



Universiteit
Leiden
The Netherlands

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

Warrens, M.J.

Citation

Warrens, M. J. (2010). Cohen's kappa can always be increased and decreased by combining categories. *Statistical Methodology*, 7, 673-677. Retrieved from <https://hdl.handle.net/1887/15982>

Version: Not Applicable (or Unknown)

License: [Leiden University Non-exclusive license](#)

Downloaded from: <https://hdl.handle.net/1887/15982>

Note: To cite this publication please use the final published version (if applicable).

Postprint. Warrens, M. J. (2010). Cohen's kappa can always be increased and decreased by combining categories. *Statistical Methodology*, 7, 673-677.

<http://dx.doi.org/10.1016/j.stamet.2010.05.003>

Author. Matthijs J. Warrens
Institute of Psychology
Unit Methodology and Statistics
Leiden University
P.O. Box 9555, 2300 RB Leiden
The Netherlands
E-mail: warrens@fsw.leidenuniv.nl

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

Matthijs J. Warrens, Leiden University

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.

Key words. Cohen's kappa; Nominal agreement; Collapsing categories; Merging categories.

Acknowledgment. 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.

1 Introduction

The kappa coefficient (Cohen, 1960; Fleiss, 1975; Kraemer, 1979; Brennan and Prediger, 1981; Zwick 1988; Warrens, 2008a,b, 2010) 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 (Visser and De Nijs, 2006) and content analysis (Krippendorff, 2004). 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} ($i, j \in \{1, \dots, k\}$) 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

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

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 Cohen (1960, 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 $\kappa = 0.492$.

Table 1: Personality descriptions of oldest child by 200 sets of fathers and mothers (Cohen, 1960).

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
$\kappa = 0.492$				

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 (1971) 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 (1980) presented a method for testing if an increase in the value of kappa is a significant change. Schouten (1986) 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 (1986) 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 Schouten (1986). The result presented in Schouten (1986) 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, \dots, b_n , at least 2 nonzero and nonidentical, and c_1, c_2, \dots, c_n be real nonnegative numbers with $c_t \neq 0$ if $b_t \neq 0$ for all $t \in \{1, \dots, n\}$. Furthermore, let $u = \sum_{r=1}^n b_r$ and $v = \sum_{r=1}^n c_r$. Then there exist indices $r, s \in \{1, \dots, n\}$ 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 $r \in \{1, \dots, n-1\}$. Since b_r and c_r for $r \in \{1, \dots, n-1\}$ are nonnegative numbers, we have $b_r v > c_r u$ for $r \in \{1, \dots, n-1\}$ and $b_r \neq 0$. Adding these inequalities we obtain $(b_1 + b_2 + \dots + b_{n-1})v > (c_1 + c_2 + \dots + c_{n-1})u$ 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. \square

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 Schouten (1986, 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 $i, j \in \{1, \dots, k\}$ 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 $i \in \{2, \dots, k\}$ and $j \in \{1, \dots, i-1\}$. 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 Lemma 1 and Lemma 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 (Warrens, 2008c) 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 $(k-1) \times (k-1)$ 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 Agresti (1990, p. 376) and Bishop, Fienberg and Holland (1976, 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

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

$\kappa = 0.560$

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 (James, 1983; Kraemer, 1992; Roberts and McNamee, 1998). These methods of analysis can be used to identify pairs of classifications between which there is substantial confusion.

References

- Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York.
- Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1976). *Discrete Multivariate Analysis. Theory and Practice*. MIT Press, Cambridge.
- Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, *41*, 687-699.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37-46.
- Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*, 378-382.
- Fleiss, J.L. (1975). Measuring agreement between two judges on the presence or absence of a trait. *Biometrics*, *31*, 651-659.
- James, I. R. (1983). Analysis of nonagreement among multiple raters. *Biometrics*, *39*, 651-657.
- Kraemer, H. C. (1979). Ramifications of a population model for κ as a coefficient of reliability. *Psychometrika*, *44*, 461-472.
- Kraemer, H. C. (1980). Extension of the kappa coefficient. *Biometrics*, *36*, 207-216.
- Kraemer, H. C. (1992). Measurement of reliability for categorical data in medical research. *Statistical Methods in Medical Research*, *1*, 183-199.
- Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, *30*, 411-433.
- Roberts, C., & McNamee, R. (1998). A matrix of kappa-type coefficients to assess the reliability of nominal scales. *Statistics in Medicine*, *17*, 471-488.
- Schouten, H. J. A. (1986). Nominal scale agreement among observers. *Psychometrika*, *51*, 453-466.
- Visser, H., & De Nijs, T. (2006). The map comparison kit. *Environmental Modelling & Software*, *21*, 346-358.
- Warrens, M. J. (2008a). On the equivalence of Cohen's kappa and the Hubert-Arabie adjusted Rand index. *Journal of Classification*, *25*, 177-183.
- Warrens, M. J. (2008b). On similarity coefficients for 2×2 tables and correction for chance. *Psychometrika*, *73*, 487-502.
- Warrens, M. J. (2008c). On association coefficients for 2×2 tables and properties that do not depend on the marginal distributions. *Psychometrika*, *73*, 777-789.
- Warrens, M.J. (2010). Inequalities between kappa and kappa-like statistics for $k \times k$ tables. *Psychometrika*, *75*, 176-185.
- Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin*, *103*, 374-378.