_{\cdot i}}{p_{i\cdot} + p_{\cdot i} - p_{i\cdot}p_{\cdot i}}).$$

As a result, $\kappa = -1$. Notice that this is a sufficient but not a necessary condition for $\kappa = -1$. κ can be -1 under some other conditions.

$\square$

3.3 Asymptotic Distributions

As before, we can apply $\delta -$ method to derive the asymptotic distribution of κ. Suppose the sample size is n. First note that for a bivariate sample of size n,

- $\hat{p}_{00} = \frac{1}{n}\sum\sum \mathrm{I}\{X=0, Y=0\}$
- $\hat{p}_{01} = \frac{1}{n}\sum\sum \mathrm{I}\{X=0, Y=1\}$
- $\hat{p}_{02} = \frac{1}{n}\sum\sum \mathrm{I}\{X=0, Y=2\}$

  $\vdots$
- $\hat{p}_{kk} = \frac{1}{n}\sum\sum \mathrm{I}\{X=k, Y=k\}$.

When n is large, we know that there is strong convergence, i.e.,

- $\hat{p}_{00} \longrightarrow \mathrm{E}(\mathrm{I}\{X=0, Y=0\}) = p_{00}$
- $\hat{p}_{01} \longrightarrow \mathrm{E}(\mathrm{I}\{X=0, Y=1\}) = p_{01}$
- $\hat{p}_{02} \longrightarrow \mathrm{E}(\mathrm{I}\{X=0, Y=2\}) = p_{02}$

  $\vdots$
- $\hat{p}_{kk} \longrightarrow \mathrm{E}(\mathrm{I}\{X=k, Y=k\}) = p_{kk}$.

First suppose that the weight w_i does not depend on the data, for $i = 0, 1, \ldots, k$.

Let $v = (p_{00}, p_{01}, p_{02}, \cdots, p_{kk})$ and $\hat{v} = (\hat{p}_{00}, \hat{p}_{01}, \hat{p}_{02}, \cdots, \hat{p}_{kk})$. Then the vector $\hat{v}$ has an asymptotic normal distribution with

$$\mathrm{E}(\hat{v}) = (p_{00}, p_{01}, p_{02}, \cdots, p_{kk})$$

and variance $n^{-1}\Sigma$, where Σ is a $(k+1)^2 \times (k+1)^2$ matrix such that

$$\Sigma = \begin{bmatrix} p_{00}(1-p_{00}) & -p_{00}p_{01} & \cdots & -p_{00}p_{0k} & -p_{00}p_{10} & \cdots & -p_{00}p_{kk} \\ -p_{00}p_{01} & p_{01}(1-p_{01}) & \cdots & -p_{01}p_{0k} & -p_{01}p_{10} & \cdots & -p_{01}p_{kk} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ -p_{kk}p_{00} & -p_{kk}p_{01} & \cdots & -p_{kk}p_{0k} & -p_{kk}p_{10} & \cdots & p_{kk}(1-p_{kk}) \end{bmatrix}. \tag{3.3}$$

Let $g(v)$ be a function of v such that

$$g(v) = \frac{\sum_{i=0}^{k} w_i\left(\frac{p_{ii}}{p_{i\cdot}+p_{\cdot i}-p_{ii}} - \frac{p_{i\cdot}p_{\cdot i}}{p_{i\cdot}+p_{\cdot i}-p_{i\cdot}p_{\cdot i}}\right)}{\sum_{i=0}^{k} w_i\left(1 - \frac{p_{i\cdot}p_{\cdot i}}{p_{i\cdot}+p_{\cdot i}-p_{i\cdot}p_{\cdot i}}\right)}.$$

Note that κ can also be written as

$$\kappa = \frac{\sum_{i=0}^{k} w_i\left[(p_{i\cdot}+p_{\cdot i})(p_{ii}-p_{i\cdot}p_{\cdot i})\right]}{\sum_{i=0}^{k} w_i\left[(p_{i\cdot}+p_{\cdot i}-p_{ii})(p_{i\cdot}+p_{\cdot i}-2p_{i\cdot}p_{\cdot i})\right]}. \tag{3.4}$$

Let

$$R = \sum_{i=0}^{k} w_i\left[(p_{i\cdot}+p_{\cdot i})(p_{ii}-p_{i\cdot}p_{\cdot i})\right]$$

and

$$Q = \sum_{i=0}^{k} w_i\left[(p_{i\cdot}+p_{\cdot i}-p_{ii})(p_{i\cdot}+p_{\cdot i}-2p_{i\cdot}p_{\cdot i})\right].$$

Using the theory on functions of asymptotically normal vectors, we know that $g(\hat{v})$ has asymptotically a normal distribution with mean $g(v)$ and variance $n^{-1}d\Sigma d'$, where d is a $(k+1)^2$ vector such that $d = (\dot{g}(p))$ with

$$d_1 = \frac{\partial g}{\partial p_{00}}\bigg|_{v=\mathrm{E}(v)} = \frac{Qt_1 - Rt_2}{Q^2},$$

$$d_2 = \frac{\partial g}{\partial p_{01}}\bigg|_{v=\mathrm{E}(v)} = \frac{Qt_3 - Rt_4}{Q^2},$$

$$d_3 = \frac{\partial g}{\partial p_{02}}\bigg|_{v=\mathrm{E}(v)} = \frac{Qt_5 - Rt_6}{Q^2},$$

$$\vdots$$

$$\vdots$$

$$d_{k+1} = \frac{\partial g}{\partial p_{0k}}\bigg|_{v=\mathrm{E}(v)} = \frac{Qt_{2k+1} - Rt_{2k+2}}{Q^2},$$

$$d_{k+2} = \frac{\partial g}{\partial p_{10}}\bigg|_{v=\mathrm{E}(v)} = \frac{Qt_{2k+3} - Rt_{2k+4}}{Q^2},$$

$$d_{k+3} = \frac{\partial g}{\partial p_{11}}\bigg|_{v=\mathrm{E}(v)} = \frac{Qt_{2k+5} - Rt_{2k+6}}{Q^2},$$

$$\vdots$$

$$\vdots$$

$$d_{(k+1)^2} = \frac{\partial g}{\partial p_{kk}}\bigg|_{v=\mathrm{E}(v)} = \frac{Qt_{2(k+1)^2-1} - Rt_{2(k+1)^2}}{Q^2},$$

where

$$t_1 = 2w_0(p_{00} - p_{0\cdot}p_{\cdot 0}) + w_0(p_{0\cdot} + p_{\cdot 0})[1 - (p_{0\cdot} + p_{\cdot 0})],$$

$$t_2 = w_0(p_{0\cdot} + p_{\cdot 0}) - 2p_{0\cdot}p_{\cdot 0} + 2w_0[(p_{0\cdot} + p_{\cdot 0}) - p_{00}][1 - (p_{0\cdot} + p_{\cdot 0})],$$

$$t_3 = w_0[p_{00} - p_{0\cdot}p_{\cdot 0} - (p_{0\cdot} + p_{\cdot 0})p_{\cdot 0}] + w_1[p_{11} - p_{1\cdot}p_{\cdot 1} - (p_{1\cdot} + p_{\cdot 1})p_{1\cdot}],$$

$t_4 = w_0[(p_{0\cdot} + p_{\cdot 0} - 2p_{0\cdot}p_{\cdot 0}) + (p_{0\cdot} + p_{\cdot 0} - p_{00})(1 - 2p_{\cdot 0})] + w_1[(p_{1\cdot} + p_{\cdot 1} - 2p_{1\cdot}p_{\cdot 1}) + (p_{1\cdot} + p_{\cdot 1} - p_{11})(1 - 2p_{1\cdot})],$

$t_5 = w_0[p_{00} - p_{0\cdot}p_{\cdot 0} - (p_{0\cdot} + p_{\cdot 0})p_{\cdot 0}] + w_2[p_{22} - p_{2\cdot}p_{\cdot 2} - (p_{2\cdot} + p_{\cdot 2})p_{2\cdot}],$

$t_6 = w_0[(p_{0\cdot} + p_{\cdot 0} - 2p_{0\cdot}p_{\cdot 0}) + (p_{0\cdot} + p_{\cdot 0} - p_{00})(1 - 2p_{\cdot 0})] + w_2[(p_{2\cdot} + p_{\cdot 2} - 2p_{2\cdot}p_{\cdot 2}) + (p_{2\cdot} + p_{\cdot 2} - p_{22})(1 - 2p_{2\cdot})],$

$\vdots$

$\vdots$

$t_{2k+1} = w_0[p_{00} - p_{0\cdot}p_{\cdot 0} - (p_{0\cdot} + p_{\cdot 0})p_{\cdot 0}] + w_k[p_{kk} - p_{k\cdot}p_{\cdot k} - (p_{k\cdot} + p_{\cdot k})p_{k\cdot}],$

$t_{2k+2} = w_0[(p_{0\cdot} + p_{\cdot 0} - 2p_{0\cdot}p_{\cdot 0}) + (p_{0\cdot} + p_{\cdot 0} - p_{00})(1 - 2p_{\cdot 0})] + w_{k+1}[(p_{k\cdot} + p_{\cdot k} - 2p_{k\cdot}p_{\cdot k}) + (p_{k\cdot} + p_{\cdot k} - p_{kk})(1 - 2p_{k\cdot})],$

$t_{2k+3} = w_0[p_{00} - p_{0\cdot}p_{\cdot 0} - (p_{0\cdot} + p_{\cdot 0})p_{0\cdot}] + w_1[p_{11} - p_{1\cdot}p_{\cdot 1} - (p_{1\cdot} + p_{\cdot 1})p_{\cdot 1}],$

$t_{2k+4} = w_0[(p_{0\cdot} + p_{\cdot 0} - 2p_{0\cdot}p_{\cdot 0}) + (p_{0\cdot} + p_{\cdot 0} - p_{00})(1 - 2p_{0\cdot})] + w_1[(p_{1\cdot} + p_{\cdot 1} - 2p_{1\cdot}p_{\cdot 1}) + (p_{1\cdot} + p_{\cdot 1} - p_{11})(1 - 2p_{\cdot 1})],$

$t_{2k+5} = 2w_1(p_{11} - p_{1\cdot}p_{\cdot 1}) + w_1(p_{1\cdot} + p_{\cdot 1})[1 - (p_{1\cdot} + p_{\cdot 1})],$

$t_{2k+6} = w_1(p_{1\cdot} + p_{\cdot 1}) - 2p_{1\cdot}p_{\cdot 1} + 2w_1[(p_{1\cdot} + p_{\cdot 1}) - p_{11}][1 - (p_{1\cdot} + p_{\cdot 1})],$

$\vdots$

$\vdots$

$t_{2(k+1)^2-1} = 2w_k(p_{kk} - p_{k\cdot}p_{\cdot k}) + w_k(p_{k\cdot} + p_{\cdot k})[1 - (p_{k\cdot} + p_{\cdot k})],$

$t_{2(k+1)^2} = w_k(p_{k\cdot} + p_{\cdot k}) - 2p_{k\cdot}p_{\cdot k} + 2w_k[(p_{k\cdot} + p_{\cdot k}) - p_{kk}][1 - (p_{k\cdot} + p_{\cdot k})].$

In other words,

$$\sqrt{n}\left[\hat{\kappa}-\kappa\right] \xrightarrow[n\to\infty]{L} \mathrm{N}\left(0, d\Sigma d'\right). \tag{3.5}$$

If, on the other hand, w_i does depend on the data, for $i = 0, 1, \ldots, k$. w_i can be considered as a function with respect to p_{00}, p_{01}, $\cdots$, p_{kk}. Then κ can be considered as a function with respect to p_{00}, p_{01}, $\cdots$, p_{kk}.

Note that

- $\hat{p}_{00} \xrightarrow{P} \mathrm{E}(\mathrm{I}\{X = 0, Y = 0\}) = p_{00}$
- $\hat{p}_{01} \xrightarrow{P} \mathrm{E}(\mathrm{I}\{X = 0, Y = 1\}) = p_{01}$
- $\hat{p}_{02} \xrightarrow{P} \mathrm{E}(\mathrm{I}\{X = 0, Y = 2\}) = p_{02}$

$\vdots$

- $\hat{p}_{kk} \xrightarrow{P} \mathrm{E}(\mathrm{I}\{X = k, Y = k\}) = p_{kk}$.

Using similar arguments as before, we also have

$$\sqrt{n}\left[\hat{\kappa}-\kappa\right] \xrightarrow[n\to\infty]{L} \mathrm{N}\left(0, d\Sigma d'\right). \tag{3.6}$$

The actual form of d depends on the specification of the w_i, for $i = 0, 1, \ldots, k$. So no matter whether w_i depends on the data or not, we can get the same form of asymptotic distributions.

3.4 Choices of Weight

There are many weighting schemes that we can consider for the conditional kappa. For example, we can use equal weight, i.e., we let w_i to be the same for all $i = 0, 1, \dots, k$. We can also use the agreement weights, i.e., using each diagonal element as the weights. Other weighting schemes include the functions of the agreement weights, such as the cube of each diagonal weights, etc..

3.5 Simulation Study

3.5.1 Simulation Design

We used the same simulation design as in the last chapter and added one more case that has 6 categories. The distribution is

Case 16

		y						
		0	1	2	3	4	5	$P_X(x)$
	0	0.02	0.01	0.03	0.001	0.001	0.002	0.064
	1	0.05	0.08	0.01	0.005	0	0	0.145
	2	0.003	0.012	0.25	0	0.002	0	0.267
x	3	0.005	0	0	0.03	0	0.01	0.3
	4	0	0.05	0	0.2	0.001	0.001	0.252
	5	0.02	0.02	0.007	0.08	0	0.1	0.227
	$P_Y(y)$	0.098	0.172	0.297	0.316	0.004	0.113	1

We have six categories and medium agreement for this case. The diagonal elements add up to 0.481. The marginal probabilities are not very different from each other.

In the simulation studies, we calculated Cohen's kappa, the proposed conditional kappa using equal weight, agreement weights and the cube of agreement weights as well as their corresponding observed variances. Since the asymptotic formula becomes increasingly complicated as there are more categories in the responses, making it practically useless, we used bootstrap to estimate the large sample variances and compared the results with the observed variances.

3.5.2 Simulation Results

Table 3.1 gives the simulation results for case 1 to case 8 and table 3.2 gives the simulation results for case 9 to case 16.

Note: Here κ means Cohen's simple kappa, rm1, rm2 and rm3 is the proposed method using equal weight, the agreement weight(using each diagonal element as a weight) and the cube of the agreement weight(the cube of each diagonal weight), respectively. The last four columns are the corresponding observed variances.

