Dependence of Weighted Kappa Coefficients on the Number of Categories

Hermann Brenner and Ulrike Kliebsch

Weighted kappa coefficients are commonly used to quantify inter- or intra-rater reliability or test-retest reliability of ordinal ratings in clinical and epidemiologic applications. In this paper, we assess the dependence of weighted kappa coefficients on the number of categories and the type of weighting scheme, which vary between applications. The most commonly used weights are weights that are proportional to the deviation of individual ratings ("linear weights") or to the square of the deviation of individual ratings ("quadratic weights"). Quadratically weighted kappa coefficients are equivalent to the intraclass correlation coefficient and to the product-moment correlation coefficient under certain conditions. We illustrate that an increase of quadratically weighted kappa coefficients with the number of categories is expected under a broad variety of conditions, whereas linearly weighted kappa coefficients appear to be less sensitive to the number of categories. Number of categories and type of weighting scheme therefore require careful consideration in the interpretation of weighted kappa coefficients. (Epidemiology 1996;7:199–202)

Keywords: kappa, reliability, statistics.

Kappa coefficients are commonly employed to quantify the level of agreement between multiple ratings of categorical variables.[1] Most applications pertain to dual ratings, but extensions to more than two ratings have also been developed.[2] Kappa coefficients may be used for both dichotomous characteristics, such as presence or absence of disease, and polytomous characteristics. The latter are often ordinally scaled in clinical and epidemiologic applications, such as classifications of the severity of symptoms or disease, or classifications of the frequency, magnitude, or duration of exposure. The number of categories used in various classification schemes varies, but it is in the range from two (the minimum possible value) to five in most practical applications.

With ordinal categories, weighted kappa coefficients are often used in which disagreements are weighted by the magnitude of the discrepancy.[3] Weighted kappa coefficients may be equivalently expressed using weights of agreement rather than disagreement. The most commonly used weights are weights that are proportional to the deviation of individual ratings ("linear weights"), such as the numbers of categories of disagreement, and weights that are proportional to the square of the deviation of individual ratings ("quadratic weights"), such as the squared numbers of categories of disagreement. Quadratic weights are often recommended since quadratically weighted kappa coefficients are equivalent to the product-moment correlation and the intraclass correlation coefficient under certain conditions.[3,4]

A kappa coefficient of 1.0 indicates maximum possible agreement, whereas a kappa coefficient of 0 indicates lack of agreement beyond agreement by chance alone (negative kappa coefficients may also occur in the case of less than chance agreement). Interpretation of the intermediate values of kappa coefficients that are typically encountered in practice is less clear. Several authors have delineated ranges of values of kappa coefficients pertaining to excellent, moderate, and poor agreement.[5–7] Such classifications are problematic, however, since values of kappa coefficients also depend on factors other than reliability, such as the marginal distributions of the ratings.[8–15]

Another factor that has to be considered in the interpretation of kappa coefficients is the number of categories. Unweighted kappa coefficients decrease with the number of categories. Although this characteristic is an obvious consequence of the fact that unweighted kappa coefficients are measures of exact agreement, it has often been overlooked in the past.[16] A much less clear issue that has received comparatively little attention is the question of how values of weighted kappa coefficients are affected by the number of categories.

In this paper, we present an investigation of the impact of the number of categories on the value of linearly and quadratically weighted kappa coefficients.

From the Department of Epidemiology, University of Ulm, Ulm, Germany.

Address correspondence to: Hermann Brenner, Department of Epidemiology, University of Ulm, Albert-Einstein-Allee 43, D-89081 Ulm, Germany.

This work was partly supported by the Federal Ministry of Research and Technology, Bonn, Germany.

Submitted May 19, 1995; final version accepted September 1, 1995.

© 1996 by Epidemiology Resources Inc.

Methods

We assumed that the classification is made on an ordinal scale on the basis of explicit measurement or subjective perception of some underlying continuous trait. This assumption applies to many classifications in clinical and epidemiologic studies. Common examples include classification of severity of disease or of levels of exposure. Lack of reliability of classification is assumed to be due to random measurement or perception error of the underlying trait.

We first addressed the situation in which the trait that underlies categorization is normally distributed with mean X^* and standard deviation σ in the population. X^* may reflect laboratory parameters, severity of symptoms, or other quantitative parameters of clinical or epidemiologic relevance (or some mathematical function of them, such as the natural logarithm). To simplify notation, we assumed that X^* is transformed into a variable X that follows the standard normal distribution with $\mu = 0$ and $\sigma = 1$. Now, let Z_1 and Z_2 be approximate measurements or perceptions of X with $Z_1 = X + e_1$ and $Z_2 = X + e_2$, where e_1 and e_2 are additive measurement or perception errors (in the case of intra-rater reliability studies or test-retest studies, e_1 and e_2 may also reflect intraindividual variability of the underlying trait). We assumed that e_1 and e_2 are independent of the true trait X and of each other, and that e_1 and e_2 each follow a normal distribution with mean 0 and standard deviation σ_e. Then, the combined measurements or perceptions Z_1 and Z_2 follow a bivariate normal distribution with mean 0 and variance $\sigma^2 + \sigma_e^2$ for each component and correlation coefficient $\rho = \sigma^2/(\sigma^2 + \sigma_e^2)$.

