Journal of Psychology & Psychotherapy

Review Article Open Access

Five Ways to Look at Cohen's Kappa

Matthijs J Warrens*

Warrens Institute of Psychology, Unit Methodology and Statistics, Leiden University, Netherlands

Abstract

The kappa statistic is commonly used for quantifying inter-rater agreement on a nominal scale. In this review article we discuss five interpretations of this popular coefficient. Kappa is a function of the proportion of observed and expected agreement, and it may be interpreted as the proportion of agreement corrected for chance. Furthermore, kappa may be interpreted as the average category reliability as well as an intraclass correlation.

Keywords: Inter-rater reliability; Inter-rater agreement; Category reliability; Average category reliability

Introduction

An important form of measurement in behavioral, social and medical sciences is nominal classification, that is, the assignment of subjects to qualitative categories, as in psychiatric diagnosis. If the rater (clinician, psychologist) did not fully understand what he or she was asked to interpret, or if the definition of the categories is ambiguous, the reliability of the ratings may be poor. A poor diagnosis will limit the possible degree of association between diagnosis and anything else.

To assess the reliability of a rating instrument researchers typically ask two raters to classify the same group of subjects independently. The pairwise ratings of a group of subjects into nominal categories are often summarized in a contingency table. Because the row labels and column labels of this contingency table are identical, the table is usually called an agreement table. Table 1 is an example of an agreement table. It contains the relative frequencies of the pairwise ratings of 100 patients by two clinicians into four categories: 1 = Schizophrenia, 2 = Bipolar disorder, 3 = Depression and 4 = other.

Table 2 and 3 are two other examples of agreement tables. Formal assessments in education nowadays consist of both numerical and contextual mathematics problems. An international reform has introduced various new solution strategies of multi digit mathematics problems.

	Clinician B				
Clinician A	1	2	3	4	Total
1 = Schizophrenia	0.23	0.01	0.01	0.00	0.25
2 = Bipolar disorder	0.00	0.20	0.01	0.02	0.23
3 = Depression	0.01	0.02	0.21	0.04	0.28
4 = Other	0.01	0.02	0.04	0.17	0.24
Total	0.25	0.25	0.27	0.23	1.00

Table 1: Relative frequencies of hypothetical diagnoses of 100 patients by two clinicians.

	Psychologist B			
Psychologist A	1	2	3	Total
1 = Long division	0.20	0.00	0.03	0.23
2 = Repeated subtraction	0.05	0.30	0.00	0.35
3 = Repeated addition	0.00	0.02	0.40	0.42
Total	0.25	0.32	0.43	1.00

Table 2: Relative frequencies of hypothetical solution strategies of 100 students coded by two psychologists.

	Algorithm B			
Algorithm A	**1**	**2**	**3**	**Total**
1 = White matter	0.45	0.02	0.00	0.47
2 = Grey matter	0.02	0.45	0.00	0.47
3 = Cerebral spinal fluid	0.00	0.01	0.05	0.06
Total	0.47	0.48	0.05	1.00

Table 3: Relative frequencies of hypothetical classifications of 8000 voxels by two algorithms.

The traditional algorithm in complex division is long division, whereas modern solution strategies are repeated subtraction and repeated addition. Table 2 contains the relative frequencies of the pairwise codings by two educational psychologists of written solution strategies of 100 sixth graders into three categories: 1 = Long division, 2 = Repeated subtraction and 3 = Repeated addition.

Magnetic Resonance Imaging (MRI) is an imaging technique that provides researchers with tools to observe noninvasively neural activity in the human brain. In MRI the brain is divided into a set of cubes, called voxels, and to interpret an image each voxel must be identified. Classification of brain tissues is usually done with software algorithms. Table 3 contains the relative frequencies of the pairwise classifications of 8000 voxels by two algorithms into three categories: 1 = White matter, 2 = Grey matter and 3 = Cerebral spinal fluid.

The agreement between the raters (or algorithms) can be used as an indicator of the quality of the categories of the rating instrument and the raters' ability to apply them. High agreement between the ratings indicates consensus in the diagnosis and interchangeability of the ratings. Cohen's kappa is the most commonly used statistic for assessing nominal agreement between two raters [1-7]. Kappa has value 1 if there is perfect agreement between the raters, and value 0 if the observed agreement is equal to agreement expected by chance. Several authors have suggested interpretation or benchmark guidelines for values between 0 and 1. The most commonly used guidelines are due to Landis and Koch [8]: 0.00 - 0.20 indicates slight agreement, 0.21- 0.40 fair agreements, 0.41-0.60 moderate agreement, 0.61-0.80 substantial

***Corresponding author:** Matthijs J, Warrens Institute of Psychology, Unit Methodology and Statistics, Leiden University, P.O. Box 9555,2300RBLeiden, Netherlands, Tel: +31 71 527 3649; Fax: +31 71 527 3619; E-mail: warrens@fsw.leidenuniv.nl

Received April 23, 2015; **Accepted** July 21, 2015; **Published** July 28, 2015

Citation: Warrens MJ (2015) Five Ways to Look at Cohen's Kappa. J Psychol Psychother 5: 197. doi: 10.4172/2161-0487.1000197

agreement, and 0.81-1.00 indicates almost perfect agreement. However, it should be noted that these guidelines are generally considered arbitrary.

