Hacettepe Journal of Mathematics and Statistics
Volume 40 (4) (2011), 609 – 616

AGREEMENT PLUS DISAGREEMENT MODEL FOR AGREEMENT DATA

Tülay Saraçbaşı*

Received 14 : 06 : 2010 : Accepted 11 : 11 : 2010

Abstract

In $R \times R$ square contingency tables where there is a one to one correspondence between the categories of the row and column variables, the agreement between the row and column classifications is of interest. Several authors modeled agreement in terms of a log-linear representation. We propose a new model by combining the agreement and disagreement models, and applying it to the cross-classification of two neurologist's rates for 149 MS patients.

Keywords: Agreement, Disagreement, Square tables, Social mobility data.

2000 AMS Classification: 62 H 17.

1. Introduction

Rater agreement is of importance in many fields such as medical, social and behavioral sciences. Subjects are classified into categories by raters. Suppose that n observations are independently assigned by the two raters according to R categories. One needs to know the agreement among the raters. The assessment of the reliability of a rating system has been considered from the perspective of inter-rater agreement. The observed proportion of agreement has been assessed in early studies. Cohen [5-6] introduced the kappa coefficient to measure the chance-corrected agreement for the nominal scale and weighted kappa for the ordinal scale. Bangdiwala [3] proposed the B statistic to quantify the agreement between the two raters. Von Eye [14] defined a new coefficient as an alternative measure of rater agreement. Cohen [6], von Eye and von Eye [13] also discussed the disagreement.

Tanner and Young [11], besides other authors, have pointed out some unsatisfactory features of kappa. Several authors, for example Schuster and von Eye [10], Agresti [1], Tanner and Young [11,12], Becker [4] proposed modeling the structure of the agreement and disagreement between raters, rather than describing it with a single summary measure. These models are in the form of log-linear models.

*Department of Statistics, Faculty of Science, Hacettepe University, 06800, Beytepe, Ankara, Turkey. E-mail: tulay.saracbasi@gmail.com Toker.hacettepe.edu.tr

2. Coefficients of Agreement and Disagreement

In this section, we consider $R \times R$ square contingency tables for two raters and give the agreement and disagreement coefficients used in the literature. Through the paper, n is the total number of observations, p_{ij} denotes the probability that an observation falls in the *ith* row and *jth* column of the table, n_{ij} is the observed frequencies, $i\cdot$ indicates the total of the *ith* row, and $\cdot j$ indicates the total of the *jth* column.

2.1. Raw agreement. Raw agreement denotes the proportion of cases in the main diagonal and indicates the probability for two and more raters' judgments matching perfectly.

Consider the rating of two raters with dichotomous ratings summarized in Table 1.

Table 1. 2×2 Agreement table

	Rater 2		
Rater 1	+	−	Total
+	a	b	$a+b$
−	c	d	$c+d$
Total	$a+c$	$b+d$	n

The raw agreement for the data is calculated as

$$p_0 = \frac{a+d}{a+b+c+d} = \frac{a+d}{n}. \tag{1}$$

2.2. Cohen's kappa (κ). The most popular measure of agreement between two nominal categorical data is Cohen's kappa given by Cohen [5] and estimated as,

$$\kappa = \frac{\sum_{i=1}^{R} p_{ii} - \sum_{i=1}^{R}\sum_{j=1}^{R} p_{i\cdot}p_{\cdot j}}{1 - \sum_{i=1}^{R}\sum_{j=1}^{R} p_{i\cdot}p_{\cdot j}}. \tag{2}$$

Cohen's kappa was originally introduced as a chance-corrected measure of agreement between two raters for nominal scales, and can be used to test the hypothesis whether the agreement is better than that expected based on the chance model of rater independence.

Fleiss *et.al* [7] showed that for sufficiently large samples, κ is normally distributed and the test statistic $z = \frac{\hat{\kappa}}{\hat{\sigma}_\kappa}$ tests the null hypothesis that $\kappa = 0$. The standard error of $\hat{\kappa}$ is

$$\sigma_{\hat{\kappa}} = \sqrt{\frac{1}{n\left(1 - \sum_{i=1}^{R} p_{i\cdot}p_{\cdot i}\right)^2}\left[\sum_{i=1}^{R} p_{i\cdot}p_{\cdot i} + \left(\sum_{i=1}^{R} p_{i\cdot}p_{\cdot i}\right)^2 - \sum_{i=1}^{R} p_{i\cdot}p_{\cdot i}\left(p_{i\cdot} + p_{\cdot i}\right)\right]}. \tag{3}$$