Table 3.1. Simulation results for conditional kappa

case	n	κ	rm1	rm2	rm3	v_1	v_2	v_3	v_4
1	20	0.5079	0.3981	0.4331	0.4588	0.0271	0.0245	0.0243	0.0285
1	40	0.5067	0.3883	0.4088	0.4318	0.0129	0.0114	0.0115	0.0136
1	80	0.5078	0.3856	0.4005	0.4211	0.0060	0.0053	0.0052	0.0062
1	200	0.5137	0.3887	0.3997	0.4173	0.0025	0.0023	0.0023	0.0026
2	100	-0.2221	-0.1110	-0.1206	-0.1259	0.0023	0.0006	0.0013	0.0031
2	200	-0.2217	-0.1113	-0.1240	-0.1337	0.0012	0.0003	0.0007	0.0020
2	400	-0.2219	-0.1118	-0.1245	-0.1343	0.0006	0.0002	0.0003	0.0011
2	800	-0.2222	-0.1121	-0.1252	-0.1356	0.0003	0.0001	0.0002	0.0006
3	100	0.8581	0.7806	0.7910	0.8055	0.0021	0.0042	0.0038	0.0037
3	200	0.8614	0.7840	0.7933	0.8074	0.0009	0.0019	0.0017	0.0017
3	400	0.8607	0.7825	0.7914	0.8052	0.0005	0.0010	0.0009	0.0008
3	800	0.8606	0.7822	0.7908	0.8043	0.0002	0.0005	0.0004	0.0004
4	100	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000
4	200	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000
4	400	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000
4	800	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000
5	100	0.4540	0.3599	0.3450	0.3294	0.0041	0.0037	0.0036	0.0044
5	200	0.4532	0.3576	0.3398	0.3205	0.0019	0.0017	0.0017	0.0021
5	400	0.4544	0.3583	0.3393	0.3188	0.0010	0.0009	0.0008	0.0010
5	800	0.4539	0.3578	0.3377	0.3157	0.0004	0.0004	0.0004	0.0005
6	100	0.0166	0.0456	0.1071	0.1022	0.0038	0.0016	0.0037	0.0046
6	200	0.0177	0.0452	0.1019	0.0984	0.0019	0.0008	0.0019	0.0024
6	400	0.0179	0.0436	0.0983	0.0978	0.0009	0.0004	0.0009	0.0011
6	800	0.0180	0.0441	0.0975	0.0974	0.0005	0.0002	0.0005	0.0006
7	100	0.4423	0.3245	0.3719	0.4313	0.0037	0.0026	0.0034	0.0064
7	200	0.4427	0.3238	0.3637	0.4208	0.0018	0.0013	0.0015	0.0028
7	400	0.4432	0.3234	0.3628	0.4215	0.0010	0.0007	0.0008	0.0015
7	800	0.4437	0.3238	0.3609	0.4178	0.0005	0.0003	0.0004	0.0008
8	100	-0.0961	-0.0470	0.0143	0.0208	0.0015	0.0004	0.0011	0.0015
8	200	-0.0981	-0.0488	0.0088	0.0132	0.0007	0.0002	0.0005	0.0006
8	400	-0.0984	-0.0492	0.0067	0.0092	0.0004	0.0001	0.0002	0.0003
8	800	-0.0973	-0.0488	0.0066	0.0080	0.0002	0.0000	0.0001	0.0001

Table 3.2. Simulation results for conditional kappa (*continued*)

case	n	κ	rm1	rm2	rm3	v_1	v_2	v_3	v_4
9	100	0.3180	0.1409	0.3788	0.3697	0.0039	0.0008	0.0040	0.0050
9	200	0.3149	0.1382	0.3732	0.3625	0.0019	0.0004	0.0019	0.0023
9	400	0.3201	0.1400	0.3772	0.3656	0.0009	0.0002	0.0009	0.0011
9	800	0.3192	0.1394	0.3756	0.3626	0.0005	0.0001	0.0005	0.0006
10	100	-0.2577	-0.1037	-0.1718	-0.1774	0.0014	0.0002	0.0011	0.0018
10	200	-0.2577	-0.1043	-0.1740	-0.1842	0.0007	0.0001	0.0005	0.0009
10	400	-0.2583	-0.1051	-0.1750	-0.1880	0.0003	0.0000	0.0002	0.0004
10	800	-0.2579	-0.1049	-0.1757	-0.1907	0.0002	0.0000	0.0001	0.0002
11	100	1.0000	1.0000	1.0000	1.0000	0*	0.0000	0.0000	0.0000
11	200	1.0000	1.0000	1.0000	1.0000	0*	0.0000	0.0000	0.0000
11	400	1.0000	1.0000	1.0000	1.0000	0*	0.0000	0.0000	0.0000
11	800	1.0000	1.0000	1.0000	1.0000	0*	0.0000	0.0000	0.0000
12	100	0.6182	0.4575	0.5259	0.5786	0.0033	0.0035	0.0038	0.0050
12	200	0.6180	0.4556	0.5177	0.5673	0.0017	0.0018	0.0019	0.0024
12	400	0.6184	0.4558	0.5142	0.5628	0.0008	0.0008	0.0009	0.0012
12	800	0.6190	0.4558	0.5133	0.5611	0.0004	0.0004	0.0004	0.0006
13	100	-0.0532	-0.0175	0.0197	0.0327	0.0008	0.0002	0.0018	0.0038
13	200	-0.0541	-0.0181	0.0130	0.0272	0.0005	0.0001	0.0009	0.0026
13	400	-0.0536	-0.0182	0.0099	0.0242	0.0002	0.0001	0.0004	0.0015
13	800	-0.0536	-0.0183	0.0088	0.0232	0.0001	0.0000	0.0002	0.0009
14	100	0.7995	0.5652	0.7814	0.8379	0.0022	0.0022	0.0023	0.0025
14	200	0.8002	0.5653	0.7791	0.8369	0.0010	0.0011	0.0011	0.0013
14	400	0.7992	0.5639	0.7760	0.8345	0.0005	0.0005	0.0006	0.0006
14	800	0.8006	0.5643	0.7761	0.8348	0.0003	0.0003	0.0003	0.0003
15	100	1.0000	1.0000	1.0000	1.0000	0	0.0000	0.0000	0.0000
15	200	1.0000	1.0000	1.0000	1.0000	0	0.0000	0.0000	0.0000
15	400	1.0000	1.0000	1.0000	1.0000	0	0.0000	0.0000	0.0000
15	800	1.0000	1.0000	1.0000	1.0000	0	0.0000	0.0000	0.0000
16	100	0.3857	0.2446	0.5215	0.7048	0.0024	0.0013	0.0039	0.0082
16	200	0.3863	0.2445	0.5157	0.7098	0.0012	0.0007	0.0020	0.0038
16	400	0.3875	0.2450	0.5121	0.7109	0.0006	0.0004	0.0010	0.0021
16	800	0.3890	0.2459	0.5116	0.7140	0.0003	0.0002	0.0004	0.0009

3.5.3 Discussion

As can be seen from the simulation results, for the first few cases, the proposed method gives values smaller than Cohen's kappa, although for many cases this difference is not very significant. And equal weight almost does not work in any case. The observed variances for Cohen's kappa and the conditional kappa are very close to each other, when the sample size is large, the differences are almost negligible.

The only significant difference between kappa and the conditional kappa appears in the last case in which the proposed method gives very large estimates.

3.5.4 Bootstrap to Estimate Variances

Since the asymptotic variance becomes increasingly complicated the categories increase, we, instead, consider using bootstrap in obtaining the estimated variances of the estimated conditional kappa. We compared Cohen's kappa with the conditional kappa using the weights $w_i = P(X = i \text{ or } Y = i)$ for $i = 0, 1, \ldots, k$. We estimated Cohen's kappa, conditional kappa, and conditional kappa estimated by using bootstrap as well as the their corresponding variances. In addition, we also included the asymptotic variances for Cohen's kappa to compare its performance with that of the observed variances of Cohen's kappa and the conditional kappa.

We did 1000 simulations for each sample size in each distribution and 800 bootstrap replicates for each simulation. The observed variances of the proposed method and the mean bootstrap variances are compared.

Table 3.3. Bootstrap simulation result for conditional kappa

Case	n	κ	rm	rm_b	v_1	v_2	v_3	v_4
1	80	0.5110	0.3897	0.3903	0.0057	0.0051	0.0054	0.0061
1	160	0.5099	0.3870	0.3873	0.0030	0.0027	0.0027	0.0031
2	80	-0.2210	-0.0779	-0.1175	0.0031	0.0001	0.0007	0.0034
2	160	-0.2193	-0.1197	-0.1182	0.0015	0.0004	0.0004	0.0015
3	80	0.8568	0.7861	0.7859	0.0024	0.0046	0.0045	0.0024
3	160	0.8596	0.7887	0.7886	0.0012	0.0023	0.0023	0.0012
4	80	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0003
4	160	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0001
5	80	0.4523	0.3291	0.3301	0.0049	0.0041	0.0037	0.0047
5	160	0.4536	0.3288	0.3293	0.0024	0.0020	0.0019	0.0023
6	80	0.0191	0.0145	0.0167	0.0050	0.0019	0.0015	0.0046
6	160	0.0193	0.0139	0.0152	0.0023	0.0008	0.0008	0.0024
7	80	0.4425	0.3134	0.3143	0.0045	0.0036	0.0032	0.0047
7	160	0.4456	0.3147	0.3151	0.0024	0.0019	0.0016	0.0024
8	80	-0.0960	-0.0506	-0.0485	0.0018	0.0005	0.0004	0.0018
8	160	-0.0944	-0.0498	-0.0488	0.0009	0.0002	0.0002	0.0009
9	80	0.3181	0.2289	0.2312	0.0049	0.0033	0.0021	0.0048
9	160	0.3152	0.2256	0.2262	0.0023	0.0015	0.0010	0.0024
10	80	-0.2558	-0.1362	-0.1335	0.0017	0.0004	0.0004	0.0017
10	160	-0.2558	-0.1363	-0.1346	0.0008	0.0002	0.0002	0.0008
11	80	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0003
11	160	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0001
12	80	0.6199	0.4813	0.4822	0.0042	0.0048	0.0041	0.0039
12	160	0.6203	0.4798	0.4800	0.0019	0.0022	0.0021	0.0020
13	80	-0.0554	-0.0291	-0.0277	0.0012	0.0003	0.0003	0.0011
13	160	-0.0543	-0.0286	-0.0279	0.0006	0.0002	0.0001	0.0006
14	80	0.7979	0.6972	0.7204	0.0027	0.0047	0.0032	0.0026
14	160	0.7985	0.6964	0.6979	0.0012	0.0021	0.0015	0.0013
15	80	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000
15	160	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000
16	80	0.3871	0.2637	0.2649	0.0032	0.0023	0.0014	0.0030
16	160	0.3863	0.2618	0.2621	0.0016	0.0011	0.0007	0.0015

Note: κ means Cohen's kappa, rm is the conditional kappa using the third weight, rm_b is the mean of the proposed method using bootstrap, v_1 is the observed variance of Cohen's kappa, v_2 is the variance of the proposed method, v_3 is the mean of the bootstrap variances and v_4 is the mean of the asymptotic variances of Cohen's kappa.

3.5.5 Discussion

The results show that the asymptotic formula for Cohen's kappa has very good performance. When the sample size is 200, variances calculated using the asymptotic formula are almost equal to the observed variances. Generally speaking, the bootstrap method also works well in that the variances obtained using bootstrap are very close to the true observed variances. Moreover, the bootstrap estimates are very close to the observed values of the conditional kappa. And the variances for the conditional kappa are smaller than those of Cohen's kappa. In other words, bootstrap is a good alternative to the complicated asymptotic distribution formula.

Chapter 4

Multivariate Kappa

4.1 Multivariate Kappa

Let X and Y be categorical variables, with the categories being designated as $0, 1, 2, \cdots, k$. Let $p_{ij} = \Pr(X = i, Y = j)$, $i, j = 0, 1, 2, \cdots, k$, denote the bivariate probability and let $p_{i\cdot} = \Pr(X = i)$ and $p_{\cdot j} = \Pr(Y = j)$ denote the marginal probabilities.

Let $I(\cdot)$ denote the indicator function, and define the $(k+1) \times 1$ vectors as

$$Z_X = \begin{bmatrix} I(X=0) \\ I(X=1) \\ I(X=2) \\ \vdots \\ I(X=k) \end{bmatrix}$$

and

$$Z_Y = \begin{bmatrix} I(Y=0) \\ I(Y=1) \\ I(Y=2) \\ \vdots \\ I(Y=k) \end{bmatrix}.$$

Define the $(k+1) \times (k+1)$ matrices P_D and P_I as

$$P_D = E[(Z_X - Z_Y)(Z_X - Z_Y)^T]$$

and

$$P_I = E_{\text{Ind}}[(Z_X - Z_Y)(Z_X - Z_Y)^T],$$

where $E_{\text{Ind}}[\cdot]$ denotes expectation under the assumption that X and Y are independent.

Suppose A is a non-zero vector such that $A \in \mathbb{R}^{k+1}$. Then

$$\begin{aligned} A^{\mathrm{T}} P_D A &= A^{\mathrm{T}} E[(Z_X - Z_Y)(Z_X - Z_Y)^T] A \\ &= E[A^{\mathrm{T}}(Z_X - Z_Y)(Z_X - Z_Y)^T A] \\ &= E[(U_A)^2] \geq 0, \end{aligned}$$

where U_A is the random variable $(Z_X - Z_Y)^{\mathrm{T}} A$. Therefore, P_D is a nonnegative definite matrix, denoted by $P_D \geq 0$. Similarly, it can be shown that P_I is also a non-negative definite matrix, denoted by $P_I \geq 0$.

Next,

$$Z_X - Z_Y = \begin{bmatrix} I(X=0) - I(Y=0) \\ I(X=1) - I(Y=1) \\ I(X=2) - I(Y=2) \\ \vdots \\ I(X=k) - I(Y=k) \end{bmatrix}.$$

Straightforward calculations lead to

$$P_D = \begin{bmatrix} (p_{0\cdot} + p_{\cdot 0} - 2p_{00}) & (-p_{01} - p_{10}) & \cdots & (-p_{0k} - p_{k0}) \\ (-p_{01} - p_{10}) & (p_{1\cdot} + p_{\cdot 1} - 2p_{11}) & \cdots & (-p_{1k} - p_{k1}) \\ \vdots & \vdots & \ddots & \vdots \\ (-p_{0k} - p_{k0}) & (-p_{1k} - p_{k1}) & \cdots & (p_{k\cdot} + p_{\cdot k} - 2p_{kk}) \end{bmatrix}$$

and

$$P_I = \begin{bmatrix} (p_{0\cdot} + p_{\cdot 0} - 2p_{0\cdot}p_{\cdot 0}) & (-p_{0\cdot}p_{\cdot 1} - p_{1\cdot}p_{\cdot 0}) & \cdots & (-p_{0\cdot}p_{\cdot k} - p_{k\cdot}p_{\cdot 0}) \\ (-p_{0\cdot}p_{\cdot 1} - p_{1\cdot}p_{\cdot 0}) & (p_{1\cdot} + p_{\cdot 1} - 2p_{1\cdot}p_{\cdot 1}) & \cdots & (-p_{1\cdot}p_{\cdot k} - p_{k\cdot}p_{\cdot 1}) \\ \vdots & \vdots & \ddots & \vdots \\ (-p_{0\cdot}p_{\cdot k} - p_{k\cdot}p_{\cdot 0}) & (-p_{1\cdot}p_{\cdot k} - p_{k\cdot}p_{\cdot 1}) & \cdots & (p_{k\cdot} + p_{\cdot k} - 2p_{k\cdot}p_{\cdot k}) \end{bmatrix}.$$

Therefore, $P_I - P_D =$

$$\begin{bmatrix} 2(p_{00} - p_{0\cdot}p_{\cdot 0}) & (p_{01} - p_{0\cdot}p_{\cdot 1}) + (p_{10} - p_{1\cdot}p_{\cdot 0}) & \cdots & (p_{0k} - p_{0\cdot}p_{\cdot k}) + (p_{k0} - p_{k\cdot}p_{\cdot 0}) \\ (p_{01} - p_{0\cdot}p_{\cdot 1}) + (p_{10} - p_{1\cdot}p_{\cdot 0}) & 2(p_{11} - p_{1\cdot}p_{\cdot 1}) & \cdots & (p_{1k} - p_{1\cdot}p_{\cdot k}) + (p_{k1} - p_{k\cdot}p_{\cdot 1}) \\ \vdots & \vdots & \ddots & \vdots \\ (p_{0k} - p_{0\cdot}p_{\cdot k}) + (p_{k0} - p_{k\cdot}p_{\cdot 0}) & (p_{1k} - p_{1\cdot}p_{\cdot k}) + (p_{k1} - p_{k\cdot}p_{\cdot 1}) & \cdots & 2(p_{kk} - p_{k\cdot}p_{\cdot k}) \end{bmatrix}.$$

Before presenting the properties of $P_I - P_D$, we first exclude one scenario in which the joint distribution is a degenerate matrix. That is, if for some $i = 0, 1, 2, \cdots, k$, $p_{ii} = 1$ and $p_{mn} = 0$ for all $m \neq i, n \neq i$, then we say that the matrix representing the joint distribution is a degenerate matrix. For example, the following joint distribution matrix is a degenerate matrix:

		y 0	1	2	$P_X(x)$
	0	1	0	0	1
x	1	0	0	0	0
	2	0	0	0	0
	$P_Y(y)$	1	0	0	1

The difference, $P_I - P_D$, satisfies the following properties:

1. $-P_I \leq P_I - P_D \leq P_I$.

2. If X and Y are independent, then $P_I - P_D$ is a null matrix.

3. $X = Y$ with probability one, i.e., $p_{ij} = 0$ for all $i \neq j$, if and only if $P_I - P_D = P_I$.

4. Excluding the set of degenerate cases, $p_{ii} = 0$ for each $i = 0, 1, 2, \cdots, k$ and $p_{ij} = p_{ji} = 0.5$ for one choice of $i \neq j$ if and only if $P_I - P_D = -P_I$.

Proof.

