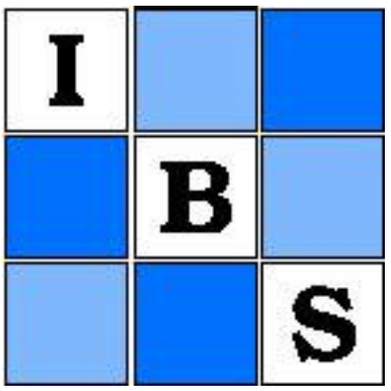

A Model for Agreement Between Ratings on an Ordinal Scale

Author(s): Alan Agresti

Source: *Biometrics*, Jun., 1988, Vol. 44, No. 2 (Jun., 1988), pp. 539-548

Published by: International Biometric Society

Stable URL: https://www.jstor.org/stable/2531866

REFERENCES
Linked references are available on JSTOR for this article:
https://www.jstor.org/stable/2531866?seq=1&cid=pdf-reference#references_tab_contents
You may need to log in to JSTOR to access the linked references.

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at https://about.jstor.org/terms

International Biometric Society is collaborating with JSTOR to digitize, preserve and extend access to *Biometrics*

JSTOR

This content downloaded from
128.227.173.42 on Wed, 20 Nov 2024 19:02:40 UTC
All use subject to https://about.jstor.org/terms

BIOMETRICS 44, 539–548
June 1988

A Model for Agreement Between Ratings on an Ordinal Scale

Alan Agresti

Department of Statistics, University of Florida,
Gainesville, Florida 32611, U.S.A.

SUMMARY

A class of models is proposed for describing agreement between raters who classify a sample on a subjective ordinal scale. The class is obtained by adding to the independence model a component describing baseline association between ratings and a main-diagonal component representing additional incidence of exact agreement. Special cases include the quasi-uniform association model introduced by Goodman (1979, *Journal of the American Statistical Association* **74,** 537–552) and a diagonal-parameter model for nominal-scale agreement proposed by Tanner and Young (1985, *Journal of the American Statistical Association* **80,** 175–180). The model having the structure of uniform association plus a main-diagonal parameter is used to describe agreement between pathologists evaluating carcinoma in situ of the uterine cervix and between neurologists diagnosing multiple sclerosis. The models presented are log-linear and can be fitted using SAS and GLIM.

1. Introduction

Suppose two raters separately classify each subject in a sample on an ordinal scale. Many ordinal scales are quite subjective, such as the scale (yes, probably, about as likely as not, probably not, no) for diagnoses about whether a subject has a certain disease. There is rarely perfect agreement between raters for such scales, partly because of differing perceptions about the meanings of the category labels and partly because of factors such as intrarater variability.

A square contingency table can be used to display joint ratings of the two raters. Two matters are traditionally considered for this table. First, one can analyze differences in the marginal distributions. For ordered response categories, there is usually interest in whether classifications by one rater tend to be higher than those by the other rater. Second, one can analyze the extent of subject-wise agreement between raters, which involves investigating the frequency of main-diagonal occurrence within the joint distribution of the ratings. This article is concerned with the second issue. For discussion of the first type of analysis, see Koch et al. (1977), for instance.

The measurement of interrater agreement has received attention primarily in the social and behavioral sciences, but it is also an important issue in the biomedical sciences. Landis and Koch (1975) presented an interesting review of several investigations of interrater error in biomedical applications. Among these were studies dealing with reliability of diagnoses based on chest radiography, diagnoses of emphysema and other respiratory diseases, diagnoses of cardiac conditions, and diagnoses of psychiatric disorders. In subsequent articles, Landis and Koch (1977a, 1977b) analyzed (i) agreement between pathologists evaluating carcinoma in situ of the uterine cervix, and (ii) agreement between diagnoses of neurologists regarding multiple sclerosis. In Section 3 we analyze these data sets using models introduced in this article.

Key words: Category distinguishability; Dependent samples; Kappa; Log-linear models; Quasi-independence; Quasi-symmetry; Uniform association.

This content downloaded from
128.227.173.42 on Wed, 20 Nov 2024 19:02:40 UTC
All use subject to https://about.jstor.org/terms

Kappa (Cohen, 1960) is the most popular measure for summarizing degree of agreement between two raters. Landis and Koch (1977a, 1977b) and Darroch and McCloud (1986) listed several articles that have dealt with properties of kappa. The latter article, as well as one by Tanner and Young (1985a), have pointed out some of kappa's unsatisfactory features. These include (i) loss of information from summarizing the table by a single number, (ii) sensitivity of value to the form of the marginal distributions, and (iii) subsequent dangers in comparing values of kappa between two tables.

