Weighted conditional kappa

TARALD O. KVÁLSETH

University of Minnesota, Minneapolis, Minnesota

In this paper, I propose a simple generalization of the so-called conditional kappa for measuring the extent of agreement between two observers categorizing a set of items. The generalization involves the incorporation of weights representing the relative seriousness of each possible disagreement between the observers. Large-sample statistical inference procedures are outlined. A numerical example is used to demonstrate the detailed computational steps required.

When two or more observers (raters or judges) independently classify n items into the same set of I mutually exclusive categories, it is often of interest to measure the degree of agreement between the observers. If the resulting nominal data for two observers are presented in terms of an $I \times I$ contingency table, the layout of the table for measuring agreement requires the row categories and the column categories to be in the same order, representing a higher level of specificity than is required for measuring association. Thus, the probability (p_{ii}) for the diagonal cell (i, i) represents the probability that both observers classify an item into the same category (i). The off-diagonal probability (p_{ij}) in the cell corresponding to row i and column j represents the probability that the first observer assigns an item to category i and the second observer assigns the same item to category j . The probabilities p_{ii} for $i = 1, 2, \dots, I$ reflect agreements between the two observers, whereas the p_{ij} for $i \neq j$ indicate their disagreements.

A number of different measures of agreement have been proposed (see, e.g., Fleiss, 1981, chap. 13). The kappa (K) and its weighted version (K_w) originally proposed by Cohen (1960, 1968) have received considerable attention and have been generalized in a number of ways (see, e.g., Fleiss, 1981). One such extension of kappa is the conditional agreement measure first proposed by Coleman (1966) and subsequently studied in more detail by Light (1971). While Cohen's kappa provides a composite measure of agreement across all categories for both observers, the Coleman-Light measure of conditional agreement is concerned with the agreement between the observers for only those items that one of the observers, say Observer 1, classified into a particular category i . Specifically, this conditional kappa was defined as (see Light, 1971; Bishop, Fienberg, & Holland, 1975, pp. 397-398; Liebetrau, 1983, pp. 37-38)

$$K_i = 1 - \frac{1 - \frac{p_{ii}}{p_{i+}}}{1 - \frac{p_{i+}}{p_{++}}} \quad (1a)$$

$$= \frac{p_{ii} - \frac{p_{i+}p_{+i}}{p_{++}}}{p_{i+} - \frac{p_{i+}p_{+i}}{p_{++}}}, \quad (1b)$$

where p_{i+} is the marginal total for row i ; that is, p_{i+} is the sum $\sum_{j=1}^I p_{ij}$ from $j=1$ to $j=I$, and p_{+i} corresponds to $p_{+j} = \sum_{i=1}^I p_{ij}$ for $j=i$.

The purpose of this paper is to suggest a generalization of the K_i that incorporates weights reflecting the fact that some disagreements may be more serious than others. Large-sample statistical inferences about the proposed weighted measure are discussed. A numerical example is also provided to exemplify the computational procedures required.

WEIGHTED MEASURE

It is easily seen that the K_i in Equation 1 may be expressed equivalently as

$$K_i = 1 - \frac{\sum_{j \neq i} p_{ij} / \sum_{j \neq i} p_{i+} p_{+j}}{\sum_{j \neq i} p_{i+} p_{+j}}, \quad (2)$$

where $\sum_{j \neq i}$ denotes the summation from $j=1$ to $j=I$, excluding the terms for $j=i$. This expression for K_i is simpler to interpret than that in Equations 1a and 1b. From Equation 2, K_i is seen to be the complement of the relative disagreement. That is, K_i is 1 minus the disagreement between the two observers for those items that the first observer placed in category i , with the disagreement being relative to (normalized by) the chance-expected disagreement given by the denominator in Equation 2. The chance-expected or random disagreement is that expected under the model of independence (i.e., when the row variable is statistically independent of the column variable).