1. The right inequality is self-evident since it means that P_D is a nonnegative matrix. To show that the left side inequality holds, we need to show that $2P_I - P_D \geq 0$. First note that

$$\begin{aligned}
P_I &= E_{\text{Ind}}[(Z_X - Z_Y)(Z_X - Z_Y)^T] \\
&= E_{\text{Ind}}[Z_X Z_X^{\mathrm{T}} - Z_X Z_Y^{\mathrm{T}} - Z_Y Z_X^{\mathrm{T}} + Z_Y Z_Y^{\mathrm{T}}] \\
&= (\Sigma_{XX} + \mu_X \mu_X^{\mathrm{T}}) - \mu_X \mu_Y^{\mathrm{T}} - \mu_Y \mu_X^{\mathrm{T}} + (\Sigma_{YY} + \mu_Y \mu_Y^{\mathrm{T}}) \\
&= (\Sigma_{XX} + \Sigma_{YY}) + (\mu_X - \mu_Y)(\mu_X - \mu_Y)^{\mathrm{T}}
\end{aligned}$$

and

$$\begin{aligned}
P_D &= E[(Z_X - Z_Y)(Z_X - Z_Y)^T] \\
&= E[Z_X Z_X^{\mathrm{T}} - Z_X Z_Y^{\mathrm{T}} - Z_Y Z_X^{\mathrm{T}} + Z_Y Z_Y^{\mathrm{T}}] \\
&= (\Sigma_{XX} + \mu_X \mu_X^{\mathrm{T}}) - (\Sigma_{XY} + \mu_X \mu_Y^{\mathrm{T}}) - (\Sigma_{YX} + \mu_Y \mu_X^{\mathrm{T}}) \\
&\quad + (\Sigma_{YY} + \mu_Y \mu_Y^{\mathrm{T}}) \\
&= (\Sigma_{XX} + \Sigma_{YY}) - \Sigma_{XY} - \Sigma_{YX} + (\mu_X - \mu_Y)(\mu_X - \mu_Y)^{\mathrm{T}}.
\end{aligned}$$

Then

$$\begin{aligned}
2P_I - P_D &= \Sigma_{XX} + \Sigma_{YY} + \Sigma_{XY} + \Sigma_{YX} + (\mu_X - \mu_Y)(\mu_X - \mu_Y)^{\mathrm{T}} \\
&= [I \quad I] \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix} \begin{bmatrix} I \\ I \end{bmatrix} + (\mu_X - \mu_Y)(\mu_X - \mu_Y)^{\mathrm{T}},
\end{aligned}$$

which is nonnegative definite because

$$\begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix}$$

is a covariance matrix and by definition, it is nonnegative definite. Since $(\mu_X - \mu_Y)(\mu_X - \mu_Y)^{\mathrm{T}}$ is also nonnegative definite, it follows that $2P_I + P_D$ is nonnegative definite because the sum of two nonnegative definite matrices is itself nonnegative definite, which completes the proof.

2. If X and Y are independent, then $p_{ij} = p_{i\cdot}p_{\cdot j}$, for $i, j = 0, 1, 2, \cdots, k$. It can easily be verified that $P_I - P_D$ is a null matrix.

Remark 2. *Note that $X = Y$ is only a sufficient but not a necessary condition for $P_I - P_D$ being a null matrix. Consider the following example:*

		y			
		0	*1*	*2*	$P_X(x)$
	0	0.04	0.08	0.08	0.2
x	*1*	0.16	0.36	0.08	0.6
	2	0	0.16	0.04	0.2
	$P_Y(y)$	0.2	0.6	0.2	*1*

Apparently, this satisfies the condition that $P_I - P_D$ is a null matrix. But X and Y are not independent.

3. If $P_I - P_D = P_I$, for each $i = 0, 1, 2, \cdots, k$, for $j = 0, 1, 2, \cdots, k$, where $j \neq i$, we have

$$p_{ij} - p_{i\cdot}p_{\cdot j} + p_{ji} - p_{j\cdot}p_{\cdot i} = -p_{i\cdot}p_{\cdot j} - p_{j\cdot}p_{\cdot i}$$

$$\Rightarrow p_{ij} + p_{ji} = 0$$

$$\Rightarrow p_{ij} = p_{ji} = 0.$$

So $p_{ij} = 0$ for $i \neq j$, that is, $X = Y$ with probability one.

Suppose, on the other hand, $p_{ij} = 0$ for $i \neq j$. It follows that $p_{ii} = p_{i\cdot} = p_{\cdot i}$. So for each $i = 0, 1, 2, \cdots, k$, $2(p_{ii} - p_{i\cdot}p_{\cdot i}) = p_{i\cdot} + p_{\cdot i} - 2p_{i\cdot}p_{\cdot i}$. Therefore, the diagonal elements of $P_I - P_D$ are the same as the corresponding ones of P_I. Since $p_{ij} = 0$ for all $i \neq j$, it can be easily verified that all the off-diagonal elements of $P_I - P_D$ are the same as the corresponding ones of P_I. Hence, $P_I - P_D = P_I$.

4. If $p_{ii} = 0$ for each $i = 0, 1, 2, \cdots, k$ and $p_{ij} = p_{ji} = 0.5$ for one choice of $i \neq j$, then $p_{i\cdot} = p_{j\cdot} = p_{\cdot i} = p_{\cdot j} = 0.5$ for some choice of $i \neq j$ and all other elements are 0. So for the specified i and j such that $p_{ij} = 0.5 = p_{ji}$, $2(p_{ii} - p_{i\cdot}p_{\cdot i}) = -0.5$. For the corresponding element in $-P_I$, it is $1[p_{i\cdot} + p_{\cdot i} - 2p_{i\cdot}p_{\cdot i}] = -0.5$. Therefore, the corresponding diagonal elements in both matrices are equal. To see that the off-diagonal elements are the same for both matrices, without loss of generality, we can assume that $p_{01} = p_{10} = 0.5$. Therefore, $p_{0\cdot} = p_{1\cdot} = p_{\cdot 0} = p_{\cdot 1}$. It can be easily verified that $(p_{01} - p_{0\cdot}p_{\cdot 1}) + (p_{10} - p_{1\cdot}p_{\cdot 0}) = p_{0\cdot}p_{\cdot 1} + p_{1\cdot}p_{\cdot 0}$. Because all the other elements are 0, we can then conclude that $P_I - P_D = -P_I$.

   Suppose, on the other hand, that $P_I - P_D = -P_I$, which implies that $2P_I = P_D$. Considering the i^{th} diagonal element of $2P_I$ and P_D, $i = 0, 1, 2, \cdots, k$, we get

$$2p_{ii} + p_{i\cdot} + p_{\cdot i} - 4p_{i\cdot}p_{\cdot i} = 0.$$

   Summing over all diagonal elements yields that

$$\sum_{i=0}^{k} p_{ii} = 2\sum_{i=0}^{k} p_{i\cdot}p_{\cdot i} - 1.$$

We want to show that

$$\sum_{i=0}^{k} p_{ii} = 0.$$

Let $A = [p_{0\cdot}\ p_{1\cdot}\ \cdots\ p_{k\cdot}]^{\mathrm{T}}$ and $B = [p_{\cdot 0}\ p_{\cdot 1}\ \cdots\ p_{\cdot k}]^{\mathrm{T}}$. Then

$$A^{\mathrm{T}}B = \sum_{i=0}^{k} p_{i\cdot}p_{\cdot i},$$

where the elements of A and B are nonnegative and $I^{\mathrm{T}}A = I^{\mathrm{T}}B = 1$. Differential calculus with a Lagrangian multiplier yields that $A^{\mathrm{T}}B$ is maximized when $A = B$, i.e., $p_{i\cdot} = p_{\cdot i}$, for $i = 0, 1, 2, \cdots, k$. If $p_{i\cdot} > 0$, then

$$\begin{aligned}
& 2p_{ii} + p_{i\cdot} + p_{\cdot i} - 4p_{i\cdot}p_{\cdot i} = 0 \\
\Rightarrow\ & 2p_{ii} + p_{i\cdot} - 4(p_{i\cdot}^{2}) = 0 \\
\Rightarrow\ & p_{ii} = -p_{i\cdot}(1 - 2p_{i\cdot}) \geq 0 \\
\Rightarrow\ & p_{i\cdot}(1 - 2p_{i\cdot}) \leq 0 \\
\Rightarrow\ & 1 - 2p_{i\cdot} \leq 0 \\
\Rightarrow\ & p_{i\cdot} \geq \frac{1}{2}.
\end{aligned}$$

Thus, $p_{i\cdot} = p_{\cdot i} = 0$ or $p_{i\cdot} = p_{\cdot i} \geq \frac{1}{2}$ for each $i = 0, 1, 2, \cdots, k$. Because $p_{0\cdot} + p_{1\cdot} + \cdots + p_{k\cdot} = 1$ and $p_{\cdot 0} + p_{\cdot 1} + \cdots + p_{\cdot k} = 1$, however, this means that there are exactly two choices of i for which $p_{i\cdot} = p_{\cdot i} = \frac{1}{2}$ and for which $p_{i\cdot} = p_{\cdot i} = 0$ for all other values of i (note that the degenerate cases of $p_{i\cdot} = p_{\cdot i} = 1$ for any i are excluded). Finally,

$$\begin{aligned}\sum_{i=0}^{k} p_{ii} &= 2\sum_{i=0}^{k} p_{i\cdot}p_{\cdot i} - 1 \\ &= 2(\frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2}) - 1 = 0,\end{aligned}$$

indicating that $p_{ii} = 0$ for each $i = 0, 1, 2, \cdots, k$.

Next, the above proof indicates that $\sum_{i=0}^{k} p_{i\cdot}p_{\cdot i} = \frac{1}{2}$. Furthermore, because $p_{ii} = 0$ for each $i = 0, 1, 2, \cdots, k$, $0 \leq p_{i\cdot} + p_{\cdot i} \leq 1$ and $2p_{ii} + p_{i\cdot} + p_{\cdot i} - 4p_{i\cdot}p_{\cdot i} = 0 \Rightarrow p_{i\cdot} + p_{\cdot i} - 4p_{i\cdot}p_{\cdot i} = 0$, this yields that $p_{i\cdot}p_{\cdot i} \leq \frac{1}{4}$. It is obvious that for each i, $(p_{i\cdot}, p_{\cdot i}) = (0, 0)$ and $(\frac{1}{2}, \frac{1}{2})$ are solutions to $p_{i\cdot} + p_{\cdot i} - 4p_{i\cdot}p_{\cdot i} = 0$. There are no other solutions $(p_{i\cdot}, p_{\cdot i})$ that satisfy $0 \leq p_{i\cdot} + p_{\cdot i} \leq 1$ and $p_{i\cdot}p_{\cdot i} \leq \frac{1}{4}$. In particular,

- if $0 \leq p_{i\cdot} \leq \frac{1}{4}$, then $p_{\cdot i} \leq 0$,
- if $\frac{1}{4} < p_{i\cdot} \leq \frac{1}{3}$, then $p_{\cdot i} \geq 1$,
- if $\frac{1}{3} < p_{i\cdot} < \frac{1}{2}$, then $p_{i\cdot}p_{\cdot i} > \frac{1}{4}$,
- if $\frac{1}{2} < p_{i\cdot} < 1$, then $p_{i\cdot} + p_{\cdot i} > 1$.

Therefore, there are only two values of i for which $(p_{i\cdot}, p_{\cdot i}) = (\frac{1}{2}, \frac{1}{2})$, indicating that $p_{ii} = 0$ for each $i = 0, 1, 2, \cdots, k$ and $p_{ij} = p_{ji} = 0.5$ for one choice of $i \neq j$, which completes the proof.

□

These results indicate that we could use the matrix $P_I - P_D$ to assess agreement between the variable X and Y. If $P_I - P_D$ is "close" to P_I, then there is strong agreement between X and Y. If $P_I - P_D$ is "close" to the null matrix, then there is no agreement beyond chance. If $P_I - P_D$ is "close" to $-P_I$, then there is strong disagreement. Instead of examining the matrix $P_I - P_D$, however, it may be simpler to construct a scalar quantity that reflects agreement. Therefore, let W be a $(k+1) \times (k+1)$ symmetric matrix of weights with $w_{ij} = 1$ for $i = 0, 1, 2, \cdots, k$ and $0 \leq w_{ij} < 1$ for $i \neq j = 0, 1, 2, \cdots, k$. Then we construct the following coefficient:

$$\begin{aligned} \kappa(W) &= \frac{\mathrm{tr}\{W(P_I - P_D)\}}{\mathrm{tr}\{WP_I\}} \\ &= 1 - \frac{\mathrm{tr}\{WP_D\}}{\mathrm{tr}\{WP_I\}}. \end{aligned} \tag{4.1}$$

If W is the $(k+1) \times (k+1)$ identity matrix, then $\kappa(W)$ reduces to Cohen's kappa coefficient and other choices of W will lead to a weighted kappa coefficient (this will be shown later). Two popular weighting schemes for weighted kappa are the Cicchetti-Allison weights:

$$w_{ij} = 1 - \frac{|i-j|}{k}, \quad \text{for } i, j = 0, 1, 2, \cdots, k,$$

and Fleiss-Cohen weights:

$$w_{ij} = 1 - \frac{(i-j)^2}{k^2}, \quad \text{for } i, j = 0, 1, 2, \cdots, k.$$

Lemma 4.1.1. *The weighting matrix W using Cicchetti-Allison weights is nonnegative definite.*

Proof. First suppose we employ Cicchetti-Allison weights. It can be seen easily that W is the correlation matrix using the following popular nonnegative linear correlation function with $\theta = k$ and $d = i - j$:

$$R(d) = 1 - \frac{1}{\theta}|d|,$$

where $|d| \leq \theta$.

Therefore, W must be nonnegative definite.

Remark 3. *If we use Fleiss-Cohen weights, however, it is not always nonnegative definite. For example, let $k = 2$, then*

$$W = \begin{bmatrix} 1 & 0.75 & 0 \\ 0.75 & 1 & 0.75 \\ 0 & 0.75 & 1 \end{bmatrix}.$$

Apparently, this is not a nonnegative definite matrix (because its determinant is -0.125).

□

The statistic, $\kappa(W)$, satisfies the following properties:

1. $-1 \leq \kappa(W) \leq 1$.

2. If X and Y are independent, then $\kappa(W) = 0$.

3. If X=Y with probability one, then $\kappa(W) = 1$.

4. If $p_{ii} = 0$ for each $i = 0, 1, 2, \cdots, k$ and $p_{ij} = p_{ji} = 0.5$ for one choice of $i \neq j$, then $\kappa(W) = -1$.

Proof.