Tanner and Young (1985a) proposed *modeling* the structure of the agreement between the raters, rather than describing it with a single summary measure. Their models are designed for use with *nominal* categorical variables. In this article, log-linear models of agreement are proposed for *ordinal* categorical scales. The primary model considered is an amalgamation of Tanner and Young's model and the uniform association model (Goodman, 1979). The models are log-linear, and they can be fitted using some of the statistical computer packages that have procedures for log-linear models, such as SAS and GLIM.

2. Models of Baseline Association Plus Agreement

Suppose each of n subjects is assigned to one of r response categories separately by two raters, A and B, and let $m_{ij} = n\pi_{ij}$ denote the expected frequency of rating i by the first rater and rating j by the second rater. Tanner and Young (1985a) suggested the log-linear model

$$\log m_{ij} = \mu + \lambda_i^A + \lambda_j^B + \delta(i, j), \tag{2.1}$$

where

$$\delta(i, j) = \begin{cases} \delta, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases}.$$

The parameter δ included for the main-diagonal cells represents agreement beyond that expected by chance—that is, beyond what would be expected if classification by rater A were statistically independent of classification by rater B. The generalization of model (2.1) in which

$$\delta(i, j) = \begin{cases} \delta_i, & i = j, \quad i = 1, \ldots, r, \\ 0 & \text{otherwise} \end{cases}$$

is the quasi-independence model, which allows differential agreement by response category. Given that the raters disagree, these models imply that their ratings are independent. Such behavior does not appear to be the norm for ordinal rating scales. In examples we have considered, there is a moderate to strong positive association between the ratings; conditional on the ratings not being identical, there is still a tendency for high (low) ratings by one rater to occur with high (low) ratings by the other rater.

For bivariate cross-classifications of ordinal variables, the linear-by-linear association model gives a simple and often adequate representation. When each classification has the same ordered categories, this model is

$$\log m_{ij} = \mu + \lambda_i^A + \lambda_j^B + \beta u_i u_j, \tag{2.2}$$

where $u_1 < \cdots < u_r$ are fixed scores assigned to the response categories. This model has only one more parameter than the independence model, and its local log-odds ratios

$$\log \theta_{ij} = \log[m_{ij} m_{i+1,j+1}/(m_{i,j+1} m_{i+1,j})] = \beta(u_{i+1} - u_i)(u_{j+1} - u_j)$$

This content downloaded from
128.227.173.42 on Wed, 20 Nov 2024 19:02:40 UTC
All use subject to https://about.jstor.org/terms

all have the same sign. For equal-interval scores (such as $\{u_i = i\}$), this is the uniform association model discussed by Goodman (1979), having uniformity in the values of these odds ratios. The importance of this model derives partly from its interpretation as a discrete analog of the bivariate normal distribution (Goodman, 1985).

Though often fine for describing association between ordinal classifications, model (2.2) is not an obvious candidate to use for modeling agreement, since it makes no allowance for behavior particular to the main diagonal. The class of models proposed next does this by allowing an extra increment of observations on the main diagonal, beyond what is predicted by an association model. In other words, overall agreement is partitioned into three parts: chance agreement (what would occur even if the classifications were independent), agreement due to a baseline association between the ratings, and an increment that reflects agreement in excess of that occurring simply from chance agreement or from the baseline association. This decomposition is represented by the model

$$\log m_{ij} = \mu + \lambda_i^A + \lambda_j^B + \beta_{ij} + \delta(i, j), \tag{2.3}$$

where the $\{\beta_{ij}\}$ are given a structural form that reflects the expected baseline association, and where $\delta(i, j) = 0$ whenever $i \neq j$ and $\{\delta(i, i) = \delta_i\}$. Generally, one would choose a form for $\{\beta_{ij}\}$ that represents a pattern of monotone association. In this article these terms are given the linear-by-linear association structure $\{\beta_{ij} = \beta u_i u_j\}$. For equal-interval scores $\{u_i\}$, model (2.3) then simplifies to the quasi-uniform association model proposed by Goodman (1979).

