Statistical Methodology 8 (2011) 268–272

Contents lists available at ScienceDirect

Statistical Methodology

journal homepage: www.elsevier.com/locate/stamet

# Weighted kappa is higher than Cohen's kappa for tridiagonal agreement tables

Matthijs J. Warrens *

*Institute of Psychology, Unit Methodology and Statistics, Leiden University, P.O. Box 9555, 2300 RB Leiden, The Netherlands*

ARTICLE INFO

*Article history:*
Received 25 May 2010
Received in revised form
8 September 2010
Accepted 10 September 2010

*Keywords:*
Cohen's kappa
Cohen's weighted kappa
Linear weights
Quadratic weights
Nominal agreement
Ordinal agreement

ABSTRACT

Cohen's kappa and weighted kappa are two popular descriptive statistics for measuring agreement between two observers on a nominal scale. It has been frequently observed in the literature that, when Cohen's kappa and weighted kappa are applied to the same agreement table, the value of weighted kappa is higher than the value of Cohen's kappa. This paper proves this phenomenon for tridiagonal agreement tables.



## 1. Introduction

The kappa coefficient [5,1,23,20–22] and weighted kappa coefficient [6,9,2,15,17,14] are popular descriptive statistics for summarizing the cross-classification of two nominal variables with $n \in \mathbb{N}_{\geq 2}$ identical categories [10]. An $n \times n$ table can for example be obtained by cross-classifying the ratings of two observers that each have classified a group of objects into $n$ categories. In this case, the $n \times n$ table can be referred to as an agreement table, since it reflects how the ratings of the two observers agree and disagree. Agreement tables occur in various fields of science, and applications of kappa and weighted kappa can therefore be found in epidemiological and clinical studies (see, for example, [16,11]), diagnostic imaging [13], map comparison [19] and content analysis [12].

It has been frequently observed in the literature that, when Cohen's kappa and weighted kappa are applied to the same agreement table, the value of weighted kappa is higher than the value of Cohen's kappa. For example, consider the data in Table 1 taken from a study in [16]. In this study two trained readers independently graded 324 iris photographs using a five-grade classification system.

* Tel.: +31 71 5273649; fax: +31 71 5273619.
*E-mail address:* warrens@fsw.leidenuniv.nl.

1572-3127/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.stamet.2010.09.004

**Table 1**
Color gradings of 324 iris photographs by two trained readers (Table 1 in [16]).

| Reader B | Reader A | | | | | Row totals |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | |
| 1 | 98 | 11 | 0 | 0 | 0 | 109 |
| 2 | 7 | 38 | 5 | 2 | 0 | 52 |
| 3 | 0 | 2 | 25 | 8 | 0 | 35 |
| 4 | 0 | 0 | 8 | 40 | 2 | 50 |
| 5 | 0 | 0 | 0 | 6 | 72 | 78 |
| Column totals | 105 | 51 | 38 | 56 | 74 | 324 |

Categories of iris color were distinguished on the basis of the predominant color (blue, gray, green, light brown, or brown) and the amount of brown or yellow pigment present in the iris. For these data Cohen's kappa equals 0.796, whereas weighted kappa using quadratic weights is equal to 0.965. A value of 1 would indicate perfect agreement between the two readers.

The value of weighted kappa does not always exceed the value of Cohen's kappa. It turns out however that the inequality holds for a special kind of agreement table. In this short paper we prove that the value of weighted kappa exceeds that of Cohen's kappa when the agreement table is tridiagonal. A tridiagonal table is a square matrix that has nonzero elements only on the main diagonal, the first diagonal below this (subdiagonal) and the first diagonal above this (superdiagonal). Note that Table 1 is almost tridiagonal. Agreement tables that are tridiagonal or approximately tridiagonal are frequently encountered in applications. Examples can be found in [18,16,8,3,7].

The paper is organized as follows. Weighted kappa is defined in the next section. The conditional inequality is proved in Section 3. Section 4 contains some conclusions.

## 2. Kappa and weighted kappa

In this section we define the weighted kappa statistic, which is usually denoted by $\kappa_w$. Cohen [6] introduced weighted kappa as a generalization of kappa [5], which is usually denoted by $\kappa$. Weighted kappa allows for assigning partial credit to the nominal categories by using weights.

