

Log-linear Models of Rater Agreement

Alexander von Eye
Michigan State University

List of Contents

1. Cohen's κ
2. A log-linear Base Model
3. Models of Rater Agreement
 - 3.1 Rating Categories have Equal Weight
 - 3.2 Rating Categories have Different Weights
 - 3.3 Models with Covariates
 - 3.4 Rater Agreement plus linear-by-linear Association in Ordinal Variables
 - 3.5 Rater Agreement, linear Association in Ordinal Variables plus Covariate
 - 3.6 More than Two Raters
 - 3.6.1 Simultaneous Agreement in all Rater Pairs
 - 3.6.2 Simultaneous Agreement of all Raters
 - 3.7 Two or More Rating Objects; Two Raters
 - 3.7 Rater-specific Trends
4. Discussion

1. Cohen's κ (1960)

Definition of κ

Proportionate reduction in error (PRE; Fleiss, 1975)

$$\kappa = \frac{\sum p_{ii} - \sum p_{i.} p_{.i}}{1 - \sum p_{i.} p_{.i}}, \quad (1)$$

where:

p_{ii} probability of the i th diagonal cell,

$p_{i.}$ und $p_{.i}$ marginal probabilities,

maximum likelihood estimator for κ (multinomial sampling;

Liebetrau, 1983)

$$\hat{\kappa} = \frac{n \sum_i f_{ii} - \sum_i f_{i.} f_{.i}}{n^2 - \sum_i f_{i.} f_{.i}}, \quad (2)$$

where:

n number of decisions made by each rater

f observed cell frequencies

Properties of κ :

1. Range of κ : $-\infty < \kappa \leq 1$;
2. Smallest possible value: $1 - \frac{n}{n - \sum_{ij} f_{ij}}$ $i \neq j$
3. $\kappa = 0$ if **the numerator = 0**, that is, only if the probability of disagreements (in off-diagonal cells) equals the probability of agreements (in diagonal cells) (κ can be 0, if the judgements are not independent). In different words, $\kappa = 0$ if the probability of agreement is identical to the expected probability for two independent raters.
4. $\kappa = 1$ only if the probability of disagreements is zero

5. κ exists, if both raters use at least two categories, that is, if the probability $p_{ij} > 0$ for at least two cells

6. If the probability of disagreements is greater than 0, the possible maximum value of κ decreases, if the marginal distributions are not uniform

7. If the probability of disagreements decreases and is **less** than the probability of agreements, κ increases monotonically. In contrast, if the probability of disagreements increases and is **greater** than the probability of agreements, then κ will not decrease monotonically (von Eye, & Sørensen, 1991).

8. $\kappa * 100$ is the percent score by which rater agreement is greater than was expected by chance

9. Data Example

In a study on the interpretation of proverbs and sentences (von Eye, Jacobson, and Wills, 1990) two raters made 129 decisions as to whether the content of the sentences/proverbs was concrete: (1) concrete; (2) between concrete and abstract; (3) abstract.

Question: Is the Agreement of the two raters beyond chance?

Table 1: The Raters' Concreteness Ratings

Rater 2 Rater 1		Concreteness			Totals
		1	2	3	
Concreteness	1	11	2	19	32
	2	1	3	3	7
	3	0	8	82	90
	Totals	12	13	104	N = 129

For the data in Table 1 we calculate:

- (1) Likelihood Ratio $X^2 = 39.03$ ($df = 4$; $p < 0.01$). The assumption of rater independence can thus be rejected.

- (2) $\kappa = 0.375$ ($se_{\kappa} = 0.079$; $p < 0.01$). The agreement between these two raters is significantly beyond chance; it exceeds chance agreement by 37.5 %

2. A log-linear Base Model

- I x J cross-classification
 - Raters A and B

Base Model

$$\log m_{ij} = \lambda_0 + \lambda_i^A + \lambda_j^B \quad (3)$$

where:

m_{ij} Expected cell frequency

λ_0 Intercept parameter

λ_i^A Main effect parameter for Rater A

λ_j^B Main effect parameter for Rater B

Analysis of rater agreement: relative to this model, that is,
relative to the model of rater independence
(Different base models can be discussed)

3. Models of Rater Agreement

3.1 Equal Weights (Tanner & Young, 1985)

$$\log m_{ij} = \lambda_0 + \lambda_i^A + \lambda_j^B + \delta(i, j),$$

where:

$\delta(i, j)$ assigns weights to the cells in the main diagonal,
specifically:

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

The resulting model is the *equal-weight agreement model* or *diagonal set model* (Wickens, 1989)