Models having $\{\delta(i, i) = \delta_i\}$ impose a perfect fit on the diagonal, with sample cell counts $\{n_{ij}\}$ and estimated expected frequencies $\{\hat{m}_{ij}\}$ satisfying $\hat{m}_{ii} = n_{ii}$ for all i. When possible, it is preferable to use more parsimonious models that are "unsaturated on the main diagonal." The simple version of model (2.3),

$$\log m_{ij} = \mu + \lambda_i^A + \lambda_j^B + \beta u_i u_j + \delta(i, j) \tag{2.4}$$

with

$$\delta(i, j) = \begin{cases} \delta, & i = j \\ 0, & \text{otherwise} \end{cases},$$

has this characteristic and yet seems to summarize agreement well for many data sets. Model (2.4) has residual degrees of freedom (df) $= (r - 1)^2 - 2$, so it is unsaturated whenever $r > 2$. The special case $\beta = 0$ is the Tanner and Young model (2.1) for nominal-scale agreement, the special case $\delta = 0$ is the linear-by-linear association model (2.2), and the special case $\beta = \delta = 0$ is the independence model. We will refer to (2.4) as the model of *agreement plus linear-by-linear association.*

Model (2.4) is a special case of the quasi-symmetry (or "symmetric association") model, since the two-factor term $\lambda_{ij} = \beta u_i u_j + \delta(i, j)$ satisfies $\lambda_{ij} = \lambda_{ji}$ for all i and j. For classification of subject a by rater b, let ρ_{abc} denote the probability the rating is in category c. In a population of S subjects, if one assumes (i) that classifications are made independently in the sense that $\pi_{ij} = S^{-1} \sum_a \rho_{a1i} \rho_{a2j}$, and (ii) that $\{\rho_{abc}\}$ satisfies the condition of no three-factor interaction, then Darroch and McCloud (1986) showed that $\{\pi_{ij}\}$ satisfy the quasi-symmetry model. In this sense, reasonable models for agreement are special cases of that model.

The model of agreement plus linear-by-linear association has simple interpretations through odds ratios. For integer-spaced scores, such as $\{u_i = i\}$,

$$\log \theta_{ij} = \begin{cases} \beta + 2\delta, & i = j \\ \beta - \delta, & |j - i| = 1. \\ \beta, & |j - i| > 1 \end{cases}$$

This content downloaded from
128.227.173.42 on Wed, 20 Nov 2024 19:02:40 UTC
All use subject to https://about.jstor.org/terms

Because $\log \theta_{ij}$ is constant whenever all four cells fall off the main diagonal, we refer to the equal-interval-scores version of (2.4) as the model of *agreement plus uniform association.* Another odds ratio, of particular interest for square tables, is

$$\tau_{ij} = m_{ii}m_{jj}/(m_{ij}m_{ji}), \quad \text{for all } i \text{ and } j.$$

Darroch and McCloud (1986) defined categories i and j to be *indistinguishable* if $\tau_{ij} = 1$ and if $\tau_{ik} = \tau_{jk}$ for all other categories k. For model (2.4),

$$\log \tau_{ij} = (u_j - u_i)^2\beta + 2\delta.$$

Thus, $\delta = \beta = 0$ (independence) implies that all categories are indistinguishable, and when $\beta > 0$ and $\delta \geq 0$, the degree of distinguishability increases as the distance between the categories increases. In particular, for integer-spaced scores, $\log \tau_{i,i+1} = \log \theta_{ii} = \beta + 2\delta$ describes the distinguishability of categories i and $i + 1$, $i = 1, \ldots, r - 1$.

Further interpretation of model (2.4) follows from its likelihood equations. Under the usual multinomial or Poisson sampling assumptions, these are

$$\hat{m}_{i+} = n_{i+}, \quad i = 1, \ldots, r;$$

$$\hat{m}_{+i} = n_{+i}, \quad i = 1, \ldots, r;$$

$$\sum\sum u_i u_j \hat{m}_{ij} = \sum\sum u_i u_j n_{ij};$$

$$\sum \hat{m}_{ii} = \sum n_{ii}.$$

The estimated joint probabilities are constrained to equal the observed joint distribution in the marginal distributions, in the correlation between the ratings, and in the proportion of exact agreement. Both for model (2.4) and nominal-scale model (2.1), the ratio of the total of the estimated expected frequencies under the model to the total expected under the independence model is $n(\sum n_{ii})/(\sum n_{i+}n_{+i})$, and the numerator of sample kappa is the difference between this measure and 1.0.