1. Straightforward calculation shows that

$$\begin{aligned} \text{tr}\{WP_D\} &= 2 - 2\sum_{i=0}^{k} p_{ii} - 2\sum\sum_{i \neq j} w_{ij} p_{ij} \\ &\geq 2 - 2\sum_{i=0}^{k} p_{ii} - 2\sum\sum_{i \neq j} p_{ij} \\ &= 2 - 2\sum_{i=0}^{k}\sum_{j=0}^{k} p_{ij} \\ &= 2 - 2 = 0. \end{aligned}$$

Similarly,

$$\begin{aligned}
\text{tr}\{WP_I\} &= 2 - 2\sum_{i=0}^{k} p_{i\cdot}p_{\cdot i} - 2\sum\sum_{i\neq j} w_{ij}p_{i\cdot}p_{\cdot j} \\
&> 2 - 2\sum_{i=0}^{k} p_{i\cdot}p_{\cdot i} - 2\sum\sum_{i\neq j} p_{i\cdot}p_{\cdot j} \\
&= 2 - 2\sum_{i=0}^{k}\sum_{j=0}^{k} p_{i\cdot}p_{\cdot j} \\
&= 2 - 2 = 0.
\end{aligned}$$

Therefore,

$$\kappa(W) = 1 - \frac{\text{tr}\{WP_D\}}{\text{tr}\{WP_I\}} \leq 1.$$

To show that the left-side inequality holds, we need to show that

$$\begin{aligned}
& \frac{\text{tr}\{WP_D\}}{\text{tr}\{WP_I\}} \geq 2 \\
\Rightarrow \quad & \frac{2 - 2\sum_{i=0}^{k} p_{ii} - 2\sum\sum_{i\neq j} w_{ij}p_{ij}}{2 - 2\sum_{i=0}^{k} p_{i\cdot}p_{\cdot i} - 2\sum\sum_{i\neq j} w_{ij}p_{i\cdot}p_{\cdot j}} \geq 2 \\
\Rightarrow \quad & 1 + \sum_{i=0}^{k} p_{ii} + \sum\sum_{i\neq j} w_{ij}p_{ij} \geq 2\sum_{i=0}^{k} p_{i\cdot}p_{\cdot i} + 2\sum\sum_{i\neq j} w_{ij}p_{i\cdot}p_{\cdot j}.
\end{aligned}$$

If we use Cicchetti-Allison weights, then

$$
\begin{aligned}
1+\sum_{i=0}^{k} p_{ii}+\sum\sum_{i\neq j} w_{ij}p_{ij} &= 1+\sum_{i=0}^{k} p_{ii}+\sum\sum_{i\neq j}(1-\frac{|i-j|}{k})p_{ij} \\
&= 2-\sum\sum_{i\neq j}\frac{|i-j|}{k}p_{ij}
\end{aligned}
$$

and

$$
\begin{aligned}
2\sum_{i=0}^{k} p_{i\cdot}p_{\cdot i}+2\sum\sum_{i\neq j} w_{ij}p_{i\cdot}p_{\cdot j} &= 2\sum_{i=0}^{k} p_{i\cdot}p_{\cdot i}+2\sum\sum_{i\neq j}(1-\frac{|i-j|}{k})p_{i\cdot}p_{\cdot j} \\
&= 2-\sum\sum_{i\neq j}\frac{|i-j|}{k}p_{i\cdot}p_{\cdot j}.
\end{aligned}
$$

Therefore, we need only to show that

$$
\begin{aligned}
& 2-\sum\sum_{i\neq j}\frac{|i-j|}{k}p_{ij} \geq 2-\sum\sum_{i\neq j}\frac{|i-j|}{k}p_{i\cdot}p_{\cdot j} \\
\Rightarrow\ & \sum\sum_{i\neq j}|i-j|p_{ij} \leq \sum\sum_{i\neq j}|i-j|p_{i\cdot}p_{\cdot j}.
\end{aligned}
$$

We can use the same statistic as constructed in chapter 1:

$$
e_d = 1-\frac{\sum\sum_{x\neq y} P_{XY}(x,y)d(x,y)}{\sum\sum_{x\neq y} P_X(x)P_Y(y)d(x,y)}.
$$

Using a method similar to Fay's (2006) argument, we can multiply the numerator and the denominator by a constant, say c, and this will not change e_d. Now denote $c \cdot d(x,y)$ by $d^*(x,y)$. The c is chosen such that $\max_{x,y} d^*(x,y) = 1$. And let $w_{xy} = 1 - d^*(x,y)$. For perfect agreement, $w_{xy} = 1$; for all $x \neq y$, $0 \leq w_{xy} < 1$.

Note that e_d can also be written as

$$
\begin{aligned}
e_d &= 1 - \frac{\sum\sum_{x \neq y} P_{XY}(x,y) d(x,y)}{\sum\sum_{x \neq y} P_X(x) P_Y(y) d(x,y)} \\
&= 1 - \frac{\sum_x \sum_y P_{XY}(x,y) d(x,y)}{\sum_x \sum_y P_X(x) P_Y(y) d(x,y)},
\end{aligned}
$$

since $d(x,y) = 0$ for all $x = y$.

Now we can rewrite e_d as

$$
e_d = 1 - \frac{\sum_x \sum_y P_{XY}(x,y) d^*(x,y)}{\sum_x \sum_y P_X(x) P_Y(y) d^*(x,y)}.
$$

Since $d^*(x,y) = 1 - w_{xy}$, so

$$
\begin{aligned}
e_d &= 1 - \frac{\sum_x \sum_y P_{XY}(x,y)(1-w_{xy})}{\sum_x \sum_y P_X(x)P_Y(y)(1-w_{xy})} \\
&= 1 - \frac{1 - \sum_x \sum_y w_{xy} P_{XY}(x,y)}{1 - \sum_x \sum_y w_{xy} P_X(x)P_Y(y)} \\
&= \frac{\Pi_o - \Pi_e}{1 - \Pi_e}.
\end{aligned}
$$

So e_d is actually the same as weighted kappa with $\Pi_o = \sum_x \sum_y w_{xy} P_{XY}(x,y)$ and $\Pi_e = \sum_x \sum_y w_{xy} P_X(x)P_Y(y)$.

Hence,

$$
\begin{aligned}
& 1 - \frac{\sum\sum_{x \neq y} P_{XY}(x,y) d(x,y)}{\sum\sum_{x \neq y} P_X(x)P_Y(y) d(x,y)} \geq -1 \\
\Rightarrow \quad & \sum\sum_{x \neq y} P_{XY}(x,y) d(x,y) \leq 2 \sum\sum_{x \neq y} P_X(x)P_Y(y) d(x,y).
\end{aligned}
$$

Take the distance function $d(x,y)$ as

$$
\text{d(x,y)=} \begin{cases} |x-y|, & \text{if } x \neq y \\ 0, & \text{if } x = y \end{cases}
$$

and this completes the proof.

Similarly, if we use Fleiss-Cohen weights, then

$$
\begin{aligned}
1+\sum_{i=0}^{k} p_{ii}+\sum\sum_{i\neq j} w_{ij}p_{ij} &= 1+\sum_{i=0}^{k} p_{ii}+\sum\sum_{i\neq j}(1-\frac{(i-j)^2}{k^2})p_{ij} \\
&= 2-\sum\sum_{i\neq j}\frac{(i-j)^2}{k^2}p_{ij}
\end{aligned}
$$

and

$$
\begin{aligned}
2\sum_{i=0}^{k} p_{i\cdot}p_{\cdot i}+2\sum\sum_{i\neq j} w_{ij}p_{i\cdot}p_{\cdot j} &= 2\sum_{i=0}^{k} p_{i\cdot}p_{\cdot i}+2\sum\sum_{i\neq j}(1-\frac{(i-j)^2}{k^2})p_{i\cdot}p_{\cdot j} \\
&= 2-\sum\sum_{i\neq j}\frac{(i-j)^2}{k^2}p_{i\cdot}p_{\cdot j}.
\end{aligned}
$$

Therefore, we need only to show that

$$
\begin{aligned}
& 2-\sum\sum_{i\neq j}\frac{(i-j)^2}{k^2}p_{ij} \geq 2-\sum\sum_{i\neq j}\frac{(i-j)^2}{k^2}p_{i\cdot}p_{\cdot j} \\
\Rightarrow \quad & \sum\sum_{i\neq j}(i-j)^2 p_{ij} \leq \sum\sum_{i\neq j}(i-j)^2 p_{i\cdot}p_{\cdot j}.
\end{aligned}
$$

Take the distance function $d(x,y)$ as

$$
\text{d(x,y)}=\begin{cases} (x-y)^2, & \text{if } x\neq y \\ 0, & \text{if } x=y \end{cases}
$$

and using a similar procedure, it can be easily verified that the inequality holds, which completes the proof.

Remark 4. *From the proof, it is obvious that if the weight function has the generic form* $w_{ij} = 1 - d(i, j)$*, for* $i, j = 1, 2, \cdots, k$*, then the inequality* $-1 \le \kappa(W) \le 1$ *always holds.*

2. Similarly, from a previous proof, we know that if X and Y are independent, then $P_I = P_D$. Hence $WP_I = WP_D$. Therefore, $\text{tr}\{WP_D\} = \text{tr}\{WP_I\}$. Hence, $\kappa(W) = 0$.

3. From a previous proof, we know that if $X = Y$ with probability one, then $P_I - P_D = P_I \Rightarrow P_D = 0 \Rightarrow WP_D = 0 \Rightarrow \text{tr}\{WP_D\} = 0 \Rightarrow \kappa(W) = 1$.

4. From a previous proof, we know that $p_{ii} = 0$ for each $i = 0, 1, 2, \cdots, k$ and $p_{ij} = p_{ji} = 0.5$ for one choice of $i \ne j$, then $P_I - P_D = -P_I \Rightarrow P_D = 2P_I \Rightarrow WP_D = 2WP_I \Rightarrow \text{tr}\{WP_D\} = 2\text{tr}\{WP_I\} \Rightarrow \kappa(W) = -1$.

□

If we set W equal to the $(k+1) \times (k+1)$ identity matrix I, then

$$\begin{aligned} \text{tr}\{P_D\} &= \sum_{i=0}^{k} p_{i\cdot} + \sum_{i=0}^{k} p_{\cdot i} - 2\sum_{i=0}^{k} p_{ii} \\ &= 2 - 2\sum_{i=0}^{k} p_{ii} \end{aligned}$$

and

$$\begin{aligned} \mathrm{tr}\{P_I\} &= \sum_{i=0}^{k} p_{i\cdot} + \sum_{i=0}^{k} p_{\cdot i} - 2\sum_{i=0}^{k} p_{i\cdot}p_{\cdot i} \\ &= 2 - 2\sum_{i=0}^{k} p_{i\cdot}p_{\cdot i}. \end{aligned}$$

Therefore,

$$\begin{aligned} \kappa(W) &= 1 - \frac{\mathrm{tr}\{P_D\}}{\mathrm{tr}\{P_I\}} \\ &= 1 - \frac{2 - 2\sum_{i=0}^{k} p_{ii}}{2 - 2\sum_{i=0}^{k} p_{i\cdot}p_{\cdot i}} \\ &= \frac{\sum_{i=0}^{k} p_{ii} - \sum_{i=0}^{k} p_{i\cdot}p_{\cdot i}}{1 - \sum_{i=0}^{k} p_{i\cdot}p_{\cdot i}} \\ &= \frac{p_o - p_c}{1 - p_c}, \end{aligned}$$

which is the standard form for Cohen's kappa.

For other choices of a weight matrix W, straightforward calculation shows that

$$\begin{aligned} \mathrm{tr}\{WP_D\} &= 2\sum_{i=0}^{k} w_{ii}(p_{i\cdot} + p_{\cdot i}) - 2\sum_{i=0}^{k}\sum_{j=0}^{k} w_{ij}p_{ij} \\ &= 2 - 2\sum_{i=0}^{k}\sum_{j=0}^{k} w_{ij}p_{ij} \end{aligned}$$

and

$$\begin{aligned}\text{tr}\{WP_I\} &= 2\sum_{i=0}^{k} w_{ii}(p_{i\cdot} + p_{\cdot i}) - 2\sum_{i=0}^{k}\sum_{j=0}^{k} w_{ij}p_{ij} \\ &= 2 - 2\sum_{i=0}^{k}\sum_{j=0}^{k} w_{ij}p_{i\cdot}p_{\cdot j}.\end{aligned}$$

Therefore,

$$\begin{aligned}\kappa(W) &= 1 - \frac{\text{tr}\{WP_D\}}{\text{tr}\{WP_I\}} \\ &= \frac{\sum\limits_{i=0}^{k}\sum\limits_{j=0}^{k} w_{ij}p_{ij} - \sum\limits_{i=0}^{k}\sum\limits_{j=0}^{k} w_{ij}p_{i\cdot}p_{\cdot j}}{1 - \sum\limits_{i=0}^{k}\sum\limits_{j=0}^{k} w_{ij}p_{i\cdot}p_{\cdot j}} \\ &= \frac{\prod_o - \prod_e}{1 - \prod_e},\end{aligned}$$

where $\prod_o = \sum\limits_{i=0}^{k}\sum\limits_{j=0}^{k} w_{ij}p_{ij}$ and $\sum\limits_{i=0}^{k}\sum\limits_{j=0}^{k} w_{ij}p_{i\cdot}p_{\cdot j}$. This is the standard form for Cohen's weighted kappa.

It is important that the weight matrix W in the construction of $\kappa(W)$ be nonnegative definite. If W is nonnegative definite, then it has a square-root decomposition, denoted as $W = W^{\frac{1}{2}}W^{\frac{1}{2}}$. Then

$$\begin{aligned}\kappa(W) &= 1 - \frac{\text{tr}(WP_D)}{\text{tr}(WP_I)} \\ &= 1 - \frac{\text{tr}(W^{\frac{1}{2}}P_DW^{\frac{1}{2}})}{\text{tr}(W^{\frac{1}{2}}P_IW^{\frac{1}{2}})}.\end{aligned}$$

Because P_D and P_I both are nonnegative definite, then $W^{\frac{1}{2}}P_DW^{\frac{1}{2}}$ and $W^{\frac{1}{2}}P_IW^{\frac{1}{2}}$ both are nonnegative definite. If, for example, $P_I \geq P_D$, then $\mathrm{tr}(WP_I) \geq \mathrm{tr}(WP_D)$. If W is not nonnegative definite, however, then $P_I \geq P_D$ does not imply that $\mathrm{tr}(WP_I) \geq \mathrm{tr}(WP_D)$.

An example is as follows. Let

$$A = \begin{bmatrix} 0.3751 & -0.5303 & 0.3750 \\ -0.5303 & 0.7501 & -0.5303 \\ 0.3750 & -0.5303 & 0.3751 \end{bmatrix}$$

and

$$B = \begin{bmatrix} 0.1250 & -0.1768 & 0.1250 \\ -0.1768 & 0.25 & -0.1768 \\ 0.1250 & -0.1768 & 0.1250 \end{bmatrix}.$$

It is straightforward to demonstrate that $A \geq B \geq 0$. Therefore, we would like $\mathrm{tr}(WA) \geq \mathrm{tr}(WB)$ for a symmetric weight matrix W. If we use the Fleiss-Cohen weight matrix, however,

$$W = \begin{bmatrix} 1 & 0.75 & 0 \\ 0.75 & 1 & 0.75 \\ 0 & 0.75 & 1 \end{bmatrix}.$$

Then

$$WA = \begin{bmatrix} -0.0226 & 0.0323 & -0.0227 \\ 0.0323 & -0.0454 & 0.0323 \\ -0.0227 & 0.0323 & -0.0226 \end{bmatrix}$$

and $\text{tr}(WA) = -0.0906$.

$$WB = \begin{bmatrix} -0.0076 & 0.0107 & -0.0076 \\ 0.0107 & -0.0152 & 0.0107 \\ -0.0076 & 0.0107 & -0.0076 \end{bmatrix}$$

and $\text{tr}(WB) = -0.0304$. Thus, although $A \geq B$, $\text{tr}(WA) \leq \text{tr}(WB)$. For this reason, the Fleiss-Cohen weight matrix is not recommended for the construction of $\kappa(W)$.

4.2 A Generalization

4.2.1 Definition

The coefficient

$$\kappa(W) = 1 - \frac{\text{tr}(WP_D)}{\text{tr}(WP_I)}$$

suggests other constructions of the form

$$\kappa_g(W) = 1 - \frac{g(WP_D)}{g(WP_I)}, \tag{4.2}$$

where g is a matrix function that satisfies the following definitions.

Definition 4.2.1. *Suppose $g(\cdot)$ is a $\mathbb{R}^{(k+1)\times(k+1)} \mapsto \mathbb{R}$ function. $g(\cdot)$ is said to be non-decreasing if $\forall$ nonnegative definite matrices $A_{(k+1)\times(k+1)}$, $B_{(k+1)\times(k+1)}$, where $A \leq B$, we have $g(A) \leq g(B)$.*

Definition 4.2.2. *Suppose $g(\cdot)$ is a $\mathbb{R}^{(k+1)\times(k+1)} \mapsto \mathbb{R}$ function. $g(\cdot)$ is said to be a scale-equivariant function if $\forall$ matrices $A_{(k+1)\times(k+1)}$ and constants c yield that $g(cA) = cg(A)$.*

Remark 5. *Definition 4.2.2 implies that $g(0) = 0$ and that $g(-A) = -g(A)$, i.e., g is an odd function.*

$g(\cdot)$ should be chosen such that it is a non-decreasing and scale-equivariant function.

The agreement coefficient $k_g(W)$ satisfies the same four properties as $\kappa(W)$ above.

Proof.

1. Since both P_I and P_D are nonnegative matrices, it follows that all the eigenvalues are nonnegative. Hence, $\sum_{i=0}^{k} v_i\lambda_i \geq 0$. Therefore, $g(WP_D) \geq 0$, $g(WP_I) > 0$. So $k_g(W) \leq 1$. From the previous proof, we know that $2P_I - P_D$ is nonnegative definite. Using similar argument, we can get $g(2P_I - P_D) \geq 0$. It then follows that $2g(WP_I) \geq g(WP_D) \Rightarrow \frac{g(WP_D)}{g(WP_I)} \leq 2 \Rightarrow k_g(W) \geq -1$.

