Journal of the Korean
Statistical Society
Vol. 21, No. 2, 1992

An Agreement Model Having Structure of Symmetry Plus Main-Diagonal Equiprobability

Sadao Tomizawa[1]

ABSTRACT

A model is proposed for describing agreement between two raters on a nominal categorical scale. This model has a structure of symmetry plus main-diagonal equiprobability, which is a special case of the quasi-symmetry model. An example is given.

1. INTRODUCTION

Suppose that two raters, A and B, separately classify each subject on an r response categories, and let π_{ij} denote the probability of rating i by the first rater and rating j by the second rater.

Cohen(1960) proposed the measure *kappa* describing degree of agreement between two raters defined as

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

[1] Department of Information Sciences, Faculty of Science & Technology, Science University of Tokyo, Noda City, Chiba, 278, Japan.

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 chance agreement that occurs if the two raters' rating are statistically independent, and the denominator is the maximum possible value for this difference; also see, e.g., Bishop et al.(1975, p395), Darroch and McClould(1986), and Agresti(1989). The kappa equals one when there is perfect agreement ($\sum \pi_{ii} = 1$) and zero when there is statistical independence.

Because as noted in, e.g., Agresti(1989), the kappa depends on marginal distributions and the loss of information is caused by reducing $\{\pi_{ij}\}$ to a single number, Tanner and Young(1985a), Darroch and McClould(1986), and Agresti(1988, 1989) proposed modeling the structure of agreement between the raters, rather than describing it with a single summary measure. Tanner and Young(1985a) and Agresti(1988) considered models having the structure of baseline association (null association and uniform association, respectively) plus a main-diagonal parameter. Agresti(1989) considered a model having the structure of symmetry plus quasi-independence with kappa as parameter. Tanner and Young(1985b) also gave some models for ordinal scale disagreement.

For classification of subject h by rater a, let ρ_{hat} denote the probability that the rating is in category t. In a population of S subjects , if one assumes (i) that classifications are made independently in the sense that $\pi_{ij} = S^{-1} \sum_h \rho_{h1i}\rho_{h2j}$, and (ii) that $\{\rho_{hat}\}$ satisfies the condition of no three-factor interaction, then Darroch and McClould(1986) showed that $\{\pi_{ij}\}$ satisfies the quasi-symmetry model. In this sense, reasonable models for agreement should be special case of the quasi-symmetry model,

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

(Also see Agresti 1988 and 1989.) Since Tanner and Young's(1985a) model and Agresti's(1988, 1989) models described above are special cases of the quasi-symmetry model, in this sense, these would be reasonable models for agreement.

The purpose of this note is to propose the other model for describing agreement. The model is also a special case of the quasi-symmetry model.

2. AN AGREEMENT MODEL

For the setting described in Section 1, consider now an agreement model defined as

$$\pi_{ij} = \begin{cases} \phi_{ij} & \text{for } i \neq j, \\ \delta & \text{for } i = j, \end{cases} \tag{2.1}$$

where $\phi_{ij} = \phi_{ji}$. This is a model having the structure of symmetry plus main-diagonal equiprobability(SDEP model). The generalization of this model in which the δ is replaced by δ_i is the usual symmetry model. The SDEP model is also a special case of the quasi-symmetry model. Also the SDEP model is equivalent to the symmetry model plus the condition that

$$\frac{\pi_{ji}}{\pi_{ii}} = \frac{\pi_{ij}}{\pi_{jj}} \qquad \text{for } i \neq j.$$

In addition, since the SDEP model implies the marginal homogeneity, under the SDEP model the probability that rater A assigns a subject to category i is equal to the probability that rater B assigns the subject to category i.

By the way, the parameter δ in the SDEP model may be replaced by kappa parameter and marginal probabilities as follows:

$$\pi_{ij} = \begin{cases} \phi_{ij} & \text{for } i \neq j \\ (\kappa + (1-\kappa)\sum_{t=1}^{r} \pi_t^2)/r & \text{for } i = j \end{cases} \tag{2.2}$$

where $\phi_{ij} = \phi_{ji}$ and $\pi_i = \pi_{i+} = \pi_{+i}$. When the SDEP model holds, κ=0 is equivalent to the condition that the probability that the two raters agree ($\sum \pi_{ii}$) is equal to the probability of chance agreement that occurs if the ratings are statistically independent ($\sum \pi_i^2$), and κ=1 is equivalent to perfect agreement. In particular, when r=3, model (2.2) with κ=0 depends only on the marginal probabilities, namely it can be expressed as

$$\pi_{ij} = \begin{cases} (1-\delta^*)/2 - \pi_k & \text{for } i \neq j, i \neq k, j \neq k,\ i = 1,2,3; j = 1,2,3; k = 1,2,3 \\ \delta^* & \text{for } i = j,\ i = 1,2,3 \end{cases} \tag{2.3}$$

where $\delta^* = (\pi_1^2 + \pi_2^2 + \pi_3^2)/3$ and $\pi_i = \pi_{i+} = \pi_{+i}$. It is easily seen that κ=0 under this model. Namely this model has the structure of symmery plus main-diagonal equiprobability plus κ=0.

3. FITTING THE MODEL

For a sample of n subjects classified by the raters, let $\{p_{ij}\}$ denote sample proportional estimates of $\{\pi_{ij}\}$. Assuming a multinomial distribution for cell counts $\{np_{ij}\}$, the maximum likelihood (ML) estimates of $\{\pi_{ij}\}$ under the SDEP model are given by

$$\hat{\pi}_{ij} = \begin{cases} (p_{ij} + p_{ji})/2 & \text{for } i \neq j \\ (\sum_{t=1}^{r} p_{tt})/r & \text{for } i = j \end{cases}$$

Also the ML estimates of $\{\pi_{ij}\}$ under the SDEP model with κ=0 applied to the data in Table 1 can be obtained by solving the likelihood equations using the Newton-Raphson method (though the detail is omitted.)