Applying the *equal-weight agreement model* to the data in

Table 1 yields the following design matrix:

$$X = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & -1 & -1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & -1 & -1 & 0 \\ 1 & -1 & -1 & 1 & 0 & 0 \\ 1 & -1 & -1 & 0 & 1 & 0 \\ 1 & -1 & -1 & -1 & -1 & -1 \end{bmatrix}$$

The model fit is:

$$\text{LR-}\chi^2 = 9.22 \text{ (df} = 3; \text{p} = 0.027)$$

→ the model fails to describe the data satisfactorily

3.2 Differentially Weighing Rating Categories

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

for example, $\delta_2 = 2\delta_1$, and $\delta_3 = 3\delta_1$ can be considered.

The choice for these weights implies that the sixth column vector in the design matrix is replaced by the vector

$$x_6' = [1, 0, 0, 0, 2, 0, 0, 0, 3].$$

The resulting model does not describe the data in Table 1 better than the equal weight agreement model:

$$\text{LR-}\chi^2 = 13.49 \text{ (df} = 3; \text{p} = 0.004).$$

Different weights may yield a better description of these data.

3.3 Models with Covariates

$$\log m_{ij} = \lambda_0 + \lambda_i^A + \lambda_j^B + \delta(i, j) + \lambda^C x_{ij},$$

with

λ^C Parameter for covariate x_{ij}

If the weights $\delta(i,j)$ are equal, this model is called the

equal-weight agreement model with covariates

or *diagonal set model with covariates*.

If the weights are unequal, this model is called

differential-weight agreement model with covariates.

To illustrate this model we use the data in Table 1 again. In addition, we use the variable *Wordiness* as a covariate. The

Interpretations of the sentences and proverbs were also evaluated in regard to wordiness with 1 = wordy, 2 = average, and 3 = not wordy. For the nine cells in Table 1 the following average wordiness ratings resulted: $x_w' = [17, 27, 3, 16, 45, 14, 1, 3, 3]$.

In combination with an *equal-weight agreement model* the model fit is excellent:

$$\text{LR-}\chi^2 = 1.85 \text{ (df} = 2; \text{p} = 0.40)$$

this improvement is due solely to the covariate.

The parameter for the covariate is estimated as $\lambda_w = -0.16$

$$(\text{se}_{\lambda_w} = 0.07 ; z = -2.20; \text{p} = 0.036).$$

Rater agreement is significant too:

$$\lambda_a = 3.65 (\text{se}_{\lambda_a} = 1.13; z = 3.23; \text{p} = 0.022).$$

We thus conclude:

- The data in Table 1 can satisfactorily explained from an *equal-weight agreement model* when the covariate of **Wordiness** of interpretation is also considered.
- A differential weight model does not improve this solution:

$$\text{LR-}\chi^2 = 2.64 \text{ (df} = 2; \text{p} = 0.267.)$$

3.4 Rater Agreement plus linear-by-linear Association in Ordinal Variables (Agresti, 1988)

$$\log m_{ij} = \lambda_0 + \lambda_i^A + \lambda_j^B + \beta u_i u_j + \delta(i, j), \quad (5)$$

with:

$u_i < \dots < u_r$ fixed or known Values of response categories,
(ranks, e. g., $u_i = i$)

β Uniform association parameter

Application of the above data example:

(1) Model without term for diagonal cells (uniform association model): $LR-\chi^2 = 13.17$ (df = 3; p = 0.004)

(2) Model with the term for rater agreement:

➤ $LR-\chi^2 = 8.90$; df = 2; p = 0.012

➤ $\Delta\chi^2 = 4.2693$; $\Delta df = 1$; p = 0.039

Both Models:

- significantly better than base model
- not good enough for data explanation

3.5 Rater Agreement, Linear Association in Ordinal Variables

plus Covariate

$$\log m_{ij} = \lambda_0 + \lambda_i^A + \lambda_j^B + \beta u_i u_j + \delta(i, j) + \lambda^C x_{ij}, \quad (6)$$

with

λ^C Parameter for covariate x_{ij}

x_{ij} j th value of the i th covariate (see model in 3.3)