2. Similarly, from the previous proof, we know that if X and Y are independent, then $P_I = P_D$. Hence $WP_I = WP_D$. Therefore, $g(WP_D) = g(WP_I)$. Hence, $\kappa_g(W) = 0$.

3. From the previous proof, we know that if $X = Y$ with probability one, then $P_I - P_D = P_I \Rightarrow P_D = 0 \Rightarrow WP_D = 0 \Rightarrow g(WP_D) = 0 \Rightarrow \kappa_g(W) = 1$.

4. From the previous proof, we know that if $p_{ii} = 0$ for each $i = 0, 1, 2, \cdots, k$ and $p_{ij} = p_{ji} = 0.5$ for one choice of $i \neq j$, then $P_I - P_D = -P_I \Rightarrow P_D = 2P_I \Rightarrow WP_D = 2WP_I \Rightarrow g(WP_D) = g(2WP_I) = 2g(WP_I) \Rightarrow \kappa_g(W) = -1$.

□

Clearly, $g(A) = \text{tr}(A)$ satisfies definitions 4.2.1 and 4.2.2. Therefore, Cohen's kappa $\kappa = 1 - \frac{\text{tr}(P_D)}{\text{tr}(P_I)}$ and Cohen's weighted kappa $\kappa(W) = 1 - \frac{\text{tr}(WP_D)}{\text{tr}(WP_I)}$, where W is nonnegative definite, are appropriate coefficients of agreement.

Another function to consider is $g(A) = \lambda_{A_1}$, the largest eigenvalue of A.

Remark 6. *The function $g(A) = \lambda_{A_1}$ satisfies definitions 4.2.1 and 4.2.2.*

Proof. 1. Suppose A and B are $(k+1) \times (k+1)$ nonnegative definite matrices with $A \leq B$. Then

$$\begin{aligned}
\lambda_{A_1} &= \sup_{\substack{d \in \mathbb{R}^{(k+1)} \\ d'd=1}} (d'Ad) \\
&\leq \sup_{\substack{d \in \mathbb{R}^{(k+1)} \\ d'd=1}} (d'Ad) + d'(B-A)d \text{ (Because B$-$A is nonnegative definite)} \\
&= \sup_{\substack{d \in \mathbb{R}^{(k+1)} \\ d'd=1}} (d'Bd) \\
&= \lambda_{B_1}.
\end{aligned}$$

Thus, $\lambda_{A_1} \leq \lambda_{B_1}$ when $A \leq B$, which means that $g(A) \leq g(B)$. So g being the largest eigenvalue is a nondecreasing matrix function.

2. Also, suppose $g(A) = \lambda_{A_1}$, where λ_{A_1} is the largest eigenvalue of A. Then $g(cA)$ is equal to the largest eigenvalue of cA, where c is a constant. And it is equal to $c \times \lambda_{A_1} = c \times g(A)$. That is, $g(cA) = cg(A)$. So it is scale-equivariant.

□

Although other matrix functions exist that satisfy definitions 4.2.1 and 4.2.2, they are extremely complex and not always applicable. Therefore, we focus on the functions $g_{\text{tr}}(A) = \text{tr}(A) =$ sum of the eigenvalues of A and $g_{\text{le}}(A) = \lambda_{A_1} =$ largest eigenvalues of A. In the context of our agreement coefficients, where we will be using $g(WP_D)$ and $g(WP_I)$, and W is nonnegative definite, our choices are:

1. the mean of the eigenvalues because we can substitute $\frac{1}{k+1}\text{tr}(A)$ for the $\text{tr}(A)$;

2. The range of the eigenvalues because the smallest eigenvalue of WP_D and WP_I is zero, so the largest eigenvalue−the smallest eigenvalue=range.

Finally, we could construct an entire class of matrix functions g that satisfy definitions 4.2.1 and 4.2.2 of weighted averages of the sum of the eigenvalues and the largest eigenvalue. Let $g_{\text{tr}}(A) = \text{tr}(A) =$ sum of the eigenvalues and $g_{\text{le}}(A) =$ largest eigenvalue of A. Let δ be a real number such that $0 \leq \delta \leq 1$. Then

$$g_{\delta}(A) = \delta g_{\text{tr}}(A) + (1 - \delta) g_{\text{le}}(A) \tag{4.3}$$

satisfies definitions 4.2.1 and 4.2.2.

In such a case,

$$\kappa_{g_\delta} = 1 - \frac{\delta * \mathrm{tr}(W * P_D) + (1-\delta) * \text{largest eigenvalue of } (W * P_D)}{\delta * \mathrm{tr}(W * P_I) + (1-\delta) * \text{largest eigenvalue of } (W * P_I)}. \tag{4.4}$$

So when $\delta = 1$, this becomes Cohen's weighted kappa coefficient. When $\delta = 0$, this becomes $\kappa_{\mathrm{le}}(W)$. For simplicity, however, we focus on the cases $\delta = 1$ (sum of the eigenvalues) and $\delta = 0$ (largest eigenvalue). The relationship between this coefficient and Cohen's kappa coefficient will be shown in the plots in the next section.

The asymptotic distribution of

$$\hat{\kappa}_{\mathrm{le}}(W) = 1 - \frac{\text{largest eigenvalue}(W\hat{P}_D)}{\text{largest eigenvalue}(W\hat{P}_I)}$$

is extremely complex, unfortunately, so we will use the bootstrap to estimate the large sample variances for $\kappa_{\mathrm{le}}(W)$.

4.2.2 Comparison of κ with κ_{g_δ}

In this section, we construct several plots to compare the Cohen's kappa agreement coefficients with the coefficients obtained using the generalization method. For your convenience, we include here again the probability distributions on which the plots are based.

The probability distribution for case 1 is:

		y			
		0	1	2	$P_X(x)$
	0	0.2	0.05	0.05	0.3
x	1	0.03	0.3	0.07	0.4
	2	0.05	0.07	0.18	0.3
	$P_Y(y)$	0.28	0.42	0.3	1

The probability distribution for case 2 is:

		y			
		0	1	2	$P_X(x)$
	0	0.05	0.1	0.05	0.2
x	1	0.22	0.05	0.03	0.3
	2	0.13	0.35	0.02	0.5
	$P_Y(y)$	0.4	0.5	0.1	1

The probability distribution for case 4 is:

		y = 0	1	2	$P_X(x)$
	0	0.2	0	0	0.2
x	1	0	0.3	0	0.3
	2	0	0	0.5	0.5
	$P_Y(y)$	0.2	0.3	0.5	1

The probability distribution for case 9 is:

		y = 0	1	2	3	$P_X(x)$
	0	0	0	0.1	0.05	0.15
x	1	0	0	0.08	0.02	0.1
	2	0.05	0.03	0.3	0.01	0.39
	3	0.06	0.02	0.03	0.25	0.36
	$P_Y(y)$	0.11	0.05	0.51	0.33	1

The probability distribution for case 12 is:

		y					
		0	1	2	3	4	$P_X(x)$
	0	0.2	0.01	0	0.01	0	0.22
	1	0	0.1	0	0.04	0.02	0.16
x	2	0.08	0	0.15	0	0.01	0.24
	3	0.01	0.04	0	0.2	0	0.25
	4	0	0.06	0	0.02	0.05	0.13
	$P_Y(y)$	0.29	0.21	0.15	0.27	0.08	1

The probability distribution for case 16 is:

		y						
		0	1	2	3	4	5	$P_X(x)$
	0	0.02	0.01	0.03	0.001	0.001	0.002	0.064
	1	0.05	0.08	0.01	0.005	0	0	0.145
	2	0.003	0.012	0.25	0	0.002	0	0.267
x	3	0.005	0	0	0.03	0	0.01	0.3
	4	0	0.05	0	0.2	0.001	0.001	0.252
	5	0.02	0.02	0.007	0.08	0	0.1	0.227
	$P_Y(y)$	0.098	0.172	0.297	0.316	0.004	0.113	1

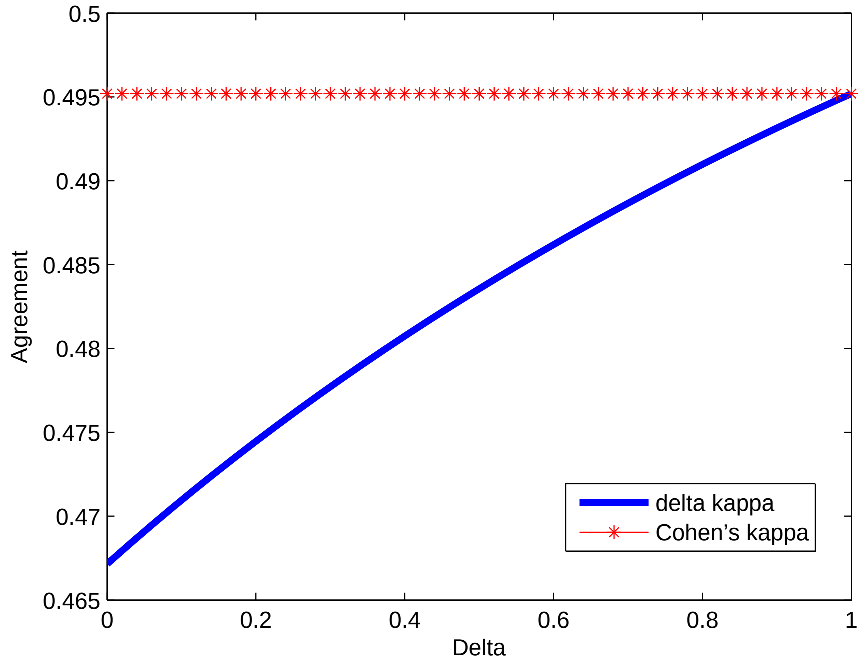

Fig. 4.1. The degree of agreement for case 1 when using kappa and the generalization method

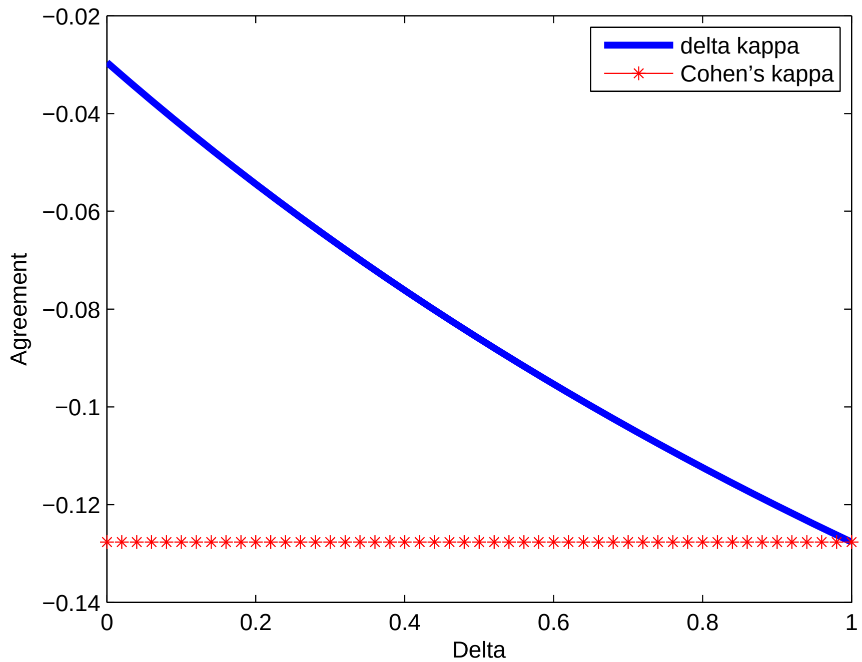

Fig. 4.2. The degree of agreement for case 2 when using kappa and the generalization method

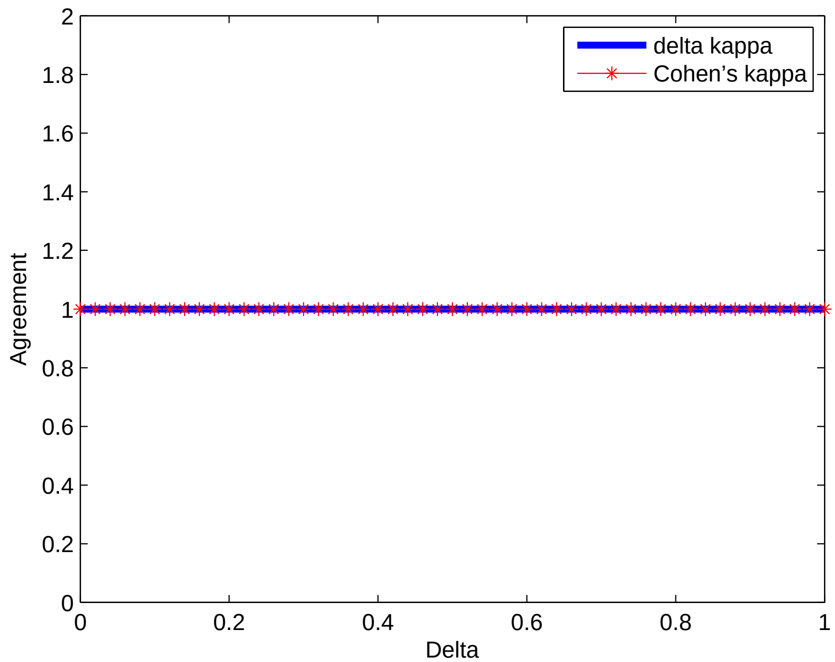

Fig. 4.3. The degree of agreement for case 4 when using kappa and the generalization method

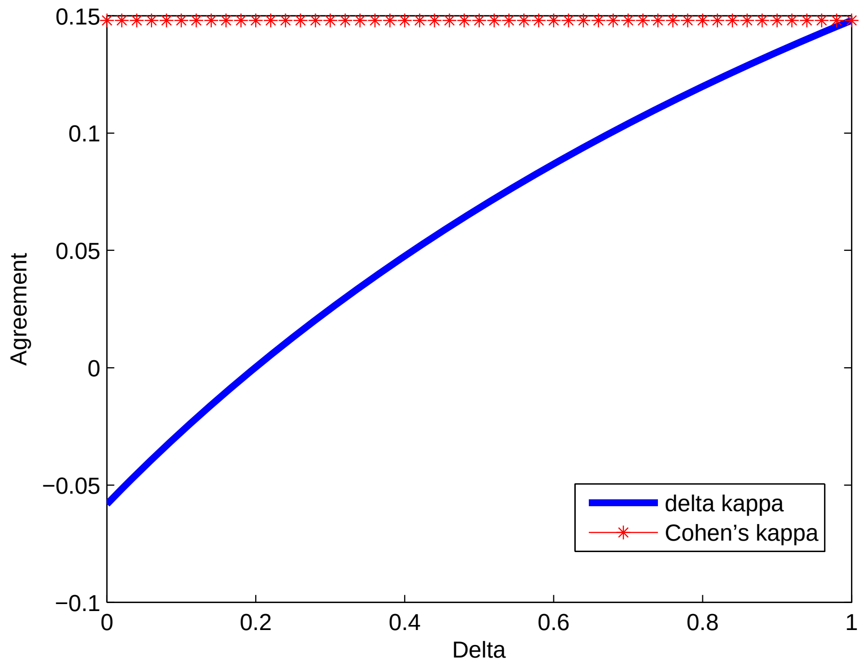

Fig. 4.4. The degree of agreement for case 9 when using kappa and the generalization method

