International Journal of Pure and Applied Mathematics
Volume 22 No. 1 2005, 13-31

EFFECT OF MISSING VALUES ON THE COHEN'S KAPPA STATISTIC FOR RATERS AGREEMENT MEASUREMENT

Adebowale Olusola Adejumo [§]
Department of Statistics
LMU, University of Munich
Ludwigstrasse 33/I, Munich, D-80539, GERMANY
e-mail: ao123adejumo@yahoo.co.uk

Abstract: In any experiment that involves measurement, counting, or diagnosis, there is the likelihood for some elements of missing observation which can be as a result of any of foreseen or unforeseen circumstances. Virtually in almost all life or social science researches, subjects are classified into categories by raters, interviewers or observers. Over the last four decades, Cohen kappa statistic has been the major statistic for measuring the overall level of agreement that exits between two raters. In this research work, we examine the effects on this statistic in a situation where certain percentages of the counts are missing, by assuming the ignorability criteria mechanism, along the main diagonal as well as off the diagonal cells of the resulting cross-classified table of the ratings by the two raters. We observed that missingness in the resulting cross-classified table has great effect on the Cohen kappa statistic by reducing the level of agreement and in some situation improving the strength of agreement depending on the structure of the missingness in such table, and also change the class of strength of agreement in some situations.

AMS Subject Classification: 62K99, 47H40, 60H25
Key Words: agreement, kappa statistic, missing at random (MAR), raters

Received: March 6, 2005

[§] Correspondence address: Department of Statistics, University of Ilorin, Ilorin, NIGERIA

1. Introduction

We consider the observation of events by two independent observers that are classified according to qualitative variables into cells of a contingency table representing corresponding population. Any cell along the diagonal of the resulting cross-classified table is being classified as the raw agreement while all other cells that are off the diagonal of the table are also regarded as raw disagreement. Any square contingency table can be used to display joint ratings of two raters. Statistical analysis with missing data is a common problem in practice. Nonresponse in a sample surveys or drop-out in a clinical trials may be two of many examples one could imagine. We establish in this paper that during the process of collecting and classifying these ratings on the subjects into classes of categories, there are possibilities of missing values. In nearly all the researches that involve ratings, measurements or diagnosis of subjects by various raters, observers or pathologists, the researchers are already aware that the most important measurement of error or bias is the raters involved in such studies. At the initial stage of the experiment reliability test or studies has to be conducted among the raters, interviewers or observers involved to assess the level of raters variability in the measurement procedure to be used in data acquisition. Categorical data are data that the response variable is classified into either nominal or ordinal categories. For nominal data, as reviewed by Banerjee et al [2], a large numbers of estimation and testing procedures like the Cohen kappa, see Cohen [5], which is our main target in this research work, the weighted kappa, see Cohen [6], the intraclass kappa block and Kraemer [4], and many other improved methods on kappa statistic, see Shoukri [16]. On the side of ordinal data, most of the medical diagnoses data often involve responses taken on an ordinal scale and many of which are very subjective. As it has been pointed out by some authors with ordinal data, an intermediate category will often be subjective to more misclassification than an extreme category because there are two directions in which to err away from the extremes. Therefore a modified kappa statistic which accounts for severity of discordance or size of discrepancy is better suited for ordinal data. The weighted kappa Cohen [6] offers such modification.

Our main objective is to examine the effect of missing observations from the diagonal, off diagonal as well as general missing that consist of both diagonal and off diagonal counts on Cohen kappa statistic. A square table can be used to display joint ratings of two raters or observers. Two matters are usually considered for this type of table. Firstly, one can analyze differences in the marginal distributions. For ordered response categories, there is usually interest on whether classifications by one rater tend to be higher than those by the other rater. Secondly, one can analyze the extent of subject-wise agreement between raters, which involves investigating the frequency of main-diagonal occurrence within the joint distribution of the ratings.

Missing value problem in statistical analysis is a common practice. Non-response in sample surveys or drop-out in clinical trials may be two of many examples one expect to happen. In an experiment that involves two or more raters, one subject may refuse to cooperate or show up for examination with one or two other raters or observers, so we regard this as missing. Also if certain number of cultures plates of bacteria samples has been examined by one rater, but before getting to the next rater, some of the plate break, we also regard to these as missing values in such an experiment. As advanced to earlier, our major goal is to examine the effects of this missing values from the table on Cohen kappa statistic. In Section 2 and Section 3 we present missing value and Cohen kappa statistic. We also present description of missing values in the ratings of raters as well as Cohen kappa with missing value in the two sections respectively. And lastly in Section 4 we have various empirical examples showing their results and conclusion are presented in Section 5.