These models can be fitted using some statistical computer packages that have log-linear model options. The Appendix shows how to use PROC CATMOD in SAS (SAS Institute, 1985) and GLIM (NAG, 1985) to fit agreement models for the first example discussed in Section 3.

When model (2.4) holds, the null hypothesis of independence between the ratings is H_0: $\beta = \delta = 0$. The null hypothesis of no extra agreement beyond that due to the baseline association between ratings is H_0: $\delta = 0$, and the null hypothesis of no extra association beyond that due to exact agreement is H_0: $\beta = 0$. These hypotheses can be tested by comparing likelihood-ratio statistics for the corresponding complete and reduced models. However, for the latter two hypotheses the relevant alternatives are usually H_a: $\delta > 0$ and H_a: $\beta > 0$. For these one can use the test statistics $z = \hat{\delta}/\hat{\sigma}(\hat{\delta})$ and $z = \hat{\beta}/\hat{\sigma}(\hat{\beta})$, where the estimated asymptotic standard errors are obtained from the estimated information matrix.

When model (2.4) is reparameterized with

$$\delta(i, j) = \begin{cases} [(r - 1)/r]\delta, & i = j \\ -\delta/r, & \text{otherwise} \end{cases},$$

and u_i is replaced by $u_i - \bar{u}$, so that $\sum_i \lambda_{ij} = \sum_j \lambda_{ij} = 0$, it is seen that the beyond-chance agreement in cell (i, i) can be described by

$$\lambda_{ii} = \beta(u_i - \bar{u})^2 + [(r - 1)/r]\delta,$$

the amount by which $\log m_{ii}$ exceeds the value corresponding to independence. Thus, the agreement plus linear-by-linear association model predicts (when $\beta > 0$) that this beyond-chance agreement is greatest at the ends of the ordinal scale. In fact, the author's experience

This content downloaded from
128.227.173.42 on Wed, 20 Nov 2024 19:02:40 UTC
All use subject to https://about.jstor.org/terms

in analyzing ordinal tables is that $[\pi_{ij}/(\pi_{i+}\pi_{+j})]$ is commonly most extreme in the corners of the table. By contrast, models having diagonals parameters and no other association parameter, such as model (2.1), permit only constant beyond-chance agreement on the main diagonal.

To summarize agreement by a single index, Darroch and McCloud (1986) argued that a measure based on the $\{\tau_{ij}\}$ is preferable to kappa. Indistinguishability of all pairs of categories is not equivalent to a kappa value of zero. However, it is equivalent to zero for averages of $\{1 - \tau_{ij}\}$ or $\{1 - \tau_{ij}^{-1}\}$ or $\{\log \tau_{ij}\}$, for the special cases of the quasi-symmetry model expected to hold for agreement modeling [e.g., model (2.4) with $\beta \geq 0$ and $\delta \geq 0$]. In this regard, the average of $\{\log \tau_{ij}\}$ is proportional to $\sum \lambda_{ii}$, the total of the category-specific agreements. For model (2.4) this total equals $\sum \lambda_{ii} = \beta \sum (u_i - \bar{u})^2 + (r - 1)\delta$. It can be overly simplistic to use such a measure to summarize or to compare levels of agreement, however, since the beyond-chance agreement has two separate components, the relative influences of which vary by response category.

In some applications the choice of scores $\{u_i\}$ for model (2.4) is questionable. Equal-interval scores give simplest interpretations, and they are a reasonable choice unless there is a more natural scoring. An alternative approach is to fit a version of the model in which the scores are replaced by parameters; for instance, $\lambda_{ij} = \beta\mu_i\mu_j + \delta(i, j)$, where the $\{\mu_i\}$ satisfy location and scale constraints. This model, which has residual df $= (r - 1)^2 - r$, is log-multiplicative rather than log-linear. It is related to a model discussed by Goodman (1979). When the model fits adequately, the distances between the estimated scores can be used in describing distinguishability of categories.

