A note on the linearly weighted kappa coefficient for ordinal scales

S. Vanbelle and A. Albert

Medical Informatics and Biostatistics, School of Public Health, University of Liège, Liège, Belgium

Abstract

A frequent criticism formulated against the use of weighted kappa coefficients is that the weights are arbitrarily defined. We show that using linear weights for a K-ordinal scale is equivalent to deriving a kappa coefficient from K-1 embedded 2×2 tables.

Key words: absolute weights, interpretation, agreement, disagreement

1 INTRODUCTION

Cohen's kappa coefficient (Cohen, 1960) is widely used to quantify agreement between two raters on a nominal scale (Ludbrook, 2002). It corrects the observed percentage of agreements between the raters for the effect of chance.

* Sophie Vanbelle, Medical Informatics and Biostatistics, School of Public Health, University of Liège, CHU Sart Tilman, 4000 Liège, Belgium, sophie.vanbelle@ulg.ac.be

A value of 0 implies no agreement beyond chance, whereas a value of 1 corresponds to a perfect agreement between the two raters. There are situations where disagreements between raters may not all be equally important. For example, on an ordinal scale, a greater "penalty" will be applied if the two categories chosen by the raters are farther apart. To account for these inequalities, Cohen (1968) introduced weights in the formulation of the agreement index leading to the weighted kappa coefficient. Although the weights are in general arbitrarily defined, those introduced by Cicchetti and Allison (1971) and by Fleiss and Cohen (1973) are the most commonly used. The former are linear and the latter have a quadratic form. Cohen (1968) showed that, under specific conditions, the weighted kappa coefficient is equivalent to the product-moment correlation coefficient. Moreover, Fleiss and Cohen (1973) and Schuster (2004) showed that the weighted kappa with a quadratic weighting scheme is equivalent to the intraclass correlation coefficient. Hereafter, we show that the weighted kappa coefficient defined with linear weights for a K-ordinal scale can be derived from (K-1) embedded 2×2 contingency tables.

2 DEFINITION OF THE WEIGHTED KAPPA COEFFICIENT

Consider two raters who classify a sample of n subjects (or objects) into K categories of an ordinal scale (see Table 1), where n_{ij} is the number of items classified into category i by rater 1 and category j by rater 2, $n_{i.}$ the number of subjects classified into category i by rater 1 and $n_{.j}$ be the number of subjects classified into category j by rater 2. Denote by p_{ij}, $p_{i.}$ and $p_{.j}$ the corresponding proportions ($i, j = 1, \cdots, K$).

Table 1

Two-way contingency table resulting from the classification of n items by 2 raters on an ordinal scale with K categories

	Rater 2					
Rater 1	1	$\ldots$	j	$\ldots$	K	Total
1	n_{11}	$\ldots$	n_{1j}	$\ldots$	n_{1K}	$n_{1.}$
i	n_{i1}	$\ldots$	n_{ij}	$\ldots$	n_{iK}	$n_{i.}$
K	n_{K1}	$\ldots$	n_{Kj}	$\ldots$	n_{KK}	$n_{K.}$
Total	$n_{.1}$	$\ldots$	$n_{.j}$	$\ldots$	$n_{.K}$	n

The weighted kappa coefficient can be defined in terms of agreement weights by

$$\kappa_w = \frac{p_o - p_e}{1 - p_e} \tag{1}$$

where $p_0 = \sum_{i=1}^{K} \sum_{j=1}^{K} w_{ij} p_{ij}$ and $p_e = \sum_{i=1}^{K} \sum_{j=1}^{K} w_{ij} p_{i.} p_{.j}$ with $0 \leq w_{ij} \leq 1$ and $w_{jj} = 1$ $(i, j = 1, \cdots, K)$, or in terms of disagreement weights by

$$\kappa_w = 1 - \frac{q_o}{q_e} \tag{2}$$

where $q_0 = \sum_{i=1}^{K} \sum_{j=1}^{K} v_{ij} p_{ij}$ and $q_e = \sum_{i=1}^{K} \sum_{j=1}^{K} v_{ij} p_{i.} p_{.j}$ with $0 \leq v_{ij} \leq 1$ and $v_{jj} = 0$ $(i, j = 1, \cdots, K)$. However, the weighted kappa coefficient can also be obtained using unscaled disagreement weights, i.e., v_{ij} not restricted to the [0,1] interval.