2.3. Weighted kappa. When the row and column classifications are ordinal categorical data, then weighted kappa is used:

$$\kappa = \frac{\sum_{i=1}^{R}\sum_{j=1}^{R} w_{ij}p_{ij} - \sum_{i=1}^{R}\sum_{j=1}^{R} w_{ij}p_{i\cdot}p_{\cdot j}}{1 - \sum_{i=1}^{R}\sum_{j=1}^{R} w_{ij}p_{i\cdot}p_{\cdot j}}, \tag{4}$$

where w_{ij} are the weight in the range $0 \le w_{ij} \le 1$ [6].

Von Eye and von Eye [13] reviewed three of the well known characteristics of Cohen's Kappa. Landis and Koch [9] defined the agreement levels as in Table 2.

Table 2. Agreement levels

Kappa Statistic	Strength of Agreement
< 0	Poor
0–0.2	Slight
0.2–0.4	Fair
0.4–0.6	Moderate
0.6–0.8	Substantial
0.8–1	Almost Perfect

2.4. Bangdiwala's B statistic. Bangdiwala [3] proposed a statistic that measures the degree of the agreement between two raters independently, and it is calculated as

$$B = \frac{\sum_{i=1}^{R} n_{ii}^2}{\sum_{i=1}^{R} n_{i\cdot} n_{\cdot i}}, \tag{5}$$

where n_{ij} is the cell entry on the main diagonal. The B statistic is a proportion ranging from zero, for no agreement, to +1 for perfect agreement.

2.5. An alternative to Cohen's kappa. Von Eye [14] proposed an alternative to Cohen's kappa. The new coefficient of rater agreement is defined as the average cell-wise proportionate reduction in error (PRE)

$$\kappa_s = \frac{1}{R} \sum_{i=1}^{R} \sum_{j=1}^{R} \frac{p_{ij} - \hat{p}_{ij}}{\min\{p_{i\cdot}, p_{\cdot j}\} - \hat{p}_{ij}}, \tag{6}$$

where $\hat{p}_{ij}$ is the probability of cell ij, that is estimated under some base model or chance model.

2.6. Disagreement. If the agreement is low, it is important to describe the nature of the departure from agreement. The investigation of disagreement may also be of interest. Cohen [6] defined the disagreement as,

$$\kappa = \frac{\sum_{i=1}^{R} \sum_{j=1}^{R} p_{i\cdot} p_{\cdot j} - \sum_{i=1}^{R} p_{ii}}{1 - \sum_{i=1}^{R} \sum_{j=1}^{R} p_{i\cdot} p_{\cdot j}}. \tag{7}$$

For the ordinal data, von Eye and von Eye [13] performed a simulation study concerning the distributional characteristics of the coefficients of disagreement and defined the disagreement cells for the weighted kappa as follows:

Case 1: All disagreement cells:

$$w_{ij} = \begin{cases} 1 & \text{if } i \neq j, \\ 0 & \text{otherwise.} \end{cases}$$

Case 2: Cells above the main diagonal:

$$w_{ij} = \begin{cases} 1 & \text{if } i < j, \\ 0 & \text{otherwise.} \end{cases}$$

Case 3: Cells right above or right below the main diagonal:

$$w_{ij} = \begin{cases} 1 & \text{if } i = j - 1 \text{ or } i = j + 1, \\ 0 & \text{otherwise.} \end{cases}$$

Positive values of κ indicate agreement better than chance. Negative values of κ indicate agreement less than chance. κ can be zero even if the raters judgments are not independent. $\kappa = 1$ only if the probability of disagreement is zero. When the probability of disagreement decreases and is smaller than the probability of agreement, κ increases monotonically; when the probability of disagreement increases and is greater than the probability of agreement, κ does not decrease monotonically [14].

In many applications, it is not sufficient to summarize the agreement by a single number. Instead of measuring the agreement by a single number, it can also be expressed in a log-linear formulation from which a corresponding parameter can be estimated. The models for ordinal rater agreement data characterize the association independently of the margins. Some authors have pointed out that Cohen's kappa statistic is insensitive to the differences between the observed and expected patterns of agreement, and some limits of this statistic. Pros and cons of kappa can be found on the web site: http://ourworld.compuserve.com/homepages/jsuebersax/kappa.htm#procon.