2. General Missing Values

As said in the introduction, statistical analysis with missing data is a common problem in practice. Little and Rubin [12] and Rubin [13] have discussed fundamental concepts for handling missing data based on decision theory and models for the mechanism of nonresponse. Various standard statistical methods have been developed to analyze rectangular data sets, that is to analyze a matrix

$$X = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & * & \ldots & x_{2p} \\ x_{31} & x_{32} & \ldots & * \\ \vdots & \vdots & & \vdots \\ x_{n1} & * & \ldots & x_{np} \end{pmatrix}. \tag{2.1}$$

The rows of this data matrix can be modelled as independent, identically distributed (iid) draws from some multivariate probability distribution. The missing values, denoted by * my occur in any pattern. The probability density function of the complete data may be be written as

$$P(X|\theta) = \prod_{i=1}^{n} f(x_i|\theta), \tag{2.2}$$

where f is the density or probability function for a single row, and θ is a vector of unknown parameters. Under missing value analysis, data on any scales can be observed.

In categorical data analysis, visualizing the structure of the data set with respect to the missing values may be the first way to get an impression of the situation on how to handle the problem. These patterns may give an impression to what extent the data are missing. If X is assumed to be missing for large values of y, the values can be ordered and a missing data pattern may describe this behavior. However, this technique may be swamped with a high level of dependencies. A way to overcome this defect consists of defining the so called missing values pattern, see Toutenburg et al [17] for details on general missing values pattern.

2.1. Missing Values Mechanisms

Let the observed part of X be represented by X_{obs}, and the missing counterpart by X_{mis}. The next focus is whether the missing data mechanism can be *ignored* or not. It is possible to make an assumption that the mechanism is ignorable or including the missing data mechanism in the statistical model. By including the missing data mechanism means including the distribution of an indicator variable R indicating if a component of the data matrix $Z = (Z_{obs}, Z_{mis})$ is observed or missing. The random variable R indicating the missingness within the data matrix Z is defined as

$$r_{ij} = \begin{cases} 1, & \text{if} \quad z_{ij} \quad \text{observed}\,, \\ 0, & \text{if} \quad z_{ij} \quad \text{missing}, \quad \forall i = 1,2,...,n, \quad j = 1,2,...,p+1\,. \end{cases} \tag{2.3}$$

The ignorability criteria of the missing data mechanism depends on whether statistical inference is based on the density $f(R, Z_{obs}|\theta, \phi)$ or on the simpler density $f(Z_{obs}, \theta)$ which is ignoring the missing mechanism, where θ is the parameter of the density of Z_{obs}, Z_{mis} and ϕ is the unknown parameter of the missing mechanism.

Therefore, the classification of missing data mechanism is thus based on the density $f(R|Z_{obs}, Z_{mis}, \phi)$, see Rubin [13], Schafer [14], Toutenburg and Nittner [18], Toutenburg et al [17] for more details. In this research work we assume the ignorability criteria mechanism for the missingness along and off the diagonal of the square table.

	Rater 1		Rater 2		
subject	resp	rating	resp	rating	status
1	1	1	1	2	obs
2	1	1	1	3	obs
3	0	-	1	2	mis
4	1	5	1	1	obs
5	1	3	0	-	mis
6	1	2	1	3	obs
7	0	-	0	-	mis*
⋮	⋮	⋮	⋮	⋮	⋮
n	1	4	1	3	obs

Table 2.1: Missing pattern for ratings of two raters

2.2. Missing Values in the Ratings of Raters

As said before in the introduction that any square contingency table can be used to display joint ratings of two raters. In this case the categories on both row and column must be the same, that is, the two raters must work with the same categorical scales. Missing observations can also be observed in the raw results for some of the subjects involve in the experiment or in the trials. For example, if certain number of cultures plates of bacteria samples have been collected and well examined by one rater, but before getting to the next rater, some of the plates got broken, misplaced, wrongly handled by the laboratory attendance or wrongly labelled or identified, we also regard to such as missing values in such experiment. To this effect, Table 2.1 described the pattern of how missing observations can occur in the raw ratings of two raters on some sets of subjects.

From the table, the rating column depends on the categorical scales agreed upon by the raters and also the combination of these two ratings determine the form of the cross-classified square contingency table for agreement. The $\{resp\}$ stands for the response status for i-th subject with j-th rater which we have defined in 2.3 as the random variable R indicating the missingness within the data matrix Z, defined in this case for two raters $(j = 1, 2)$ as

$$r_{ij} = \begin{cases} 1, & \text{if} \quad z_{ij} \quad \text{observed}\,, \\ 0, & \text{if} \quad z_{ij} \quad \text{missing}, \quad \forall i = 1, 2, ..., n, \quad j = 1, 2\,, \end{cases} \tag{2.4}$$

such that the matrix for R base on Table 2.1 can be of the form

$$R = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 0 & 0 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix} . \tag{2.5}$$

Also each subject combined response status S_i is defined as

$$S_i = \begin{cases} obs, & \text{if} \quad r_{i1} = r_{i2} = 1 \quad \text{as observed}\,, \\ mis, & \text{if} \quad (r_{i1} = 0 \text{ and } r_{i2} = 1) \text{ or } (r_{i1} = 1 \text{ and } r_{i2} = 0) \\ & \hfill \text{as missing,} \\ mis*, & \text{if} \quad r_{i1} = r_{i2} = 0 \quad \text{as totally missing} \quad \forall i = 1, 2, ..., n. \end{cases} \tag{2.6}$$

