Biom. J. 37 (1995) 7, 811–820 Akademie Verlag

Raking Kappa: Describing Potential Impact of Marginal Distributions on Measures of Agreement

ALAN AGRESTI and ATALANTA GHOSH
Department of Statistics, University of Florida
Gainesville, Florida, U.S.A. 32611

MATILDE BINI
Department of Statistics, University of Firenze
Firenze, Italy

Summary

Several authors have noted the dependence of kappa measures of inter-rater agreement on the marginal distributions of contingency tables displaying the joint ratings. This paper introduces a smoothed version of kappa computed after raking the table to achieve pre-specified marginal distributions. A comparison of kappa with raked kappa for various margins can indicate the extent of the dependence on the margins, and can indicate how much of the lack of agreement is due to marginal heterogeneity.

Key words: Cohen's kappa; Raked contingency tables; Rater agreement; Table standardization; Weighted kappa.

1. Introduction

The need to assess inter-rater agreement occurs frequently in many applications, particularly in biomedical areas such as pathology (classifying tumors according to type or stage of development; see LANDIS and KOCH (1977)), radiology (classifying results of chest x-rays; see FLETCHER and OLDHAM (1949)), and psychiatry (classifying subjects according to type of mental illness; see FLEISS et al. (1965)). Cohen's kappa (see COHEN (1960)) is the most popular summary measure for describing agreement between two raters on a nominal scale.

Recently, several authors have pointed out that a potential difficulty with kappa is that its value depends strongly on the marginal distributions. Two studies having the same intrinsic agreement, in terms of a probabilistic mechanism for classification, may produce quite different values of kappa because of having different prevalences of cases of the various types. Hence, it can be misleading to compare kappa values from studies having substantially different marginal distributions. Literature discussing this potential disadvantage of kappa

includes DARROCH and MCCLOUD (1986), FEINSTEIN and CICCHETTI (1990), KRAUTH (1984), SPITZNAGEL and HELZER (1985), THOMPSON and WALTER (1988), UEBERSAX (1987), and ZWICK (1988).

This paper presents an adjustment of kappa that can help reveal the impact of the marginal distributions on the value of kappa. We define a measure, *raked kappa*, that is the value of Cohen's kappa computed for an adjustment of the sample table that has certain pre-specified marginal distributions. The adjusted (or "raked") table is a standardized form of the sample table having the same association structure but having the targeted marginal distributions. Raked kappa represents the value that kappa would have taken under certain conditions. For instance, one might compute raked kappa to describe the impact of the raters calibrating themselves to achieve marginal homogeneity, with distribution corresponding to cases in a population of interest. We show how to obtain an estimated standard error for raked kappa, and also for raked weighted kappa for ordinal classifications.

2. Raked Kappa

We first introduce notation for kappa. Let m denote the number of categories for a categorical response variable, let N denote the number of subjects rated, and denote two observers by A and B. For the sample, let p_{ij} denote the proportion of cases that observer A classifies in category i and observer B classifies in category j, $i = 1, \ldots, m$, $j = 1, \ldots, m$. In this table, $\sum_i p_{ii}$ is the total proportion of cases of agreement. Perfect agreement in the sample occurs when $\sum_i p_{ii} = 1$.

The sample value of Cohen's kappa is defined as

$$\hat{\kappa} = \frac{\sum p_{ii} - \sum p_{i+}\, p_{+i}}{1 - \sum p_{i+}\, p_{+i}}. \tag{1}$$

The numerator in (1) compares the observed agreement to that expected if the ratings were statistically independent, referred to as "chance" agreement, whereas the denominator is the maximum possible value of the numerator.

Let $\{r_{ij}\}$ denote proportions having the same odds ratios as $\{p_{ij}\}$ for all 2×2 subtables formed using pairs of rows and pairs of columns, yet having marginal distributions given by some pre-specified values $\{r_{i+} = a_i\}$ and $\{r_{+j} = b_j\}$. This table, called the *raked* table, can be calculated using iterative proportional fitting. One re-scales the cell entries to satisfy successively the row and column target values. The process of marginal raking of tables while preserving certain association patterns dates at least to DEMING and STEPHAN (1940). See also MOSTELLER (1968), FREEMAN and KOCH (1976), and AGRESTI (1981). LITTLE and WU (1991) noted that the raked proportions are maximum likelihood estimates of the cell probabilities in the population version of the raked table.