Since its introduction kappa has been applied in thousands of research applications. Several authors have identified difficulties with the interpretation of kappa [4,9,10]. For example, kappa depends on the base rates of the categories. The base rates reflect how often the categories were used by the raters. Kappas from samples with different base rates are therefore not comparable [6,9,10]. However, since kappa has several useful interpretations, it is likely that it continues to be a standard tool for summarizing inter-rater agreement on a nominal scale in the near future [1,3,4,11,12].

The kappa statistic was introduced by Cohen [2] in 1960. However, the basic idea of an agreement measure was anticipated substantially before 1960. For example, decades earlier Corrado Gini already considered measures for assessing agreement on a nominal scale [11,13,14]. Furthermore, Cohen's paper [2] was a response to Bennett et al. [15] and Scott [16] a few years before. The measure in Scott [16] is closely related to kappa, and is also-known as the intraclass kappa [17]. Although it was not the first measure of inter-rater agreement on a nominal scale, kappa is the most widely used agreement measure [1,3,4,11,18].

In this article we discuss five ways to look at kappa. Following Cohen and others our focus is on kappa as a computational index. We present several algebraic interpretations. Since computation of a sample kappa requires no assumptions about a population, the interpretations are distribution free. Several interpretations have been around for quite a while, while others were discovered more recently. It is not claimed here that all possible interpretations of Cohen's kappa are discussed. For example, additional interpretations of kappa can be found in [7,17,18].

Kappa as a Function of the Proportion Observed and Expected Agreement

Cohen's kappa is a dimensionless index that can be used to express the agreement between two raters in a single number. Let p_{ii} denote the proportion of patients classified into category *i* according to both raters. The sum of these proportions is called the proportion of observed agreement, which is given by

$$P_o = \sum p_{ii}. \tag{1}$$

For the data in Table 1 we have

P_o=0.23+0.20+0.21+0.17=0.81

Furthermore, let p_{i+} and p_{+i} denote the proportion of patients classified into category *i* by the first and second rater, respectively. The numbers p_{i+} and p_{+i} are the base rates; they reflect how often the raters used category *i*. Using the base rates the proportion of expected agreement is given by

$$P_e = \sum p_i + p_{+i}. \tag{2}$$

For the data in Table 1 we have

P_e=$(0.25)^2$+(0.23)(0.25)+(0.28)(0.27)+(0.24)(0.23)=0.25

Next, the kappa statistic is usually defined as a function of the observed and expected agreement, given by

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{3}$$

Equation (3) is the usual formula found in introductory statistics textbooks. For the data in Table 1 we have $\kappa = (0.81 - 0.25)/(1 - 0.25) = 0.75$, which, according to the guidelines in Landis and Koch [7], indicates a substantial level of agreement. For the data in Table 2 we have p_o = 0.90, p_e = 0.35 and $\kappa = 0.85$, whereas for the data in Table 3 we have p_o = 0.95, p_e = 0.45 and $\kappa = 0.91$. In the latter two cases agreement is almost perfect.

Kappa as the Proportion of Agreement Corrected for Chance

It is sometimes desirable that the theoretical value of an agreement measure is zero if the classifications made by the raters are statistically independent [19]. For example, kappa has zero value under statistical independence, but the proportion of observed agreement has not. Furthermore, since raters may agree on the diagnoses simply by chance, the value of the proportion of observed agreement is generally considered to be artificially high.

If a coefficient does not have zero value under statistical independence of the raters, it can be corrected for agreement due to chance [4,19,20]. After correction for chance agreement, a measure *M* has a form

$$\frac{M - E(M)}{1 - E(M)'} \tag{4}$$

where expectation *E*(*M*) is the value of *M* under statistical independence.

The value of the observed agreement P_o under statistical independence is given by E(P_o) = P_e. Using the latter formula together with M = Po in equation (4) yields equation (3). Hence, kappa is the chance-corrected version of the proportion of observed agreement. We thus have the following alternative interpretation of kappa: $\kappa = 0.75$ is the proportion of agreement P_o corrected for chance. Note that to calculate the value of kappa in practice one can simply use equation (3).