We considered two types of classification schemes: (1) classification by quantiles of perceived levels of the underlying trait, and (2) classification by *a priori* defined cutpoints. Fixed cutpoints were arbitrarily determined using the following algorithm: $c_i = -2 + 4 \times i/k$, where k denotes the total number of categories and c_i denotes the cutpoint between the ith and the $(i + 1)$th category $[1 \leq i \leq (k - 1)]$. With this algorithm, cutpoints are equally spaced in the range from -2 to $+2$, which covers more than 95% of true values of the underlying trait X. With both classification schemes, we varied the number of categories between 2 and 8, and we varied the standard deviation of the measurement or perception error σ_e between 0.25, 0.50, 1.00, and 2.00 to reflect a broad range of classification accuracy. The categories were numbered in ascending order, and we employed both linearly and quadratically weighted kappa coefficients on the categorical ratings (using as weights the numbers of categories of disagreement between both ratings and the squared numbers of categories of disagreement, respectively).

Derivation of expected weighted kappa coefficients for the scenarios assuming normal distribution of the underlying trait is outlined in Appendix 1.

Although the underlying trait and the measurement or perception error can often be assumed to be approximately normally distributed, other distributional forms are also of interest. In particular, skewed distributions of the underlying traits and variation of measurement or perception error according to the true value are common. We therefore carried out additional analyses in which we assumed the underlying trait X to follow the exponential distribution with mean 1, and in which we assumed that the measurement or perception error of the natural logarithm of X follows the normal distribution with mean 0 and standard deviation σ_e. These assumptions reflect situations in which the distribution of the underlying trait is strongly skewed to the right (with a lower limit of 0) and in which measurement or perception error increases with the absolute true levels of the trait.

Again, we considered classifications by quantiles and by fixed cutpoints. The latter were determined using the following algorithm: $c_i = 3 \times i/k$, where k denotes the total number of categories and c_i denotes the cutpoint between the ith and the $(i + 1)$th category $[1 \leq i \leq (k - 1)]$. This algorithm produces equally spaced cutpoints in the range from 0 to +3 (which covers more than 95% of true values of the underlying trait). As before, we varied the number of categories between 2 and 8, and we varied σ_e between 0.25 and 2.0.

All calculations were made by numerical integration with the software package SAS using the cumulative distribution function PROBNORM of the standard normal distribution.[17]

Numerical Illustration

Figure 1 shows the expected kappa coefficients for the scenarios assuming normal distribution of the underlying trait and classification of individuals by quantiles of observed values. The four graphs pertain to four levels of measurement or perception error (σ_e = 0.25, 0.50, 1.00, and 2.00). In each graph, expected linearly weighted kappa coefficients (*black bars*) and quadratically weighted kappa coefficients (*hatched bars*) are depicted for increasing numbers of categories up to a maximum of eight.

As expected, kappa coefficients are in the range of 0 to 1 and decrease with the standard deviation of measurement or perception error. Whereas coefficients are around 0.80 for $\sigma_e = 0.25$ (*upper left graph*), coefficients are below 0.20 for $\sigma_e = 2.00$ (*lower right graph*). Kappa coefficients are generally somewhat lower than the correlation coefficient of the bivariate normal distribution of the pairs of observation of the continuous trait (Z_1, Z_2), which is given as $1/(1 + \sigma_e^2)$ and depicted as a *horizontal line* in the graphs. If there are only two categories, weighting is obviously without influence on the kappa coefficient. The kappa coefficient tends to decrease slightly with the number of categories if linear weights are used. In contrast, a more pronounced increase of kappa coefficients is expected with increasing numbers of categories if quadratic weights are used. In that case, the kappa coefficients are close to the correlation coefficient for the underlying continuous trait if there are five or more categories. Results for the scenar-