We define *raked kappa*, denoted by $\hat{\kappa}^*$, to be Cohen's kappa computed for the raked proportions $\{r_{ij}\}$. It corresponds to the value of kappa for a table having the target margins, but maintaining the same association structure as the observed table. We denote raked kappa computed for the raking having uniform target values $\{a_i = b_i = 1/m\}$ by $\hat{\kappa}_u^*$. This raking of the table helps to clarify patterns of agreement and disagreement. If the original table displays independence, then this raked table has $\{r_{ij} = 1/m^2\}$ and $\hat{\kappa}_u^* = 0$; if the original table shows perfect agreement, then this raked table has elements $\{r_{ii} = 1/m,\ r_{ij} = 0 \text{ for } i \neq j\}$ and $\hat{\kappa}_u^* = 1$.

3. Example

The two cross classifications in Table 1 are taken from KRAUTH (1984). The row and column marginal distributions in the first one are quite different, caused by observer A frequently choosing category 3 when observer B choses category 2. Thus, the overall agreement is rather poor, with sample kappa value of $\hat{\kappa}_1 = 0.310$ and estimated standard error of 0.040. One might consider how much stronger the agreement might have been if the observers had calibrated their ratings so as to have identical marginal distributions. Indeed, many statisticians believe that kappa only makes sense as a summary measure of agreement when the margins are practically the same.

If the margins of the first cross classification were uniform, but the odds ratios were preserved, we would obtain the raked proportions shown in Table 2. For these, $\hat{\kappa}_u^* = 0.696$, a quite dramatic improvement. If both margins were the same as the observed row margin, then $\hat{\kappa}^* = 0.649$, whereas if both margins were the same as the observed column margin, then $\hat{\kappa}^* = 0.640$. If both margins were the average of these margins, then $\hat{\kappa}^* = 0.632$. The use of raked kappa points out that the lack of good agreement is explained by the severe marginal heterogeneity. Even with marginal homogeneity, however, the agreement would be considered moderate, rather than strong, because of the preserved tendency for A to select category 3 when B choses 2.

The second cross classification in Table 1 has somewhat stronger sample agreement, with $\hat{\kappa}_2 = 0.429$, based on an estimated standard error of 0.054. In this case, the margins are similar. Thus, sample kappa is similar to versions of raked kappa for rakings in which both margins equal the observed row margin, both margins equal the observed column margin, or both margins are the average of the observed row and column margins. The sample odds ratios within the body of the second cross classification are weak, and the agreement is relatively weak regardless of the choice of marginal distributions. Table 2 shows the raking for uniform marginal proportions. The values of raked kappa are shown in Table 3, for various target margins for the cross classifications of Table 1.

4. Inference for Raked Kappa

For inference, we assume multinomial sampling for the original cell counts. Let p and r denote vectors containing the elements of $\{p_{ij}\}$ and $\{r_{ij}\}$, respectively. Let D and D_r denote $m^2 \times m^2$ diagonal matrices with the respective elements of p and r on the main diagonal. Let K denote an $m^2 \times (m-1)^2$ marix that describes the log odds ratios preserved in the raked table. That is, the equation $K' \log p = K' \log r$ represents the preserved contrasts

$$\log p_{ij} - \log p_{im} - \log p_{mj} + \log p_{mm} = \log r_{ij} - \log r_{im} - \log r_{mj} + \log r_{mm},$$
$$i = 1, \ldots, m-1, j = 1, \ldots, m-1.$$

Following FREEMAN and KOCH (1976), the asymptotic covariance matrix for r equals

$$V_r = K[K' D_r^{-1} K]^{-1} K' D^{-1} K [K' D_r^{-1} K]^{-1} K'/N. \tag{2}$$

Let d denote the $m^2 \times 1$ vector that contains partial derivatives of kappa with respect to the cell proportions, evaluated at the raked sample proportions. Expressing $\hat{\kappa}^* = \hat{v}/\hat{\delta}$ and letting $I_{(i=j)}$ denote the indicator of whether $i=j$, the element of d for the cell in row i and column j equals

$$d_{ij} = [\hat{\delta} I_{(i=j)} + (\hat{v} - \hat{\delta})(r_{j+} + r_{+i})]/\hat{\delta}^2.$$

By the delta method (e.g., AGRESTI (1990, Sec. 12.1)), the estimated asymptotic variance of raked kappa is $d' V_r d$.