Cohen's kappa coefficient is a particular case of the weighted kappa coefficient where $w_{ij} = 1$ $(v_{ij} = 0)$ for $i = j$ and $w_{ij} = 0$ $(v_{ij} = 1)$ for $i \neq j$

$(i, j = 1, \cdots, K)$. Cicchetti and Allison (1971) proposed "linear" weights of the form $w_{ij} = 1 - |i-j|/(K-1)$, whereas Fleiss and Cohen (1973) used the quadratic weights $w_{ij} = 1 - (i-j)^2/(K-1)^2$. The disagreement weights $v_{ij} = (i-j)^2$ are also commonly used (Ludbrook (2002); Agresti (2002)) as are the linear disagreement weights $v_{ij} = |i-j|$.

Cohen (1968) showed that if the marginal distributions of the 2 raters are the same and if the weights of disagreement are defined as $v_{ij} = (i-j)^2$, the weighted kappa coefficient is equivalent to the product-moment correlation coefficient. Furthermore, Fleiss and Cohen (1973) showed that using the weights v_{ij}, the weighted kappa coefficient has the same interpretation as the intraclass correlation coefficient of reliability when systematic variability between raters is included as a component of total variation. More recently, Schuster (2004) explicitly decomposed the weighted kappa coefficient defined with the quadratic disagreement weights in terms of rater means, rater variances and rater covariance in the context of a two-way analysis of variance. To the best of our knowledge, no interpretation was given for the weighted agreement coefficient with linear agreement or disagreement weights.

3 THE REVISITED WEIGHTED KAPPA COEFFICIENT

Hereafter, we shall focus on the linear weights introduced by Cicchetti and Allison (1971) ($w_{ij} = 1 - |i-j|/(K-1)$) and revisit the weighted kappa coefficient for an ordinal scale. The interpretation of the agreement index obtained with the linear disagreement weights ($v_{ij} = |i-j|$) will follow straightforwardly

Table 2

Reduction of the $K \times K$ contingency table into a 2×2 classification table by selecting a cut-off level k ($k = 1, \cdots, K$) on the ordinal scale (see text)

	Rater 2		
Rater 1	$\leq k$	$> k$	Total
$\leq k$	$N_{11}(k)$	$N_{12}(k)$	$N_{1.}(k)$
$> k$	$N_{21}(k)$	$N_{22}(k)$	$N_{2.}(k)$
Total	$N_{.1}(k)$	$N_{.2}(k)$	n

since

$$w_{ij} = 1 - \frac{v_{ij}}{K-1}. \tag{3}$$

For any "cut-off" value k ($k = 1, \cdots, K-1$), the $K \times K$ contingency table (see Table 1) can be reduced into a 2×2 classification table by summing up all observations below and above the first k rows and first k columns (see Table 2) where

$$N_{11}(k) = \textstyle\sum_{i=1}^{k} \sum_{j=1}^{k} n_{ij} \qquad N_{12}(k) = \sum_{i=1}^{k} \sum_{j=k+1}^{K} n_{ij}$$

$$N_{21}(k) = \textstyle\sum_{i=k+1}^{K} \sum_{j=1}^{k} n_{ij} \quad N_{22}(k) = \sum_{i=k+1}^{K} \sum_{j=k+1}^{K} n_{ij}$$

Let $F_{lm}(k) = \dfrac{1}{n} N_{lm}(k)$, $F_{l.} = \dfrac{1}{n} N_{l.}(k)$ and $F_{.m} = \dfrac{1}{n} N_{.m}(k)$ be the corresponding joint and marginal frequencies ($l, m = 1, 2; k = 1, \cdots, K-1$). Finally, denote by

$$p_o(k) = F_{11}(k) + F_{22}(k) \tag{4}$$

and

$$p_e(k) = F_{1.}(k)F_{.1}(k) + F_{2.}(k)F_{.2}(k) \tag{5}$$

the observed and expected weighted agreements corresponding to Table 2.

Now, consider the quantities

$$p_o^* = \frac{1}{K-1}\sum_{k=1}^{K-1} p_o(k) \tag{6}$$

and

$$p_e^* = \frac{1}{K-1}\sum_{k=1}^{K-1} p_e(k) \tag{7}$$

We show that $p_o^* = p_o$ and $p_e^* = p_e$ where p_o and p_e are respectively the "linearly" weighted proportions of observed and expected agreement, as defined by Cicchetti & Allison (1971).