3. Examples

Table 1 is based on data presented in Landis and Koch (1977b) and originally reported by Holmquist, McMahon, and Williams (1967). Seven pathologists classified each of 118 slides in terms of carcinoma in situ of the uterine cervix, based on the most involved lesion, using the ordered categories (1) negative; (2) atypical squamous hyperplasia; (3) carcinoma in situ; (4) squamous carcinoma with early stromal invasion; (5) invasive carcinoma. The analysis given here ignores issues dealing with comparing marginal distributions of responses; see Landis and Koch (1977b) for such an analysis. Instead, it focuses on illustrating models for the degree of agreement using data provided for the first two pathologists,

Table 1
Cross-classification of pathologist ratings, with expected frequencies in parentheses for model of agreement plus uniform association

Pathologist A	Pathologist B				
	1	2	3	4	5
1	22 (22.1)	2 (1.7)	2 (2.2)	0 (0.0)	0 (0.0)
2	5 (4.4)	7 (8.9)	14 (12.5)	0 (.1)	0 (0.0)
3	0 (.4)	2 (1.0)	36 (36.1)	0 (.5)	0 (0.0)
4	0 (.1)	1 (.4)	14 (15.6)	7 (5.4)	0 (.5)
5	0 (0.0)	0 (0.0)	3 (2.6)	0 (1.0)	3 (2.4)

This content downloaded from
128.227.173.42 on Wed, 20 Nov 2024 19:02:40 UTC
All use subject to https://about.jstor.org/terms

labeled A and B. The 5 × 5 cross-classification of their ratings contains 12 empty cells, indicative of the sparseness off the main diagonal that commonly occurs for such data.

When goodness-of-fit tests are applied to sparse data, the distribution of the likelihood-ratio statistic G^2 is not well approximated by the chi-squared distribution. However, this statistic does serve well for comparing unsaturated models, as do statistics based directly on model parameter estimates (Haberman, 1977; Agresti and Yang, 1986). For Table 1, the diagonal-parameter model (2.1) has $G^2 = 30.9$ based on degrees of freedom (df) = 15. The addition to the model of a baseline uniform association between ratings yields a dramatic improvement, as model (2.4) with $\{u_i = i\}$ has $G^2 = 8.4$ with df = 14. Estimated expected frequencies for that model are presented in Table 1. Goodness-of-fit statistics for these and other models are summarized in Table 2.

Table 2
Summary of models of agreement fitted to Table 1

Model	Goodness of fit	df
Independence	131.2	16
Diagonal-parameter	30.9	15
Uniform association	16.2	15
Agreement plus uniform association	8.4	14
Quasi-uniform association	1.3	10
Quasi-symmetry	1.0	6

The maximum likelihood parameter estimates for model (2.4), with asymptotic standard errors given in parentheses, are $\hat{\delta} = 1.067$ (a.s.e. = .404) and $\hat{\beta} = 1.150$ (a.s.e. = .342). There is strong evidence of extra agreement beyond that due to the baseline association, and there is strong evidence of extra association beyond that due to the exact agreement. Using these estimates, the beyond-chance agreement can be summarized as follows: For $i = 1, 2, 3, 4$, the odds that the diagnosis of pathologist A is $i + 1$ rather than i is estimated to be $\exp(\hat{\beta} + 2\hat{\delta}) = 26.7$ times higher when the diagnosis of pathologist B is $i + 1$ than when it is i. Similarly, one can summarize the baseline association: For $|i - j| > 1$, the odds that the diagnosis of pathologist A is $i + 1$ rather than i is estimated to be $\exp(\hat{\beta}) = 3.2$ times higher when the diagnosis of pathologist B is $j + 1$ than when it is j.

The parameter-scores version of model (2.4) was also fitted, giving $G^2 = 8.0$ with df = 11. The estimated scores, scaled so that $\hat{\mu}_1 = 1$ and $\hat{\mu}_5 = 5$ for ease of comparison with fixed-integer scores $\{u_i = i\}$, are (1, 2.03, 2.91, 4.17, 5). This shows why the uniform association version of the model fits adequately.