2.3. Missing Value Pattern in the Ratings of Raters

The missing pattern depends on the nature of missingness as we have presented in Table 2.1. Take for instance if the categorical scale for the raters is numbered 1 to 5, if rater 1's response to one subject is 1, and the rater 2's response to that same subject is missing, then if we assume that this missing response is also 1, this implies that the missing is in the diagonal cell, but if we assume it is not 1, that means it will be any of letters 2 to 5, then the missing of that particular subject is in the off-diagonal cells. Generally, if there are many responses, says 'a,b,c...', that are missing, all may fall along the diagonal or in the off-diagonal cells and sometimes there may be the combination of the two missing patterns, that is, off and along the diagonal combined. Examples of these tables with different patterns of missing for 2×2 contingency tables are as given in Table 2.3 to Table 2.5.

Now assume that we have a 2×2 contingency table for the ratings of two raters in Table 2.3.

3. Cohen Kappa Statistic

Cohen [5] proposed a standardized coefficient of raw agreement for nominal scales in terms of the proportion of the subjects classified into the same category

Category	Rater 2				
Rater 1	1	2	$\dots$	I	total
1	n_{11}	n_{12}	$\dots$	n_{1I}	n_{1+}
2	n_{21}	n_{22}	$\dots$	n_{2I}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
I	n_{I1}	n_{I2}	$\dots$	n_{II}	n_{I+}
total	n_{+1}	n_{+2}	$\dots$	n_{+I}	n_{++}

Table 2.2: Complete cross-classified table for two raters

Category	rater 2			
rater 1	1	2	total	missing
1	n^*_{11}	n_{12}	n_{1+}	a_1
2	n_{21}	n^*_{22}	n_{2+}	a_2
total	n_{+1}	n_{+2}	n_{++}	
missing	a_1	a_2		a

Table 2.3: 2×2 table for two raters with missing along the diagonal

by the two observers, which is estimated as

$$\pi_o = \sum_{i=1}^{I} \pi_{ii} \tag{3.1}$$

and under the baseline constraints of complete independence between ratings by the two observers,which is the expected agreement proportion estimated as

$$\pi_e = \sum_{i=1}^{I} \pi_{i.} \pi_{.i} \tag{3.2}$$

The kappa statistic, which measures the overall agreement level under observed value compared with the overall agreement level under expected value, can now be estimated by

$$\widehat{k} = \frac{\widehat{\pi}_o - \widehat{\pi}_e}{1 - \widehat{\pi}_e}, \tag{3.3}$$

where $\widehat{\pi}_o$ and $\widehat{\pi}_e$ are as defined above.

Category	rater 2			
rater 1	1	2	total	missing
1	n_{11}	n_{12}^{**}	n_{1+}	b_1
2	n_{21}^{**}	n_{22}	n_{2+}	b_2
total	n_{+1}	n_{+2}	n_{++}	
missing	b_2	b_1		b

Table 2.4: 2×2 table for two raters with missing off the diagonal

Category	rater 2			
rater 1	1	2	total	missing
1	n_{11}^{***}	n_{12}^{***}	n_{1+}	$a_1 + b_1$
2	n_{21}^{***}	n_{22}^{***}	n_{2+}	$a_2 + b_2$
total	n_{+1}	n_{+2}	n_{++}	
missing	$a_1 + b_2$	$a_2 + b_1$		a+b

Table 2.5: 2×2 table for two raters with missing in both along and off the diagonal

Earlier approaches to this problem have been focused on the observed proportion of agreement Goodman and Kruskal [8], suggesting that chance agreement can be ignored. Later Cohen's kappa was introduced for measuring nominal scale chance-corrected agreement. Scott [15] defined π_e using the underlying assumption that the distribution of proportions over the I categories for the population is known, and is equal for the two raters. Therefore if the two raters are interchangeable, in the sense that the marginal distributions are identical, then Cohen's and Scott's measures are equivalent because Cohen's kappa is an extension of Scott's index of chance-corrected measure. To determine whether $\widehat{k}$ differs significantly from zero, one could use the asymptotic variance formulae given by Fleiss et al [7] for the general I×I tables. Under the hypothesis of only chance agreement, the estimated large-sample variance of $\widehat{k}$ is given by

$$\widehat{var_o}(\widehat{k}) = \frac{\pi_e + \pi_e^2 - \sum_{i=1}^{I} \pi_{i.}\pi_{.i}(\pi_{i.} + \pi_{.i})}{n(1-\pi_e)^2}. \tag{3.4}$$

Assuming that

$$\frac{\widehat{k}}{\sqrt{\widehat{var_o}(\widehat{k})}}, \tag{3.5}$$

Kappa statistic	*Strength of agreement*
< 0.00	poor
0.00-0.20	slight
0.21-0.40	fair
0.41-0.60	moderate
0.61-0.80	substantial
0.81-1.00	almost perfect.