To illustrate, we consider raked kappa for the uniform rakings of the two cross classifications in Table 1. For the first table, $\hat{\kappa}_u^* = 0.696$ has estimated standard error 0.085; for the second table, $\hat{\kappa}_u^* = 0.356$ has estimated standard error 0.073. An approximate 95% confidence interval for the difference in population raked kappa values equals

$$(0.696 - 0.356) \pm 1.96[(0.085)^2 + (0.073)^2]^{1/2}, \quad \text{or} \quad 0.340 \pm 0.220.$$

Though the sample kappa values indicated better agreement in the second table, the uniform raked value is significantly higher for the first table. A cursory inspection of the raked proportions in Table 2 shows the stronger agreement in the first raked table. If the population contained equal numbers of cases of the three types, and if the raters calibrated themselves to achieve marginal homogeneity while maintaining their association structure, the raters from the first table might show stronger agreement than those from the second. Of course, calibrating while maintaining associations is a big "IF," since the marginal distributions are so different in the first table. Such a comparison is more tenuous than comparing raked values for two cross classifications in which each table had

approximate marginal homogeneity, but in which the margins were somewhat different in the two tables. In such a case, one might rake to common margins for the two tables (perhaps representing a known distribution of cases in a population of interest) in order to better compare the kappa agreement values.

Table 3 shows values of the estimated asymptotic standard error (ASE) for raked kappa applied to the two cross classifications in Table 1, for various target margins. These are provided for illustrative purposes, but one should note that this ASE applies when the targets for the raking were pre-specified.

Table 1

Sample tables for rater agreement

	Rater B				Rater B			
Rater A	1	2	3	Total	1	2	3	Total
1	31	1	1	33	106	10	4	120
2	1	30	1	32	22	28	10	60
3	1	97	37	135	2	12	6	20
Total	33	128	39	200	130	50	20	200

Source: KRAUTH (1984)

Table 2

Uniform raked proportions for Table 1

	Rater B				Rater B			
Rater A	1	2	3	Total	1	2	3	Total
1	.306	.003	.025	.333	.253	.041	.039	.333
2	.025	.246	.063	.333	.066	.145	.122	.333
3	.003	.084	.246	.333	.014	.147	.172	.333
Total	.333	.333	.333	1.000	.333	.333	.333	1.000

Table 3

Raked kappa for various target margins

Row	Column	Kappa-1	ASE-1	Kappa-2	ASE-2
Observed	Observed	0.310	0.019	0.429	0.053
Uniform	Uniform	0.696	0.085	0.356	0.073
Average	Average	0.632	0.112	0.438	0.054
Row	Row	0.649	0.093	0.439	0.055
Column	Column	0.640	0.100	0.437	0.054

5. Raking Model-Fitted Proportions

The raking process preserves empty cells. If a cell count equals zero, then the raked entry in that cell equals zero. Empty cells can potentially cause problems with the raking process, in the sense that raked estimates may not exist having the target margins. Because of this, one might add a small constant (e.g., 10^{-6}) to empty cells before raking a table. When the empty cells fall off the main diagonal, as is usually the case, the effect of this on Cohen's kappa is minor since it depends only on the total count on the main diagonal and on the marginal counts of the original table. However, different small constants can result in quite different raked tables, and one should use the raking process on sparse tables only with caution.

When a table contains several empty cells, a less *ad hoc* approach fits a reasonable model to the data before computing the raked table. This reflects the fact that the empty cells are sampling zeroes rather than structural zeroes, and that one expects the true cell probabilities to be non-zero. If one fits a quasi-independence or quasi-symmetric type of model that preserves the marginal totals and the main diagonal counts, then kappa is the same for the model fit as it is for the sample data; on the other hand, the table of fitted values and the corresponding raked table will not contain empty cells, and would often provide better estimates of corresponding population tables.