Since

$$\begin{aligned}
p_o^* &= \frac{1}{K-1}\sum_{k=1}^{K-1}\left(\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij} + \sum_{i=k+1}^{K}\sum_{j=k+1}^{K} p_{ij}\right)\\
&= \frac{1}{K-1}\sum_{k=1}^{K-1}\left(\sum_{i=1}^{K}\sum_{j=1}^{K} p_{ij} - \sum_{i=1}^{k}\sum_{j=k+1}^{K} p_{ij} - \sum_{i=k+1}^{K}\sum_{j=1}^{k} p_{ij}\right)\\
&= \sum_{i=1}^{K}\sum_{j=1}^{K} p_{ij} - \frac{1}{K-1}\sum_{k=1}^{K-1}\left(\sum_{i=1}^{k}\sum_{j=k+1}^{K} p_{ij} + \sum_{i=k+1}^{K}\sum_{j=1}^{k} p_{ij}\right)
\end{aligned} \tag{8}$$

and

$$\begin{aligned}
p_o &= \sum_{i=1}^{K}\sum_{j=1}^{K}\left(1 - \frac{|i-j|}{K-1}\right) p_{ij}\\
&= \sum_{i=1}^{K}\sum_{j=1}^{K} p_{ij} - \frac{1}{K-1}\sum_{i=1}^{K}\sum_{j=1}^{K} |i-j| p_{ij}\\
&= \sum_{i=1}^{K}\sum_{j=1}^{K} p_{ij} - \frac{1}{K-1}\sum_{i=1}^{K}\sum_{j=1}^{i} (i-j) p_{ij} - \frac{1}{K-1}\sum_{i=1}^{K-1}\sum_{j=i+1}^{K} (j-i) p_{ij},
\end{aligned} \tag{9}$$

it suffices to prove that

$$\sum_{k=1}^{K-1}\left(\sum_{i=1}^{k}\sum_{j=k+1}^{K}p_{ij}+\sum_{i=k+1}^{K}\sum_{j=1}^{k}p_{ij}\right)=\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}(j-i)p_{ij}+\sum_{i=1}^{K}\sum_{j=1}^{i}(i-j)p_{ij} \quad (10)$$

We have successively,

$$\begin{aligned}
&\sum_{k=1}^{K-1}\left(\sum_{i=1}^{k}\sum_{j=k+1}^{K}p_{ij}+\sum_{i=k+1}^{K}\sum_{j=1}^{k}p_{ij}\right)=\sum_{k=1}^{K-1}\sum_{i=1}^{k}\sum_{j=k+1}^{K}p_{ij}+\sum_{k=1}^{K-1}\sum_{i=k+1}^{K}\sum_{j=1}^{k}p_{ij}\\
&=\sum_{i=1}^{1}\sum_{j=2}^{K}p_{ij}+\sum_{i=1}^{2}\sum_{j=3}^{K}p_{ij}+\cdots+\sum_{i=1}^{K-1}\sum_{j=K}^{K}p_{ij}\\
&+\sum_{i=2}^{K}\sum_{j=1}^{1}p_{ij}+\sum_{i=3}^{K}\sum_{j=1}^{2}p_{ij}+\cdots+\sum_{i=K}^{K}\sum_{j=1}^{K-1}p_{ij}\\
=&\sum_{i=1}^{1}\sum_{j=2}^{K}p_{ij}+\sum_{i=1}^{1}\sum_{j=3}^{K}p_{ij}+\sum_{i=2}^{2}\sum_{j=3}^{K}p_{ij}+\cdots+\sum_{i=1}^{1}\sum_{j=K}^{K}p_{ij}+\sum_{i=2}^{K-1}\sum_{j=K}^{K}p_{ij}\\
+&\sum_{i=K}^{K}\sum_{j=1}^{1}p_{ij}+\sum_{i=2}^{K-1}\sum_{j=1}^{1}p_{ij}+\sum_{i=K}^{K}\sum_{j=1}^{2}p_{ij}+\sum_{i=3}^{K-1}\sum_{j=1}^{2}p_{ij}+\cdots+\sum_{i=K}^{K}\sum_{j=1}^{K-1}p_{ij}\\
=&\sum_{j=2}^{K}(j-1)p_{1j}+\sum_{i=2}^{2}\sum_{j=3}^{K}p_{ij}+\cdots+\sum_{i=2}^{K-1}\sum_{j=K}^{K}p_{ij}\\
+&\sum_{j=1}^{K-1}(K-j)p_{Kj}+\sum_{i=2}^{K-1}\sum_{j=1}^{1}p_{ij}+\cdots+\sum_{i=3}^{K-1}\sum_{j=1}^{2}p_{ij}\\
=&\sum_{j=2}^{K}(j-1)p_{1j}+\sum_{j=3}^{K}(j-2)p_{2j}+\cdots+\sum_{j=K}^{K}(j-(K-1))p_{K-1,j}\\
+&\sum_{j=1}^{K-1}(K-j)p_{Kj}+\sum_{j=1}^{K-2}(K-1-j)p_{K-1,j}+\cdots+\sum_{j=1}^{K-(K-1)}(K-(K-1)-j)p_{K-(K-1),j}\\
=&\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}(j-i)p_{ij}+\sum_{i=1}^{K}\sum_{j=1}^{i}(i-j)p_{ij} \qquad (11)
\end{aligned}$$