Table 3.1: Range of kappa statistic with the respective strength of agreement

Category	rater 2		
rater 1	1	2	total
1	π_{11}	π_{12}	π_{1+}
2	π_{21}	π_{22}	π_{2+}
total	π_{+1}	π_{+2}	π_{++}

Table 3.2: 2×2 probability distribution table

a normal distribution follows, one can test the hypothesis of chance agreement by reference to the standard normal distribution. In the context of reliability studies, however, this test of hypothesis is of little interest, since generally the raters are trained to be reliable. In this case, a lower bound on kappa is more appropriate.

Landis and Koch [11] have characterized different ranges of arbitrary values for kappa with respect to the degree of agreement they suggest and these have become a standard in all the literatures, see the ranges of kappa statistic with the respective strength of agreement as presented in Table 3.1.

For more on Cohen kappa statistic see also Jolayemi [10]; Barnhart and Williamson [3] and others.

3.1. Kappa Statistic with Missing Observations

In a given I×I contingency table as in Table 2.2, we assume that there are missing observation in some of the cells as given in Tables 2.3, 2.4, and 2.5.

Now the joint probability distribution table for raters 1 and 2 for a 2×2 table will be as in Table 3.2.

Cohen kappa statistic was given in equation (3.3) as

$$\widehat{k} = \frac{\widehat{\pi}_o - \widehat{\pi}_e}{1 - \widehat{\pi}_e}\,,$$

where

$$\pi_o = \sum_{i=1}^{I} \pi_{ii}$$

and

$$\pi_e = \sum_{i=1}^{I} \pi_{i.}\pi_{.i}\,.$$

In terms of cell counts, Cohen kappa statistic can as well be expressed as

$$\widehat{k} = \frac{n_{++}\sum_{i=1}^{I} n_{ii} - \sum_{i=1}^{I} n_{i+}n + i}{n_{++}^2 - \sum_{i=1}^{I} n_{i+}n_{+i}}\,, \tag{3.6}$$

if $I = 2$, Cohen kappa is

$$\widehat{k} = \frac{(\pi_{11} + \pi_{22}) - \{(\pi_{1+}\pi_{+1}) + (\pi_{2+}\pi_{+2})\}}{1 - \{(\pi_{1+}\pi_{+1}) + (\pi_{2+}\pi_{+2})\}}\,, \tag{3.7}$$

where

$$\pi_{11} = \frac{n_{11}}{n_{++}},\ \pi_{22} = \frac{n_{22}}{n_{++}},\ \pi_{1+} = \frac{n_{11} + n_{12}}{n_{++}},\ \pi_{+1} = \frac{n_{11} + n_{21}}{n_{++}},$$

and

$$\pi_{2+} = \frac{n_{21} + n_{22}}{n_{++}},\ \pi_{+2} = \frac{n_{12} + n_{22}}{n_{++}}.$$

Now, we assume that missing mechanism is ignorable as we have in Table 2.3 and Table 2.4 in some of the cells of the table. However, for simplicity in checking the effect of this missingness on Cohen kappa statistic, we substitute n_{11} with $(n_{11} - a_1)$, n_{1+} with $(n_{1+} - a_1)$, n_{+1} with $(n_{+1} - a_1)$ and n_{++} with $(n_{++} - a_1)$ for a_1 missing observations along the diagonal as we have in Table 2.3, such that,

$$\pi_{11}^* = \frac{n_{11} - a_1}{(n_{++} - a_1)} \neq \frac{n_{11}}{n_{++}} \quad \text{if } a_1 \neq 0,$$

$$\pi_{1+}^* = \frac{(n_{11} - a_1) + n_{12}}{(n_{++} - a_1)} \neq \frac{(n_{11} + n_{12})}{n_{++}} \quad \text{if } a_1 \neq 0,$$

$$\pi_{+1}^* = \frac{(n_{11} - a_1) + n_{21}}{(n_{++} - a_1)} \neq \frac{(n_{11} + n_{21})}{n_{++}} \quad \text{if } a_1 \neq 0.$$