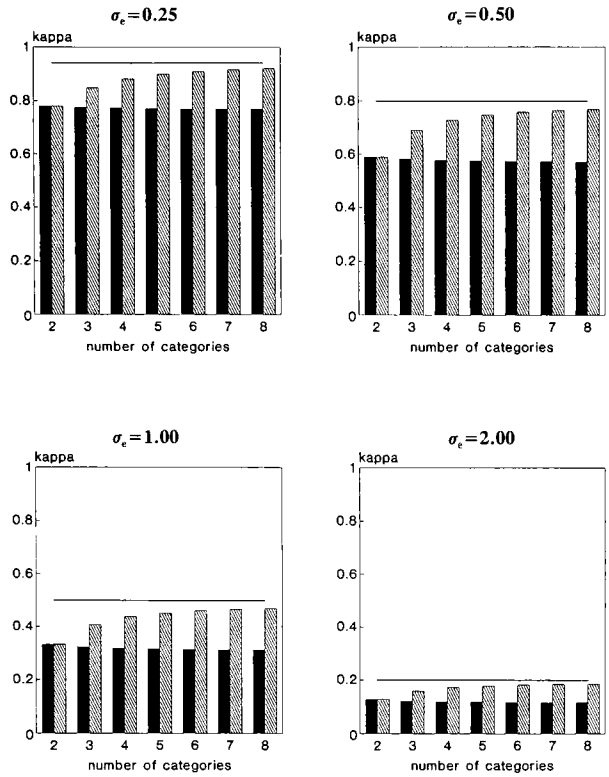

FIGURE 1. Expected kappa coefficients for the scenarios assuming normal distribution of the underlying trait and classification of individuals by quantiles of observed values. *Black bars* = linearly weighted kappa coefficients; *hatched bars* = quadratically weighted kappa coefficients; *horizontal line* = correlation coefficient of the continuous trait.

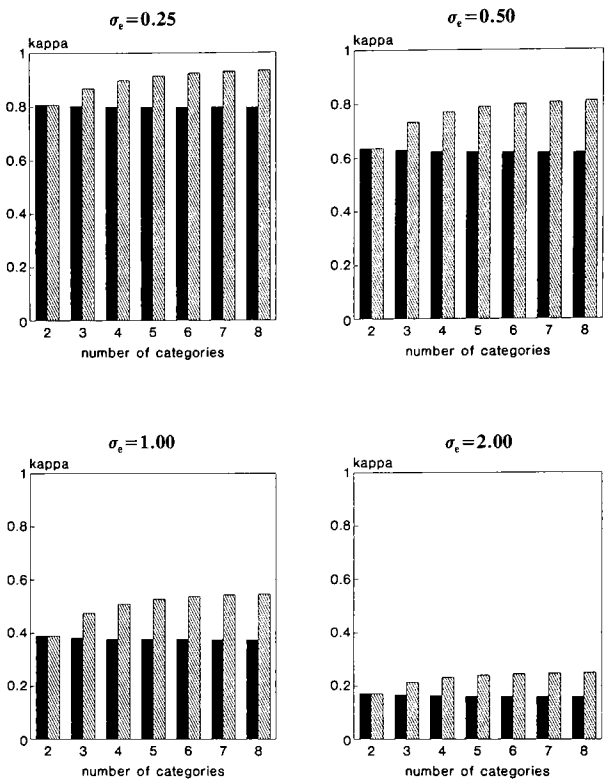

FIGURE 2. Expected kappa coefficients for the scenarios assuming exponential distribution of the underlying trait and classification of individuals by quantiles of observed values. *Black bars* = linearly weighted kappa coefficients; *hatched bars* = quadratically weighted kappa coefficients.

ios assuming normal distribution of the underlying trait and classification of individuals by fixed cutpoints were essentially equivalent and are therefore not illustrated separately.

Results for the scenarios assuming exponential distribution of the underlying trait and classification of individuals by quantiles of observed values are shown in Figure 2. Although overall levels of kappa were higher in these scenarios, the dependence of kappa coefficients on the number of categories was very similar to the corresponding scenarios assuming normal distribution of the underlying trait (Figure 1). We also found an increase of quadratically weighted kappa coefficients with the number of categories in the scenarios assuming exponential distribution of the underlying trait and classification of individuals by fixed cutpoints. In these scenarios (which are not illustrated separately to save space), expected linearly weighted kappa coefficients were also increasing with the number of categories, but the increase was much less pronounced than the increase of quadratically weighted coefficients.

Discussion

The dependence of kappa coefficients on the weighting scheme has been emphasized previously.[16,18] This paper demonstrates that the number of categories is an additional determinant of the magnitude of weighted kappa coefficients. Our findings indicate that quadratically weighted kappa coefficients tend to increase with the number of categories in many instances. This result contrasts with findings for unweighted kappa coefficients, which decrease with the number of categories.[16] Variation of the quadratically weighted kappa coefficient with the number of categories appears to be strongest in the range from two to five categories, the range of categories that is most frequently used in practical applications. In contrast, linearly weighted kappa coefficients tend to be less affected by the number of categories and might therefore eventually be preferred in special situations in which the focus is on comparing reliability between items with different numbers of categories. Nevertheless, the increase of quadratically weighted kappa coefficients with the number of categories can also be considered to be a desirable property, since, after all, as the number of categories increases, so does the proportion of the variability in the true variable captured by the imperfect ordinal variable.