Let V denote the estimated asymptotic covariance matrix for the model-fitted proportions. Then, FREEMAN and KOCH (1976) showed (though the expression is printed incorrectly in their article) that the estimated covariance matrix for the proportions based on raking the model-fitted values equals

$$V_r = K[K'D_r^{-1}K]^{-1}K'D^{-1}VD^{-1}K[K'D_r^{-1}K]^{-1}K'. \tag{3}$$

Thus, the asymptotic variance of raked kappa for the raking of the model-fit equals $d'V_r d$.

6. Raking Weighted Kappa

Cohen's kappa is designed for nominal classifications. For ordinal classifications, *weighted kappa* (SPITZER et al., 1967) used weights $\{w_{ij}\}$ to allow disagreements to be more serious as the distance between the two categories selected by the two raters increases. For $0 \leqq w_{ij} \leqq 1$, with all $w_{ii} = 1$, weighted kappa is given by

$$\hat{\kappa}_w = \frac{\sum w_{ij}p_{ij} - \sum w_{ij}p_{i+}p_{+j}}{1 - \sum w_{ij}p_{i+}p_{+j}}. \tag{4}$$

Using $\{r_{ij}\}$ in place of $\{p_{ij}\}$ in this definition yields weighted kappa for raked proportions. Expressing raked weighted kappa as $\hat{\kappa}_w^* = \hat{v}_w/\hat{\delta}_w$, its estimated

asymptotic variance is $d_w' V_r d_w$, where the elements of the partial derivative vector equal

$$d_{ij,w} = \left[w_{ij}\hat{\delta}_w + (\hat{v}_w - \hat{\delta}_w)\left(\sum_a w_{aj} r_{a+} + \sum_b w_{ib} r_{+b}\right)\right] / \hat{\delta}^2 .$$

In inter-rater agreement studies with ordinal response scales having several categories, it is not uncommon for several of the off-diagonal counts to equal zero, particularly for cells that are far from the main diagonal. In such cases, it is sensible to compute raked weighted kappa for model-fitted values.

We illustrate raked weighted kappa using Table 4, taken from a study (CONFORTINI et al., 1993) on cervical cancer from the Center for Cancer Study and Prevention, in Florence, Italy. Table 4 shows the results of analysis on a set of 100 slides carried out by a cytologist at the laboratory using seven ordered categories, compared with an 'expert' diagnosis. For these data, the sample value of Cohen's kappa equals $\hat{\kappa} = 0.497$. Since the classification scale is ordinal, we might alternatively use weighted kappa. Using weights $w_{ij} = 1 - (i-j)^2/(m-1)^2$, we obtain $\hat{\kappa}_w = 0.600$.

In this study, one might compute raked versions of kappa after raking the rater's margin to match the observed one for the 'expert' rater. This would enable us to describe the potential agreement if the rater shared the expert's distribution of classifications. The large number of empty cells implies, however, that the raked proportions do not exist. For instance, the sixth row contains only one non-empty cell. For a raked version of this row to have a row total of 9, all 9 observations must fall in the cell in column 6, because other cells in the row are empty. But then column 6 would necessarily contain more than 9 observations, which is a contradiction.

To remedy this problem, one could collapse the table to fewer rows and columns, or add small constants to the empty cells. Instead, we smoothed the counts by fitting a model. Not surprisingly, the independence model fits poorly, having a likelihood-ratio goodness-of-fit statistic of $G^2 = 161.1$, based on $df = 36$. The quasi-symmetry model, which adds symmetric association terms to the independence model, is a useful model for agreement data that usually gives a much better fit. It is implied by a Rasch model structure whereby logits for response outcomes depend additively on effects of raters and subjects rated (TJUR, 1982, DARROCH and MCCLOUD, 1986). However, its additional sufficient statistics are the totals $p_{ij} + p_{ji}$, many of which equal zero for these data. Thus, this model smooths only some of the empty cells. It gives a good fit to the set of cells for which $p_{ij} + p_{ji} > 0$, having $G^2 = 6.3$, based on $df = 6$.