Table 3, taken from Landis and Koch (1977a), displays diagnoses of multiple sclerosis for two neurologists who classified patients in two sites, Winnipeg and New Orleans. The diagnostic classes are (1) certain multiple sclerosis; (2) probable multiple sclerosis; (3) possible multiple sclerosis; (4) doubtful, unlikely, or definitely not multiple sclerosis. For the Winnipeg patients, the agreement plus uniform association model has $G^2 = 9.4$ based on df = 7, with $\hat{\beta} = .804$ (a.s.e. = .155) and $\hat{\delta} = -.028$ (a.s.e. = .243). For the New Orleans patients, the corresponding results are $G^2 = 8.8$ with df = 7, $\hat{\beta} = 1.041$ (a.s.e. = .296), and $\hat{\delta} = .028$ (a.s.e. = .348). The results are similar and suggest a single model for the full 4 × 4 × 2 cross-classification of neurologist A rating-by-neurologist B

This content downloaded from
128.227.173.42 on Wed, 20 Nov 2024 19:02:40 UTC
All use subject to https://about.jstor.org/terms

Table 3
Diagnostic classifications regarding multiple sclerosis

New Orleans neurologist	Winnipeg neurologist: Winnipeg patients 1	2	3	4	Winnipeg neurologist: New Orleans patients 1	2	3	4
1	38	5	0	1	5	3	0	0
2	33	11	3	0	3	11	4	0
3	10	14	5	6	2	13	3	4
4	3	7	3	10	1	2	4	14

Table 4
Summary of models of agreement fitted to Table 3

Model	Goodness of fit	df
Independence for each site	115.4	18
Homogeneous diagonal-parameter	79.4	17
Homogeneous uniform association	19.2	17
Homogeneous agreement plus uniform association	19.2	16
Homogeneous quasi-uniform association	16.8	13
Homogeneous quasi-symmetry	13.4	12

rating-by-site. For the model in which a uniform association parameter and a main-diagonal parameter (homogeneous for the two sites) are added to the model of conditional independence between ratings, given site, $G^2 = 19.2$ based on df = 16, with $\hat{\beta} = .864$ (a.s.e. = .138) and $\hat{\delta} = .017$ (a.s.e. = .197). This result suggests the further simplification $\delta = 0$, for which there is homogeneous uniform association between the neurologists' ratings for each set of patients. This model fits adequately, with $G^2 = 19.2$ and df = 17. The goodness of fit of these and other models is summarized in Table 4.

For the data sets considered here, and for most others the authors has considered having ordered categories, much of the beyond-chance agreement is explained by the baseline association between the ratings. For ordinal rating scales, therefore, it is rarely adequate to use models for nominal-scale agreement that imply independence off the main diagonal.

4. Comments

Other models proposed for square tables with ordered categories may also be useful for describing agreement. These include a model proposed by Haber (1985) expressed as $\log \tau_{ij} = \alpha + |j - i|\beta$, other diagonals-parameter models (Goodman, 1972; Tanner and Young, 1985b; Hout, Duncan, and Sobel, 1987), and other log-linear or log-multiplicative models (Hauser and Massagli, 1983; Cox, Przepiora, and Plackett, 1982; Jørgensen, 1985). The agreement plus linear-by-linear association model has several positive features. These include the following:

(i) It utilizes the ordering of the response categories.
(ii) Given that the raters disagree, it does not assume that the ratings are independent.
(iii) The model is unsaturated on the main diagonal.
(iv) The baseline association and extra agreement parameters are easily interpreted.
(v) It is a quasi-symmetry model.
(vi) It is a simple model that, in a wide variety of situations, explains most of the variation that remains after one has fitted the independence model.

This content downloaded from
128.227.173.42 on Wed, 20 Nov 2024 19:02:40 UTC
All use subject to https://about.jstor.org/terms

Hence, this is a reasonable model to use in a first attempt to fit the data. Although it is *too* simple to give an adequate fit to some tables, the author's experience is that it almost always fits much better than quasi-independence models [such as (2.1)]. Even when it does not fit adequately, the pattern of the larger residuals provides interesting information about the structure of agreement.

Multirater generalizations of the models discussed in this article can be formulated directly using the approach suggested in Tanner and Young (1985a). These generalized models can be useful for comparing levels and patterns of agreement for various pairs of raters. They can be difficult to implement unless the number of raters is small, however, since the relevant cross-classifications are very large and sparse. This author plans to present alternative ways of modeling pairwise and multiway agreement among several raters, both for nominal and ordinal rating schemes, in a future article.