Some of those are:

(i) "Loss of information from summarizing the table by a single number.
(ii) It really does not distinguish the disagreement.
(iii) Kappa may be low even though there are high levels of agreement.
(iv) With ordered category data, one must select weights arbitrarily to calculate weighted kappa.
(v) Kappa requires that two raters use the same rating categories. There are situations where one is interested in measuring the consistency of ratings for raters that use different categories (e.g., one uses a scale of 1 to 3, another uses a scale of 1 to 5)."

Most authors developed log-linear models arguing that these models provide more information on the pattern of agreement. Tanner and Young [11], Agresti [1], and Becker [4], Schuster and von Eye [10] suggested some agreement models. The agreement model proposed by Tanner and Young [11] is defined as

$$\log \mathrm{m_{ij}} = \mu + \lambda_i^A + \lambda_j^B + \delta_{ij},\ i, j = 1, \ldots, R, \tag{8}$$

where m_{ij} denotes the expected frequency, A is the row variable, B the column variable, μ the main effect, λ_i represents the row parameter and λ_j represents the column parameter. In the model, the parameters satisfy the restrictions, $\sum_{i=1}^{R} \lambda_i^A = \sum_{j=1}^{R} \lambda_j^B = 0$. Also, δ_{ij} denotes the agreement parameter,

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

The above model is expressed in terms of the odds ratios

$$\log \theta_{ij} = \begin{cases} 2\delta & \text{if } |\, i - j \,| = 0, \\ -\delta & \text{if } |\, i - j \,| = 1, \\ 1 & \text{if } |\, i - j \,| > 1, \end{cases}$$

and this model has $(R-1)^2 - 1$ residual degrees of freedom. Tanner and Young [12] also described the disagreement model for ordinal rater agreement data as follows,

$$\log m_{ij} = \mu + \lambda_i^A + \lambda_j^B + \delta_{ij},\ i, j = 1, \ldots, R, \tag{9}$$

where δ_{ij} is given by

$$\delta_{ij} = \begin{cases} \delta & \text{if } i \neq j, \\ 0 & \text{otherwise,} \end{cases}$$

and this model also has $(R-1)^2-1$ residual degrees of freedom. Tanner and Young [12] also characterize the symmetric band disagreement that can be illustrated by replacing δ_{ij} in model (9) by the definition below:

$$\delta_{ij} = \begin{cases} \delta_1 & \text{if } |i-j|=1, \\ \delta_2 & \text{if } |i-j|=2 \\ \cdots & \cdots \\ \cdots & \cdots \\ \cdots & \cdots \\ \delta_{R-1} & \text{if } |i-j|=R-1, \\ 0 & \text{otherwise.} \end{cases}$$

This model hypothesizes that the chance-corrected frequencies are constant within each pair of bands, but possibly different across the different pairs [12]. Parameters in the model can be estimated by the maximum likelihood method.

Agresti [1] suggested the model of agreement plus linear-by-linear association for ordinal categorical data,

$$\log m_{ij} = \mu + \lambda_i^A + \lambda_j^B + \beta u_i v_j + \delta_{ij},\ i,j = 1,\ldots,R, \tag{10}$$

where β is the association parameter between the row and the column variables, u_i and v_j are the row and column scores respectively; $u_1 < u_2 < u_3 < \cdots < u_R$, $v_1 < v_2 < v_3 < \cdots < v_C$. The model has $(R-1)^2-2$ residual degrees of freedom and can be expressed in terms of the odds ratios

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

Schuster and von Eye [10] have proposed the new log-multiplicative agreement model for the agreement data. All agreement models are based on the Goodman's association models [8].

3. Agreement plus disagreement (AD) model

We can generalize the class of models that represent the agreement and disagreement models together as follows

$$\log m_{ij} = \mu + \lambda_i^A + \lambda_j^B + \gamma_{ij} + \delta_{ij}, \tag{11}$$

where γ_{ij} indicates the agreement parameter for $i=j$ and δ_{ij} denotes the symmetric band disagreement parameters for $i \neq j$ in the proposed model.

The parameters in Equation (11) are defined as:

$$\gamma_{ij} = \begin{cases} \gamma_0 & \text{if } i=j, \\ 0 & \text{otherwise} \end{cases}$$

$$\delta_{ij} = \begin{cases} \delta_1 & \text{if } |i-j|=1, \\ \delta_2 & \text{if } |i-j|=2, \\ \cdots & \cdots \\ \cdots & \cdots \\ \cdots & \cdots \\ \delta_{R-1} & \text{if } |i-j|=R-1. \end{cases}$$

Since model (11) has more $(R-1)$ parameters than the agreement and disagreement models, the residual degree of freedom for this model is $(R-1)(R-2)$. The model also includes the agreement part.

