Ilorin Journal of Science
Volume 2, Number 1, 2015, pp. 110 – 129 (Printed in Nigeria)
ISSN: 2408 – 4840 © 2014 Faculty of Physical Sciences, University of Ilorin
https://doi.org/10.54908/iljs.2015.02.01.009

ILORIN
JOURNAL OF SCEINCE

ILJS-15-037

A Study of Conventional Missing at Random on Kappa-like Statistics for Raters Agreement Measurement

Adejumo*, A. O.

Department of Statistics, University of Ilorin, Ilorin, Nigeria.

Abstract

Kappa-like statistics such as Cohen Kappa and Intraclass Kappa have been in existence in the last four decades. They are mainly for the measurement of overall level of agreement that may exist between two raters. Subjects are being classified into categories by raters, interviewers or observers in all life or social science researches. In any experiment that involves measurement, counting, treatments or diagnosis, irrespective of the level of reliability, there could be likelihood for some elements of missing observations. In this research work, we examine the effects of these missing observations with different missing patterns from the square table of cross-classified ratings of raters on these two Kappa-like statistics. We assumed a conventional missing at random (CMAR) which is a special case of missing at random (MAR) mechanism criteria. We observed that value from Intraclass Kappa statistic are consistently lower than their respective estimates from Cohen Kappa statistic and missingness improved or worsened the strength of agreement as the missing percentage increases depending on the missing pattern under consideration.

Keywords: Agreement, Kappa statistic, intraclass, conventional missing at random (CMAR), raters.

1. Introduction

Statistical analysis with missing data is a common problem in practice. Nonresponse in a sample survey or drop-out in clinical trials may be two of many examples one could imagine. Any square contingency table can be used to display joint ratings from two raters. 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. Categorical data are data that the response variable is classified into either nominal or ordinal categories. For nominal data, which is our focus in this paper, as reviewed by Banerjee *et al.* (1999) a large numbers of estimation and testing procedures like the Cohen Kappa (Cohen, 1960), the intraclass Kappa (Block and Kraemer, 1988).

*Corresponding Author: Adejumo, A. O.
E-mail: aodejumo@unilorin.edu.ng

For other improved methods on Kappa statistic, see Shoukri (2004), Thomas *et al*. (2008), Manuel *et al.* (2011), Martin and Femia (2008) and Rueben (2015). 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 called weighted Kappa (Cohen, 1968) which accounts for severity of discordance or size of discrepancy is better suited for ordinal data.

Missing value problem in statistical analysis has being in existence over a long decades. For instance, in a medical screening study, an inexpensive or easily administered test can be given to a large number of subjects. For the purpose of calibration a second, or more expensive, and more reliable test is administered to the same subjects. Missing values are likely to occur in this study because not all will be available on the second occasion. In clinical trials of long-time studies, some individuals may not cooperate or do not participate over the whole period and drop out may be experienced. In the analysis of lifetime data, these individuals are called censored. Censoring is a mechanism causing nonrandomly missing data (Toutenburg and Nittner, 2002).

In this paper we compare the two Kappa-like statistics and also examine their behaviour on raters agreement measure where there are missing observations with different patterns. In the next two sections we present missing value and description of the two Kappa-like statistics with and without missing observation. Some empirical examples are in section 4 and results and conclusion are in section 5.

2. Materials and Methods

Little and Rubin (1987) and Rubin (1976) 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 analyse rectangular data sets, that is to analyze a matrix

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

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

$$P(X \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta), \quad (2)$$

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 of 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 behaviour. 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.* (2002) for details on general missing values pattern.

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:

$$r_{ij} = \begin{cases} 1, & if \quad z_{ij} \quad observed \\ 0, & if \quad z_{ij} \quad missing \end{cases}, \tag{3}$$

for all i=1,2,...,n, j=1,2,...,p+1.

The ignorability criteria of the missing data mechanism depends on whether statistical inference is based on the density $f(R,Z_{obs}|\theta,\varphi)$ 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 mechanisms is thus based on the density $f(R|\ Z_{obs},Z_{mis}\ \varphi)$ see Schafer (1997), Toutenburg and Nittner (2002), Toutenburg *et al.* (2002) for more details.

In this research work we assume a conventional missing at random (CMAR) missing criteria mechanism for the missingness in such region of the square table (Adejumo *et al.* 2004). These sectional parts may be observations in cells along the diagonal, lower diagonal, upper diagonal, off diagonal of the square table and so on. CMAR is assumed because of the special attention given to such specific region of the table under consideration before assuming further missing at random (MAR) in such region rather than considering the complete parts of the table together.