Résumé

Une classe de modèles est proposée pour la description de l'accord entre 2 juges qui classent des sujets selon une échelle subjective ordinale. Cette classe est obtenue par l'addition au modèle d'indépendance de deux composantes, l'une décrivant l'association entre les deux classements et l'autre, dite de "diagonale principale," représentant l'incidence additionnelle des accords réels. La classe englobe le modèle d'association quasi uniforme de Goodman (1979, *Journal of the American Statistical Association* **74,** 537–552) et le modèle avec "paramètre diagonal" proposé par Tanner et Young (1985, *Journal of the American Statistical Association* **80,** 175–180). Un modèle particulier (association uniforme plus paramètre de diagonale principale) est utilisé pour décrire l'accord entre anatomopathologistes évaluant le cancer in situ du col de l'utérus et entre neurologues diagnostiquant la sclérose en plaques. Les modèles présentés sont log linéaires et peuvent être estimés et ajustés grâce aux logiciels SAS et GLIM.

References

Agresti, A. and Yang, M. (1986). An empirical investigation of some effects of sparseness in contingency tables. *Computational Statistics and Data Analysis* **5,** 9–21.

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

Cox, M. A. A., Przepiora, P., and Plackett, R. L. (1982). Multivariate contingency tables with ordinal data. *Utilitas Mathematica* **21A,** 29–42.

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

Goodman, L. A. (1972). Some multiplicative models for the analysis of cross-classified data. In *Proceedings of 6th Berkeley Symposium on Mathematical Statistics and Probability, Vol.* 1, 649–696. Berkeley: University of California Press.

Goodman, L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association* **74,** 537–552.

Goodman, L. A. (1985). The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries. *Annals of Statistics* **13,** 10–69.

Haber, M. (1985). Maximum likelihood methods for linear and log-linear models in categorical data. *Computational Statistics and Data Analysis* **3,** 1–10.

Haberman, S. J. (1977). Log-linear models and frequency tables with small expected cell counts. *Annals of Statistics* **5,** 1148–1169.

Hauser, R. M. and Massagli, M. P. (1983). Some models of agreement and disagreement in repeated measures of occupation. *Demography* **20,** 449–460.

Holmquist, N. S., McMahon, C. A., and Williams, O. D. (1967). Variability in classification of carcinoma in situ of the uterine cervix. *Archives of Pathology* **84,** 334–345.

Hout, M., Duncan, O. D., and Sobel, M. E. (1987). Association and heterogeneity: Structural models of similarities and differences. *Sociological Methodology* **17,** 145–184.

Jørgensen, B. (1985). Estimation of interobserver variation for ordinal rating scales. *Lecture Notes in Statistics: Generalized Linear Models* **32,** 93–104.

Koch, G. G., Landis, J. R., Freeman, J. L., Freeman, D. H., Jr., and Lehnen, R. G. (1977). A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics* **33,** 133–158.

This content downloaded from
128.227.173.42 on Wed, 20 Nov 2024 19:02:40 UTC
All use subject to https://about.jstor.org/terms

Landis, J. R. and Koch, G. G. (1975). A review of statistical methods in the analysis of data arising from observer reliability studies (part I). *Statistica Neerlandica* **29,** 101–123.

Landis, J. R. and Koch, G. G. (1977a). The measurement of observer agreement for categorical data. *Biometrics* **33,** 159–175.

Landis, J. R. and Koch, G. G. (1977b). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* **33,** 363–374.

NAG (1985). *The GLIM Release 3.77 Manual.* Downers Grove, Illinois: Numerical Algorithms Group, Inc.

SAS Institute, Inc. (1985). *SAS User's Guide*: *Statistics, Version 5 Edition.* Cary, North Carolina: SAS Institute, Inc.

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

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

Received November 1986; *revised September* 1987.

APPENDIX

Using SAS and GLIM to Fit Agreement Models

PROC CATMOD in SAS can be used to fit a wide variety of models for categorical data. The default model is a generalized logit model in which each response category is paired with the last category. For instance, if the column variable is the response variable, then the generalized logit model corresponding to model (2.4) is

$$\begin{aligned}\log(m_{ij}/m_{ir}) &= \log(m_{ij}) - \log(m_{ir}) \\ &= (\lambda_j^B - \lambda_r^B) + \beta u_i(u_j - u_r) + [\delta(i,j) - \delta(i,r)] \\ &= \alpha_j + \beta x_{ij} + \delta y_{ij}, \quad i = 1, \ldots, r, \quad j = 1, \ldots, r-1,\end{aligned}$$