Suppose that two observers each distribute $m \in \mathbb{N}_{\geq 1}$ given objects (individuals) among a set of $n \in \mathbb{N}_{\geq 2}$ mutually exclusive categories, that are defined in advance. Let the agreement table $T$ with entries $t_{ij}$ $(i, j \in \{1, 2, \ldots, n\})$ be the cross-classification of the ratings of the observers, where $t_{ij}$ indicates the number of objects placed in category $i$ by the first observer and in category $j$ by the second observer. The elements on the main diagonal of $T$, $t_{ii}$ for $i \in \{1, 2, \ldots, n\}$, are usually called the agreements because they reflect the number of objects that the observers placed in the same categories. All other elements, $t_{ij}$ for $i \neq j$, are usually called the disagreements.

For notational convenience, let $P$ be the agreement table of the same size as $T$ $(n \times n)$ with entries $p_{ij} = t_{ij}/m$. Row and column totals

$$p_i = \sum_{j=1}^{n} p_{ij} \quad \text{and} \quad q_j = \sum_{i=1}^{n} p_{ij}$$

are the marginal totals of $P$. The weighted kappa statistic can be defined as

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

where

$$p_o^w = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} p_{ij} \quad \text{and} \quad p_e^w = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} p_i q_j$$

with $w_{ij} \in [0, 1]$ and $w_{ii} = 1$ for $i, j \in \{1, 2, \ldots, n\}$. In (1) we assume that $p_e^w < 1$ to avoid the indeterminate case $p_e^w = 1$.

**Table 2**
The form of a tridiagonal matrix of size $n \times n$. The $a_i$ for $i \in \{1, 2, \ldots, n\}$ are the elements of the main diagonal, whereas the $b_i$ and $c_i$ for $i \in \{1, 2, \ldots, n-1\}$ are, respectively, the elements of the superdiagonal and subdiagonal. All other elements are 0.

| Reader B | Reader A | | | | | | Row totals |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | $\cdots$ | $\cdots$ | $n-1$ | $n$ | |
| 1 | $a_1$ | $b_1$ | | | | | $p_1$ |
| 2 | $c_1$ | $a_2$ | $b_2$ | | | | $p_2$ |
| $\vdots$ | | $\ddots$ | $\ddots$ | $\ddots$ | | | $\vdots$ |
| $\vdots$ | | | $\ddots$ | $\ddots$ | $\ddots$ | | $\vdots$ |
| $n-1$ | | | | $c_{n-2}$ | $a_{n-1}$ | $b_{n-1}$ | $p_{n-1}$ |
| $n$ | | | | | $c_{n-1}$ | $a_n$ | $p_n$ |
| Column totals | $q_1$ | $q_2$ | $\cdots$ | $\cdots$ | $q_{n-1}$ | $q_n$ | 1 |

Examples of weights for $\kappa_w$ that have been proposed in the literature are the linear weights [4,17,14] given by

$$w_{ij}^{L} = 1 - \frac{|i-j|}{n-1}, \tag{2}$$

and the quadratic weights [9,15] given by

$$w_{ij}^{Q} = 1 - \left(\frac{i-j}{n-1}\right)^2. \tag{3}$$

Using the weights in (2) we have $p_o^w = 0.959$, $p_e^w = 0.555$ and $\kappa_w = 0.908$ for the data in Table 1. Furthermore, using the weights in (3) we have $p_o^w = 0.989$, $p_e^w = 0.682$ and $\kappa_w = 0.965$ for the data in Table 1.

If $w_{ij} = 0$ for $i, j \in \{1, 2, \ldots, n\}$ and $i \neq j$, then $p_o^w$ and $p_e^w$ become, respectively,

$$p_o = \sum_{i=1}^{n} p_{ii} \quad \text{and} \quad p_e = \sum_{i=1}^{n} p_i q_i.$$

In this case, (1) is equivalent to

$$\kappa = \frac{p_o - p_e}{1 - p_e},$$

which is the ordinary or unweighted kappa statistic [5]. For the data in Table 1, we have $p_o = 0.843$, $p_e = 0.229$ and $\kappa = 0.796$. Using the weights (2) or (3), the statistics $\kappa$ and $\kappa_w$ are equivalent if $n = 2$. Statistics $\kappa$ and $\kappa_w$ are also equivalent if $p_o = 1$.

## 3. A conditional inequality

In the theorem below we prove an inequality between $\kappa$ and $\kappa_w$. We first consider a restriction on the weights of $\kappa_w$. In general, we have $w_{ij} \in [0, 1]$ for $i, j \in \{1, 2, \ldots, n\}$ and $w_{ii} = 1$ for $i \in \{1, 2, \ldots, n\}$ for the elements on the main diagonal of $P$. Note that, if we were to use the weights in (2) or (3), the weights would be a decreasing function of the distance $|i-j|$, that is, disagreements corresponding to adjacent categories would have higher weights than disagreements corresponding to categories that are further apart.