Then Cohen kappa statistic in equation (3.7) will become

$$\begin{aligned}
\widehat{k} &= \frac{(\pi^*_{11}+\pi_{22})-\{(\pi^*_{1+}\pi^*_{+1})+(\pi_{2+}\pi_{+2})\}}{1-\{(\pi^*_{1+}\pi^*_{+1})+(\pi_{2+}\pi_{+2})\}}\\
&= \frac{((\frac{(n_{11}-a_1)}{(n_{++}-a_1)})+\pi_{22})-\{((\frac{(n_{11}-a_1)+n_{12}}{(n_{++}-a_1)})(\frac{(n_{11}-a_1)+n_{21}}{(n_{++}-a_1)}))+(\pi_{2+}\pi_{+2})\}}{1-\{((\frac{(n_{11}-a_1)+n_{12}}{(n_{++}-a_1)})(\frac{(n_{11}-a_1)+n_{21}}{(n_{++}-a_1)}))+(\pi_{2+}\pi_{+2})\}}\\
&= \frac{((\frac{(n_{11}-a_1)}{(n_{++}-a_1)})+\frac{n_{22}}{(n_{++}-a_1)})-\{((\frac{(n_{11}-a_1)+n_{12}}{(n_{++}-a_1)})(\frac{(n_{11}-a_1)+n_{21}}{(n_{++}-a_1)}))+(\frac{n_{2+}}{(n_{++}-a_1)}\frac{n_{+2}}{(n_{++}-a_1)})\}}{1-\{((\frac{(n_{11}-a_1)+n_{12}}{(n_{++}-a_1)})(\frac{(n_{11}-a_1)+n_{21}}{(n_{++}-a_1)}))+(\frac{n_{2+}}{(n_{++}-a_1)}\frac{n_{+2}}{(n_{++}-a_1)})\}}\\
&= \frac{\{(n_{++}-a_1)(n_{11}+n_{22}-a_1)\}-(n_{11}-a_1)^2-\{(n_{12}+n_{21})(n_{11}-a_1)\}-n_{12}n_{21}-n_{2+}n_{+2}}{(n_{++}-a_1)^2-(n_{11}-a_1)^2-\{(n_{12}+n_{21})(n_{11}-a_1)\}-n_{12}n_{21}-n_{2+}n_{+2}}\\
&= \frac{\{(n_{++}-a_1)(n_{11}+n_{22}-a_1)\}-\{(n_{11}-a_1)^2+\{C(n_{11}-a_1)\}+D\}}{(n_{++}-a_1)^2-\{(n_{11}-a_1)^2+\{C(n_{11}-a_1)\}+D\}}\\
&\neq \frac{\{n_{++}(n_{11}+n_{22})\}-\{n^2_{11}+\{C(n_{11})\}+D\}}{n^2_{++}-\{n^2_{11}+\{C(n_{11}\}+D\}} \quad \text{if } a_1\neq 0, \qquad (3.8)
\end{aligned}$$

where

$$c=(n_{12}+n_{21}),\ D=(n_{12}n_{21}+n_{2+}n_{+2}).$$

For $'b_1'$ missing observations in off the diagonal of the table as we have in Table 2.4, we also substitute n_{12} with $(n_{12}-b_1)$, n_{1+} with $(n_{1+}-b_1)$, n_{+2} with $(n_{+2}-b_1)$ and n_{++} with $(n_{++}-b_1)$, such that,

$$\pi^{**}_{12}=\frac{n_{12}-b_1}{(n_{++}-b_1)}\neq\frac{n_{12}}{n_{++}} \quad \text{if } b_1\neq 0,$$

$$\pi^{**}_{1+}=\frac{n_{11}+(n_{12}-b_1)}{(n_{++}-b_1)}\neq\frac{(n_{11}+n_{12})}{n_{++}} \quad \text{if } b_1\neq 0,$$

$$\pi^{**}_{+2}=\frac{(n_{12}-b_1)+n_{22}}{(n_{++}-b_1)}\neq\frac{(n_{12}+n_{22})}{n_{++}} \quad \text{if } b_1\neq 0.$$

Cohen kappa statistic will then become

$$\begin{aligned}
\widehat{k} &= \frac{(\pi_{11}+\pi_{22})-\{(\pi^{**}_{1+}\pi_{+1})+(\pi_{2+}\pi^{**}_{+2})\}}{1-\{(\pi^{**}_{1+}\pi_{+1})+(\pi_{2+}\pi^{**}_{+2})\}}\\
&= \frac{(\pi_{11}+\pi_{22})-\{(\frac{n_{11}+(n_{12}-b_1)}{(n_{++}-b_1)})(\frac{n_{11}+n_{21}}{(n_{++}-b_1)})+(\frac{n_{21}+n_{22}}{(n_{++}-b_1)})(\frac{(n_{12}-b_1)+n_{22}}{(n_{++}-b_1)})\}}{1-\{((\frac{n_{11}+(n_{12}-b_1)}{(n_{++}-b_1)})(\frac{n_{11}+n_{21}}{(n_{++}-b_1)}))+(\frac{n_{21}+n_{22}}{(n_{++}-b_1)})(\frac{(n_{12}-b_1)+n_{22}}{(n_{++}-b_1)})\}}\\
&= \frac{\{\frac{n_{11}}{(n_{++}-b_1)}+\frac{n_{22}}{(n_{++}-b_1)}\}-\{(\frac{n_{11}+(n_{12}-b_1)}{(n_{++}-b_1)})(\frac{n_{11}+n_{21}}{(n_{++}-b_1)})+(\frac{n_{21}+n_{22}}{(n_{++}-b_1)})(\frac{(n_{12}-b_1)+n_{22}}{(n_{++}-b_1)})\}}{1-\{((\frac{n_{11}+(n_{12}-b_1)}{(n_{++}-b_1)})(\frac{n_{11}+n_{21}}{(n_{++}-b_1)}))+(\frac{n_{21}+n_{22}}{(n_{++}-b_1)})(\frac{(n_{12}-b_1)+n_{22}}{(n_{++}-b_1)})\}}\\
&= \frac{\{(n_{++}-b_1)(n_{11}+n_{22})\}-n^2_{11}-n^2_{22}-\{(n_{+1}+n_{2+})(n_{12}-b_1)\}-n_{11}n_{21}-n_{21}n_{22}}{\{(n_{++}-b_1)^2\}-n^2_{11}-n^2_{22}-\{(n_{+1}+n_{2+})(n_{12}-b_1)\}-n_{11}n_{21}-n_{21}n_{22}}\\
&= \frac{\{(n_{++}-b_1)(n_{11}+n_{22})\}-\{P+\{Q(n_{12}-b_1)\}+R\}}{\{(n_{++}-b_1)^2\}-\{P+\{Q(n_{12}-b_1)\}+R\}}
\end{aligned}$$