The number of degrees of freedom (df) for testing the goodness-of-fit of the SDEP model is $(r-1)(r+2)/2$.

4. AN EXAMPLE

The data in Table 1 taken directly from Bishop et al.(1975, p397) are based on two supervisors who were asked to rate independently the classroom styles of 72 student teachers as authoritarian, democratic, or permissive. Agresti(1989) fitted a model having the structure of symmetry plus quasi-independence with kappa as parameter (SQI model) to these data and gave Pearson's chi-squared statistic χ^2=7.7 based on df=5. Now the SDEP model proposed here fits the data in Table 1 well, yielding χ^2=4.3($P \simeq 0.5$) and likelihood-ratio statistic G^2=5.5 ($P \simeq 0.4$) based on df=5. Since the χ^2 value for the SDEP model is less than that for the SQI model with the same number of df, we would prefer the SDEP model to the SQI model for the data in Table 1. In addition, under the SQI model the ML estimate of κ is $\hat{\kappa}$ =0.37 (from Agresti 1989), and under the SDEP model it is $\hat{\kappa}$=0.370, and also the estimate of κ calculated directly from $\{p_{ij}\}$ is $\hat{\kappa}$=0.362 (from Bishop et al. 1975, p.396). Therefore we see that these three estimated κ are quite close.

For testing the hypothesis that κ=0 under the assumption that the SDEP model holds true, the difference between the G^2 values for the SDEP model with κ=0 [i.e., (2.3)] and the SDEP model is 21.9-5.5=16.4 ($P < 0.001$) based on df=6-5=1. Therefore this value is statistically highly significant. Thus we see the strong effect of kappa parameter in the SDEP model with form (2.2), namely the strong evidence of a difference between the probability that the two supervisors agree and the probability of chance agreement that occurs if the two raters' ratings are statistically independent. [Note that the SDEP model with κ=0 fits the data in Table 1 very poorly, yielding G^2=21.9 ($P \simeq 0.001$) based on df=6.]

Since the SDEP model is a special case of the quasi-symmetry model, we shall apply the quasi-symmetry model to the data in Table 1. Then this model has G^2=3.1 ($P > 0.05$) based on df=1. For testing the hypothesis that the SDEP model holds under the assumption that the quasi-symmetry model holds true, the difference beteween the G^2 values for the SDEP model and the quasi-symmetry model is 2.4 ($P > 0.6$) based on 5-1=4 df. Therefore this hypothesis is accepted and hence the SDEP model is preferable to the quasi-symmetry model for these data.

5. REMARKS

The SDEP model proposed here should be applied for nominal-scale agreement because it is invariant under arbitrary permutation of the categories applied to both rows and columns of the square table displaying joint ratings of the two raters. Also the SDEP model has positive feature as follows: (i) it is unsaturated on the main diagonal, and (ii) it is a quasi-symmetry model.

6. NOTE

The purpose of this paper is to describe the agreement model when the number of raters is *two*. However, the readers may also be interseted to discuss it when the number of raters is greater than two.

Suppose that three raters classify each subject on an γ response categories; and let π_{ijk} denote the probability of rating i by the first rater, rating j by the second rater and rating k by the third rater. Bishop et al. (1975, p.301) describes the model of symmetry for the $\gamma \times \gamma \times \gamma$ contingency table. Using it we can extend model (2.1) into the case where the number of raters is three, as follows:

$$\pi_{ijk} = \begin{cases} \phi_{ijk} & \text{for } i \neq j \text{ or } j \neq k \text{ or } i \neq k \\ \delta & \text{for } i = j = k \end{cases}$$

(6.1)

where $\phi_{ijk} = \phi_{ikj} = \phi_{jik} = \phi_{jki} = \phi_{kij} = \phi_{kji}$. However, it seems difficult to give the suitable interpretation of model (6.1), and to express model (6.1) using kappas for many raters, given by Fleiss(1971) and Conger(1980). [See Fleiss(1971) and Conger(1980) for the details of kappas for many raters.] In addition, we now do not get the suitable data for applying model (6.1). Finally, note that model (6.1) could also be extended into the case where the number of raters is greater than three, although the details are omitted.

ACKNOWLEDGEMENTS

The author is grateful to two anonymous referees for their valuable suggestions on improving this paper.

REFERENCES

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

(2) Agresti, A. (1989). An agreement model with kappa as parameter. *Statistics and Probability Letters*, 7, 271-273.

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

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

(5) Conger, A.J. (1980). Intergration and generalization of kappas for multiple raters. *Psychological Bulletin*, 88, 322-328.

(6) Darroch, J.N. and McClould, P.I. (1986). Category distinguishability and observer agreement. *Australian Journal of Statistics*, 28, 371-388.

(7) Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76, 378-382.

(8) Tanner, M.A. and Young, M.A. (1985a). Modeling agreement among raters. *Journal of the American Statatistical Association*, 80, 175-180.

(9) Tanner, M.A. and Young, M.A. (1985b). Modeling ordinal scale disagreement. *Psychological Bulletin*, 98, 408-415.

Table 1. Student teachers rated by supervisors

Rating by Supervisor 1	Rating by Supervisor 2			Total
	Authoritarian	Democratic	Permissive	
Authoritarian	17 (14.0)	4 (4.5)	8 (9.0)	29
Dermocratic	5 (4.5)	12 (14.0)	0 (1.5)	17
Permissive	10 (9.0)	3 (1.5)	13 (14.0)	26
Total	32	19	21	72

Note : Parenthesized values are estimated expected frequencies for the SDEP model.