Consider the structure of the agreement table presented in Table 2. Let $v \in (0, 1]$ and let $w(a_i)$ denote the weight corresponding to the element $a_i$. In the theorem below we require that the elements on the main diagonal have weight 1, the elements on the superdiagonal and the subdiagonal have the same weight $v$, and that all other weights are between 0 and $v$. Examples of weights that satisfy these conditions are the weights presented in (2) and (3).

**Theorem.** *Suppose the agreement table has the form presented in Table 2, and suppose that not all $b_i$ and $c_i$ are 0. Let $v \in (0, 1]$ and let the weights of $\kappa_w$ be given by*

$$w(a_i) = 1 \quad \text{for } i \in \{1, 2, \ldots, n\},$$
$$w(b_i) = w(c_i) = v \quad \text{for } i \in \{1, 2, \ldots, n-1\},$$
$$w_{ij} \in [0, v) \quad \text{for } i, j \in \{1, 2, \ldots, n\} \text{ and } |i - j| \geq 2.$$

*Then $\kappa_w > \kappa$.*

**Proof.** We first show that (4) is equivalent to (6). Since $1 - p_e$ and $1 - p_e^w$ are positive numbers, we have $\kappa_w > \kappa$ if and only if

$$\frac{p_o^w - p_e^w}{1 - p_e^w} > \frac{p_o - p_e}{1 - p_e} \tag{4}$$
$$\Updownarrow$$
$$(p_o^w - p_e^w)(1 - p_e) > (p_o - p_e)(1 - p_e^w)$$
$$\Updownarrow$$
$$p_o^w - p_e^w - p_o^w p_e + p_e^w p_e > p_o - p_e - p_o p_e^w + p_e^w p_e. \tag{5}$$

Under the conditions of the theorem, $p_e^w > p_e$, that is, $p_e^w - p_e$ is a positive number. Subtracting $p_e^w p_e$ from and adding $p_o p_e$ to both sides of (5), we obtain

$$(p_o^w - p_o)(1 - p_e) > (p_e^w - p_e)(1 - p_o)$$
$$\Updownarrow$$
$$\frac{p_o^w - p_o}{p_e^w - p_e} > \frac{1 - p_o}{1 - p_e}. \tag{6}$$

Next, consider Table 2. Since $v$ is the common weight of all elements on the superdiagonal and subdiagonal, we have

$$p_o = \sum_{i=1}^{n} a_i \quad \text{and} \quad p_o^w = \sum_{i=1}^{n} a_i + v \sum_{i=1}^{n-1} (b_i + c_i)$$

and hence

$$p_o^w - p_o = v \sum_{i=1}^{n-1} (b_i + c_i) \quad \text{and} \quad 1 - p_o = \sum_{i=1}^{n-1} (b_i + c_i).$$

Thus, $p_o^w - p_o = v(1 - p_o)$, and since $1 - p_o$ (not all $b_i$ and $c_i$ are 0), $p_e^w - p_e$, $1 - p_e$, and $v$ are positive numbers, (6) holds if and only if

$$v(1 - p_e) > p_e^w - p_e. \tag{7}$$

Because

$$\sum_{i=1}^{n} \sum_{j=1}^{n} p_i q_j = 1,$$

(7) is equal to

$$v \left( \sum_{i=1}^{n} \sum_{j=1}^{n} p_i q_j - \sum_{i=1}^{n} p_i q_i \right) > \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} p_i q_j - \sum_{i=1}^{n} p_i q_i. \tag{8}$$

Inequality (8) holds since $v > w_{ij}$ for $i, j \in \{1, 2, \ldots, n\}$ and $|i - j| \geq 2$. This completes the proof. □

## 4. Conclusions

It has been frequently observed in the literature that, when Cohen's kappa and weighted kappa are applied to the same agreement table, the value of weighted kappa is higher than the value of Cohen's kappa. In this short paper we proved this phenomenon for tridiagonal agreement tables. A tridiagonal table is a square matrix that has nonzero elements only on the main diagonal, the first diagonal below this and the first diagonal above this. Agreement tables that are tridiagonal or almost tridiagonal (see for example Table 1) are frequently encountered in applications. Hence, tridiagonal agreement tables are general enough to make this result useful.