In terms of the local odds ratios the model can be expressed for a $R \times R$ table as,

$$\log \theta_{ij} = \begin{cases} 2\gamma_0 - 2\delta_1 & \text{if } i = j, \\ -\gamma_0 + 2\delta_1 - \delta_2 & \text{if } |i-j| = 1, \\ -\delta_1 + 2\delta_2 - \delta_3 & \text{if } |i-j| = 2, \\ -\delta_2 + 2\delta_3 - \delta_4 & \text{if } |i-j| = 3, \\ \dots & \dots \\ -\delta_{k-1} + 2\delta_k - \delta_{k+1} & \text{if } |i-j| = R-1. \end{cases} \tag{12}$$

The usual formula of the local odds ratio in terms of the expected frequencies for the underlying model is, for example for θ_0 $(k = i - j = 0)$ and setting $\theta_{ij} = \theta_k$,

$$\theta_0 = \frac{m_{11} m_{22}}{m_{12} m_{21}},$$

where the expected frequencies can be expressed as a logarithmic model equation,

$$\begin{aligned} \log\, m_{11} &= \mu + \lambda_1^A + \lambda_1^B + \gamma_0, \\ \log\, m_{22} &= \mu + \lambda_2^A + \lambda_2^B + \gamma_0, \\ \log\, m_{21} &= \mu + \lambda_2^A + \lambda_1^B + \delta_1, \text{ and} \\ \log\, m_{12} &= \mu + \lambda_1^A + \lambda_2^B + \delta_1. \end{aligned}$$

Substituting these expected values into the odds ratio formula and simplifying the equations, we get Equation (12). All the odds ratios can be found in a similar way. The agreement plus disagreement model can be fitted using any statistical software that has log-linear options.

The advantages of using AD model are:

(i) Investigates both agreement and disagreement components at the same time,
(ii) Parameter interpretations give detailed results,
(iii) Fits data better than the agreement and disagreement models.

4. An example

A real data set directly taken from Landis and Koch [9] is displayed in Table 3. In this data set two neurologists independently classified 149 MS patients into one of the following classes: 1: Certain MS, 2: Probable MS, 3: Possible MS, 4: Doubtful, unlikely, or definitely not MS. The Results of neurologist 1 and 2 are used for the analysis.

Table 3. Cross-classification table of two neurologists' rates

	Neurologist 2			
Neurologist 1	1	2	3	4
1	38 (36.48)	5 (7.32)	0 (0.49)	1 (0.21)
2	33 (31.71)	11 (12.48)	3 (2.25)	0 (1.06)
3	10 (12.01)	14 (12.76)	5 (4.52)	6 (5.71)
4	3 (3.79)	7 (4.44)	3 (4.24)	10 (10.52)

The independence, agreement, disagreement and AD models were fitted to the rater agreement data. The independence model was rejected ($L^2 = 59.285$, $P = 0.000$), thus there is strong evidence of association in the data. Weighted kappa is found to be 0.38 with 0.052 standard error. The models of agreement and disagreement were applied to Table 3 and yielded associated likelihood ratio chi-square statistics of 44.194 ($P = 0.000$) with 8 degrees of freedom. These models do not fit the data well. On the other hand, the AD model fits the data well, with a likelihood ratio statistic $L^2 = 5.672$ ($P = 0.461$) based on 6 degrees of freedom.

The parenthesized values in Table 3 are the maximum likelihood estimates of the expected frequencies under the AD model. Parameter estimates with their standard errors are displayed in Table 4. These results are based on sampling zero corrections (0.5 is added to zero cells).

Table 4. Parameter Estimates under the AD model

Parameter	Estimate	St.Error	Z-value
γ_0	3.094	0.623	4.962*
δ_1	2.757	0.622	4.436*
δ_2	1.427	0.602	2.371*

*Significant at 5%

Here δ_3 is a redundant parameter, it is not computed. In Table 4, γ_0 indicates the agreement on the main diagonal, δ_1, δ_2 are the symmetric disagreement parameters indicating the first order difference and subsequent differences, respectively.

It is noted that γ_0, δ_1 and δ_2 are statistically significant parameters. After estimating the parameters, the local odds ratios can be obtained. From Equation (12), the odds ratios can be estimated for $k = 0, 1$ and 2. The significance test can be performed for the ln (odds) ratios. The $H_0 : \ln(\theta_{ij}) = 0$ hypothesis is tested against the alternative $H_A : \ln(\theta_{ij}) \neq 0$.