Missing values in the ratings of raters

Missing observations can also be observed in the raw results for some of the subjects involve in an experiment or in trials. For example, consider a survey on the choice of decaffeinated Coffee at two different purchase dates for five different brands of Coffee. Somebody may be involved in the first purchase but not available again on the date for the second purchase or died before that date. Also if certain numbers of slides are to be examined for variability in the classification of carcinoma in situ of the uterine cervix by two different pathologists, if one pathologist has successfully classified the slides, but before getting to the next pathologist, some of the slides got broken, misplaced, or wrongly labelled or identified. We refer to these two cases as missing values in such experiments. Tables 7 and 8 presents the

complete tables of such experiments. To this effect, Table 1 described the pattern of how missing observations can occur in the raw ratings of two raters on some sets of subjects.

Table 1: Missing pattern for ratings of two raters.

	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*
.	.	.	.	.	.
.	.	.	.	.	.
.	1	4	.	.	.
n			1	3	Obs

The {*resp}* stands for the response status for i^{th} subject with j^{th} rater which we have defined in equation 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:

$$r_{ij} = \begin{cases} 1, & if \quad z_{ij} \quad observed \\ \\ 0, & if \quad z_{ij} \quad missing \end{cases}, \tag{4}$$

for all i=1,2,...,n, j=1,2,...,p+1, such that the matrix for R base on Table 1 can be of the form

$$R = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 0 & 0 \\ . & . \\ . & . \\ . & . \\ 1 & 1 \end{pmatrix} . \tag{5}$$

Also each subject combined response status S_i is defined as:

$$S_i = \begin{cases} obs, \quad if \quad r_{i1} = r_{i2} = 1 \quad as \quad observed \\ mis, \quad if \quad (r_{i1} = 0 \quad and \quad r_{i2} = 1 \quad or \quad r_{i1} = 1 \& \quad r_{i2} = 0) \quad as \quad missing \\ mis^*, \quad if \quad r_{i1} = r_{i2} = 0 \quad as \quad totally \quad missing \quad \forall i = 1,2,...,n. \end{cases} \tag{6}$$

Missing value pattern in the ratings of raters

The missing pattern depends on the nature of missingness as we have presented in Table 1. Take for instance if the categorical scale for the raters is numbered 1 to 5, if the response of rater 1 to one subject is 1, and that of rater 2 to the 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 x 2 contingency tables are as given in Tables 3 to 5.

Table 2: Complete cross-classified table for two raters.

Category	Rater 2	
Rater 1	1 2 ... I	total
1	n_{11} n_{12} ... n_{1I}	n_{1+}
2	n_{21} n_{22} ... n_{2I}	n_{2+}
.	. . . .	.
.	. . . .	.
.	. . ..	.
I	n_{I1} n_{I2} ... n_{II}	n_{I+}
total	N_{+1} n_{+2} ... n_{+I}	n_{++}

Now assume that we have a 2 x 2 contingency table for the ratings of two raters in Table 3.

Table 3: 2 x 2 table for two raters with missing along the diagonal.

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 4: 2 x 2 table for two raters with missing off the diagonal.

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

Table 5: 2 x 2 table for two raters with missing in both along and off the diagonal.

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

Kappa-like statistics

We present two Kappa-like statistics (Cohen Kappa and Intraclass Kappa statistics) that are often used to measure agreement without necessarily attaching weights as in the case of weighted Kappa statistic which was proposed by Cohen (1968).

Cohen's Kappa coefficient

Cohen (1960) proposed a standardized coefficient of raw agreement for nominal scales in terms of the proportion of the subjects classified into the same category by the two observers, which is estimated as

$$\pi_o = \sum_{i=1}^{I} \pi_{ii} \qquad (7)$$

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

$$\pi_e = \sum_{i=1}^{I} \pi_{i.}\pi_{.i}. \qquad (8)$$

The Kappa statistic can now be estimated by:

$$\hat{k}_c = \frac{\hat{\pi}_o - \hat{\pi}_e}{1 - \hat{\pi}_e} \quad , \tag{9}$$

where $\hat{\pi}_o$ and $\hat{\pi}_e$ are as defined above. Cohen's Kappa was introduced for measuring nominal scale chance-corrected agreement. To determine whether $\hat{k}_c$ differs significantly from zero, one could use the asymptotic variance formulae given by Fleiss *et al.* (1969) for the general I x I tables. Under the hypothesis of only chance agreement, the estimated large-sample variance of $\hat{k}_c$ is given by

$$\hat{\text{var}}_o(\hat{k}_c) = \frac{\pi_e + \pi_e^2 - \sum_{i=1}^{I} \pi_{i.}\pi_{.i}(\pi_{i.} + \pi_{.i})}{n(1-\pi_e)^2}. \tag{10}$$