If a set of disagreement weights $\{w_{ij}\}$ can be assigned on some rational grounds, with w_{ij} denoting the weight for cell (i, j), the following weighted conditional agreement measure is proposed:

$$K_{wi} = 1 - \frac{\phi_{1i}}{\phi_{2i}}; \phi_{1i} = \sum_{j \neq i} w_{ij} p_{ij}, \phi_{2i} = \sum_{j \neq i} w_{ij} p_{i+} p_{+j}. \quad (3)$$

The value of K_{wi} equals 1 in the case of perfect (conditional) agreement, equals 0 when the actual disagreement

Requests for reprints should be addressed to T. O. Kvålseth, Department of Mechanical Engineering, University of Minnesota, Minneapolis, MN 55455.

is equal to the chance-expected disagreement (i.e., for the independence situation when $p_{ij} = p_i + p_j$ for each i and j), and is negative if the actual disagreement exceeds that expected by chance.

The measure K_{wi} is a population quantity about which it may be desirable to make statistical inferences, starting with point estimation. Under the multinomial sampling model, where only the total number of classified items (n) is fixed, the maximum likelihood estimator $\hat{K}_{wi}$ of K_{wi} is obtained by simply substituting the observed proportions $\hat{p}_{ij} = n_{ij}/n$ for the population probabilities p_{ij} in Equation 3. Thus, $\hat{K}_{wi}$ is given by

$$\hat{K}_{wi} = 1 - \frac{\hat{\phi}_{1i}}{\hat{\phi}_{2i}}; \hat{\phi}_{1i} = \sum_{j \neq i} w_{ij} \hat{p}_{ij}, \quad \hat{\phi}_{2i} = \sum_{j \neq i} w_{ij} \hat{p}_{+i} \hat{p}_{+j}. \quad (4)$$

By using the so-called multivariate δ method (e.g., Bishop et al., 1975, pp. 492-497), it can be shown that the estimator $\hat{K}_{wi}$ in Equation 4 is asymptotically normally distributed with mean K_{wi} and variance $\sigma_{\infty}^2(\hat{K}_{wi})$ given by

$$\sigma_{\infty}^2(\hat{K}_{wi}) = \frac{1}{n\hat{\phi}_{2i}^2} (\sum_{ij} p_{ij} A_{ij}^2 - B_i^2), \quad (5a)$$

where

$$A_{ij} = 1 - v_{ij} + (1 - K_{wi})(\bar{v}_{i+} + \bar{v}_{+j} - 2), \quad (5b)$$

$$B_i = (1 - K_{wi})(1 - \phi_{2i}) - K_{wi}, \quad (5c)$$

$$\bar{v}_{i+} = \sum_j v_{ij} p_{+j}; \quad \bar{v}_{+j} = \sum_i v_{ij} p_{i+}, \quad (5d)$$

with $v_{ij} = w_{ij}$ for row i (i.e., the table row on which K_{wi} is based), $v_{ij} = 0$ for $j = i$ and $v_{ij} = 0$ for all cells in the other rows. This variance expression can also be derived from that given by Fleiss, Cohen, and Everitt (1969) for the agreement-weighted overall kappa estimator $\hat{K}_w$. Thus, for large n , $\hat{K}_{wi}$ has an approximate normal distribution with mean K_{wi} given by Equation 3 and variance given by Equations 5a-5d. This result may be used for constructing confidence intervals and testing hypotheses regarding the population quantity K_{wi} .

Because the statistic

$$Z_i = \frac{\hat{K}_{wi} - K_{wi}}{\sigma_{\infty}(\hat{K}_{wi})} \quad (6)$$

has an approximate standard normal distribution, it follows that an approximate $(1 - \alpha)$ 100% confidence interval for K_{wi} is given by

$$\hat{K}_{wi} - z_{\alpha/2} \hat{\sigma}_{\infty}(\hat{K}_{wi}) < K_{wi} < \hat{K}_{wi} + z_{\alpha/2} \hat{\sigma}_{\infty}(\hat{K}_{wi}) \quad (7)$$

where $z_{\alpha/2}$ is obtained from standard normal tables (e.g.,

$z_{\alpha/2} = 1.96$ for $\alpha = .05$) and where the hat over the σ_{∞} term signifies that $\hat{\sigma}_{\infty}(\hat{K}_{wi})$ is an estimate for $\sigma_{\infty}(\hat{K}_{wi})$ obtained by substituting (a) the observed probabilities $\hat{p}_{ij}$ for the population probabilities p_{ij} , and (b) the estimated $\hat{K}_{wi}$ for K_{wi} in Equations 5a-5d. For testing some null hypothesis $H_0: K_{wi} = K_0$ (specified), the statistic in Equation 6 needs to be formulated under H_0 as

