Statistics & Probability Letters 7 (1989) 271–273
North-Holland

February 1989

AN AGREEMENT MODEL WITH KAPPA AS PARAMETER

Alan AGRESTI

Department of Statistics, University of Florida, Gainesville, FL 32611, USA

Received March 1988
Revised May 1988

Abstract Cohen's (1960) kappa summarizes agreement between classifications of two fixed raters on a categorical scale. This note exhibits a simple quasi-symmetry model for which kappa contains all relevant information about the structure of agreement and disagreement

Keywords categorical data, correlation model, marginal homogeneity, quasi independence, quasi symmetry, symmetry

1. Introduction

Suppose two observers rate subjects on a categorical scale. The raters and the categorical scale are fixed. For a population of subjects, let π_{ij} denote the proportion classified into category i by rater A and category j by rater B. Denote the number of categories by r.

The most popular measure for describing degree of agreement between two fixed raters is Cohen's (1960) *kappa*, defined by

$$\kappa = \frac{\sum \pi_{ii} - \sum \pi_{i+}\pi_{+i}}{1 - \sum \pi_{i+}\pi_{+i}} \tag{1.1}$$

where $\pi_{i+} = \sum_j \pi_{ij}$ and $\pi_{+i} = \sum_j \pi_{ji}$. The numerator of kappa is the difference between the actual probability of agreement and the probability of agreement corresponding to statistical independence of the observers' ratings; the denominator gives the maximum possible value for this difference. Both numerator and denominator are based on the given marginal distributions $\{\pi_{i+}\}$ and $\{\pi_{+i}\}$ for the two observers.

Hanley (1987) remarked on kappa's popularity, noting that more than 800 articles in the social sciences cited Cohen's paper between 1960 and 1985. Fleiss (1981) surveyed the literature on kappa and many of its generalizations. Because of kappa's dependence on marginal distributions and the loss of information caused by reducing $\{\pi_{ij}\}$ to a single number, Tanner and Young (1985), Darroch and McCloud (1986), and others have criticized kappa and proposed instead *modeling* the structure of agreement.

Our intent here is to note a case in which reconciliation is possible between the approaches of model-building and summarization using kappa. We exhibit a simple model for cell probabilities in which kappa is the parameter. Thus, taking the model-building approach does not preclude using kappa to describe the structure of agreement and disagreement. The model described is quite restrictive, however, indicating that kappa is unlikely to be sufficient for summarizing agreement for most data sets.

2. Kappa and quasi symmetry

In a population of S subjects, let ρ_{abc} denote the probability that subject a is classified by rater b into category c. If one assumes (1) that the two raters' classifications are made blindly, in the sense that $\pi_{ij} = S^{-1} \sum_a \rho_{ai1}\rho_{aj2}$, and (2) that $\{\rho_{abc}\}$ satisfies the loglinear model of no three-factor interaction, then Darroch and McCloud (1986) showed that $\{\pi_{ij}\}$ satisfies the quasi-symmetry

model. Thus, when these assumptions seem reasonable, models for agreement should be special cases of the quasisymmetry model,

$$\pi_{ij} = a_i b_j c_{ij} \quad \text{where } c_{ij} = c_{ji} \text{ for all } i \text{ and } j. \tag{2.1}$$

Substituting (2.1) into (1.1), we see that when quasi symmetry holds, kappa depends on both main-effect and association parameters. Considerable simplification occurs when the main-effect parameters are the marginal probabilities; that is, when $a_i = \pi_{i+}$ and $b_j = \pi_{+j}$ for all i and j. Model (2.1) then has the same form for main-effect parameters as the correlation model and the correspondence analysis model discussed by Goodman (1986). In this case, $\{c_{ii}\}$ are determined by $\{c_{ij}$ for $i \neq j\}$ from the constraints $\sum_i \pi_{i+} c_{ij} = 1$ or $\sum_j \pi_{+j} c_{ij} = 1$, and kappa can be expressed as

$$\kappa = \frac{\sum \pi_{i+}\pi_{+i}(c_{ii} - 1)}{1 - \sum \pi_{i+}\pi_{+i}} = \frac{-\sum_{i \neq j} \pi_{i+}\pi_{+j}(c_{ij} - 1)}{\sum_{i \neq j} \pi_{i+}\pi_{+j}}.$$

