Statistical Methodology 7 (2010) 673–677

Contents lists available at ScienceDirect

Statistical Methodology

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

Cohen's kappa can always be increased and decreased by combining categories

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 24 February 2010
Received in revised form
6 May 2010
Accepted 7 May 2010

Keywords:
Cohen's kappa
Nominal agreement
Collapsing categories
Merging categories

ABSTRACT

The kappa coefficient is a popular descriptive statistic for summarizing the cross classification of two nominal variables with identical categories. It has been frequently observed in the literature that combining two categories increases the value of kappa. In this note we prove the following existence theorem for kappa: For any nontrivial $k \times k$ agreement table with $k \in \mathbb{N}_{\geq 3}$ categories, there exist two categories such that, when combined, the kappa value of the collapsed $(k-1) \times (k-1)$ agreement table is higher than the original kappa value. In addition, there exist two categories such that, when combined, the kappa value of the collapsed table is smaller than the original kappa value.

1. Introduction

The kappa coefficient [4,6,8,3,19,15,16,18] is a popular descriptive statistic for summarizing the cross classification of two nominal variables with $k \in \mathbb{N}_{\geq 2}$ identical categories. These $k \times k$ tables occur in various fields of science, including psychometrics, educational measurement, epidemiology, map comparison [14] and content analysis [11]. Suppose that two observers each distribute $m \in \mathbb{N}_{\geq 1}$ objects (individuals) among a set of $k \in \mathbb{N}_{\geq 2}$ mutually exclusive categories, that are defined in advance. Let the agreement table A with elements a_{ij} $(i, j \in \{1, \ldots, k\})$ be the cross classification of the ratings of the observers, where a_{ij} indicates the number of objects placed in category i by the first observer and in category j by the second observer. For notational convenience, let P be the agreement table of the same size as A $(k \times k)$ with elements $p_{ij} = a_{ij}/m$. Row and column totals

$$p_{i+} = \sum_{j=1}^{k} p_{ij} \quad \text{and} \quad p_{+j} = \sum_{i=1}^{k} p_{ij}$$

* 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.2010.05.003

Table 1
Personality descriptions of oldest child by 200 sets of fathers and mothers [4].

Father	Mother			Row totals
	Type 1	Type 2	Type 3	
Type 1	0.44	0.05	0.01	0.50
Type 2	0.07	0.20	0.03	0.30
Type 3	0.09	0.05	0.06	0.20
Column totals	0.60	0.30	0.10	1.00

$\kappa = 0.492$.

are the marginal proportions of P. The kappa coefficient is defined as

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

where

$$p_o = \sum_{i=1}^{k} p_{ii} \quad \text{and} \quad p_e = \sum_{i=1}^{k} p_{i+}p_{+i}.$$

Table A is also called an agreement table. As an example, consider the data in Table 1 taken from [4], p. 37. In this study, 200 sets of fathers and mothers were asked to identify which of three personality descriptions best describes their oldest child. Table 1 is the proportion table of the cross classification of the fathers description and mothers description of the oldest child. We have $p_o = 0.44 + 0.20 + 0.06 = 0.70$, $p_e = (0.50)(0.60) + (0.30)^2 + (0.20)(0.10) = 0.41$ and $\kappa = 0.492$.

The number of categories used in various classification schemes varies from the minimum number of two to five in many practical applications. It is sometimes desirable to combine some of the k categories, for example, when two categories are easily confused, and then calculate the kappa value of the collapsed $(k-1) \times (k-1)$ agreement table. It has been frequently observed in applications that this increases the value of kappa. Furthermore, Fleiss [5] considered all mergers of four of the five categories of a data set and showed that combining categories can both increase and decrease the value of kappa. Kraemer [9] presented a method for testing if an increase in the value of kappa is a significant change. Schouten [13] presented a necessary and sufficient condition for the value of kappa to increase when two categories are combined, and showed that the result can be used to detect categories that are easily confused.

Schouten [13] showed that it depends on which categories are combined whether the value of kappa increases or decreases. These categories can be found by trial and error, or by the procedures proposed in [13]. The result presented in [13] gives rise to the following question: Is it always possible to increase or decrease the value of kappa by merging two categories. The answer is affirmative. In the following we show that for any nontrivial table with $k \in \mathbb{N}_{\geq 3}$ categories there exist two categories such that, when the two are merged, the kappa value of the collapsed $(k-1) \times (k-1)$ agreement table is higher than the original kappa value, and that there exist two categories such that, when combined, the kappa value of the collapsed table is smaller than the original kappa value.

2. Results

The main result is the theorem below. We first present two auxiliary results.