$$Z_{i0} = \frac{\hat{K}_{wi} - K_0}{\hat{\sigma}_{\infty,0}(\hat{K}_{wi})}, \quad (8)$$

where $\hat{\sigma}_{\infty,0}^2(\hat{K}_{wi})$ denotes the value of $\sigma_{\infty}^2(\hat{K}_{wi})$ in Equations 5a-5d under H_0 , that is, when K_0 is substituted for K_{wi} and when the sample probabilities $\hat{p}_{ij}$ are substituted for the p_{ij} . The test of $H_0: K_{wi} = K_0$ against the two-sided alternative $H_1: K_{wi} \neq K_0$ can then be based on the following decision rule:

$$\text{reject } H_0: K_{wi} = K_0 \text{ if } |Z_{i0}| > z_{\alpha/2}. \quad (9)$$

Under $H_0: K_{wi} = K_0 = 0$, the appropriate form of $\hat{\sigma}_{\infty,0}^2(\hat{K}_{wi})$ to be used in Equation 8 is seen from Equations 5a-5d to be

$$\hat{\sigma}_{\infty,0}^2(\hat{K}_{wi}) = \frac{1}{n\hat{\phi}_{2i}^2} \{ \sum_{ij} \hat{p}_{i+} \hat{p}_{+j} [1 + v_{ij} - (\bar{v}_{i+} + \bar{v}_{+j})]^2 - (1 - \hat{\phi}_{2i})^2 \} \quad (10)$$

where $\hat{\phi}_{2i}$ is defined in Equation 4 and where the p_{i+} and p_{+j} in Equation 5d are replaced by their sample estimates $\hat{p}_{i+}$ and $\hat{p}_{+j}$.

NUMERICAL EXAMPLE

In order to exemplify the use of the developments outlined above, consider the data given in Table 1 and representing the results when a professor and a teaching assistant graded the examinations for the same sample of 100 students according to the letter grades A, B, and C. The entries in this table are the observed proportions $\hat{p}_{ij}$ and the chosen disagreement weights w_{ij} (in parentheses).

A casual look at Table 1 shows that the professor and the teaching assistant agreed completely in their grading for 65% of the students since the sum of the diagonal entries in the table equals .65. It may be of interest to analyze the agreement between the two graders for those students whom the professor gave one of the specific grades A, B, or C. Consider first the estimated agreement $\hat{K}_{w1}$ for the students given the grade A by the professor. From Equation 4 and the data in Table 1, it follows that

$$\hat{\phi}_{11} = .5(.03) + 1(0) = .0150$$

$$\hat{\phi}_{21} = .5(.18) + .51(.18) + 1(.18) = .0801$$

$$\hat{K}_{w1} = 1 - (.0150)/(.0801) = .813.$$

Table 1
Joint Proportions of Grades Assigned by Two Individuals
(and Disagreement Weights) When Grading 100 Students

Professor	Teaching Assistant			Totals
	A	B	C	
A	.15 (0)	.03 (.5)	0 (1)	.18
B	.10 (.5)	.40 (0)	.09 (.5)	.59
C	.05 (1)	.08 (.5)	.10 (0)	.23
Totals	.30	.51	.19	1

Similarly, it is found that

$$\hat{K}_{w2} = .343 \text{ and } \hat{K}_{w3} = .295.$$

By comparison, the unweighted $\hat{K}_i$ from Equation 1 or 2 (substituting the $\hat{p}_{ij}$ for the p_{ij}) is found to be

$$\hat{K}_1 = .762, \hat{K}_2 = .343, \hat{K}_3 = .302.$$

The estimated variance of $\hat{K}_{w1}$ may be computed as follows: Since $v_{12} = .5$, $v_{13} = 1$, and all other $v_{ij} = 0$, it follows from Equation 5d that $\bar{v}_{1+} = .4450$, $\bar{v}_{2+} = \bar{v}_{3+} = 0$ and $\bar{v}_{+1} = 0$, $\bar{v}_{+2} = .0900$ and $\bar{v}_{+3} = .1800$. The $\hat{A}_{ij}$ can be computed from Equation 5b where, for example,