These formulas highlight kappa's dependence on the marginal distributions as well as the structure of agreement or disagreement. Kappa is independent of the marginal distributions when c_{ij} is constant for $i \neq j$. If $c_{ij} = c$ for $i \neq j$, then $\kappa = 1 - c$. From the constraints on $\{c_{ij}\}$, it follows that the $\{\pi_{ij}\}$ necessarily exhibit marginal homogeneity in this case, with $\pi_{i+} = \pi_{+i} = \pi_i$, say, for $i = 1, \ldots, r$. Thus, this quasi-symmetry model is

$$\pi_{ij} = \pi_i \pi_j (1 - \kappa) \quad \text{for } i \neq j$$

and

$$\pi_{ii} = \pi_i^2 + \kappa \pi_i (1 - \pi_i) \tag{2.2}$$

a model of "uniform disagreement." For model (2.2), the $\{\pi_{ij}\}$ are determined by κ and the marginal probabilities. When the model holds, $\kappa = 0$ is equivalent to statistical independence of the ratings and $\kappa = 1$ is equivalent to perfect agreement ($\Sigma \pi_{ii} = 1$). The joint distribution is a weighted average of a distribution satisfying independence and a distribution having perfect agreement, with weights $(1 - \kappa)$ and κ.

Tallis (1962) proposed a multivariate multinomial distribution, of which (2.2) is a special case. Tallis studied square tables with integer scores assigned to the categories, in which case κ is also the correlation for distribution (2.2). Kraemer (1979) obtained (2.2) for a model for 2×2 tables. Model (2.2) has extremely simple structure, being a special case of both the symmetry and quasi-independence models.

3. Fitting the model

For a sample of n subjects rated by the observers, let $\{p_{ij}\}$ denote sample proportion estimates of $\{\pi_{ij}\}$. Assuming a multinomial distribution for cell counts $\{np_{ij}\}$, the likelihood equations for (2.2) are

$$1 - \sum p_{ii} - \sum \frac{p_{ii}(1 - \hat{\pi}_i)}{\hat{\pi}_i + \hat{\kappa}/(1 - \hat{\kappa})} = 0,$$

$$\hat{\lambda}\hat{\pi}_i\hat{\pi}_{ii} - (p_{i+} + p_{+i})\hat{\pi}_{ii} + \hat{\kappa} p_{ii} \hat{\pi}_i = 0, \quad i = 1, \ldots, r,$$

where

$$\hat{\lambda} = \left[(1 - \hat{\kappa})\left(\sum \hat{\pi}_i (p_{i+} + p_{+i})\right) + \hat{\kappa}\left(2 - \sum p_{ii}\right)\right] \Big/ \left(\sum \hat{\pi}_{ii}\right)$$

and

$$\hat{\pi}_{ii} = \hat{\pi}_i^2 + \hat{\pi}_i (1 - \hat{\pi}_i)\hat{\kappa}.$$

These equations can be solved using iterative methods. In examples we have considered, good initial approximations to the ML estimates are obtained by taking $\hat{\pi}_i = (p_{i+} + p_{+i})/2$ and solving the first equation for $\hat{\kappa}$. Since we are using a form of (2.1) for which the $\{c_{ij}$ for $i \neq j\}$ determine the $\{c_{ii}\}$, this model cannot have an independent set of parameters for the main diagonal, and it will not give a perfect fit for cells on that diagonal.

The degrees of freedom for goodness-of-fit tests of the model are $df = r^2 - r - 1$. When the model holds, the estimate $\hat{\kappa}$ for the model estimates the same characteristic (1.1) as does the usual sample kappa. Model-based $\hat{\kappa}$ is a better estimator, since the $\{\hat{\pi}_{ij}\}$ are better than $\{p_{ij}\}$ as estimators of $\{\pi_{ij}\}$. When the model does not hold but fits fairly well, it follows (e.g., from Bishop et al., 1975, Section 9.2) that model-based $\hat{\kappa}$ is still a better estimator unless the sample size is quite