$$\neq \quad \frac{\{n_{++}(n_{11}+n_{22})\} - \{P + \{Q(n_{12})\} + R\}}{\{n_{++}^2\} - \{P + \{Q(n_{12})\} + R\}} \quad \text{if } b_1 \neq 0 \tag{3.9}$$

where

$$P = (n_{11}^2 + n_{22}^2),\ Q = (n_{+1} + n_{2+}),\ \text{and } R = (n_{11}n_{21} + n_{21}n_{22}).$$

Also for a_1 and b_1 respectively being missing observations along and off the diagonal of the table as we have in Table 2.5, we substitute n_{11} with $(n_{11} - a_1)$, n_{12} with $(n_{12} - b_1)$, n_{1+} with $(n_{1+} - (a_1 + b_1))$, n_{+1} with $(n_{+1} - a_1)$, n_{+2} with $(n_{+2} - b_1)$ and n_{++} with $(n_{++} - (a_1 + b_1))$, such that,

$$\pi_{11}^{***} = \frac{n_{11} - a_1}{(n_{++} - (a_1 + b_1))} \neq \frac{n_{11}}{n_{++}} \quad \text{if } a_1, b_1 \neq 0,$$

$$\pi_{12}^{***} = \frac{n_{12} - b_1}{(n_{++} - (a_1 + b_1))} \neq \frac{n_{12}}{n_{++}} \quad \text{if } a_1, b_1 \neq 0,$$

$$\pi_{1+}^{***} = \frac{(n_{11} - a_1) + (n_{12} - b_1)}{(n_{++} - (a_1 + b_1))} \neq \frac{(n_{11} + n_{12})}{n_{++}} \quad \text{if } a_1, b_1 \neq 0,$$

$$\pi_{+1}^{***} = \frac{(n_{11} - a_1) + n_{21}}{(n_{++} - (a_1 + b_1))} \neq \frac{(n_{11} + n_{21})}{n_{++}} \quad \text{if } a_1, b_1 \neq 0,$$

$$\pi_{+2}^{***} = \frac{(n_{12} - b_1) + n_{22}}{(n_{++} - (a_1 + b_1))} \neq \frac{(n_{12} + n_{22})}{n_{++}} \quad \text{if } a_1, b_1 \neq 0.$$

Cohen kappa statistic will then become

$$\widehat{k} = \frac{(\pi_{11}^{***} + \pi_{22}) - \{(\pi_{1+}^{***}\pi_{+1}^{***}) + (\pi_{2+}\pi_{+2}^{***})\}}{1 - \{(\pi_{1+}^{***}\pi_{+1}^{***}) + (\pi_{2+}\pi_{+2}^{***})\}}$$

$$= \frac{(\frac{(n_{11}-a_1)+n_{22}}{(n_{++}-(a_1+b_1))}) - \{\frac{(n_{1+}-(a_1+b_1))(n_{+1}-a_1)+(n_{2+})(n_{+2}-b_1)}{(n_{++}-(a_1+b_1))^2}\}}{1 - \{\frac{(n_{1+}-(a_1+b_1))(n_{+1}-a_1)+(n_{2+})(n_{+2}-b_1)}{(n_{++}-(a_1+b_1))^2}\}} =$$

$$\frac{\{(n_{++} - (a_1 + b_1))((n_{11} - a_1) + n_{22})\} - \{(n_{+1} - a_1)(n_{1+} - (a_1 + b_1))\} - \{n_{2+}(n_{+2} - b_1)\}}{\{(n_{++} - (a_1 + b_1))^2\} - \{(n_{+1} - a_1)(n_{1+} - (a_1 + b_1))\} - \{n_{2+}(n_{+2} - b_1)\}}$$

$$\neq \frac{\{(n_{++})(n_{11} + n_{22})\} - \{(n_{+1}n_{1+}) + (n_{2+}n_{+2})\}}{\{n_{++}^2\} - \{(n_{+1}n_{1+}) + (n_{2+}n_{+2})\}} \text{ if } a_1, b_1 \neq 0. \tag{3.10}$$

3.2. Case-by-Case Formulations

So in order to examine further the effects of the missing observation along the diagonal, off the diagonal or both combined, for a given table of ratings of raters on n_{++} subjects in a Cohen kappa statistic, we consider some appropriate and relevant empirical examples with certain percentages missing in the next section by considering the following pattern cases.

For each of the cases below we consider 5% and 10% missing observations of the total counts n_{++}.

— *Case 1.* When there are certain percentages missing along the diagonal.

— *Case 2.* When there are certain percentages missing off the diagonal.