$$\hat{A}_{12} = 1 - .5 + (1 - .8127) (.4450 + .0900 - 2) = .2256.$$

Based on all the $\hat{A}_{ij}$ and the $\hat{B}_1$ estimated from Equation 5c as

$$\hat{B}_1 = (1 - .8127) (1 - .0801) - .8127 = -.6404,$$

it is found from Equation 5a that

$$\hat{\sigma}_{\infty}^2(\hat{K}_{w1}) = \frac{1}{100(.0801)^2} [4.161 - (-.6404)^2] = .0093.$$

Similarly, it is found that

$$\hat{\sigma}_{\infty}^2(\hat{K}_{w2}) = .0079 \text{ and } \hat{\sigma}_{\infty}^2(\hat{K}_{w3}) = .0164.$$

From Equation 7 and the preceding sample estimates of K_{w1} and $\hat{\sigma}_{\infty}^2(\hat{K}_{w1})$, a 95% confidence interval for K_{w1} is [.62, 1.00]. A null hypothesis such as $H_0: K_{w3} = K_0 = 0$ may then be tested against the two-sided alternative $H_1: K_{w3} \neq 0$ by inserting into Equation 8 $\hat{K}_{w3} = .295$, $K_0 = 0$, and $\hat{\sigma}_{\infty,0}(\hat{K}_{w3}) = .114$ as computed from Equation 10, resulting in $Z_{30} = 2.588$. Since $2.588 > z_{\alpha/2} = 1.96$ for $\alpha = .05$, it is concluded on the basis of Equation 9 that this H_0 is rejected at the .05 level of significance.

CONCLUDING COMMENTS

Whereas the Coleman-Light measure K_i of conditional agreement considers all disagreements to be equally serious (i.e., the w_{ij} in Equation 3 are all 1), the new measure K_{wi} allows for the assignment of different degrees of seriousness for different disagreements. Of course, the utility of this weighted measure rests on the condition that the weights can be assigned a priori in a meaningful way. The weights used in the preceding example were taken as

$$w_{ij} = |i - j| / (I - 1)$$

with the number of categories $I = 3$, so that $0 \leq w_{ij} \leq 1$. However, other weighting schemes, with possibly some $w_{ij} > 1$, may also be used.

It may also be noted that a weighted measure of agreement across all row and column categories may be defined by considering the overall weighted disagreement as being the sum of numerators ϕ_{1i} of the individual conditional kappas K_{wi} divided by the sum of their denominators ϕ_{2i} . That is, an overall weighted kappa may be defined as

$$K_w = 1 - \sum_i \phi_{1i} / \sum_i \phi_{2i},$$

where the ϕ_{1i} and ϕ_{2i} are defined in Equation 3. This K_w , which is equivalent to the disagreement-weighted Cohen's kappa (Cohen, 1968), assumes the estimated value $\hat{K}_w = .432$ for the numerical example given above. According to Fleiss (1981, pp. 218, 223), this $\hat{K}_w$ value may be taken to represent "fair" agreement beyond chance, whereas the previously computed conditional $\hat{K}_{w1}$ represents "excellent" agreement and the $\hat{K}_{w2}$ and $\hat{K}_{w3}$ both represent "poor" agreement beyond chance.

REFERENCES

- BISHOP, Y. M. M., FIENBERG, S. E., & HOLLAND, P. W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: MIT Press.
- COHEN, J. (1960). A coefficient of agreement for nominal scales. *Educational & Psychological Measurement*, **20**, 37-46.
- COHEN, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, **70**, 213-220.
- COLEMAN, J. S. (1966). *Measuring concordance in attitude*. Unpublished manuscript, Johns Hopkins University, Department of Social Relations, Baltimore.
- FLEISS, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: Wiley.
- FLEISS, J. L., COHEN, J., & EVERITT, B. S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, **72**, 323-327.
- LIEBETRAU, A. M. (1983). *Measures of association*. (Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-001). Beverly Hills and London: Sage Publications.
- LIGHT, R. J. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin*, **76**, 365-377.