Table 1
Student teachers rated by supervisors

Rating by Supervisor 1	Rating by Supervisor 2			Total
	Authoritarian	Democratic	Permissive	
Authoritarian	17	4	8	29
	(20 3)	(4 6)	(6.5)	(31.4)
Democratic	5	12	0	17
	(4 6)	(8 7)	(3 5)	(16 8)
Permissive	10	3	13	26
	(6.5)	(3 5)	(13.7)	(23.7)
Total	32	19	21	72
	(31 4)	(16 8)	(23 7)	

Note Parenthesized values are estimated expected frequencies for model 2 2

large. When the model does not hold, $\hat{\kappa}$ is a consistent estimator of the value of κ in model (2.2) for which the best fit occurs for the population table. This κ parameter describes agreement in a smoothed table in which raters are forced to have the same distributions for their ratings.

To illustrate the model, we refer to Table 1, used by Bishop et al. (1975, p. 397) to illustrate kappa. Two supervisors rated the classroom styles of a sample of 72 teachers as authoritarian, democratic, or permissive. The sample kappa of 0.36 indicates fairly weak agreement. The ML estimates for (2.2) are $\hat{\kappa} = 0.37$, $\hat{\pi}_1 = 0.44$, $\hat{\pi}_2 = 0.23$, and $\hat{\pi}_3 = 0.33$, with estimated expected frequencies as given in Table 1. The Pearson statistic equals 7.7, based on df = 5. The model does not provide a tight fit to Table 1, but $\hat{\kappa}$ gives a rough description for its structure: Any particular disagreement is about 63% as likely to occur as if the ratings were statistically independent (i.e., $\hat{\pi}_{ij}/\hat{\pi}_i\hat{\pi}_j = 1 - \hat{\kappa} = 0.63$ for all $i \neq j$), whereas the probability the raters agree that a teacher is in category i exceeds the value corresponding to independence by about $0.37\hat{\pi}_i(1 - \hat{\pi}_i)$. This is a more detailed interpretation than the usual one for kappa that the difference between observed and chance agreement is 0.36 times its maximum possible value.

4. Summary

The *symmetry plus quasi-independence model* is $\pi_{ij} = a_i a_j c_{ij}$ with c_{ij} constant for $i \neq j$. We have seen that kappa is a natural parameter for describing agreement when this model holds with $\{a_i = \pi_i\}$. When model (2.2) fits reasonably well, it provides an additional interpretation for Cohen's kappa in terms of agreement and disagreement structure. We do not believe this model has very broad scope, however, since it is so simplistic. When it fits poorly, various agreement patterns can have the same kappa value, and kappa alone is unlikely to be adequate for describing agreement. One can then employ a more general model-building approach, for instance one using category-specific kappas, or one using alternative parameters in a loglinear model, such as presented by Tanner and Young (1985) and by Agresti (1988). The result of the model-building process may be that a single number can be used (as in model (2.2)), or several indices may be needed to describe the agreement adequately.

Acknowledgement

I am thankful to Dr. Phyllis Gimotty for pointing out the Tallis reference and for other helpful comments.

References

Agresti, A. (1988), A model for agreement between ratings on an ordinal scale, *Biometrics* **44**, 539–548.

Bishop, Y.V.V., S.E. Fienberg and P.W. Holland (1975), *Discrete Multivariate Analysis* (MIT Press, Cambridge, MA)

Cohen, J. (1960), A coefficient of agreement for nominal scales, *Educational and Psychological Measurement* **20**, 37–46.

Darroch, J.N. and P.I. McCloud (1986), Category distinguishability and observer agreement, *Austral J Statist* **28**, 371–388

Fleiss, J. (1981), *Statistical Methods for Rates and Proportions*, 2nd ed. (Wiley, New York).

Goodman, L.A. (1986), Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables, *International Statistical Review* **54**, 243–309.

Hanley, J.A (1987), Standard error of the kappa statistic, *Psychological Bulletin* **102**, 315–321.

Kraemer, H.C. (1979) Ramifications of a population model for κ as a coefficient of reliability, *Psychometrika* **44**, 461–472

Tallis, G.M. (1962), The use of a generalized multinomial distribution in the estimation of correlation in discrete data, *J Royal Statist. Soc B* **24**, 530–534

Tanner, M.A. and M.A. Young (1985), Modeling agreement among raters, *J Amer Statist Assoc.* **80**, 175–180