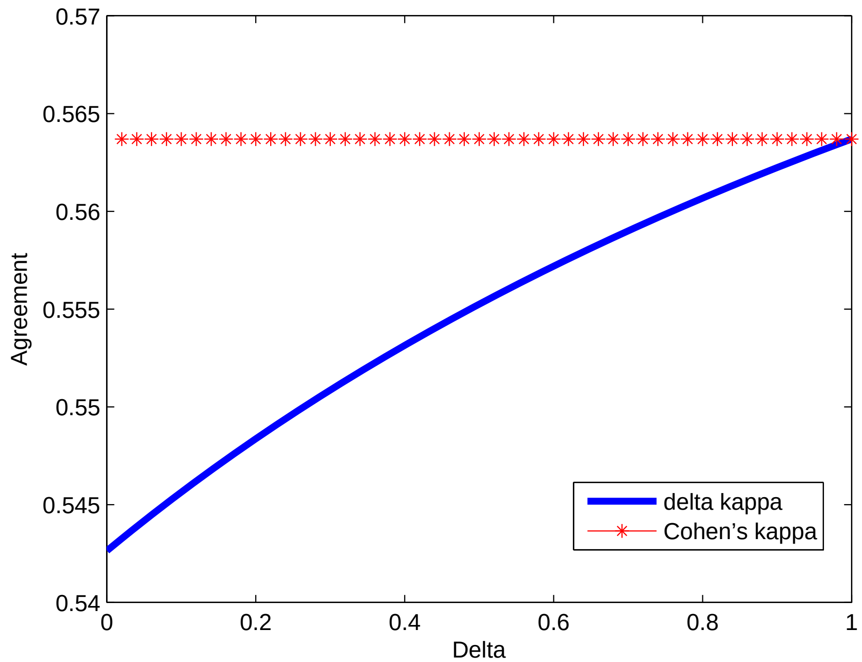

Fig. 4.5. The degree of agreement for case 12 when using kappa and the generalization method

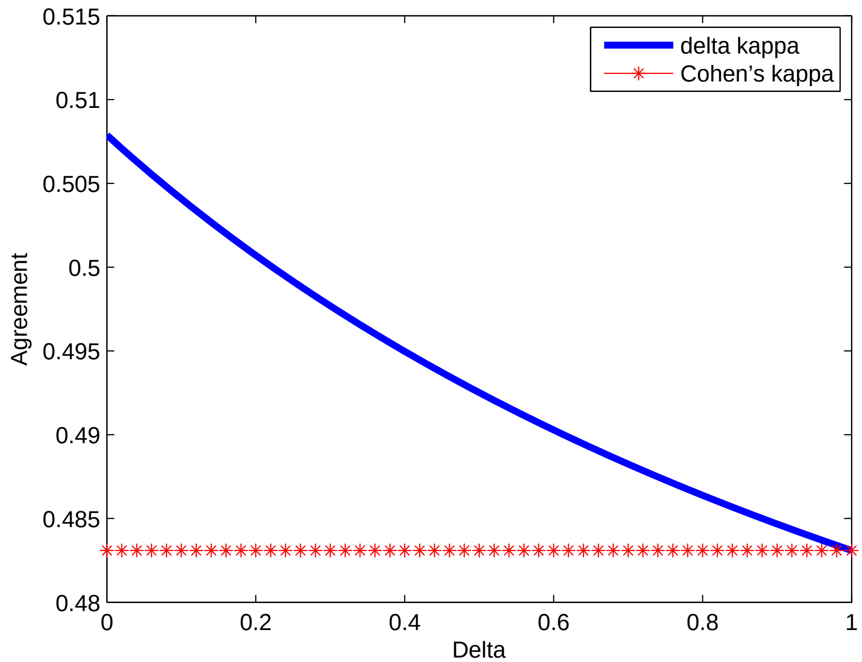

Fig. 4.6. The degree of agreement for case 16 when using kappa and the generalization method

The plots show that there are not substantial differences between Cohen's weighted kappa coefficient and κ_{g_δ}. And they confirm our statement that when $\delta = 1$, this becomes Cohen's weighted kappa coefficient and when $\delta = 0$, this becomes $\kappa_{\text{le}}(W)$.

4.3 Alternative Method

We can generate a different approach to the construction of agreement coefficients in the following manner. Note that P_I and P_D are $(k+1) \times (k+1)$ matrices of rank k because $P_I \cdot 1 = 0$ and $P_D \cdot 1 = 0$, where 1 is a $(k+1) \times 1$ vector of unit values. Thus, $P_I \cdot 1 = 0 \cdot 1$, indicating that 0 is an eigenvalue for P_I with corresponding (standardized) eigenvector $e_0 = \frac{1}{\sqrt{k+1}}$. The same is true for P_D as well. Let $E_I = [e_0 \; e_1 \; \cdots \; e_k]$ denote the $(k+1) \times (k+1)$ matrix of eigenvectors for P_I. By definition, E_I is orthogonal, so $E_I(E_I)^{\text{T}} = (E_I)^{\text{T}} E_I = I$, which yields that $1^{\text{T}} e_1 = 0, \cdots, 1^{\text{T}} e_k = 0$.

Because P_I is not of full rank, its inverse does not exist, but it does have a Moore-Penrose generalized inverse, denoted by $(P_I)^+$. If $\Lambda_I = \text{Diag}(0, \lambda_1, \cdots, \lambda_k)$ denotes the $(k+1) \times (k+1)$ diagonal matrix of eigenvalues for P_I, then the Moore-Penrose generalized inverse of Λ_I is $(\Lambda_I)^+ = \text{Diag}(0, (\lambda_1)^{-1}, \cdots, (\lambda_k)^{-1})$ and the Moore-Penrose generalized inverse of P_I is $(P_I)^+ = E_I(\Lambda_I)^+(E_I)^{\text{T}}$.

Returning to $-P_I \leq P_I - P_D \leq P_I$, this is equivalent to

$$-((P_I)^+)^{\frac{1}{2}} P_I ((P_I)^+)^{\frac{1}{2}} \leq ((P_I)^+)^{\frac{1}{2}} P_I ((P_I)^+)^{\frac{1}{2}} - ((P_I)^+)^{\frac{1}{2}} P_D ((P_I)^+)^{\frac{1}{2}} \leq ((P_I)^+)^{\frac{1}{2}} P_I ((P_I)^+)^{\frac{1}{2}},$$

where $((P_I)^+)^{\frac{1}{2}} = E_I((\Lambda_I)^+)^{\frac{1}{2}}(E_I)^{\text{T}}$.

But

$$
\begin{aligned}
((P_I)^+)^{\frac{1}{2}} P_I ((P_I)^+)^{\frac{1}{2}} &= E_I((\Lambda_I)^+)^{\frac{1}{2}} (E_I)^{\mathrm{T}} E_I \Lambda_I (E_I)^{\mathrm{T}} E_I ((\Lambda_I)^+)^{\frac{1}{2}} (E_I)^{\mathrm{T}} \\
&= E_I((\Lambda_I)^+)^{\frac{1}{2}} \Lambda_I ((\Lambda_I)^+)^{\frac{1}{2}} (E_I)^{\mathrm{T}} \\
&= E_I \mathrm{Diag}(0, 1, \cdots, 1)(E_I)^{\mathrm{T}} \\
&= e_1(e_1)^{\mathrm{T}} + \cdots + e_k(e_k)^{\mathrm{T}} \\
&= I - 1 \cdot 1^{\mathrm{T}}/(k+1).
\end{aligned}
$$

Therefore, we get that

$$
-\{I - 1 \cdot 1^{\mathrm{T}}/(k+1)\} \leq \{I - 1 \cdot 1^{\mathrm{T}}/(k+1)\} - ((P_I)^+)^{\frac{1}{2}} P_D ((P_I)^+)^{\frac{1}{2}} \leq \{I - 1 \cdot 1^{\mathrm{T}}/(k+1)\}.
$$

If $g(A)$ is a function of the symmetric matrix A, and $g(\cdot)$ satisfies the properties defined previously, then a class of agreement coefficients is given by

$$
\kappa_g^* = 1 - \frac{g\{(P_I^+)^{\frac{1}{2}} P_D (P_I^+)^{\frac{1}{2}}\}}{g\{(I - \frac{1}{k+1} 1 \cdot 1^{\mathrm{T}})\}}.
$$

If $g(A) = \mathrm{tr}(A)$, then

$$
\kappa_{\mathrm{tr}}^* = 1 - \frac{1}{k} \mathrm{tr}(P_D (P_I)^+). \tag{4.5}
$$

Properties of κ_g^*:

1. $-1 \leq \kappa_{\mathrm{tr}}^* \leq 1$

2. $\kappa^*_{\text{tr}} = 1$ if $X = Y$ with probability one, i.e., $p_{ij} = 0$ for all $i \neq j$.

3. $\kappa^*_{\text{tr}} = -1$ if and only if $p_{ii} = 0$ for each $i = 0, 1, 2, \cdots, k$ and $p_{ij} = p_{ji} = 0.5$ for one choice of $i \neq j$, excluding the set of degenerate cases.

4. $\kappa^*_{\text{tr}} = 0$ if X and Y are independent.

Proof.

1. Using the above inequality, we know that

$$0 \leq (P_I^+)^{\frac{1}{2}} P_D (P_I^+)^{\frac{1}{2}} \leq 2(I - \frac{1}{k+1} 1 \cdot 1^{\mathrm{T}}).$$

Therefore,

$$\begin{aligned} & 0 \leq \operatorname{tr}((P_I^+)^{\frac{1}{2}} P_D (P_I^+)^{\frac{1}{2}}) \leq \operatorname{tr}\{2(I - \frac{1}{k+1} 1 \cdot 1^{\mathrm{T}})\} \\ & \Rightarrow 0 \leq \operatorname{tr}(P_D (P_I)^+) \leq 2k \\ & \Rightarrow -1 \leq \kappa^*_{\text{tr}} \leq 1. \end{aligned}$$

2. If $X = Y$ with probability one, i.e., $p_{ij} = 0$ for all $i \neq j$, from previous proof, we know that $P_I - P_D = P_I$. That is, $P_D = 0$. Thus $tr(P_D(P_I)^+) = 0 \Rightarrow \kappa^*_{\text{tr}} = 1$

3. From the previous proof, we know that, excluding the set of degenerate cases, if $p_{ii} = 0$ for each $i = 0, 1, 2, \cdots, k$ and $p_{ij} = p_{ji} = 0.5$ for one choice of $i \neq j$, then

$P_I - P_D = -P_I \Rightarrow P_I = \frac{1}{2}P_D$. Therefore, $(P_I)^+ = (\frac{1}{2}P_D)^+ = 2(P_D)^+$. Hence,

$$\begin{aligned}
\mathrm{tr}(P_D(P_I)^+) &= 2\mathrm{tr}(P_D(P_D)^+) \\
&= 2\mathrm{tr}((P_D^+)^{\frac{1}{2}} P_D (P_D^+)^{\frac{1}{2}}) \\
&= 2\mathrm{tr}(I - 1 \cdot 1^{\mathrm{T}}/(k+1)) \\
&= 2k.
\end{aligned}$$

Therefore, $\kappa^*_{\mathrm{tr}} = -1$.

4. From the previous proof, we know that if X and Y are independent, then $P_I = P_D$.

   Therefore,

$$\begin{aligned}
\mathrm{tr}(P_D(P_I)^+) &= \mathrm{tr}(P_D(P_D)^+) \\
&= \mathrm{tr}((P_D^+)^{\frac{1}{2}} P_D (P_D^+)^{\frac{1}{2}}) \\
&= \mathrm{tr}(I - 1 \cdot 1^{\mathrm{T}}/(k+1)) \\
&= k.
\end{aligned}$$

   Therefore, $\kappa^*_{\mathrm{tr}} = 0$.

□

If we set $g(A) = \mathrm{tr}(W * A)$, then

$$\begin{aligned}\kappa_{\mathrm{tr}}^*(W) &= 1 - \frac{\mathrm{tr}(W * P_D * (P_I)^+)}{\mathrm{tr}\{W * (I - \frac{1}{k+1} 1 \cdot 1^{\mathrm{T}})\}} \\ &= 1 - \frac{\mathrm{tr}(W * P_D * (P_I)^+)}{\mathrm{tr}(W) - \frac{1}{k+1}\mathrm{tr}(W * 1 \cdot 1^{\mathrm{T}})} \\ &= 1 - \frac{\mathrm{tr}(W * P_D * (P_I)^+)}{\mathrm{tr}(W) - \frac{1}{k+1}\sum_{i=0}^k \sum_{j=0}^k W_{ij}}. \end{aligned} \tag{4.6}$$

In general, if $g(\cdot)$ satisfies definitions 4.2.1 and 4.2.2, then $\kappa_g^*(W)$ has the following properties:

1. $-1 \leq \kappa_g^*(W) \leq 1$
2. $\kappa_g^*(W) = 1$ if $X = Y$ with probability one, i.e., $p_{ij} = 0$ for all $i \neq j$.
3. $\kappa_g^*(W) = -1$ if and only if $p_{ii} = 0$ for each $i = 0, 1, 2, \cdots, k$ and $p_{ij} = p_{ji} = 0.5$ for one choice of $i \neq j$, excluding the set of degenerate cases.
4. $\kappa_g^*(W) = 0$ if X and Y are independent.

Proof.

Without loss of generality, we set $W = I$. The proofs for other weight matrices W are slightly more complex.

1. From the previous proof, we know that

$$0 \leq (P_I^+)^{\frac{1}{2}} P_D (P_I^+)^{\frac{1}{2}} \leq 2(I - \frac{1}{k+1} 1 \cdot 1^{\mathrm{T}}).$$

Since $g(\cdot)$ is a non-decreasing and scale-equivalent function, we have

$$0 \leq g\{(P_I^+)^{\frac{1}{2}} P_D (P_I^+)^{\frac{1}{2}}\} \leq 2g\{(I - \frac{1}{k+1} 1 \cdot 1^{\mathrm{T}})\}.$$

It is then straightforward to see that $-1 \leq \kappa_g^* \leq 1$

2. If $X = Y$ with probability one, i.e., $p_{ij} = 0$ for all $i \neq j$, from previous proof, we know that $P_D = 0$. It is then straightforward to see that $\kappa_g^* = 0$.

3. From the previous proof, we know that, excluding the set of degenerate cases, if $p_{ii} = 0$ for each $i = 0, 1, 2, \cdots, k$ and $p_{ij} = p_{ji} = 0.5$ for one choice of $i \neq j$, then $P_I - P_D = -P_I \Rightarrow P_I = \frac{1}{2} P_D$. Hence $(P_I)^+ = (\frac{1}{2} P_D)^+ = 2(P_D)^+$ and $((P_I)^+)^{\frac{1}{2}} = \sqrt{2}((P_D)^+)^{\frac{1}{2}}$. Therefore, $(P_I^+)^{\frac{1}{2}} P_D (P_I^+)^{\frac{1}{2}} = 2(P_D^+)^{\frac{1}{2}} P_D (P_D^+)^{\frac{1}{2}}$ and $(P_I^+)^{\frac{1}{2}} P_I (P_I^+)^{\frac{1}{2}} = (P_D^+)^{\frac{1}{2}} P_D (P_D^+)^{\frac{1}{2}}$. Hence, $\kappa_g^* = -1$.

4. From the previous proof, we know that if X and Y are independent, then $P_I = P_D$. It is straightforward to see that $\kappa_g^* = 0$.

□

As before, we can also use $g(A) = \lambda_{A_1}$, the largest eigenvalue of A as another function for $g(\cdot)$.

Thus, there are four coefficients we propose to assess agreement with categorical data: (1) $\kappa_{\mathrm{tr}}(W)$, which is equivalent to Cohen's weighted kappa, (2) $\kappa_{\mathrm{le}}(W)$, (3) $\kappa_{\mathrm{tr}}^*(W)$ and (4) $\kappa_{\mathrm{le}}^*(W)$.

Lemma 4.3.1. *When $k = 1$, that is, when there are only two categories, the four coefficients are equivalent and they all reduce to Cohen's kappa.*

Proof. 1. When $k = 1$, the Cicchetti-Allision weights become an identity matrix. Therefore,

$$\begin{aligned}\kappa_{\text{tr}}(W) &= 1 - \frac{\text{tr}(WP_D)}{WP_I} \\ &= 1 - \frac{\text{tr}(P_D)}{\text{tr}(P_I)} \\ &= \kappa.\end{aligned}$$

2. Since the weighting matrix is a 2×2 matrix, so the two eigenvalues of P_D can be denoted by 0 and λ_1. Therefore, the largest eigenvalue is equal to $\lambda_1 = \text{tr}(P_D)$. Similarly, it can be shown that the largest eigenvalue of P_I is also equal to $\text{tr}(P_I)$. Thus, $\kappa_{\text{tr}}(W) = \kappa_{\text{le}}(W)$.