In the theorem we require that the elements on the main diagonal have weight 1, the elements on the first diagonals above and below the main diagonal have a weight $v \in (0, 1]$, and all other weights are between 0 and $v$. Examples of weights that satisfy these conditions are the linear weights [4,17, 14] and the quadratic weights [9,13,15]. In particular, the latter weights are frequently used in applications, although the weighted kappa statistic allows the use of weights of other types [6].

## References

[1] R.L. Brennan, D.J. Prediger, Coefficient kappa: some uses, misuses, and alternatives, Educational and Psychological Measurement 41 (1981) 687–699.
[2] H. Brenner, U. Kliebsch, Dependence of weighted kappa coefficients on the number of categories, Epidemiology 7 (1996) 199–202.
[3] J.-M. Cai, T.S. Hatsukami, M.S. Ferguson, R. Small, N.L. Polissar, C. Yuan, Classification of human carotid atherosclerotic lesions with in vivo multicontrast magnetic resonance imaging, Circulation 106 (2002) 1368–1373.
[4] D. Cicchetti, T. Allison, A new procedure for assessing reliability of scoring EEG sleep recordings, The American Journal of EEG Technology 11 (1971) 101–109.
[5] J. Cohen, A coefficient of agreement for nominal scales, Educational and Psychological Measurement 20 (1960) 37–46.
[6] J. Cohen, Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit, Psychological Bulletin 70 (1968) 213–220.
[7] M.S. Dirksen, J.J. Bax, A. De Roos, J.W. Jukema, R.J. Van der Geest, K. Geleijns, E. Boersma, E.E. Van der Wall, H.J. Lamb, Usefulness of dynamic multislice computed tomography of left ventricular function in unstable angina pectoris and comparison with echocardiography, The American Journal of Cardiology 90 (2002) 1157–1160.
[8] W.W. Eaton, K. Neufeld, L.-S. Chen, G. Cai, A comparison of self-report and clinical diagnostic interviews for depression, Archives of General Psychiatry 57 (2000) 217–222.
[9] J.L. Fleiss, J. Cohen, The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability, Educational and Psychological Measurement 33 (1973) 613–619.
[10] J.L. Fleiss, J. Cohen, B.S. Everitt, Large sample standard errors of kappa and weighted kappa, Psychological Bulletin 72 (1969) 323–327.
[11] U. Jakobsson, A. Westergren, Statistical methods for assessing agreement for ordinal data, Scandinavian Journal of Caring Sciences 19 (2005) 427–431.
[12] K. Krippendorff, Reliability in content analysis: some common misconceptions and recommendations, Human Communication Research 30 (2004) 411–433.
[13] H.L. Kundel, M. Polansky, Measurement of observer agreement, Radiology 288 (2003) 303–308.
[14] P.W. Mielke, K.J. Berry, A note on Cohen's weighted kappa coefficient of agreement with linear weights, Statistical Methodology 6 (2009) 439–446.
[15] C. Schuster, A note on the interpretation of weighted kappa and its relations to other rater agreement statistics for metric scales, Educational and Psychological Measurement 64 (2004) 243–253.
[16] J.M. Seddon, C.R. Sahagian, R.J. Glynn, R.D. Sperduto, E.S. Gragoudas, The Eye Disorders Case-Control Study Group, Evaluation of an iris color classification system, Investigative Ophthalmology & Visual Science 31 (1990) 1592–1598.
[17] S. Vanbelle, A. Albert, A note on the linearly weighted kappa coefficient for ordinal scales, Statistical Methodology 6 (2009) 157–163.
[18] J.C. Van Swieten, P.J. Koudstaal, M.C. Visser, H.J.A. Schouten, J. Van Gijn, Interobserver agreement for the assessment of handicap in stroke patients, Stroke 19 (1987) 604–607.
[19] H. Visser, T. De Nijs, The map comparison kit, Environmental Modelling & Software 21 (2006) 346–358.
[20] M.J. Warrens, On the equivalence of Cohen's kappa and the Hubert–Arabie adjusted rand index, Journal of Classification 25 (2008) 177–183.
[21] M.J. Warrens, Inequalities between kappa and kappa-like statistics for $k \times k$ tables, Psychometrika 75 (2010) 176–185.
[22] M.J. Warrens, Cohen's kappa can always be increased and decreased by combining categories, Statistical Methodology 7 (2010) 673–677.
[23] R. Zwick, Another look at interrater agreement, Psychological Bulletin 103 (1988) 374–378.