Assuming that

$$\frac{\hat{k}_c}{\sqrt{\hat{\text{var}}_o(\hat{k}_c)}} \tag{11}$$

follows a normal distribution, one can test the hypothesis of chance agreement by reference to the standard normal distribution. And the confidence interval (CI) of size

$100(1-\alpha)\% = \hat{k}_c \pm z_{1-\frac{\alpha}{2}} SE(\hat{k}_c)$ can be obtained for $\hat{k}_c$, where *SE* is the standard error.

Intraclass Kappa

Intraclass Kappa was defined for data consisting of blind dichotomous ratings on each of n subjects by two fixed raters. It is assumed that the ratings on a subject are interchangeable; that is in the population of subjects, the two ratings for each subject have a distribution that is invariant under permutations of the raters to ensure that there is no rater bias (Bloch and Kraemer, 1989; Barnhart and Williamson, 2002; and others). Let X_{ij} denote the rating for the i^{th} subject by the j^{th} rater, i=1,2,...,n, j=1,2. For each subject 'i', let $\pi_i = P(X_{ij} = 1)$, be the probability that the rating is a success. Over the population of subjects, let $E(\hat{\pi}_i) = \Pi$, $\Pi' = (1 - \Pi)$ and $\text{var}(\hat{\pi}_i) = \sigma_\pi^2$. The intraclass Kappa as defined by Bloch and Kraemer (1989) is then

$$k_I = \frac{\sigma_\pi^2}{\Pi\Pi'}. \tag{12}$$

Thus, the log-likelihood function is given as

$$\log L(\Pi, k_I \setminus \mathrm{n}_{11}, n_{12}, n_{21}, n_{22}) = n_{11}\log(\pi^2 + k_I\Pi\Pi') + (n_{12}+n_{21}) \times \log\{\Pi\Pi'(1-k_I)\} + n_{22}\log(\pi'^2 + k_I\Pi\Pi').$$

The maximum likelihood estimators $\hat{\pi}$ and $\hat{k}_I$ for Π and k_I are obtained as

$$\hat{\pi} = \frac{2n_{11} + n_{12} + n_{21}}{2n} \tag{13}$$

and

$$\hat{k}_I = \frac{4(n_{11}n_{22} - n_{12}n_{21}) - (n_{12} - n_{21})^2}{(2n_{11} + n_{12} + n_{21})(2n_{22} + n_{12} + n_{21})}, \tag{14}$$

with the estimated standard error for $\hat{k}_I$ given as

$$SE(\hat{k}_I) = \{\frac{1-\hat{k}_I}{n}[(1-\hat{k}_I)(1-2\hat{k}_I) + \frac{\hat{k}_I(2-\hat{k}_I)}{2\hat{\pi}(1-\hat{\pi})}]\}^{\frac{1}{2}}. \tag{15}$$

This is only restricted to a 2 x 2 table.

Barnhart and Williamson (2002) considered intraclass Kappa for measuring agreement between two readings for a categorical response with I categories if the two readings are replicated measurements. It assumes no bias because the probability of a positive rating is the same for the two readings due to replication, and it is given as:

$$\hat{k}_{In} = \frac{\sum_{i=1}^{I}\pi_{ii} - \sum_{i=1}^{I}((\pi_{i.} + \pi_{.i})/2)^2}{1 - \sum_{i=1}^{I}((\pi_{i.} + \pi_{.i})/2)^2}, \tag{16}$$

with variance for the estimated value as:

$$\hat{\mathrm{var}}_o(\hat{k}_{In}) = \frac{\pi_{\exp} + (\pi_{\exp})^2 - \pi_{\exp}(\sum_{i=1}^{I}((\pi_{i.}\pi_{.i})/2)^2)}{n(1-\pi_{\exp})^2}, \tag{17}$$

where

$$\pi_{\exp} = \sum_{i=1}^{I}((\pi_{i.} + \pi_{.i})/2)^2 \quad .$$

We also assumed that:

$$\frac{\hat{k}_{In}}{\sqrt{\hat{\mathrm{var}}_o(\hat{k}_{In})}} \tag{18}$$

follows a normal distribution. Also the confidence interval (CI) of size

$100(1-\alpha)\% = \hat{k}_{In} \pm z_{1-\frac{\alpha}{2}} SE(\hat{k}_{In})$ can be obtained for $\hat{k}_{In}$.