3. From the previous proof,we know that $(P_I)^+ = E_I(\Lambda_I)^+ E_I^{\text{T}}$. When $k = 1$, (Λ_I) is

$$(\Lambda_I) = \begin{bmatrix} 0 & 0 \\ 0 & \lambda_{I1} \end{bmatrix}$$

and $(\Lambda_I)^+$ is

$$(\Lambda_I)^+ = \begin{bmatrix} 0 & 0 \\ 0 & 1/(\lambda_{I1}) \end{bmatrix}.$$

Because $k = 1$,

$$\begin{aligned}
\kappa^*_{\text{tr}}(W) &= 1 - \text{tr}(P_D(P_I)^+) \\
&= 1 - \text{tr}(P_D \frac{1}{\lambda_{I1}} e_{I1} e_{I1}^{'}) \\
&= 1 - \frac{\text{tr}(P_D e_{I1} e_{I1}^{'})}{\lambda_{I1}} \\
&= 1 - \frac{\text{tr}(P_D e_{I1} e_{I1}^{'})}{\text{tr}(P_I)},
\end{aligned}$$

where e_{I1} is the eigenvector of P_I associated with the largest eigenvalue λ_{I1} of P_I. But for the case $k = 1$, because $e_{I1} = e_{D1} = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}$, e_{I1} also is the eigenvector of P_D associated with the largest eigenvalue λ_{D1} of P_D. Thus,

$$\begin{aligned}
\kappa^*_{\text{tr}}(W) &= 1 - \frac{\text{tr}(P_D e_{D1} e_{D1}^{'})}{\text{tr}(P_I)} \\
&= 1 - \frac{\text{tr}(e_{D1}^{'} P_D e_{D1})}{\text{tr}(P_I)} \\
&= 1 - \frac{\lambda_{D1}}{\text{tr}(P_I)} \\
&= 1 - \frac{\text{tr}(P_D)}{\text{tr}(P_I)} \\
&= \kappa.
\end{aligned}$$

4. From the derivation of $\kappa^*_{\text{le}}(W)$, where $W = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, we know that it can be written as

$$\kappa^*_{\text{le}}(W) = 1 - \frac{\text{largest eigenvalue of } [((P_I)^+)^{\frac{1}{2}} P_D ((P_I)^+)^{\frac{1}{2}}]}{\text{largest eigenvalue of } [I - \frac{1}{2} 1 \cdot 1^{\text{T}}]}.$$

But largest eigenvalue for the case $k = 1$ is equal to the trace, so $\kappa^*_{\text{le}}(W) = \kappa^*_{\text{tr}}(W) = \kappa$.

□

4.4 Simulation Study

In this section, we compare the performance of the four proposed coefficients with that of Cohen's kappa or weighted kappa.

4.4.1 Simulation Result for $\kappa_{\mathbf{le}}(W)$

Table 4.1 gives the simulation results for for the proposed alternative method with the function g being the largest eigenvalue.

Note: $\kappa_{tr}(W)$ is Cohen's weighted kappa, $\kappa_{le}(W)$ is the proposed method with the g function being the largest eigenvalue, $\kappa_{leb}(W)$ is the corresponding bootstrap estimate. v_1, v_2 are the observed variances for the estimate of Cohen's weighted kappa and the estimate of $\kappa_{le}(W)$, respectively and v_3 is the estimated variance using bootstrap.

Table 4.1. Simulation results for $\kappa_{\text{le}}(W)$

case	n	κ_W	$\kappa_{\text{le}}(W)$	$\kappa_{\text{leb}}(W)$	v_1	v_2	v_3
1	80	0.4914	0.4600	0.4522	0.0075	0.0117	0.0113
1	160	0.4912	0.4622	0.4582	0.0036	0.0056	0.0058
2	80	-0.1282	-0.0356	-0.0411	0.0034	0.0065	0.0068
2	160	-0.1270	-0.0319	-0.0347	0.0019	0.0035	0.0034
3	80	0.8516	0.8365	0.8292	0.0032	0.0048	0.0046
3	160	0.8489	0.8359	0.8321	0.0015	0.0023	0.0024
4	80	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000
4	160	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000
5	80	0.5331	0.5510	0.5448	0.0044	0.0069	0.0067
5	160	0.5355	0.5526	0.5510	0.0023	0.0038	0.0036
6	80	0.0536	-0.0191	-0.0223	0.0075	0.0143	0.0147
6	160	0.0544	-0.0186	-0.0203	0.0037	0.0078	0.0077
7	80	0.3939	0.2983	0.2939	0.0080	0.0140	0.0138
7	160	0.3948	0.3003	0.2980	0.0040	0.0074	0.0070
8	80	-0.1716	-0.2514	-0.2504	0.0038	0.0081	0.0076
8	160	-0.1725	-0.2535	-0.2530	0.0020	0.0040	0.0039
9	80	0.1442	-0.0597	-0.0553	0.0059	0.0097	0.0098
9	160	0.1454	-0.0612	-0.0590	0.0029	0.0051	0.0050
10	80	-0.3122	-0.4027	-0.3976	0.0035	0.0080	0.0079
10	160	-0.3160	-0.4076	-0.4051	0.0017	0.0040	0.0040
11	80	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000
11	160	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000
12	80	0.5605	0.5331	0.5228	0.0063	0.0092	0.0091
12	160	0.5624	0.5373	0.5322	0.0028	0.0045	0.0045
13	80	-0.0742	-0.0882	-0.0899	0.0025	0.0070	0.0072
13	160	-0.0736	-0.0861	-0.0870	0.0012	0.0035	0.0037
14	80	0.7849	0.7599	0.7555	0.0044	0.0072	0.0068
14	160	0.7900	0.7675	0.7655	0.0021	0.0036	0.0034
15	80	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000
15	160	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000
16	80	0.4823	0.4981	0.4877	0.0042	0.0064	0.0063
16	160	0.4816	0.5020	0.4965	0.0021	0.0032	0.0031

4.4.2 Simulation Result for $\kappa^*_{\mathbf{tr}}(W)$ With W Being the Identity Matrix

Table 4.2 gives the simulation result for the proposed alternative method with the function g being the sum of the eigenvalues and W being the identity matrix (thus we will compare Cohen's kappa coefficient and the proposed method).

Note: κ is estimate of Cohen's kappa, κ_{tr} is the estimate of proposed method with the g function being the sum of the eigenvalues, κ^*_{trb} is the corresponding bootstrap estimate. v_1, v_2 are the observed variances for the estimate of Cohen's kappa and the estimate of κ_{tr}, respectively and v_3 is the estimated variance using bootstrap.

4.4.3 Simulation Result for $\kappa^*_{\mathbf{le}}(W)$

Table 4.3 gives the simulation result for the proposed method with the function g being the largest eigenvalue when the generalized inverse matrix is used.

Note: κ_W is the estimate for Cohen's weighted kappa. $\kappa^*_{le}(W)$ is the estimate of the proposed alternative method with g being the largest eigenvalue. $\kappa^*_{leb}(W)$ is the bootstrap estimate. v_1, v_2 are the observed variances for Cohen's weighted kappa and $\kappa^*_{le}(W)$, respectively and v_3 is the estimated variance using bootstrap.

4.4.4 Discussion

In most cases, all the agreement coefficients we have proposed give very close results to Cohen's kappa or weighted kappa. Bootstrap turns out to be a good alternative

Table 4.2. Simulation results for $\kappa^*_{\text{tr}}(W)$

case	n	κ	κ^*_{tr}	$\kappa^*_{\text{tr}b}$	v_1	v_2	v_3
1	80	0.5072	0.5028	0.4982	0.0059	0.0060	0.0062
1	160	0.5099	0.5063	0.5040	0.0030	0.0030	0.0031
2	80	-0.2210	-0.2134	-0.2101	0.0031	0.0030	0.0029
2	160	-0.2212	-0.2144	-0.2128	0.0015	0.0014	0.0015
3	80	0.8599	0.8544	0.8517	0.0023	0.0026	0.0028
3	160	0.8594	0.8550	0.8537	0.0011	0.0012	0.0013
4	80	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000
4	160	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000
5	80	0.4513	0.4950	0.4933	0.0024	0.0022	0.0021
5	160	0.4542	0.4984	0.4951	0.0051	0.0045	0.0043
6	80	0.0150	0.0498	0.0483	0.0047	0.0045	0.0041
6	160	0.0176	0.0541	0.0532	0.0024	0.0022	0.0021
7	80	0.4378	0.4339	0.4297	0.0049	0.0048	0.0044
7	160	0.4427	0.4397	0.4376	0.0025	0.0024	0.0022
8	80	-0.0968	-0.0873	-0.0856	0.0020	0.0017	0.0014
8	160	-0.0985	-0.0895	-0.0887	0.0009	0.0008	0.0007
9	80	0.3172	0.1401	0.1427	0.0046	0.0011	0.0011
9	160	0.3173	0.1385	0.1399	0.0021	0.0005	0.0005
10	80	-0.2553	-0.2048	-0.1995	0.0017	0.0009	0.0010
10	160	-0.2578	-0.2071	-0.2047	0.0008	0.0004	0.0004
11	80	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000
11	160	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000
12	80	0.6152	0.5776	0.5702	0.0040	0.0047	0.0044
12	160	0.6192	0.5851	0.5813	0.0021	0.0024	0.0022
13	80	-0.0552	-0.0350	-0.0347	0.0011	0.0012	0.0012
13	160	-0.0525	-0.0317	-0.0317	0.0006	0.0006	0.0006
14	80	0.8026	0.6167	0.6197	0.0026	0.0018	0.0031
14	160	0.7992	0.6156	0.6144	0.0014	0.0009	0.0008
15	80	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000
15	160	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000
16	80	0.3856	0.3333	0.3278	0.0033	0.0029	0.0025
16	160	0.3867	0.3389	0.3360	0.0015	0.0014	0.0013

Table 4.3. Simulation results for $\kappa^*_{\text{le}}(W)$

case	n	κ_W	$\kappa^*_{\text{le}}(W)$	$\kappa^*_{\text{leb}}(W)$	v_1	v_2	v_3
1	80	0.4880	0.4768	0.4801	0.0073	0.0107	0.0100
1	160	0.4881	0.4755	0.4762	0.0039	0.0057	0.0055
2	80	-0.1289	0.0234	0.0388	0.0038	0.0014	0.0021
2	160	-0.1263	0.0149	0.0250	0.0019	0.0008	0.0010
3	80	0.8518	0.8397	0.8377	0.0030	0.0049	0.0051
3	160	0.8500	0.8366	0.8356	0.0017	0.0028	0.0026
4	80	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000
4	160	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000
5	80	0.5340	0.6591	0.6569	0.0048	0.0048	0.0046
5	160	0.5338	0.6593	0.6583	0.0024	0.0022	0.0022
6	120	0.0548	0.3028	0.3038	0.0049	0.0076	0.0073
6	160	0.0560	0.3065	0.3070	0.0040	0.0065	0.0057
7	80	0.3869	0.4081	0.4151	0.0076	0.0060	0.0062
7	160	0.3943	0.4094	0.4130	0.0041	0.0035	0.0033
8	80	-0.1716	0.0692	0.0757	0.0038	0.0007	0.0007
8	160	-0.1721	0.0676	0.0707	0.0020	0.0004	0.0003
9	160	0.1497	0.2111	0.2146	0.0029	0.0007	0.0007
9	320	0.1490	0.2085	0.2103	0.0015	0.0003	0.0003
10	160	-0.3172	0.0624	0.0631	0.0017	0.0002	0.0002
10	320	-0.3180	0.0617	0.0620	0.0009	0.0001	0.0001
11	160	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000
11	320	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000
12	160	0.5641	0.5159	0.5172	0.0032	0.0036	0.0033
12	320	0.5611	0.5104	0.5111	0.0015	0.0017	0.0017
13	160	-0.0739	0.0392	0.0476	0.0012	0.0003	0.0003
13	320	-0.0748	0.0352	0.0390	0.0006	0.0001	0.0001
14	160	0.7912	0.7555	0.7541	0.0020	0.0038	0.0040
14	320	0.7906	0.7542	0.7534	0.0010	0.0018	0.0020
15	160	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000
15	320	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000
16	160	0.4798	0.5444	0.5438	0.0021	0.0040	0.0039
16	320	0.4831	0.5490	0.5486	0.0011	0.0022	0.0020

as an estimation of the large sample variances, which , in many cases, are smaller than the variances of kappa or weighted kappa. For $\kappa_{\mathrm{le}}(W)$, the only significant difference appear in case 9 (even so, they both indicate poor agreement). For $\kappa^*_{\mathrm{tr}}(W)$, the only substantial difference appears in case 9 (even so, they both indicate poor agreement) and case 14 (Cohen's kappa indicates that there is excellent agreement while the proposed method indicates that there is good agreement). For $\kappa^*_{\mathrm{le}}(W)$, the only significant difference appears in case 5 (even so, they both indicate fair agreement). Note that for the last method we use relatively large sample size since if the sample size is too small, drawing bootstrap samples, it often happens that, when the probabilities are small, some rows and columns have no observation. In such a case, the actual dimension of the estimated P_I and P_D are reduced by 1, resulting in a null eigenvalue. This makes sense in that when we have a relatively large number of categories in response, say $k = 5$, then the total number of all possible pairs of ratings would be 36. Even if we have $n = 160$, in each cell, on average there are less than 5 observations. Too few observations make it hard to estimate the actual probability distribution.

The proposed method is a very promising method for measurement of agreement for categorical data due to the following reason: (1). It is a generalization of Cohen's kappa and weighted kappa coefficient. (2). It provides very consistent and reasonable results. (3). In many cases, it gives variances smaller than those of Cohen's kappa and weighted kappa, meaning that we are more precise in drawing statistical inferences. (4). It also can be adapted to accommodate other matrix functions that might have better performance.

4.5 Examples for Ordinal Data

In this section, we use several ordinal categorial examples from the literature to see the application of the proposed method and compare it with that of Cohen's weighted kappa.

The first example is from Agresti (2002), in which two pathologists, A and B, separately rated 118 subjects regarding the presence and extent of carcinoma of the uterine cervix. The rating scale has the ordered categories (0) negative, (1) atypical squamous hyperplasia, (2) carcinoma *in situ* and (3) squamous or invasive carcinoma. The data is summarized in the following table:

		B				
		0	1	2	3	$P_X(x)$
	0	0.1864	0.0169	0.0169	0	0.2203
A	1	0.0424	0.0593	0.1186	0	0.2203
	2	0	0.0169	0.3051	0	0.3220
	3	0	0.0085	0.1441	0.0848	0.2374
	$P_Y(y)$	0.2288	0.1017	0.5847	0.0848	1

The following tables shows the result calculated from the above example:

Cohen's kappa is 0.4930. All the coefficients using trace functions work well. The proposed methods using the largest eigenvalue function , when applying the Cicchetti-Allison weights, seem to give larger agreement coefficient, indicating good agreement

Table 4.4. Results for example 1

Weight	κ_W	$\kappa_{\text{tr}}(W)$	$\kappa_{\text{le}}(W)$	$\kappa^*_{\text{tr}}(W)$	$\kappa^*_{\text{le}}(W)$
Cicchetti-Allison	0.6489	0.6489	0.7716	0.6410	0.7624
Fleiss-Cohen	0.7839	0.7839	0.7839	0.7574	0.7510

vs fair agreement implied by Cohen's weighted kappa coefficients and our two other agreement coefficients.

The second example is a hypothetical one from Fleiss et al (2002), in which subjects are classified into one of 3 categories by rater A and into another by rather B. The three categories are: (0) Psychotic, (1) Neurotic and (2) Organic. The data is summarized in the following table:

		y = 0	1	2	$P_X(x)$
	0	0.75	0.01	0.04	0.8
x	1	0.05	0.04	0.01	0.1
	2	0	0	0.10	0.1
	$P_Y(y)$	0.80	0.05	0.15	1

The following tables shows the result calculated from the above example:

Cohen's kappa is 0.6765 for this example. And we can see that $\kappa_{\text{tr}}(W)$ is equal to Cohen's weighted kappa, as we have shown before. And all the other coefficients gives results that are very close to corresponding Cohen's weighted kappa.