Lemma 1. *Let $n \in \mathbb{N}_{\geq 2}$ and let $b_1, b_2, \ldots, b_n$, at least 2 nonzero and nonidentical, and $c_1, c_2, \ldots, c_n$ be real nonnegative numbers with $c_t \neq 0$ if $b_t \neq 0$ for all $t \in \{1, \ldots, n\}$. Furthermore, let $u = \sum_{r=1}^{n} b_r$ and $v = \sum_{r=1}^{n} c_r$. Then there exist indices $r, s \in \{1, \ldots, n\}$ with $r \neq s$ such that*

$$\frac{b_r}{c_r} > \frac{u}{v} \quad \text{and} \quad \frac{b_s}{c_s} < \frac{u}{v}.$$

Proof. Without loss of generality, suppose $b_1/c_1 > u/v$. If $b_2 \neq 0$ and $b_2/c_2 < u/v$ then we are finished. Instead, suppose that $b_r = 0$ or $b_r/c_r > u/v$ for $r \in \{1, \ldots, n-1\}$. Since b_r and c_r

for $r \in \{1, \ldots, n-1\}$ are nonnegative numbers, we have $b_r v > c_r u$ for $r \in \{1, \ldots, n-1\}$ and $b_r \neq 0$. Adding these inequalities we obtain $(b_1 + b_2 + \cdots + b_{n-1})v > (c_1 + c_2 + \cdots + c_{n-1})u$ or $(u - b_n)v > (v - c_n)u$. The latter inequality is equivalent to $b_n/c_n < u/v$, since $u > b_n$ and $v > c_n$, and because b_n and c_n are nonnegative numbers. This completes the proof. □

For the lemma and theorem below, we assume the following situation. Let P be any $k \times k$ agreement table and let κ denote the corresponding kappa value. Let κ^* denote the kappa value corresponding to the $(k-1) \times (k-1)$ agreement table we obtain by combining categories i and j of P. Lemma 2 is a slightly adapted version of the result in [13], p. 455.

Lemma 2. *We have $\kappa^* > \kappa$ if and only if*

$$\frac{p_{ij} + p_{ji}}{p_{i+}p_{+j} + p_{j+}p_{+i}} > \frac{1 - p_o}{1 - p_e}.$$

Similarly, we have $\kappa^* < \kappa$ if and only if

$$\frac{p_{ij} + p_{ji}}{p_{i+}p_{+j} + p_{j+}p_{+i}} < \frac{1 - p_o}{1 - p_e}.$$

Theorem. *Assume that P has at least 2 nonidentical and nonzero elements that are not on the main diagonal. Then there exist categories i and j such that $\kappa^* > \kappa$ if i and j are combined. Furthermore, there exist categories i' and j', with not both $i = i'$ and $j = j'$, such that $\kappa^* < \kappa$ if i' and j' are combined.*

Proof. Note that the p_{ij} and the $p_{i+}p_{j+}$ for $i, j \in \{1, \ldots, k\}$ satisfy the criteria of the b_r and c_r of Lemma 1. Let $n = k(k-1)/2$, let $b_r = p_{ij} + p_{ji}$ and let $c_r = p_{i+}p_{+j} + p_{j+}p_{+i}$ with

$$r = \frac{(i-1)(i-2)}{2} + j,$$

for $i \in \{2, \ldots, k\}$ and $j \in \{1, \ldots, i-1\}$. We have

$$u = \sum_{r=1}^{n} b_r = \sum_{j<i}^{k} (p_{ij} + p_{ji}) = 1 - p_o$$

and

$$v = \sum_{r=1}^{n} c_r = \sum_{j<i}^{k} (p_{i+}p_{+j} + p_{j+}p_{+i}) = 1 - p_e.$$

The result then follows from application of Lemmas 1 and 2. □

3. Discussion

In the previous section we proved that for any agreement table with $k \in \mathbb{N}_{\geq 3}$ categories, there exist two categories such that, when the two are combined, the kappa value of the collapsed $(k-1) \times (k-1)$ agreement table is higher than the original kappa value, and that there exist two categories such that, when combined, the kappa value of the collapsed table is smaller than the original kappa value. Especially an increase in the value of kappa when merging categories has been frequently observed in applications. The theorem is an existence theorem. It states that there exist categories for increasing (decreasing) the kappa value, but it does not specify which categories these are.

As an example, consider the data in Table 1. The 3×3 table has a kappa value of 0.492. With 3 categories there are 3 pairs of categories that can be combined. Table 2 contains the 2×2 tables [17] that we obtain if, respectively, categories 1 and 2, 1 and 3, and 2 and 3 are combined. If categories 1 and 2 are merged the kappa value decreases to 0.308. If categories 1 and 3 or 2 and 3 are combined the kappa value increases to 0.524 and 0.560 respectively.

The existence theorem is applicable if the elements of the agreement table are not all equal (any nontrivial $k \times k$ table). As a result it is possible, after combining two categories, to inspect the collapsed

Table 2
The three 2 × 2 proportion tables that are obtained if 2 categories of Table 1 are merged.