Kappa-like statistics with missing observations

In a given I x I contingency table as in Table 2, we assume that the missingness mechanism criterion is a conventional missing at random (CMAR) with special missing patterns in the cells as given in Tables 3, 4, and 5.

Now the joint probability distribution table for raters 1 and 2 for a 2 x 2 table will be as in Table 6.

Table 6: 2 x 2 probability table

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_{++}

Cohen Kappa statistic was given in equation 9 as:

$$\hat{k}_c = \frac{\hat{\pi}_o - \hat{\pi}_e}{1-\hat{\pi}_e} \quad ,$$

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 becomes

$$\hat{k}_c = \frac{n_{++}\sum_{i=1}^{I} n_{ii} - \sum_{i=1}^{I} n_{i+}n_{+i}}{n_{++}^2 - \sum_{i=1}^{I} n_{i+}n_{+i}}. \qquad (19)$$

If *I=2*, Cohen Kappa is

$$\hat{k}_c = \frac{(\pi_{11}+\pi_{22}) - (\pi_{1+}\pi_{+1} + \pi_{2+}\pi_{+2})}{1-(\pi_{1+}\pi_{+1} + \pi_{2+}\pi_{+2})}, \qquad (20)$$

where

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

$$\pi_{1+} = \frac{n_{11}+n_{12}}{n_{++}}, \pi_{+2} = \frac{n_{12}+n_{22}}{n_{++}}, \quad \pi_{+1} = \frac{n_{11}+n_{21}}{n_{++}}, \ \pi_{2+} = \frac{n_{21}+n_{22}}{n_{++}}.$$

Intraclass Kappa statistic was given as:

$$k_{In} = \frac{\sum_{i=1}^{I}\pi_{ii} - \sum_{i=1}^{I}((\pi_{i+}+\pi_{+i})/2)^2}{1-\sum_{i=1}^{I}((\pi_{i+}+\pi_{+i})/2)^2}.$$

This can also be expressed in terms of counts data rather than proportion as follows:

$$k_{In} = \frac{\sum_{i=1}^{I}\frac{n_{ii}}{n_{++}} - \sum_{i=1}^{I}((\frac{n_{i+}}{n_{++}}+\frac{n_{+i}}{n_{++}})/2)^2}{1-\sum_{i=1}^{I}((\frac{n_{i+}}{n_{++}}+\frac{n_{+i}}{n_{++}})/2)^2}.$$

For I=2, Intraclass Kappa becomes

$$k_{In} = \left(\frac{\frac{n_{11}}{n_{++}}+\frac{n_{22}}{n_{++}} - ((\frac{n_{1+}}{n_{++}}+\frac{n_{2+}}{n_{++}}+\frac{n_{+1}}{n_{++}}+\frac{n_{+2}}{n_{++}})/2)^2}{1-((\frac{n_{1+}}{n_{++}}+\frac{n_{2+}}{n_{++}}+\frac{n_{+1}}{n_{++}}+\frac{n_{+2}}{n_{++}})/2)^2} \right).$$

Now, as earlier mentioned we assume that the missing mechanism is a conventional missing at random (CMAR) with special missing patterns as in Tables 3 and 4 in some of the cells of the table. However, for simplicity in checking the effect of this missingness on the two Kappa-like statistics under consideration, 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 in Table 3, such that,

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

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

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

For 'b_1' missing observations in the off diagonal of the table as in Table 4 by substituting 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

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

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

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

Also for 'a_1' and 'b_1' respectively being the combinations of missing observations along and off the diagonal of the table as in Table 5, i.e, missing from the entire parts of the table, 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,

$$\pi_{11} = \frac{(n_{11} - a_1)}{(n_{++} - (a_1 + b_1))} \neq \frac{n_{11}}{n_{++}} \qquad if \qquad a_1, b_1 \neq 0,$$

$$\pi_{12} = \frac{(n_{12} - b_1)}{(n_{++} - (a_1 + b_1))} \neq \frac{n_{12}}{n_{++}} \qquad if \qquad a_1, b_1 \neq 0,$$

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

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

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

In order to generalize, we show mathematically the effects of missing observations on the two Kappa-like statistics by using only the last case which is combinations of the other two missing patterns as in Table 5.