Application to the above data example (equal weight model)

$$\text{LR-}\chi^2 = 1.64 \text{ (df} = 1; \text{p} = 0.201);$$

Parameter for covariate wordiness:

$$\lambda^W = -0.17; \text{se}_\lambda = 0.08; z = -2.10; \text{p} = 0.018$$

Parameter for rater agreement:

$$\lambda_\delta = 3.51; \text{se}_\lambda = 1.25, z = 2.82; \text{p} = 0.002$$

Parameter for linear interaction:

$$\beta = 0.23; \text{se}_\beta = 0.50; z = 0.45; \text{p} = 0.325$$

3.6 More than two Raters

3.6.1 Simultaneous Agreement of all Rater Pairs

Example: three Raters:

$$\log m_{ijk} = \lambda_0 + \lambda_i^A + \lambda_j^B + \lambda_k^C + \delta(i, j, \cdot) + \delta(i, \cdot, k) + \delta(\cdot, j, k) \quad (7)$$

with

$$i, j, k = 1, \dots, I$$

$$I = I_1 = I_2 = I_3$$

Table 2: Three rater of concreteness of interpretations

	1	2	3	Total
1	4	3	6	13
2	2	1	3	6
3	2	2	17	21
Total	8	6	26	40

	1	2	3	Total
1	0	1	2	3
2	1	1	1	3
3	0	0	4	4
Total	1	2	7	10

	1	2	3	Total
1	0	1	3	4
2	0	1	8	9
3	0	4	96	100
Total	0	6	107	113

for the model in (7) (agreement cells **shaded red**) we obtain:

- $LR-\chi^2 = 17.97; df = 17; p = 0.373$
- $\delta(1,2,.) = 0.99; se_{12.} = 0.23; z = 4.26; p < 0.01$
- $\delta(1,.,3) = 1.10; se_{1.3} = 0.31; z = 3.56; p < 0.01$
- $\delta(.,2,3) = 0.71; se_{.23} = 0.28; z = 2.54; p = 0.016$

➔ model can be retained

3.6.2 Simultaneous Agreement of all Raters

Instead of (7) we now use the model:

$$\log m_{ijk} = \lambda_0 + \lambda_i^A + \lambda_j^B + \lambda_k^C + \delta(i, j, k) \quad (8)$$

For the model in (8) (agreement **striped red**) we obtain for the sample data:

- $\text{LR-}\chi^2 = 20.90; \text{df} = 19; \text{p} = 0.35$
- $\delta(1,2,3) = 1.92; \text{se}_{123} = 0.25; z = 7.72; \text{p} < 0.01$
 - $\Delta\chi^2 = 3.34; \text{df} = 3; \text{p} = 0.34$

➔ the model in (8) is more parsimonious and is, therefore, retained

3.7 Two or more Rating Objects; two Raters

$$\log m_{ijk} = \lambda_0 + \lambda_i^A + \lambda_j^B + \lambda_k^C + \delta_1(i, j) + \delta_2(i, j) \quad (9)$$

Data example

Table 3: Wordiness Ratings of two Raters

Rater 2 Rater 1	Wordiness				
Wordiness		1	2	3	Totals
1	17	27	3	47	
2	16	45	14	75	
3	1	3	3	7	
Totals	34	75	20	N = 129	

$$\kappa = 0.11$$

The Data in Table 3 and in Table 1 will now be analyzed simultaneously

Two comparison models:

(1) $\delta_1(i, j) = \delta_2(i, j)$:

- $\text{LR-}\chi^2 = 231.23; \text{df} = 10; p < 0.01$

Model must be rejected

(2) no constraints on δ -values

- $\text{LR-}\chi^2 = 215.14; \text{df} = 10; p < 0.01$

Model must also be rejected

3.7 Rater-Specific Trends

- can be cast as model with covariates
- Example: one rater uses higher ratings than the other

Application to data in Table 3: Assumption that Rater 2 uses higher scores. That is, Rater 2 tends to view interpretations as “not wordy.”

- $LR-\chi^2 = 3.46; df = 2; p = 0.18$
 - Rater agreement:
 $\delta_m = 0.37; se_{\delta_m} = 0.20; z = 1.80; p = 0.04$
 - Trend:
 $\delta_t = 0.82; se_{\delta_t} = 0.48; z = 1.70; p = 0.047$

Model can be retained

4. Discussion

- The models decompose decisions of two or more raters into the three parts
 - (1) Rater agreement (δ parameter)
 - (2) Association model (e.g., β parameter, see Agresti, 1988)
 - (3) Covariates (λ^C parameter)

- More complete description of data becomes possible than with κ alone
 - Additional models and classes of models can be conceptualized, e.g., LCA models

 - Additional generalized Versions of κ become possible (Schuster & von Eye, in preparation)

- Models for rater agreement are useful in particular if the

structure of the probabilities that underlie the cross-classified judgements of two or more raters is complex and can, therefore, not adequately be summarized using a single coefficient such as κ .