Kappa as an Average Kappa When Two Categories Are Combined

The number of categories used for various rating instruments varies from the minimum number of two to five in many practical applications. It is sometimes desirable to combine some of the categories, for example, when two categories are easily confused [5,7]. With *m* categories there are *m*(*m*-1)/2 pairs, and thus *m*(*m*-1)/2 ways to combine two categories. For example, Table 1 has four categories and there are 4(4-1)/2=6 ways to combine two categories.

Although the value of kappa may increase if two categories are combined, it is a common misconception that this is always the case [5,7]. In fact, kappa may also decrease. For example, in Table 1 there is little disagreement between the raters on the categories 1 = Schizophrenia and 2 = Bipolar. The two categories can be clearly distinguished from one another. If we combine the two categories the kappa value decreases from $\kappa = 0.75$ to $\kappa_{12} = 0.17$ (the subscripts 12 in κ_{12} denotes that categories 1 and 2 are combined). On the other hand, there is some disagreement between the raters on categories 3 = Depression and 4 = Other. If we combine these two categories kappa increase to $\kappa_{34} = 0.82$, which is a substantial increase. The remaining four values of kappa that can be obtained if we combine two categories are, $\kappa_{13} = \kappa_{14} = 0.72$, $\kappa_{23} = 0.74$ and $\kappa_{24} = 0.76$. Thus, kappa can either increase or decrease if we combine categories, and both cases are always possible [5].

By combining categories the value of kappa may increase. Hence, the reliability of a rating instrument can be increased by combining appropriate categories. By combining categories one can assess the quality of the categories of the rating instrument and the raters' ability to apply them. If two categories are combined in practice it is important to consider the (possible) substantial meaning of the new category.

The overall kappa is an average of all kappas that are obtained if we combine two categories. More precisely, the overall kappa is a weighted average of these kappas if we use the denominators of the kappas as weights. A weighted average is a mean value just like an ordinary average. For example, for Table 1 the weights are 0.63, 0.61, 0.63, 0.62, 0.64 and 0.62, respectively, and the weighted average is equal to

$$\frac{0.63(0.71)+0.61(0.72)+0.63(0.72)+0.62(0.74)+0.64(0.76)+0.62(0.82)}{0.63+0.61+0.63+0.62+0.64+0.62}=0.75=\kappa$$

Thus, the overall kappa is an average of the kappas corresponding to all possible tables that can be obtained by combining two categories. We thus have the following alternative interpretation of kappa: if we combine two categories the average kappa is $\kappa=0.75$.

Kappa as the Average Category Reliability

Kappa summarizes the agreement between two raters over all categories. It is frequently more informative to assess the agreement between the raters on the individual categories [7,18]. The category reliability of category *i* is given by

$$\kappa_i=\frac{p_{ii}-p_{i+}p_{+i}}{(p_{i+}+p_{+i})/2-p_{i+}p_{+i}} \qquad (5)$$

The category reliability in equation (5) reflects the agreement between the raters on category *i*. The statistic can also be obtained in the following way. An agreement table (for example Table 1) can be collapsed into a smaller table of size 2×2 by combining all categories other than the one of current interest (category *i*) into a single ``all others" category. The kappa value corresponding to this 2×2 table is the category reliability given in equation (5).

With *m* categories there are *m* category reliabilities, one for each category. Consider for example Table 1. The category reliability of Schizophrenia is $\kappa_i=0.89$, indicating almost perfect agreement, whereas the reliability of Bipolar is $\kappa_2=0.78$, indicating good agreement. The reliabilities of Depression and Other are $\kappa_3=0.67$ and $\kappa_4=0.64$, respectively, indicating only moderate to good agreement. The category reliabilities illustrate that agreement is much better on the first two categories than the last two categories.

The overall kappa is an average of the individual category reliabilities. More precisely, kappa is a weighted average of the category reliabilities using the denominator of the category reliabilities as weights [19]. For example, for Table 1 the weights are 0.19, 0.18, 0.20 and 0.18, respectively, and the weighted average is equal to

$$\frac{\sum w_i\kappa_i}{\sum w_i}=\frac{0.19(0.89)+0.18(0.78)+0.20(0.67)+0.18(0.64)}{0.19+0.18+0.20+0.18}=0.75=\kappa$$

We thus have the following alternative interpretation of kappa: the average category reliability is $\kappa=0.75$.

The interpretation of the overall kappa as the average category reliability has the following consequence. If the category reliabilities are quite different, for example, high agreement between the raters on one category (Schizophrenia, $\kappa_i=0.89$) but low agreement on another category (Depression, $\kappa_3=0.67$), the overall kappa cannot fully reflect the complexity of the patterns of agreement between the raters. It would therefore be good practice to report both the overall kappa and the category reliabilities of an agreement table. Such practice would provide substantially more information than reporting only a single number [19].