Thus, $p_o^* = p_o$. The proof for $p_e^* = p_e$ proceeds similarly. Thus, using the linear agreement weights introduced by Cicchetti and Allison (1971), the observed and expected weighted agreements are merely the mean values of the corresponding proportions of all 2×2 tables obtained by collapsing the first k categories and last $K - k$ categories $(k = 1, \cdots, K - 1)$ of the original $K \times K$ classification table. When considering the linear disagreement weights, the observed and expected weighted disagreements correspond to the sum of the observed and expected proportions of disagreement of the $K-1$ embedded 2×2 tables, respectively.

4 EXAMPLE

Gilmour et al. (1997) conducted an agreement study to compare two methods for assessing cervical ectopy, defined as the presence of endocervical-type columnar epithelium on the portio surface of the cervix. A computerized planimetry method was developed for measuring cervical ectopy, and the reliability of that method was compared with direct visual assessment. Photographs of the cervix of 85 women without cervical disease were assessed for cervical ectopy by three medical raters who used both assessment methods. The response of interest, cervical ectopy size, was an ordinal variable with four categories: (1) minimal, (2) moderate, (3) large and (4) excessive. The contingency table for two of the three raters using the visual method is displayed in Table 3. In each cell, the first term corresponds to the cell count, the second term to the linear agreement weight and the third one to the linear disagreement weight.

When computing the weighted observed and expected agreements, we obtain $p_o = 0.800$, $p_e = 0.583$, yielding $\kappa_w = 0.520$. Since $K = 4$, three "embedded" 2×2 tables can be constructed as described before (see Table 4). From these tables, we calculate $p_o^* = \frac{1}{3}\sum_{k=1}^{3} p_o(k) = (0.812 + 0.788 + 0.800)/3 = 0.800$ and $p_e^* = \frac{1}{3}\sum_{k=1}^{3} p_e(k) = (0.618 + 0.506 + 0.626)/3 = 0.583$. These are as expected equal to p_o and p_e, respectively. It should be remarked that the av-

Table 3

Two-way contingency table resulting from cervical ectopy ratings using the visual method by two raters

	Rater 2				
Rater 1	1	2	3	4	Total
1	13 [a]	2	0	0	15
	1.0 [b]	0.67	0.33	0.0	
	0.0 [c]	1.0	2.0	3.0	
2	10	16	3	0	29
	0.67	1.0	0.67	0.33	
	1.0	0.0	1.0	2.0	
3	3	7	3	0	13
	0.33	0.67	1.0	0.67	
	2.0	1.0	0.0	1.0	
4	1	4	12	11	28
	0.0	0.33	0.67	1.0	
	3.0	2.0	1.0	0.0	
Total	27	29	18	11	85

[a] Observed counts

[b] Linear agreement weights $w_{ij} = 1 - |i - j|/(K - 1)$

[c] Linear disagreement weights $v_{ij} = |i - j|$

erage kappa coefficient derived from the tables, namely $\overline{\kappa} = \frac{1}{3}\sum_{k=1}^{3}\kappa(k) = (0.507 + 0.572 + 0.465)/3 = 0.515$, differs from κ_w. The weighted observed and expected disagreements are equal to $q_o = 0.600$ and $q_e = 1.25$, respectively, yielding a weighted kappa coefficient of $\kappa_w = 0.52$. From the embedded tables, we have $q_o^* = \sum_{k=1}^{3} q_o(k) = 0.188 + 0.212 + 0.200 = 0.600$ and $p_e^* = \sum_{k=1}^{3} q_e(k) = 0.382 + 0.494 + 0.374 = 1.25$, as expected.