Cohen Kappa statistic will then become

$$\hat{k} = \left(\frac{\{\pi_{11}^m + \pi_{22}^m\} - \{\pi_{1+}^m \pi_{+1}^m + \pi_{2+}^m \pi_{+2}^m\}}{1 - \{\pi_{1+}^m \pi_{+1}^m + \pi_{2+}^m \pi_{+2}^m\}} \right),$$

$$= \left(\frac{\dfrac{\{n_{11}^m + n_{22}^m)}{n_{++}^m} - \dfrac{\{n_{1+}^m n_{+1}^m + n_{2+}^m n_{+2}^m\}}{{n_{++}^m}^2}}{1 - \dfrac{\{n_{1+}^m n_{+1}^m + n_{2+}^m n_{+2}^m\}}{{n_{++}^m}^2}} \right),$$

$$= \left(\frac{\dfrac{(n_{11} - a_1) + (n_{22} - a_2)}{n_{++} - (a + b)} - \dfrac{\{(n_{1+} - A_{11})(n_{+1} - A_{12}) + (n_{2+} - A_{22})(n_{+2} - A_{21})\}}{(n_{++} - (a + b))^2}}{1 - \dfrac{\{(n_{1+} - A_{11})(n_{+1} - A_{12}) + (n_{2+} - A_{22})(n_{+2} - A_{21})\}}{(n_{++} - (a + b))^2}} \right),$$

$$= \left(\frac{(n_{++} - (a+b))\{(n_{11} - a_1) + (n_{22} - a_2)\} - \{(n_{1+} - A_{11})(n_{+1} - A_{12}) + (n_{2+} - A_{22})(n_{+2} - A_{21})\}}{(n_{++} - (a+b))^2 - \{(n_{1+} - A_{11})(n_{+1} - A_{12}) + (n_{2+} - A_{22})(n_{+2} - A_{21})\}} \right)$$

,

$$\neq \left(\frac{n_{++}\{n_{11} + n_{22}\} - \{n_{1+}n_{+1} + n_{2+}n_{+2}\}}{n_{++}^2 - \{n_{1+}n_{+1} + n_{2+}n_{+2}\}} \right), if \quad a_1, a_2, b_1, b_2 \quad \neq 0.$$

Intraclass Kappa statistic also will be

$$k_{In} = \left(\frac{\pi_{11}^m + \pi_{22}^m - ((\pi_{1+}^m + \pi_{2+}^m + \pi_{+1}^m + \pi_{+2}^m)/2)^2}{1 - ((\pi_{1+}^m + \pi_{2+}^m + \pi_{+1}^m + \pi_{22}^m)/2)^2} \right),$$

$$= \left(\frac{\frac{n_{11}^m}{n_{++}^m} + \frac{n_{22}^m}{n_{++}^m} - ((\frac{n_{1+}^m}{n_{++}^m} + \frac{n_{2+}^m}{n_{++}^m} + \frac{n_{+1}^m}{n_{++}^m} + \frac{n_{+2}^m}{n_{++}^m})/2)^2}{1 - ((\frac{n_{1+}^m}{n_{++}^m} + \frac{n_{2+}^m}{n_{++}^m} + \frac{n_{+1}^m}{n_{++}^m} + \frac{n_{+2}^m}{n_{++}^m})/2)^2} \right),$$

$$= \left(\frac{\frac{n_{11} - a_1}{(n_{++} - A)} + \frac{n_{22} - a_2}{(n_{++} - A)} - ((\frac{n_{1+} - A_{11}}{(n_{++} - A)} + \frac{n_{2+} - A_{22}}{(n_{++} - A)} + \frac{n_{+1} - A_{12}}{(n_{++} - A)} + \frac{n_{+2} - A_{21}}{(n_{++} - A)})/2)^2}{1 - ((\frac{n_{1+} - A_{11}}{(n_{++} - A)} + \frac{n_{2+} - A_{22}}{(n_{++} - A)} + \frac{n_{+1} - A_{12}}{(n_{++} - A)} + \frac{n_{+2} - A_{21}}{(n_{++} - A)})/2)^2} \right),$$