Kappa as an Intraclass Correlation

While kappa is commonly used to assess agreement between two raters when subjects are classified on a nominal scale, intraclass correlations are often used when two or more raters classify the same group of subjects on a numerical scale [21-24]. An intraclass correlation describes how strongly subjects in the same group or category resemble each other.

Rae [22] showed that if we use the Gini-Light-Margolin concept of partitioning variance for qualitative data, then Cohen's kappa may be interpreted as an intraclass correlation [21,23]. Let σ_r^2, σ_s^2 and σ_e^2 denote, respectively, the rater variance, the subjects variance, and the error variance. Furthermore, let *N* denote the total number of subjects (= sample size). Using these definitions kappa can be written as

$$\kappa=\frac{\sigma_r^2}{\sigma_s^2+\sigma_r^2+\sigma_e^2+\frac{1}{N-1}(\sigma_r^2+\sigma_e^2)} \qquad (6)$$

For large *N* equation (6) approximates

$$\frac{\sigma_s^2}{\sigma_s^2+\sigma_r^2+\sigma_e^2}, \qquad (7)$$

which is interpretable as the intraclass correlation of reliability when systematic variability among the raters is included as a component of the total variation. We thus have the following alternative interpretation of kappa: in terms of variance, the degree of resemblance of the subjects is $\kappa=0.75$.

Conclusion

In this article we reviewed five ways to look at Cohen's kappa. Certainly there are other ways to interpret kappa [7,17,18]. We do not presume here to have summarized all useful and interesting approaches of kappa. Nevertheless, the five approaches illustrate the diversity of interpretations available to researchers who use kappa.

The various interpretations of Cohen's kappa show the growth of this popular statistic over the past decades. Its popularity has led to the development of many extensions, including, kappas for three or more raters [20,22] and kappas for ordinal categories [23,24]. However, Cohen's statistic in (3) is surprisingly unchanged from the one originally proposed.

References

1. Brennan RL, Prediger DJ (1981) Coefficient kappa: Some uses, misuses, and alternatives. Educational and Psychological Measurement 41: 687–699.
2. Cohen J (1960) A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20: 37–46.
3. Sim J, Wright CC (2005) The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. Phys Ther 85: 257-268.
4. Warrens MJ (2010) Inequalities between kappa and kappa-like statistics for k×k tables. Psychometrika 75: 176-185.
5. Warrens MJ (2010) Cohen's kappa can always be increased and decreased by combining categories. Statistical Methodology 7: 673-677.
6. Warrens MJ (2010) A formal proof of a paradox associated with Cohen's kappa. Journal of Classification 27: 322-332.
7. Warrens MJ (2011) Cohen's kappa is a weighted average. Statistical Methodology 8: 473–484.
8. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33: 159-174.

9. Uebersax JS (1987) Diversity of decision-making models and the measurement of interrater agreement. Psychological Bulletin 101: 140–146.
10. Vach W (2005) The dependence of Cohen's kappa on the prevalence does not matter. J Clin Epidemiol 58: 655-661.
11. Warrens MJ (2013) A comparison of Cohen's kappa and agreement coefficients by Corrado Gini. International Journal of Research and Reviews in Applied Sciences 16: 345–351.
12. Warrens MJ (2014) New interpretations of Cohen's kappa. Journal of Mathematics. ID 203907.
13. Goodman LA, Kruskal WH (1959) Measures of association for cross classifications II: Further discussion and references. Journal of the American Statistical Association 54: 123–163.
14. Weida FM (1928) On various concepts of correlation. Annals of Mathematics 29: 276–312.
15. Bennett EM, Alpert E, Goldstein AC (1954) Communications through limited-response questioning. Public Opinion Quarterly 18: 303–308.
16. Scott WA (1955) Reliability of content analysis: The case of nominal scale coding. Public Opinion Quarterly 19: 321–325.
17. Bloch DA, Kraemer HC (1989) 2 x 2 kappa coefficients: Measures of agreement or association. Biometrics 45: 269-287.
18. Warrens MJ1 (2008) On Similarity Coefficients for 2x2 Tables and Correction for Chance. Psychometrika 73: 487-502.
19. Warrens MJ (2010) Inequalities between multi-rater kappas. Advances in Data Analysis and Classification 4: 271-286.
20. Fleiss JL (1975) Measuring agreement between two judges on the presence or absence of a trait. Biometrics 31: 651-659.
21. Rae G (1988) The equivalence of multiple rater kappa statistics and intraclass correlation coefficients. Educational and Psychological Measurement 48: 367–374.
22. Shrout PE, Fleiss JL (1979) Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin 86: 420–428.
23. Warrens MJ (2014) Corrected Zegers-ten Berge coefficients are special cases of Cohen's weighted kappa. Journal of Classification 31: 179–193.
24. Warrens MJ (2013) Conditional inequalities between Cohen's kappa and weighted kappas. Statistical Methodology 10: 14–22.