5 DISCUSSION

The weighted kappa coefficient is widely used to quantify the agreement between 2 raters on an ordinal scale. The weights are generally given a priori and defined arbitrarily. Graham and Jackson (1993) observed that the value of the weighted kappa coefficient can vary considerably according to the weighting scheme used and henceforth may lead to different conclusions. In practice, the linear (Cicchetti and Allison, 1971) and quadratic (Fleiss and Cohen, 1973) weighting schemes are the most widely used. Quadratic weights have received much attention in the literature because of their practical interpretation. For instance, Fleiss and Cohen (1973) and Schuster (2004) showed that using the weights $v_{ij} = (i-j)^2$, the weighted kappa coefficient can be interpreted as an intraclass correlation coefficient in a two-way analysis of variance setting. In this article, we focused on the linearly weighted kappa coefficient defined by Cicchetti and Allison (1971) or equivalently defined by the linear disagreement weights $v_{ij} = |i-j|$ and strove to give an intuitive interpretation of it. Specifically, we showed that the observed and expected weighted agreements are merely the mean values of the corresponding proportions of all 2×2

Table 4

All possible embedded 2×2 classification tables ($k = 1, 2, 3$) derived from the original 4×4 contingency table for cervical ectopy ratings by two raters

	Rater 2		
Rater 1	≤ 1	> 1	Total
≤ 1	13	2	15
> 1	14	56	70
Total	27	58	85

$p_o(1) = 0.812; q_o(1) = 0.188$

$p_e(1) = 0.618; q_e(1) = 0.382$

$\kappa(1) = 0.507$

	Rater 2		
Rater 1	≤ 2	> 2	Total
≤ 2	41	3	44
> 2	15	26	41
Total	56	29	85

$p_o(2) = 0.788; q_o(2) = 0.212$

$p_e(2) = 0.506; q_e(2) = 0.494$

$\kappa(2) = 0.572$

	Rater 2		
Rater 1	≤ 3	> 3	Total
≤ 3	57	0	57
> 3	17	11	28
Total	74	11	85

$p_o(3) = 0.800; q_o(3) = 0.200$

$p_e(3) = 0.626; q_e(3) = 0.374$

$\kappa(3) = 0.465$

tables obtained by collapsing the first k categories and last $K-k$ categories ($k=1,\cdots,K-1$) of the original $K\times K$ classification table. It should be noted, however, that the weighted agreement coefficient derived from the original table is not equal to the mean value of the non-weighted $K-1$ κ coefficients obtained from the 2×2 collapsed tables. When using linear disagreement weights, the weighted observed and expected disagreements are obtained by the sum rather than the average of the corresponding elements of the 2×2 tables. In other words, the linearly weighted kappa coefficient can simply be derived from $K-1$ embedded 2×2 classification tables. The linear form of the kappa coefficient, besides its simplicity, presents some advantages over the quadratic version. As demonstrated by Brenner and Kliebsch (1996), it is less sensitive to the number of categories and should therefore be preferred when the number of categories of the ordinal scale is large. As a conclusion, we have shown that the linearly weighted kappa coefficient for a K-ordinal table can be naturally derived from non-weighted observed and expected agreements (disagreements) computed from $K-1$ embedded 2×2 classification tables.

References

[1] Agresti, A., 2002. Categorical data analysis (2nd ed.). New York: John Wiley and Sons.

[2] Brenner, H., Kliebsch, U., 1996. Dependence of weighed kappa coefficients on the number of categories. Epidemiology 7, 199–202.

[3] Cicchetti, D., Allison, T., 1971. A new procedure for assessing reliability of scoring eeg sleep recordings. American Journal EEG Technology 11, 101–109.

[4] Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational

and Psychological Measurement 20, 37–46.

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

[6] Fleiss, J. L., Cohen, J., 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measure of reliability. Educational and Psychological Measurement 33, 613–619.

[7] Gilmour, E., Ellerbrock, T., Koulos, J., Chiasson, M., Williamson, J., Kubn, L., Wright, T. J., 1997. Measuring cervical ectopy: direct visual assessment versus computerized planimetry. American Journal of Obstetrics and Gynecology 176, 108–111.

[8] Graham, P., Jackson, R., 1993. The analysis of ordinal agreement data:beyond weighted kappa. Journal of Clinical Epidemiology 46, 1055–1062.

[9] Ludbrook, J., 2002. Statistical techniques for comparing measurers and methods of measurement: a critical review. Clinical and Experimental Pharmacology and Physiology 29, 527–536.

[10] Schuster, C., 2004. A note on the interpretation of weighted kappa and its relation to other rater agreement statistics for metric scales. Educational and Psychological Measurement 64, 243–253.