The hypothesis can be tested by the z statistic, and $z = \dfrac{\ln(odds)}{\text{ASE}(\ln(odds))}$ is asymptotically normally distributed [2]. The odds ratios with their standard errors and the z values are given in Table 5.

Table 5. Estimated values of the odds ratios for $k = 0, 1, 2, 3$

k	Odds Ratios	ln (odds R.)	SE (ln(odds R.))	Z-value
0	1.96	0.674	0.5249	1.2840
1	2.69	0.993	0.5227	1.89*
2	1.099	0.097	0.8079	0.12

*Significant at 10%

The local odds ratios estimated from Equation (12) can be interpreted as, for example, for k=0: the local odds ratio that the probability that neurologist 1 and neurologist 2 agree is 1.96 times higher than that neurologist 1 and neurologist 2 disagree (that is, $\theta_0 = m_{ij}m_{i+1j+1}/m_{ij+1}m_{i+1j}$, $(i = j = 1, 2, 3)$.

From the results in Table 5, while the odds ratio for $i = j$ appears to be statistically not significant, the other estimates for $k = 1$ is significant and that for $k = 2$ is not significant.

5. Conclusions

Even though Cohen's kappa has a great importance in medical, psychological, behavioral, educational sciences and the like, many authors have pointed out some unsatisfactory features and difficulties of kappa. Kappa is insensitive to differences between the observed and expected patterns of agreement. There will be a loss of information from summarizing the table by a single number, and it does not distinguish the disagreement. Kappa may be low even though there are high levels of agreement. Log-linear models have become an important tool in the analysis of these type of data. Therefore, the log-linear agreement and disagreement models have been preferred over kappa. Rather than summarizing the agreement, one may wish to analyze the structure of the agreement in the data. Thus, modeling the agreement becomes of interest.

In this paper, a new model (the AD model) is proposed. This model can be easily applied to square contingency tables having ordered categories. Parameter estimates can be easily interpreted. Instead of investigating the agreement and disagreement separately, one can use this new model to explore them together. The parameters based on the symmetric disagreement show on which diagonal the disagreement is stronger.

The proposed model gives the ability for the raters to distinguish between categories and to compare rating scales for two raters. In an example, agreement between neurologist 1 and neurologist 2 was found to be significant for the rates $k = i - j = j - i = 1$ $[(i, j), (i + 1, j + 1)]$ versus $(i, j + 1)$ or $(i + 1, j)$.

References

[1] Agresti, A. *A model for agreement between ratings on an ordinal scale*, Biometrics **44**, 539–548, 1988.
[2] Agresti, A. *Categorical Data Analysis* (Wiley, New York, 2002).
[3] Bangwidala, S. *A graphical test for observer agreement*, Proceedings of the 45^{th} International Statistical Institute Meeting, Amsterdam, **1**, 307–308, 1985.
[4] Becker, M. P. *Using association models to analyse agreement data: two examples*, Statistics in Medicine **8**, 1199–1207, 1989.
[5] Cohen, J. *A coefficient of agreement for nominal scales*, Educational and Psychological Measurement **20**, 37–46, 1960.
[6] Cohen, J. *Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit*, Psychological Bulletin **70**, 213–220, 1968.
[7] Fleiss, J. L., Cohen, J. and Everitt, B. S. *Large sample standard errors of κ and weighted κ*, Psychological Bulletin **72**, 323–327, 1969.
[8] Goodman L. A. *Simple models for the analysis of association in cross-classifications having ordered categories*, Journal of American Statistical Association **74**, 537–553, 1979.
[9] Landis, J. R. and Koch, G. G. *The measurement of observer agreement for categorical data*, Biometrics **33**, 159–174, 1977.
[10] Schuster C. and Von Eye, A. *Models for ordinal agreement data*, Biometrical Journal **43** (7), 795–808, 2001.
[11] Tanner, M. A. and Young, M. A. *Modeling agreement among raters*, Journal of American Statistical Association **80**, 175–180, 1985.
[12] Tanner, M. A. and Young, M. A. *Modeling ordinal scale disagreement*, Psychological Bulletin **98** (2), 408–415, 1985.
[13] von Eye, A. and Von Eye, M. *Can one use Cohen's kappa to examine disagreement?* Methodology **1** (4), 129–142, 2005.
[14] von Eye, A. *An alternative to Cohen's κ*, European Psychologist **11** (1), 12–24, 2006.