$$= \left(\frac{\frac{(n_{11} - a_1) + (n_{22} - a_2)}{(n_{++} - (a+b))} - ((\frac{G}{(n_{++} - (a+b))})/2)^2}{1 - ((\frac{G}{(n_{++} - (a+b))})/2)^2} \right),$$

$$= \left(\frac{\frac{(n_{11} - a_1) + (n_{22} - a_2)}{(n_{++} - (a+b))} - ((\frac{G}{2(n_{++} - (a+b))}))^2}{1 - ((\frac{(n_{1+} - (a_1 + b_1)) + G}{2(n_{++} - (a+b))}))^2} \right),$$

$$= \left(\frac{4\{(n_{++} - (a+b))(n_{11} - a_1) + (n_{22} - a_2)\} - \{G\}^2}{4\{(n_{++} - (a+b))\}^2 - \{G\}^2} \right),$$

$$\neq \left(\frac{4\{n_{++}(n_{11} + n_{22})\} - \{n_{1+} + n_{2+} + n_{+1} + n_{+2}\}^2}{4n_{++}^2 - \{n_{1+} + n_{2+} + n_{+1} + n_{+2}\}^2} \right) \quad if \quad a_1, a_2, b_1, b_2 \neq 0.$$

where

$A_{11} = (a_1 + b_1), \quad A_{22} = (a_2 + b_2),$

$A_{12} = (a_1 + b_2), \quad A_{21} = (a_2 + b_1),$

$A = (a + b),$

$G = (n_{1+} - (a_1 + b_1)) + (n_{2+} - (a_2 + b_2)) + (n_{+1} - (a_1 + b_2)) + (n_{+2} - (a_2 + b_1)).$

In a given I x I contingency table as in Table 2, we assume that the missingness mechanism criteria is a conventional missing at random (CMAR) with special missing patterns in the cells as given in Tables 3, 4, and 5.

Case-by-case formulations

We examine the effects of the missing observation with different missing patterns for a given table of ratings of raters on n_{++} subjects on the two Kappa-like statistics. We achieve this by considering the following cases with certain percentages missing.

On each of the following cases we consider 0%, 5% and 10% missing observations of the total counts n_{++} from their respective region.

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 the entire table (along and off diagonal combined).

For each of the cases we obtain the estimates for the two Kappa-like statistics, with their respective standard errors, Z-values based on equations (11) and (18). We create a bound for each statistic ($\hat{k}_c, \hat{k}_{In}$) by obtaining the 95% confidence interval of size

$100(1-\alpha)\% = \hat{k} \pm z_{1-\frac{\alpha}{2}} SE(\hat{k}).$

3. Empirical examples and Results

Example 1.

Consider the data arising from the study reported in Holmquist *et al.* (1967) 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 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. See the Appendix for the Kappa-like statistics estimates with their respective confidence intervals for the following table on the cross-classification for pathologists D and F.

Table 7: Cross-classification of pathologists D and F on carcinoma in situ of the uterine cervix of 118 slides.

Category	Pathologist F					Total
Pathologist D	1	2	3	4	5	
1	35	3	0	0	0	38
2	25	17	5	1	0	48
3	2	9	11	0	1	23
4	0	2	4	0	2	8
5	0	0	0	0	1	1
Total	62	31	20	1	4	118

Example 2.

Consider a set of data on the choice of decaffeinated Coffee at two different purchase dates for five different brands of Coffee taken from Agresti (1996). See the Appendix for the Kappa-like statistics estimates with their respective confidence interval for the following table on choice of decaffeinated Coffee at two purchase dates; “Hp=High point”, “T=Taster”, “S=Sanka”, “N=Nescafe”, “B=Brim”.

Table 8: Choice of decaffeinated Coffee at two purchase dates.

Category	Second purchase					Total
First purchase	Hp	T	S	N	B	
Hp	93	17	44	7	10	171
T	9	46	11	0	9	75
S	17	11	155	9	12	204
N	6	4	9	15	2	36
B	10	4	12	2	27	55
Total	135	82	231	33	60	541

From the tables in the Appendix, we observed that the estimates given by Cohen Kappa statistic for agreement measure are always bigger than the estimates given by Intraclass Kappa statistic for the same sets of tables. We also observed that the bounds of 95%CI($\hat{k}_c$) for Cohen Kappa statistic are consistently fall within the lower and the upper bounds of 95%CI($\hat{k}_{In}$)for Intraclass Kappa statistic. This is true for all the cases considered with different missing percentages as seen in the tables. Also from the tables, under case 1, we observed that the higher the missing value percentages the worse the strength of agreement become. But for cases 2 and 3 the higher the percentages of missing observations the better the strength of agreement between the two raters become, these results justified the previous studies by Jolayemi, (1990 and 1991).