Table 4 contains the raking of the fitted values for the quasi-symmetry model. The raking produces a clear ridge along the main diagonal, with some of the serious disagreements in the sample table eliminated (such as when the expert choses category 5 and the rater choses category 2). The raked table has both

kappa and weighted kappa equal to 0.748, compared to the respective values of 0.497 and 0.600 for the unraked table. An alternative set of weights $w_{ij} = 1 - |i-j|/|m-1|$ yields $\hat{\kappa}_w = 0.598$ and $\hat{\kappa}_w^* = 0.785$. For kappa and both versions of weighted kappa, there is potential for moderate improvement in agreement under marginal homogeneity. The increases in kappa and weighted kappa values for the raked table are perhaps surprising, considering that the original margins are not dramatically different. This illustrates that kappa or weighted kappa can potentially be quite sensitive to fairly minor changes in marginal distributions.

Table 4

Data with raking of quasi-symmetry model fit to match margin for expert rater

	Expert Rater							
Rater A	1	2	3	4	5	6	7	Total
1	12	5	0	0	0	0	0	17
	(12.1)	(3.5)	(0.0)	(0.0)	(0.0)	(0.1)	(1.3)	(17)
2	2	16	4	1	6	1	1	31
	(3.6)	(17.0)	(3.1)	(0.0)	(0.0)	(0.1)	(1.3)	(25)
3	0	2	7	3	0	0	1	13
	(0.0)	(3.2)	(7.3)	(0.0)	(0.0)	(0.1)	(0.4)	(11)
4	0	0	0	2	3	0	0	5
	(0.0)	(0.1)	(0.2)	(5.7)	(0.0)	(0.0)	(0.0)	(6)
5	0	0	0	0	16	5	0	21
	(0.0)	(0.0)	(0.0)	(0.3)	(24.7)	(0.0)	(0.0)	(25)
6	0	0	0	0	0	1	0	1
	(0.0)	(0.0)	(0.0)	(0.0)	(0.3)	(8.7)	(0.0)	(9)
7	3	2	0	0	0	2	5	12
	(1.3)	(1.2)	(0.4)	(0.0)	(0.0)	(0.1)	(4.1)	(7)
Total	17	25	11	6	25	9	7	100

Source: CONFORTINI et al., (1993).

7. Comments

As shown above for the first cross classification in Table 1, one use of raked kappa is to indicate how strong agreement could have been if the raters had the same targets, corresponding to a lack of inter-observer bias. In particular, one could compute raked kappa for various targets satisfying marginal homogeneity. This helps to reveal how much of the lack of agreement may be due to a lack of calibration of the two raters.

Even if raters are calibrated to achieve marginal homogeneity, however, values of kappa from different studies may be incomparable because of markedly different prevalence of the various types. To illustrate, consider Table 5. Each cross classification exhibits marginal homogeneity, but the incidence in the first

category is much less in the first table than the second (.05 vs. .40). The sample kappa values of 0.244 and 0.513 suggest that agreement is better for the raters of the second set of subjects. In fact, the odds ratios are the same (10) in the two tables, so the raked kappa values are equal for any target margins that are identical for the two studies. For the 2×2 case with each margin equal to $(\pi, 1-\pi)$, kappa is related to the odds ratio $\hat{\theta}$ by $\hat{\theta} - 1 = \hat{\kappa}/[\pi(1-\pi)(1-\hat{\kappa})^2]$. For instance, for the uniform margins (i.e., $\pi = 0.5$) with $\hat{\theta} = 10$, $\hat{\kappa} = 0.520$. Thus, a second use of raked kappa is to compare agreements from two studies having approximate marginal homogeneity within studies but severe marginal heterogeneity between studies.

Table 5

Tables having identical values of raked kappa

Table A		Table B	
141	359	2830	1170
359	9149	1170	4830

Finally, though raked kappa is useful for the purposes just described, we do not believe that it is a panacea. In practice, a calibration process may result in a change in association. For instance, in the 2×2 case, consider the latent class model in which there is independence between ratings, given the 'true' response category; that is, the observed table is a mixture of two 2×2 tables. For given sensitivity and specificity, the odds ratio between the ratings in the observed 2×2 table changes as the distribution of cases of the two types changes. Generally, how the odds ratio behaves depends on a variety of factors that could all potentially themselves change as the distribution of cases change, including the conditional distribution for the response outcome given the true outcome, and misclassification probabilities. In $m \times m$ tables or in their multi-rater generalizations, we believe it is more informative to utilize modeling strategies that investigate the structure of the agreement and that analyze residual departures from that structure. AGRESTI (1992) provided a recent survey of modeling strategies.