In the interpretation of kappa coefficients, a variety of factors other than weighting scheme and number of categories have to be taken into account. The importance of the marginal distributions of ratings on the

values of kappa coefficients has previously been addressed by several authors.[8-15] In particular, the importance of imbalance and asymmetry of marginal ratings has been emphasized. Feinstein and Cicchetti[12] coined the terms balanced marginals and symmetry in the context of dual binary classifications for ratings with equal proportions of positive and negative classifications, and for situations with identical distributions of marginals in both ratings, respectively. Marginal distributions are symmetrical in all scenarios assessed in this paper, which enables one to assess the impact of the number of categories independent of the influence of asymmetry. In contrast to the scenarios assuming classification of individuals by quantiles of observed values, marginal distributions are strongly imbalanced in the scenarios assuming fixed cutpoints. The striking similarities of results for the very different assumptions with respect to the distribution of the underlying trait, to the distribution of measurement or perception error, and to the balance of marginal classifications indicate that the observed patterns might apply to a broad variety of conditions.

References

1. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20:37–46.
2. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull 1971;76:378–382.
3. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968;70:213–220.
4. Fleiss JL, Cohen J. The equivalence of weighted Kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Meas 1973; 33:613–619.
5. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–174.
6. Fleiss JL. Statistical Methods for Rates and Proportions. 2nd ed. New York: John Wiley and Sons, 1981.
7. Altman DG. Practical Statistics for Medical Research. London: Chapman and Hall, 1991.
8. Kraemer HC. Ramifications of a population model for κ as a coefficient of reliability. Psychometrika 1979;44:461–472.
9. Spitznagel EL, Helzer JE. A proposed solution to the base rate problem in the kappa statistic. Arch Gen Psychiatry 1985;42:725–728.
10. Thompson WD, Walter WD. A reappraisal of the kappa coefficient. J Clin Epidemiol 1988;41:949–958.
11. Gjørup T. The kappa coefficient and the prevalence of a diagnosis. Methods Inf Med 1988;27:184–186.
12. Feinstein AR, Cicchetti DV. High agreement but low kappa. I. The problems of two paradoxes. J Clin Epidemiol 1990;43:543–549.
13. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol 1993;46:423–429.
14. Guggenmoos-Holzmann I. How reliable are chance-corrected measures of agreement? Stat Med 1993;12:2191–2205.
15. Donker DK, Hasman A, van Geijn HP. Interpretation of low kappa values. Int J Biomed Comput 1993;33:55–64.
16. Maclure M, Willett WC. Misinterpretation and misuse of the kappa statistic. Am J Epidemiol 1987;126:161–169.
17. SAS Institute, Inc. SAS Language: Reference. Version 6. 1st ed. Cary, NC: SAS Institute, Inc., 1990.
18. Graham P, Jackson R. The analysis of ordinal agreement data: beyond weighted kappa. J Clin Epidemiol 1993;46:1055–1062.

Appendix 1

Let $f(.)$ denote the density function, and let $g(.)$ denote the cumulative distribution function of the standard normal distribution. Let k be the total number of categories, and let c_i denote the cutpoint between the ith and the $(i + 1)$th category for $1 \leq i \leq (k - 1)$. Then, the probability that an individual with true underlying trait x is classified into category i in a single rating, denoted p_i, equals:

$$g\left(\frac{c_i - x}{\sigma_e}\right) \quad \text{for} \quad i = 1,$$

$$g\left(\frac{c_i - x}{\sigma_e}\right) - g\left(\frac{c_{i-1} - x}{\sigma_e}\right), \quad \text{for} \quad 2 \leq i \leq (k - 1)$$

and

$$1 - g\left(\frac{c_i - x}{\sigma_e}\right) \quad \text{for} \quad i = k,$$

assuming normal distribution of measurement error with mean 0 and standard deviation σ_e. Let p_{ij} denote the proportion of individuals classified in categories i and j in the first and second rating, respectively. This proportion is given as:

$$p_{ij} = \int_{-\infty}^{+\infty} [f(x) \times p_i(x) \times p_i(x)]dx,$$

assuming standard normal distribution of the true underlying trait and independence of both ratings conditional on the true value. Expected values of the kappa coefficients can be calculated from the p_{ij} $(1 \leq i \leq k,\ 1 \leq j \leq k)$ using standard equations.[6]