4. Conclusion

Intraclass Kappa statistic estimates for the strength of agreement are consistently lower than the respective estimates obtained with the Cohen Kappa statistic. We also observed that increment in the percentages of missingness along the diagonal worsen the strength of the agreement between the two raters. However, increment in the percentages of missingness in the off diagonal part of the table of the ratings for two raters improves the strength of the agreement between them. This implies that the two statistics under considerations encourage missing in the off diagonal part of the table. To this effect, modelling the structure of agreement will be a better way of monitoring the effects of missing observation on the agreement measure.

Acknowledgements

The author acknowledges the comments of the reviewers for their suggestions and comments in improving the quality of the manuscripts.

References

Adejumo, A. O., Heumann, C., and Toutenburg, H. (2004): On the effects of missing values on comparison of Symmetry model and Quasi-symmetry model for raters agreement measure. *SFB386-Discussion Paper 370, Ludwig-Maximilians- Universität*, München.

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

Banerjee, M., Capozzoli, M., Mcsweeney, L. and Sinha, D. (1999): Beyond Kappa: A review of interrater agreement measure. *The Can. J. Stat.* **27**(1), 03-23.

Barnhart, H. X. and Williamson, J. M. (2002): Weighted Least squares approach for comparing correlated Kappa. *Biometric*., **58**, 1012-1019

Bloch, D. A. and Kraemer, H. C. (1989): 2 x 2 Kappa coefficients: Measures of agreement or association. *Biometrics*. **45**, 269-287.

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

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

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

Hernaez, R. (2015). Reliability and agreement studies: a guide for clinical investigators. GUT **64** (7), 1018 – 1027.

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

Jolayemi, E.T., (1990): On the measure of Agreement between two raters, *Biometrical Journal*. **32**:87-93.

Jolayemi, E.T., (1991): A Multiraters Agreement Index for Ordinal Classification, *Biometrical Journal*. **4**:485-492.

Manuel, A., Juan J. L. and Anaha, B. (2011): A simulation study of rater agreement measures with 2x2 contingency tables. *Psicológica*. **32**, 385-402.

Martin, A. and Femia, P. (2008): Chance-corrected measures of reliability and validity in 2x 2 tables. *Communications in Statistics. Theory and Methods*, **37**, 760-772.

Little, R, J. A. and Rubin, D. B. (1987): *Statistical analysis with missing Data.* New York, Wiley.

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

Schafer J. L. (1997): *Analysis of incomplete multivariate data*. Chapman and Hall.

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

Fanshawe, T. R., Lynch, A. G., Ellis, I. O., Green, A. R. and Hanka, R. (2008): Assessing Agreement between Multiple Raters with Missing Rating Information, Applied to Breast Cancer Tumour Grading *PLoS ONE.* **3**(8), e2925-1-e2925-2.

Toutenburg, H., Heumann, C., Nittner, T. and Scheid, S. (2002): Parametric and

Nonparametric Regression with Missing X'S- A review. *J. Iranian Statist. Soc.* **1** (1, 2), 79-109.

Toutenburg, H. and Nittner, T. (2002): Linear regression models with incomplete categorical covariates. *Computational Statist.* **17** (2), 215-232.

Appendix

Table 9: Cohen Kappa statistic estimates for pathologists D and F on carcinoma in situ when there are missing with different patterns.

		Cohen			95% CI for $\hat{k}_c$	
Cases	%missing	Kappa($\hat{k}_c$)	Stand. Error	z-value	Lower bound	Upper bound
Complete	0%	0.3368	0.0565	5.9668	0.2262	0.4474
Case 1	5% 10%	0.3060 0.2676	0.0572 0.0582	5.3516 4.5990	0.1939 0.1535	0.4181 0.3816
Case 2	5% 10%	0.3773 0.4136	0.0588 0.0611	6.3461 6.7724	0.2578 0.2939	0.4883 0.5333
Case 3	5% 10%	0.3380 0.3390	0.0578 0.0592	5.8506 5.7246	0.2248 0.2230	0.4513 0.4551

Table 10: Intraclass Kappa statistic estimates for pathologists D and F on carcinoma in situ when there are missing with different patterns.