— *Case 3.* When there are certain percentages missing in both along and off the diagonal combined.

We obtain the Cohen kappa statistic, the standard error based on the square root of variance of kappa given in equation (3.4) as

$$\widehat{var_o}(\widehat{k}) = \frac{\pi_e + \pi_e^2 - \sum_{i=1}^{I} \pi_{i.}\pi_{.i}(\pi_{i.} + \pi_{.i})}{n(1-\pi_e)^2}.$$

By assuming that

$$\frac{\widehat{k}}{\sqrt{\widehat{var_o}(\widehat{k})}},$$

follows a normal distribution which was given in equation (3.5) for each of the tables with different sizes of missing observation.

4. Empirical Examples

Example 1. Consider the sample of juvenile convicted of a felony in Florida in 1987 (Agresti, [1]). Matched pairs were formed using criteria such as age and the number of prior offenses. For each pair, one subject was handled in the juvenile courts and the other was transferred to the adult courts. The response of interest was whether the juvenile was rearrested by the end of 1988.

Here:

— k_c is Cohen kappa estimate for the complete table without missing,

— k_{5d} and k_{10d} are Cohen kappa estimates for 5% and 10% missing from n_{++} for Case 1,

— k_{5ofd} and k_{10ofd} are Cohen kappa estimates for 5% and 10% missing from n_{++} for Case 2 and

— k_{5b} and k_{10b} are Cohen kappa estimates for 5% and 10% missing from n_{++} for Case 3.

Category	juvenile court		
Adult court	rearrested	no rearrested	total
rearrested	158	515	673
no rearrested	290	1134	1424
total	448	1649	2097

Table 4.1: Cross-classification of trials of juvenile convicted of a felony in juvenile and adult courts

	type	Cohen Kappa	stand.error	z-value
Complete	k_c	0.03412657	0.02102658	1.623021
Case 1	k_{5d}	-0.02358988	0.021469	-1.098788
	k_{10d}	-0.08869	0.02192879	-4.044739.
Case 2	k_{5ofd}	0.08728164	0.02163029	4.035159
	k_{10ofd}	0.1490395	0.0228779	6.687046
Case 3	k_{5b}	0.03087266	0.02155309	1.4324
	k_{10b}	0.02706453	0.02211314	1.223912

Table 4.2: Cohen kappa statistic estimates for juvenile convicted of a felony in juvenile and adult courts

Example 2. Consider the data arising from the study reported in Holmquist et al [9] that investigated the variability in the classification of carcinoma in situ of the uterine cervix in which seven pathologists were requested to separately evaluate and classify 118 slides into one of the the following five categorical scales based on the most involved lesion: 1=negative; 2=atypical squamous hyperplasia; 3=carcinoma in situ; 4=squamous carcinoma with early stromal invasion; 5=invasive carcinoma. These pathologists are labelled with letters A, B, C, D, E, F and G. We have the cross-classification for pathologists A and G as follows.

Example 3. Consider the data on journal citation among four statistical theory and methods journals during 1987-1989 (Agresti, [1]). The more often that articles in a particular journal are cited, the more prestige that journal accrues. For citations involving a pair of journals X and Y, view it as a "victory" for X if it is cited by Y and a "defeat" for X if it is cites Y. The categories used are BIOM=*Biometrika*, COMM=*Communications in Statistics*, JASA=*Journal of the American Statistical Association*, JRSSB=*Journal of the Royal Statistical Society Series B.*

Category	Pathologist G					
Pathologist A	1	2	3	4	5	Total
1	24	2	0	0	0	26
2	7	13	6	0	0	26
3	1	4	32	1	0	38
4	0	1	20	1	0	22
5	0	0	3	1	2	6
Total	32	20	61	3	2	118

Table 4.3: Cross-classification of pathologists A and G on carcinoma in situ of the uterine cervix of 118 slides

	type	Cohen Kappa	stand.error	z-value
Complete	k_c	0.4666	0.04991	9.349308
Case 1	k_{5d}	0.4425	0.05041	8.7772
	k_{10d}	0.4094	0.0510682	8.019383
Case 2	k_{5ofd}	0.5060095	0.052471	9.643692
	k_{10ofd}	0.5452338	0.05432	10.03808
Case 3	k_{5b}	0.4728546	0.05137662	9.203693
	k_{10b}	0.47988	0.05295	9.062188

Table 4.4: Cohen kappa statistic estimates for pathologists A and G on carcinoma in situ

5. Results and Conclusion

5.1. Results

We present the summary results for the empirical examples in the previous section for Cohen kappa under various cases with the appropriate strength as proposed by Landis and Koch [11].