Table 4.5. Results for example 2

Weight	κ_W	$\kappa_{\text{tr}}(W)$	$\kappa_{\text{le}}(W)$	$\kappa^*_{\text{tr}}(W)$	$\kappa^*_{\text{le}}(W)$
Cicchetti-Allison	0.7222	0.7222	0.7434	0.7072	0.7698
Fleiss-Cohen	0.7553	0.7553	0.7553	0.7763	0.7686

The third example is by Fay (2005), who proposed the Random Marginal Agreement Coefficient. In the example, 149 patients were classified into four categories by two neurologists. The four categories are: (0) Certain multiple sclerosis, (1) probable MS and (2) possible MS (50:50 odds) and (3) doubtful. The data is summarized in the following table:

		B 0	1	2	3	$P_X(x)$
	0	0.2550	0.0336	0	0.0067	0.2953
A	1	0.2215	0.0738	0.0201	0	0.3154
	2	0.0671	0.0940	0.0336	0.0403	0.2350
	3	0.0201	0.2484	0.0201	0.0671	0.1543
	$P_Y(y)$	0.5637	0.2484	0.0738	0.1141	1

Fay then modified the data by supposing that 10 patients that were rated '2' by Neurologist 1 and '0' by Neurologist 2, were instead rated '0' by Neurologist 1 and '2' by Neurologist 2. Hence the new data is:

		B				
		0	1	2	3	$P_X(x)$
	0	0.2550	0.0336	0.0671	0.0067	0.3624
A	1	0.2215	0.0738	0.0201	0	0.3154
	2	0	0.0940	0.0336	0.0403	0.1679
	3	0.0201	0.2484	0.0201	0.0671	0.1543
	$P_Y(y)$	0.4966	0.2484	0.1409	0.1141	1

The following tables shows the result calculated from the first data set:

Table 4.6. Results for the first data set of example 3

Weight	κ_W	$\kappa_{\text{tr}}(W)$	$\kappa_{\text{le}}(W)$	$\kappa^*_{\text{tr}}(W)$	$\kappa^*_{\text{le}}(W)$
Cicchetti-Allison	0.3797	0.3797	0.4974	0.3920	0.5267
Fleiss-Cohen	0.5246	0.5246	0.5246	0.5443	0.5321

For the second data set, the results are:

Table 4.7. Results for the second data set of example 3

Weight	κ_W	$\kappa_{\text{tr}}(W)$	$\kappa_{\text{le}}(W)$	$\kappa^*_{\text{tr}}(W)$	$\kappa^*_{\text{le}}(W)$
Cicchetti-Allison	0.3553	0.3553	0.4706	0.3734	0.5099
Fleiss-Cohen	0.5035	0.5035	0.5035	0.5266	0.5159

In this example, for the first data set, Cohen's kappa is 0.2079 and for the second data set, it is 0.1855, which means that there is less agreement in the second data set. Fay argues that Cohen's kappa fails to work in this case since the two data sets have identical diagonal elements (exact matches) and the second data set has closer matching marginal probabilities. This appears to be true at the first glance. However, a more in-depth investigation shows that it is not as simple at it looks. We know that Cohen's kappa is chance excluded, which means it measures the agreement which is exclusive of the agreement expected by chance. Changing the data set changes the marginal values and hence the agreement expected by chance. In both cases, the observed agreement is the same, what is different is the expected agreement. A simple algebraic calculation shows that the second data set has more agreement that is expected by chance (0.0198 more). Although the marginal probabilities are close to each other, more agreement are obtained by chance. So if we want an index that is chance-corrected, then Cohen's kappa still makes sense. The result also shows that all the agreement coefficients show a decrease of the degree of agreement in the second example.

The above examples confirm that $\kappa_{\text{tr}}(W)$ is the same as Cohen's weighted kappa coefficient, that all agreement coefficients perform well with $\kappa^*_{\text{tr}}(W)$ being the best, giving results very close to Cohen's weighted kappa. $\kappa_{\text{le}}(W)$ and $\kappa^*_{\text{le}}(W)$, in some cases, tend to give slightly larger agreement coefficients.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

In this thesis, several alternative methods to Cohen's kappa and weighted kappa coefficients are developed to measure agreement when responses are categorical.

In chapter 2, we establish a pairwise agreement minus disagreement coefficient. An overall agreement coefficient can be obtained as a weighted sum of each pairwise agreement minus disagreement coefficients. The properties of this overall agreement coefficient are investigated and its large sample distribution is developed. Its performance is exhibited through simulation studies, which shows that it is particularly good at detecting poor agreement.

In chapter 3, we develop a conditional kappa, an agreement coefficient that is obtained by measuring agreement conditioned on each possible category. We examine the properties of the conditional kappa and develop its large sample distribution. Due to the complexity of its large sample distribution, we choose to use bootstrap instead in estimating the large sample variances of the estimator, which turns out to be very close to the observed values. Simulation study shows that overall, the conditional kappa has good performance and bootstrap can be used to estimate the large sample variances, although it is more computation intensive.

In chapter 4, we develop the multivariate kappa by constructing two matrices P_I and P_D, the difference of which, we believe, contains all the information concerning agreement. Different matrix functions can be applied to the two matrices to construct a scalar measurement of the degree of agreement. We consider two functions, the trace function (sum of eigenvalues) and the largest eigenvalue. An alternative method is to apply the two functions to the generalized inverse of the two matrices. Altogether, this gives us four agreement coefficients. When there are only two categories, all the four agreement coefficients reduce to Cohen's kappa coefficient. The properties of these four coefficients are investigated and through simulation studies, their performance is compared with that of Cohen's kappa or weighted kappa.

5.2 Some Additional Thoughts

The measurement of agreement for categorical data, although at first glance seeming to be simple, turns out to be a very complex topic. One worrisome issue is whether we should exclude the so-called agreement obtained by chance. We already discussed this briefly in the example given by Fay (2005). Here is an extreme example:

		Y		
		0	1	$P_X(x)$
X	0	0.81	0.09	0.9
	1	0.09	0.01	0.1
	$P_Y(y)$	0.9	0.1	1

Although we have high diagonal values and the same marginal values, Cohen's kappa coefficient is 0 since all the agreement is obtained when we assume that X and Y are independent. Fay's RMAC also gives 0.

However, if we change the data slightly in the following way:

		Y 0	1	$P_X(x)$
X	0	0.81	0.18	0.99
	1	0	0.01	0.01
	$P_Y(y)$	0.81	0.19	1

The kappa is 0.082, implying that there is more agreement in the second data set, despite that in the second case we have the same diagonal values and more imbalanced marginal values. Fay's RMAC remains at 0, which implies insensitivity to the change of the marginal probabilities. Fay argues that this is an advantage of his proposed method. We really doubt this argument since if his argument holds, we should expect a smaller agreement coefficient instead of being the same for both cases. In the first table, Cohen's kappa gives 0 because by its definition, all the agreements we get are observed by chance. But in reality, assuming that rater A is giving evaluation that is independent of rater B does not mean that they should not be close to each other, as is the case in the first table. So the question is, should these agreements be removed when measuring the overall agreement?

Another problem is the reproducibility of the measurement of agreement coefficient itself. Suppose we are measuring agreement between two raters A and B. We don't want our conclusion of the degree of agreement between A and B to vary depending on which evaluation method the two raters are using. In other words, we wish that our agreement coefficients agree with each other no matter what kind of evaluation method the two raters are using. This topic has been covered in the simulation part of chapter 2.

5.3 Future Work

The multivariate kappa seems to be a very promising method as a generalization to Cohen's kappa and weighted kappa coefficients. Future research would include:

1. Investigate whether there are others forms of matrix functions that can be applied to the multivariate kappa.

2. If such functions exist, investigate which function has better performance.

3. Investigate whether chance should be included in constructing agreement coefficients.

4. Investigate whether this multivariate kappa can be generalized to deal with multi-rater agreement problems.

5. Investigate the agreement issue with respect to the employment of different evaluation methods. Investigate whether the number of categories in the responses can be taken into account in constructing the agreement coefficients or whether suggestions can be made concerning the number of categories that should be used in the responses

Bibliography

[1] Aickin, M., Maximum Likelihood Estimation of Agreement in the Constant Predictive Probability Model, and Its Relation to Cohen's Kappa, *Biometrics*, Vol. 46, No. 2 (Jun., 1990), 293-302.

[2] Agresti, A., *Categorical Data Analysis*, 2nd edition, Wiley, 2002.

[3] Agresti, A., A Model for Agreement Between Ratings on an Ordinal Scale, *Biometrics*, Vol. 44, No. 2 (Jun., 1988), 539-548.

[4] Banerjee, M., Capozzoli, M., McSweeney, L. and Sinha, D., Beyond Kappa: A Review of Interrater Agreement Measures, *The Canadian Journal of Statistics*, Vol. 27, No. 1. (Mar., 1999), 3-23.

[5] Barnhart, H. X., Haber, M., and Song, J., Overall Concordance Correlation Coefficient for Evaluating Agreement Among Multiple Observers, *Biometrics* 58, 1020-1027, December, 2002.

[6] Barnhart, H. X. and Williamson, J. M., Modeling Concordance Correlation Coeffcient via GEE to Evaluate Reproducibility, *Biometrics*, 57, 2001, 931-940.

[7] Basu S., Basu, A., and Raychaudhuri, A., Measuring Agreement between Two Raters for Ordinal Response: A Model-Based Approach, *The Statistician*, Vol. 48, No. 3 (1999), 339-348.

[8] Berman A. and Plemmons, R.J., Monotonicity and the Generalized Inverse, *SIAM Journal on Applied Mathematics*, Vol. 22. No. 2. (Mar., 1972), 155-161.

[9] Carrasco, J. L. and Jover, L., Estimating the Generalized Concordance Correlation Coefficient throught Variance Components, *Biometrics*, 59, 2003, 849-858.

[10] Carrasco, J. L. and Jover, L., Concordance Correlation Coefficient Applied to Discrete Data, *Statistics in Medicine*, 2005; 24; 4021-4034.

[11] Chinchilli, V. M., Martel, J. K., Kumanyika, S., and Lloyd, T., A Weighted Concordance Correlation Coefficient for Repeated Measurement Desgins, *Biometrics*, 52, 1996, 341-353.

[12] Cohen, J., A Coefficient of Agreement for Nominal Scales, *Educational and Psychological Measurement* 1960, 20, 37-46.

[13] Cohen, J., Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit, *Psychological Bulletin*, 1968, 70, 213-220.

[14] Davies, M. and Fleiss, J. L., Measuring Agreement for Multinomial Data, *Biometrics*, Vol. 38, No. 4. (Dec., 1982), 1047-1051.

[15] Eaton, M. L. and Tyler D. E., On Wielandt's Inequality and Its Application to the Asymptotic Distribution of the Eigenvalues of a Random Symmetric Matrix, *The Annals of Statistics*, Vol. 19, No. 1. (Mar., 1991), 260-271.

[16] Fay, M. P., Random Marginal Agreement Coefficients: Rethinking the Adjustment for Chance When Measuring Agreement, *Biostatistics* 2005, Vol.6, No.1, 171-180.

[17] Fay, M. P., Supplementary Material for ”Random Marginal Agreement Coefficients: Rethinking the Adjusted for Chance when Measuring Agreement”, technical report, 2006.

[18] Feinstein, A. R., Cicchetti, D.V., High Agreement but Low Kappa: The Problems of Two Paradoxes, *Journal of Clinical Epidemiology*, Vol. 43, No. 6 (1990), 539-632.

[19] Fisher, A. G., On Construction and Properties of the Generalized Inverse, *SIAM Journal on Applied Mathematics*, Vol. 15, No. 2. (Mar., 1967), 269-272.

[20] Fleiss, J. L. and Cohen, J., Large Sample Standard Errors of Kappa and Weighted Kappa, *Psychological Bulletin*,Vol. 72, No. 5, 1969, 323-327.

[28] Fleiss, J. L., Levin, B. and Paik, M. C., *Statistical Methods for Rates and Proportions*, 3rd edition, New York: Wiley, 2002.

[22] Gonin, R., Lipsitz, S. R., Fitzmaurice, G. M. and Molenberghs, G., Regression Modelling of Weighted κ by Using Generalized Estimating Equations, *Applied Statistics*, Vol. 49, No. 1 (2000), 1-18.

[23] Gross, S. T., The Kappa Coefficient of Agreement for Multiple Observers When the Number of Subjects is Small, *Biometrics*, Vol. 42, No. 4 (Dec., 1986), 883-893.

[24] King, T. S., Chinchilli, V. M., Robust Estimator of the Concordance Correlation Coefficient, *Journal of Biopharmaceutical Statistics*, 11(3), 83-105, 2001.

[25] King, T. S., Chinchilli, V. M., A Generalized Concordance Correlation Coefficient for Continuous and Categorical Data, *Statist. Med.* 2001; 20:2131-2147.

[26] Kraemer, H. C., Extension of the Kappa Coefficient, *Biometrics*, Vol. 36, No. 2 (Jun., 1980), 207-216.

[27] Kupper, L. L. and Hafner, K. B., On Assessing Interrater Agreement for Multiple Attribute Responses, *Biometrics*, Vol. 45, No. 3. (Sep., 1989), 957-967.

[28] Landis, J. R. and Koch, G. G., The Measurement of Observer Agreement for Categorical Data, *Biometrics*, Vol. 33, 1977, 159-174.

[29] Laurent, R. T. St., Evaluating Agreement with a Gold Standard in Method Comparison Studies, *Biometrics*, Vol. 54 (Jun., 1998), 537-545.

[30] Lee, J. and Tu, Z., A Better Confidence Interval for Kappa on Measuring Agreement between Two Raters with Binary Outcomes, *J. of Computational and Graphical Statistics*,Vol. 3, No. 3 (Sep., 1994), 301-321.

[31] Lehmann, E. L., *Theory of Point Estimation*, 2nd edition, New York: Springer, 1998.

[32] Lewis, G. H. and Johnson, R. G., Kendall's Coefficient of Concordance for Sociometric Rankings with Self Excluded, *Sociometry*, Vol. 34, No. 4 (Dec., 1971), 496-503.

[33] Lin, L. I., A Concordance Correlation Coefficient to Evaluate Reproducibility, *Biometrics*, Vol.45, No.1 (Mar.,1989), 255-268.

[34] Lin, L. I., Assay Validation Using the Concordance Correlation Coefficient, *Biometrics*, Vol.48, No.2 (Jun.,1992), 599-604.

[35] Lin, L. I., Hedayat, A. S., Sinha, B., and Yang, M., Statistical Methods in Assessing Agreement: Models, Issues, And Tools, *Journal of Statistical Association*, March 2002, Vol. 97, No. 457, Theory and Methods.

[36] Li, R. and Chow, M., Evaluation of Reproducibility for Paired Functional Data, *Journal of Multivariate Analysis* 93, 81-101, 2005.

[37] Liu, X., Du, Y., Teresi, J., and Hasin, D. S., Concordance Correlation in the Measurement of Time to Event, *Statist. Med.*, 2005; 24: 1409-1420.

[38] Longford, N. T., Multivariate Shrinkage Estimation of Small Area Means and Proportions, *Journal of the Royal Statistical Society: Series A* (*Statistics in Society*), Vol. 162, No. 2. (1999), 227-245

[39] Lumley, T., Generalized Estimating Equations For Ordinal Data: A Note on Working Correlation Structure, *Biometrics*, Vol. 52, March 1996, 354-361.

[40] Multivariate, R. J., Aspects of Multivariate Statistical Theory, Wiley, 1982.

[41] Plemmons, R. J. and Cline, R.E., The Generalized Inverse of a Nonnegative Matrix, *Proceedings of the American Mathematical Society*, Vol. 31, No. 1. (Jan., 1972), 46-50.

[42] Rao, C. R., A Note on a Generalized Inverse of a Matrix with Applications to Problems in Mathematical Statistics, *J. of the Royal Statistical Society: Series B* (*Methodological*), Vol. 24, No. 1. (1962), 152-158.

[43] Stewart, G. W., On the Continuity of the Generalized Inverse, *SIAM Journal on Applied Mathematics*, Vol. 17, No. 1. (Jan., 1969), 33-45.

[44] Vonesh, E. F., Chinchilli, V. M. and Pu, K., Goodness-of-Fit in Generalized Nonlinear Mixed-Effects Models, *Biometrics*, Vol. 52 (Jun., 1996), 572-587.

[45] Wet, D. and Mandl, F., On the Asymptotic Distribution of Eigenvalues, *Proceedings of the Royal Society of London, Series A, Mathematical and Physical Sciences*, Vol. 200, No. 1063. (Feb. 22, 1950), 572-580.

Vita

Jingyun Yang was born in Laizhou, Shandong, P. R. China, on February 5th, 1978. He received his Bachelor of Law, with departmental and general honors, in sociology from the People's University of China in July, 2000. He received his M.A. in sociology from University of Oregon in June, 2002. He enrolled in the Ph.D. program in Statistics at The Pennsylvania State University in August, 2002. He has been a teaching assistant, research assistant and instructor during his study at The Pennsylvania State University.

He is a member of the American Statistical Association.