		Intraclass			95% CI for $\hat{k}_{In}$	
Cases	%missing	Kappa($\hat{k}_{In}$)	Stand. Error	z-value	Lower bound	Upper bound
Complete	0%	0.3203	0.0899	3.5652	0.1442	0.4964
Case 1	5% 10%	0.2870 0.2450	0.0913 0.0943	3.1419 2.5982	0.1079 0.0602	0.4660 0.4299
Case 2	5% 10%	0.3600 0.4022	0.0933 0.0982	3.8573 4.0965	0.1771 0.2098	0.5430 0.5946
Case 3	5% 10%	0.3321 0.3220	0.0930 0.0965	3.4565 3.3355	0.1391 0.1328	0.5035 0.5111

Table 11: Cohen Kappa statistic estimates for choice of decaffeinated Coffee at two purchase dates when there are missing with different patterns.

		Cohen			95% CI for $\hat{k}_c$	
Cases	%missing	Kappa($\hat{k}_c$)	Stand. Error	z-value	Lower bound	Upper bound
complete	0%	0.4765	0.0245	19.485	0.4285	0.5244
Case 1	5% 10%	0.4473 0.4147	0.0252 0.0260	17.752 15.939	0.3979 0.3637	0.4967 0.4657
Case 2	5% 10%	0.5195 0.5725	0.0252 0.0262	20.619 21.880	0.4701 0.5212	0.5689 0.6237
Case 3	5% 10%	0.4839 0.4922	0.0252 0.0261	19.176 18.859	0.4344 0.4410	0.5333 0.5433

Table 12: Intraclass Kappa statistic estimates for choice of decaffeinated Coffee at two purchase dates when there are missing with different patterns.

		Intraclass			95% CI for $\hat{k}_{In}$	
Cases	%missing	Kappa($\hat{k}_{In}$)	Stand. Error	z-value	Lower bound	Upper bound
complete	0%	0.4751	0.0354	13.405	0.4057	0.5446
Case 1	5% 10%	0.4458 0.4129	0.0367 0.0380	12.160 10.861	0.3739 0.3384	0.5176 0.4874
Case 2	5% 10%	0.5181 0.5709	0.0368 0.0387	14.077 14.738	0.4459 0.4949	0.5902 0.6468
Case 3	5% 10%	0.4824 0.4904	0.0368 0.0382	13.127 12.834	0.4103 0.4155	0.5544 0.5653

Table 13: Summary of Kappa-like statistics estimates for pathologists D and F on carcinoma in situ when there are missing with different patterns.

			95% CI for $\hat{k}_c$			95% CI for $\hat{k}_{In}$	
Cases	%missing	Kappa($\hat{k}_c$)	Lower bound	Upper bound	Kappa($\hat{k}_{In}$)	Lower bound	Upper bound
complete	0%	0.3368	0.2262	0.4474	0.3203	0.1442	0.4964
Case 1	5% 10%	0.3060 0.2676	0.1939 0.1535	0.4181 0.3816	0.2870 0.2450	0.1079 0.0602	0.4660 0.4299
Case 2	5% 10%	0.3773 0.4136	0.2578 0.2939	0.4883 0.5333	0.3600 0.4022	0.1771 0.2098	0.5430 0.5946
Case 3	5% 10%	0.3380 0.3390	0.2248 0.2230	0.4513 0.4551	0.3321 0.3220	0.1391 0.1328	0.5035 0.5111

Table 14: Summary of Kappa-like statistics estimates for choice of decaffeinated Coffee at two purchase dates when there are missing with different patterns.

			95% CI for $\hat{k}_c$			95% CI for $\hat{k}_{In}$	
Cases	%missing	Kappa($\hat{k}_c$)	Lower bound	Upper bound	Kappa($\hat{k}_{In}$)	Lower bound	Upper bound
complete	0%	0.4765	0.4285	0.5244	0.4751	0.4057	0.5446
Case 1	5%	0.4473	0.3979	0.4967	0.4458	0.3739	0.5176
	10%	0.4147	0.3637	0.4657	0.4129	0.3384	0.4874
Case 2	5%	0.5195	0.4701	0.5689	0.5181	0.4459	0.5902
	10%	0.5725	0.5212	0.6237	0.5709	0.4949	0.6468
Case 3	5%	0.4839	0.4344	0.5333	0.4824	0.4103	0.5544
	10%	0.4922	0.4410	0.5433	0.4904	0.4155	0.5653