From the summary Table 5.1 to Table 5.3 which are for different cases stated in Section 3.2, we observed that the higher the percentage of missing observations along the diagonal the lower the Cohen kappa statistic estimate become, that is, the value reduces as the percentage of missing value increases along the diagonal to the extend that the class of strength of agreement for some examples change. However, we also observed that the higher the percentage of missing observations off the diagonal cells, the higher the Cohen kappa statistic

Category	Cited journal				
Citing journal	BIOM	COMM	JASA	JRSSB	Total
BIOM	714	33	320	284	1351
COMM	730	425	513	276	1944
JASA	498	68	1072	325	1963
JRSSB	221	17	142	188	568
Total	2163	543	2047	1073	5826

Table 4.5: Cited and citing journals of four statistical theory and methods journals

	type	Cohen Kappa	stand.error	z-value
Complete	k_c	0.2119863	0.007152349	29.9387
Case 1	k_{5d}	0.1719391	0.007152349	23.62517
	k_{10d}	0.1277029	0.007404541	17.24657
Case 2	k_{5ofd}	0.2370145	0.007454658	31.79414
	k_{10ofd}	0.007776741	0.06107083	34.15594
Case 3	k_{5b}	0.2041593	0.00736901	27.70513
	k_{10b}	0.1956437	0.0076062	25.72161

Table 4.6: Cohen kappa statistic estimates for cited and citing journals

	Complete		5% missing		10% missing	
Example	k_c	strength	$k_{5\%}$	strength	$k_{10\%}$	strength
1	0.03413	slight	-0.02359	poor	-0.08869	poor
2	0.4666	moderate	0.4425	moderate	0.4094	moderate
3	0.2120	fair	0.1719	slight	0.1277	slight

Table 5.1: Summary table of results for Case 1

estimate. For missing observation in both along and off diagonal, we observed that the higher the percentage, the higher the Cohen kappa if the table is sparse in nature as we have in Example 2. But if the table is not sparse as we have in Example 1 and Example 3, the higher the percentage of missing the lower the Cohen kappa statistic estimate. We also tried one more sparse data which was the cross-classification of pathologists D and F on carcinoma in situ of the uterine cervix of 118 slides as presented by Holmquist et al [9] and the same conclusion as we have for Example 2 was reached.

	Complete		5% missing		10% missing	
Example	k_c	strength	$k_{5\%}$	strength	$k_{10\%}$	strength
1	0.03413	slight	0.08728	slight	0.1490	slight
2	0.4666	moderate	0.5060	moderate	0.5452	moderate
3	0.2120	fair	0.2370	fair	0.2656	fair

Table 5.2: Summary table of results for Case 2

	Complete		5% missing		10% missing	
Example	k_c	strength	$k_{5\%}$	strength	$k_{10\%}$	strength
1	0.03413	slight	0.03087	slight	0.02706	slight
2	0.4666	moderate	0.4729	moderate	0.4799	moderate
3	0.2120	fair	0.2042	fair	0.1956	slight

Table 5.3: Summary table of results for Case 3

5.2. Conclusion

Cohen kappa has been one of the major statistics for measuring the strength of agreement between raters as we have demonstrated in this paper. However, in a situation where there are missing observations in the cells along the diagonal, the estimate of this statistic become worse as the number of missing observations increases. Also this statistic become better as the number of missing observations in cells off diagonal increases. In summary, missingness in the cells off the diagonal improved the strength of agreement while missingness in the cells along the diagonal reduced the strength of agreement that would be given by Cohen kappa statistic.

References

[1] A. Agresti, *Introduction to Categorical data analysis*, Wiley (1996).

[2] M. Banerjee, M. Capozzoli, L. Mcsweeney, D. Sinha, Beyond Kappa: A review of interrater agreement measure, *The Cana. J. of. Statist.*, **27**, No. 1 (1999), 3-23.

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

[4] D.A. Bloch, H.C. Kraemer, 2×2 kappa coefficients: Measures of agreement or association, *Biometrics*, **45** (1988), 269-287.

[5] J. Cohen, A coefficient of agreement for nominal scales, *Edu. and Psych. Meas.*, **20** (1960), 37-46.

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

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

[8] L.A. Goodman, W.H. Kruskal, Measuring of association for cross classifications, *J. Amer. Statist. Assoc.*, **49** (1954), 732-768.

[9] N.S. Holmquist, C.A. McMahon, O.D. Williams, Variability in classification of carcinoma in situ of the uterine cervix, *Archives of Pathology*, **84** (1967), 334-345.

[10] E.T. Jolayemi, On the measure of agreement between two raters, *Biometrika*, **32**, No. 1 (1990), 87-93.

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

[12] R.J.A. Little, D.B. Bubin, *Statistical Analysis with Missing Data*, New York, Wiley (1987).

[13] D.B. Rubin, Inference and missing data, *Biometrika*, **63** (1976), 581-592.

[14] J.L. Schafer, *Analysis of Incomplete Multivariate Data*, Chapman and Hall (1997).

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

[16] M.M. Shoukri, *Measures of Interobserver Agreement*, Chapman and Hall (2004).

[17] H. Toutenburg, C. Heumann, T. Nittner, S. Scheid, Parametric and Nonparametric Regression with Missing $X'S$ – A review, *J. Iranian Statist. Socie.*, **1**, No. 1, 2 (2002), 79-109.

[18] H. Toutenburg, T. Nittner, Linear regression models with incomplete categorical covariates, *Computational Statist.*, **17**, No. 2 (2002), 215-232.