where $x_{ij} = u_i(u_j - u_r)$ and $y_{ij} = 1$ for $i = j$, $y_{ij} = -1$ for $i = r$, and $y_{ij} = 0$ otherwise. PROC CATMOD allows the user to specify the design matrix. For model (2.4), the design matrix has $r(r-1)$ rows, one for each of the $(r-1)$ logits in the r rows of the table. The parameter vector has elements $(\alpha_1, \ldots, \alpha_{r-1}, \beta, \delta)$, where β and δ are the parameters of interest. Figure 1 gives the code for fitting this model to Table 1. The rows of the 20×6 design matrix are specified in the MODEL statement. For instance, the fifth entry in each row contains the values of x_{ij} for the 20 logits. The ML option in this statement requests the maximum likelihood fit.

```
DATA AGREE;
INPUT A $ B $ COUNT @@;
CARDS;
1 1 22  1 2 2  1 3 2   1 4 0  1 5 0
2 1 5   2 2 7  2 3 14  2 4 0  2 5 0
3 1 0   3 2 2  3 3 36  3 4 0  3 5 0
4 1 0   4 2 1  4 3 14  4 4 7  4 5 0
5 1 0   5 2 0  5 3 3   5 4 0  5 5 3
PROC CATMOD ORDER = DATA;
  WEIGHT COUNT;
POPULATION A;
MODEL B = (1 0 0 0  8  1,  0 1 0 0  6  0,  0 0 1 0  4  0,  0 0 0 1  2  0,
           1 0 0 0  4  0,  0 1 0 0  3  1,  0 0 1 0  2  0,  0 0 0 1  1  0,
           1 0 0 0  0  0,  0 1 0 0  0  0,  0 0 1 0  0  1,  0 0 0 1  0  0,
           1 0 0 0 -4  0,  0 1 0 0 -3  0,  0 0 1 0 -2  0,  0 0 0 1 -1  1,
           1 0 0 0 -8 -1,  0 1 0 0 -6 -1,  0 0 1 0 -4 -1,  0 0 0 1 -2 -1)
          /ML NOGLS PRED = FREQ;
TITLE 'MODELING AGREEMENT';
```

Figure 1. SAS code for fitting model (2.4) to Table 1.

This content downloaded from
128.227.173.42 on Wed, 20 Nov 2024 19:02:40 UTC
All use subject to https://about.jstor.org/terms

GLIM is a versatile interactive program for fitting generalized linear models. The response variable must have a distribution in the linear exponential family. Models are specified for monotone transformations (called *links*) of the mean of the response variable. Log-linear models treat the cell counts as independent Poisson responses, and they use the log-link function. Figure 2 contains code for fitting several agreement models using version 3.77 of GLIM on a personal computer. The first CALC directive sets up a vector of cross-products of integer scores, which is the coefficient of β in models of uniform association. The second CALC directive defines a vector with components equal to 1.0 when the two factors are at the same level and 0.0 otherwise. This is the coefficient of δ in models (2.1) and (2.4). The YVAR directive specifies that the cell count is the response. The ERROR directive specifies the Poisson response distribution, for which the log-link is the default. The FIT directive specifies the independent variables for the model. This particular statement requests four models: the independence model, model (2.1), model (2.2), and model (2.4). From this directive, GLIM fits the model and reports the likelihood-ratio statistic as the "scaled deviance." Parameter estimates, standard errors, and estimated expected frequencies are given as a result of the DIS directive.

```
$units 25
$factor a 5  b 5
$data a b count
$read
1 1 22   1 2 2   1 3 2    1 4 0   1 5 0
2 1 5    2 2 7   2 3 14   2 4 0   2 5 0
3 1 0    3 2 2   3 3 36   3 4 0   3 5 0
4 1 0    4 2 1   4 3 14   4 4 7   4 5 0
5 1 0    5 2 0   5 3 3    5 4 0   5 5 3
$calc unif = a*b $
$calc delta = %eq (a, b) $
$yvar count
$error p $
$fit a + b: + delta: − delta + unif: + delta $
$dis e r $
```

Figure 2. GLIM code for fitting agreement models to Table 1.

This content downloaded from
128.227.173.42 on Wed, 20 Nov 2024 19:02:40 UTC
All use subject to https://about.jstor.org/terms