An S-Plus function (BECKER et al., 1988) for computing kappa, raked kappa for supplied target margins, the weighted kappa analogs, and their estimated standard errors is available free from the authors by e-mail or upon receipt of a formatted $3^1/_2$ inch floppy diskette.

Acknowledgements

This research was partially supported by a grant from the National Institutes of Health.

References

AGRESTI, A., 1981: A hierarchical system of interaction measures for multidimensional contingency tables. J. Roy. Statist. Soc. **B 43,** 293–301.

AGRESTI, A., 1990: *Categorical Data Analysis.* New York: Wiley.

AGRESTI, A., 1992: Modeling patterns of agreement and disagreement. Statist. Meth. Med. Res. **1,** 201–218.

BECKER, R. A., CHAMBERS, J. M. and WILKS, A. R., 1988: *The New S Language: A Programming Environment for Data Analysis and Graphics.* Pacific Grove, CA: Wadsworth & Brooks Cole.

COHEN, J., 1960: A coefficient of agreement for nominal scales. Educ. Psych. Meas. **20,** 213–220.

CONFORTINI, M., BIGGERI, A., CARRIAGI, M. P., CAROZZI, F. M., MINUTI, P. A., RUSSO, A., PALLI, D., 1993: Intralaboratory reproducibility in cervical cytology. Acta Cytologica **37,** 49–54.

DARROCH, J. N. and MCCLOUD, P. I., 1986: Category distinguishability and observer agreement. Australian J. Statist. **28,** 371–388.

DEMING, W. E. and STEPHAN, F. F., 1940: On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. Ann. Math. Statist. **11,** 427–444.

FEINSTEIN, A. R. and CICCHETTI, D. V., 1990: High agreement but low kappa: I. The problems of two paradoxes. J. Clin. Epidemiology **43,** 543–549.

FLEISS, J. L., SPITZER, R. and BURDOCK, E., 1965: Estimating accuracy of judgment using recorded interviews. Arch. Gen. Psychiatry **12,** 562–567.

FLETCHER, C. M. and OLDHAM, P. D., 1949: Problem of consistent radiological diagnosis in coalminers' pneumonoconiosis. Brit. J. Industr. Med. **6,** 168–183.

FREEMAN, D. H., Jr. and KOCH, G. G., 1976: An asymptotic covariance structure for testing hypotheses on raked contingency tables from complex sample surveys. Proc. Amer. Statist. Assoc., Social Statist. Sec., 330–335.

KRAUTH, J., 1984: A modification of kappa for interobserver bias. Biom. J. **26,** 435–445.

LANDIS, J. R. and KOCH, G. G., 1977: An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics **33,** 363–374.

LITTLE, R. J. A. and WU, M-M., 1991: Models for contingency tables with known margins when target and sampled populations differ. J. Amer. Statist. Assoc. **86,** 87–95.

MOSTELLER, F., 1968: Association and estimation in contingency tables. J. Amer. Statist. Assoc. **63,** 1–28.

SPITZER, R. L., COHEN, J., FLEISS, J. L. and ENDICOTT, J., 1967: Quantification of agreement in psychiatric diagnosis. Arch. Gen. Psychiatry **17,** 83–87.

SPITZNAGEL, E. L. and HELZER, J. E., 1985: A proposed solution to the base rate problem in the kappa statistic. Arch. Gen. Psychiatry **42,** 725–728.

THOMPSON, W. D. and WALTER, S. D., 1988: A reappraisal of the kappa coefficient. J. Clin. Epidemiology **41,** 949–958.

TJUR, T., 1982: A connection between Rasch's item analysis model and a multiplicative Poisson model. Scand. J. Statist. **9,** 23–30.

UEBERSAX, J. S., 1987: Diversity of decision-making models and the measurement of interrater agreement. Psychol. Bull. **101,** 140–146.

ZWICK, R., 1988: Another look at inter-rater agreement. Psychol. Bull. **103,** 374–378.

Received, April, 1994
Revised, Dec., 1994
Accepted, Jan., 1995

ALAN AGRESTI
Department of Statistics, University of Florida
204 Griffin-Floyd Hall
Gainesville, Florida 32611-8545
U.S.A
e-mail AA@STAT.UFL.EDU