Father	Mother		Row totals
	Types 1+2	Type 3	
Types 1+2	0.76	0.04	0.80
Type 3	0.14	0.06	0.20
Column totals	0.90	0.10	1.00
$\kappa = 0.308$			
Father	Mother		Row totals
	Types 1+3	Type 2	
Types 1+3	0.60	0.10	0.70
Type 2	0.10	0.20	0.30
Column totals	0.70	0.30	1.00
$\kappa = 0.524$			
Father	Mother		Row totals
	Type 1	Types 2 +3	
Type 1	0.44	0.06	0.50
Types 2+3	0.16	0.34	0.50
Column totals	0.60	0.40	1.00
$\kappa = 0.560$			

$(k-1) \times (k-1)$ table for two new categories so that, when the new categories are combined, the value of kappa is again increased (decreased). This process can be repeated until there are only $k = 2$ categories left. As an example, consider the 5×5 table presented in [1], p. 376 and [2], p. 206 on occupational status for 3500 British father–son pairs. The 5×5 table, denoted by (1)(2)(3)(4)(5), has a kappa value of 0.182. Combining categories 4 and 5 we obtain a 4×4 table denoted by (1)(2)(3)(4, 5). This table has a kappa value of 0.255. The tables (1)(2)(3, 4, 5) and (1)(2, 3, 4, 5) have kappa values of 0.329 and 0.406 respectively, illustrating that kappa can be increased by successively merging categories. The tables (1, 3)(2)(4)(5), (1, 3)(2, 4)(5) and (1, 3, 5)(2, 4) have kappa values of 0.179, 0.162 and −0.282, which illustrates that kappa can also be decreased by successively combining categories.

Although the existence theorem states that for $k \in \mathbb{N}_{\geq 3}$ there exist two categories for increasing the kappa value, it is perhaps not methodologically sound to improve a nominal scale using only statistical criteria. For example, after combining categories the resultant scale may have higher reliability but lack face validity. Furthermore, if merging two categories raises the value of kappa, this may indicate that the two categories are easily confused. Several authors have developed measures of confusion between pairs of categories [7,10,12]. These methods of analysis can be used to identify pairs of classifications between which there is substantial confusion.

Acknowledgements

The author thanks the associate editor, Sophie Vanbelle and an anonymous reviewer for their helpful comments and valuable suggestions on an earlier version of this note.

References

[1] A. Agresti, Categorical Data Analysis, Wiley, New York, 1990.
[2] Y.M.M. Bishop, S.E. Fienberg, P.W. Holland, Discrete Multivariate Analysis. Theory and Practice, MIT Press, Cambridge, 1976.
[3] R.L. Brennan, D.J. Prediger, Coefficient kappa: some uses, misuses, and alternatives, Educational and Psychological Measurement 41 (1981) 687–699.
[4] J. Cohen, A coefficient of agreement for nominal scales, Educational and Psychological Measurement 20 (1960) 213–220.
[5] J.L. Fleiss, Measuring nominal scale agreement among many raters, Psychological Bulletin 76 (1971) 378–382.
[6] J.L. Fleiss, Measuring agreement between two judges on the presence or absence of a trait, Biometrics 31 (1975) 651–659.
[7] I.R. James, Analysis of nonagreement among multiple raters, Biometrics 39 (1983) 651–657.
[8] H.C. Kraemer, Ramifications of a population model for κ as a coefficient of reliability, Psychometrika 44 (1979) 461–472.

[9] H.C. Kraemer, Extension of the kappa coefficient, Biometrics 36 (1980) 207–216.
[10] H.C. Kraemer, Measurement of reliability for categorical data in medical research, Statistical Methods in Medical Research 1 (1992) 183–199.
[11] K. Krippendorff, Reliability in content analysis: some common misconceptions and recommendations, Human Communication Research 30 (2004) 411–433.
[12] C. Roberts, R. McNamee, A matrix of kappa-type coefficients to assess the reliability of nominal scales, Statistics in Medicine 17 (1998) 471–488.
[13] H.J.A. Schouten, Nominal scale agreement among observers, Psychometrika 51 (1986) 453–466.
[14] H. Visser, T. De Nijs, The map comparison kit, Environmental Modelling & Software 21 (2006) 346–358.
[15] M.J. Warrens, On the equivalence of Cohen's kappa and the Hubert–Arabie adjusted Rand index, Journal of Classification 25 (2008a) 177–183.
[16] M.J. Warrens, On similarity coefficients for 2×2 tables and correction for chance, Psychometrika 73 (2008b) 487–502.
[17] M.J. Warrens, On association coefficients for 2×2 tables and properties that do not depend on the marginal distributions, Psychometrika 73 (2008c) 777–789.
[18] M.J. Warrens, Inequalities between kappa and kappa-like statistics for $k \times k$ tables, Psychometrika 75 (2010) 176–185.
[19] R. Zwick, Another look at interrater agreement, Psychological Bulletin 103 (1988) 374–378.