Rater agreement - ordinal ratings

Karl Bang Christensen
Dept. of Biostatistics,
Univ. of Copenhagen

NORDSTAT, 2012

http://biostat.ku.dk/~kach/
Rater agreement - ordinal ratings

Methods for analyzing rater agreement are well-established when ratings are dichotomous or if they can be assumed to be normally distributed.

\[ r = 1, \ldots, R \]

\[ s = 1, \ldots, S \]

\[ X_{rs} \in \{0, 1, \ldots, K\} \]
Two raters (R=2)

\[ \begin{array}{ccc}
A & A & A \\
A & A & A \\
\end{array} \]

\( Pr(a) \) observed proportion of agreement, \( \kappa \) coefficient

\[
\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}
\]  

(1)

where

\[
Pr(e) = \sum_k Pr(X_1 = k)Pr(X_2 = k)
\]

is the expected proportion of agreement under independence.

Two raters (R=2)

κ coefficient (1) is widely used

(i) the value depends on the margins and thus on the sample
(ii) if ratings are ordinal this is not taken into account
(iii) no rational for saying that, e.g., κ > 0.7 is good.

κ coefficient (1) is a marginal (population average) measure.

(i) value depends on margins

three tables with agreement about \( \frac{70}{100} \) of the subjects:

\[
\begin{array}{ccc}
X_2 & + & - \\
X_1 & + & 20 & 20 \\
& - & 10 & 50 \\
\kappa = 0.35 \\
\end{array}
\quad
\begin{array}{ccc}
X_2 & + & - \\
X_1 & + & 10 & 20 \\
& - & 10 & 60 \\
\kappa = 0.21 \\
\end{array}
\quad
\begin{array}{ccc}
X_2 & + & - \\
X_1 & + & 5 & 20 \\
& - & 10 & 65 \\
\kappa = 0.08 \\
\end{array}
\]
(ii) if ratings are ordinal this is not taken into account

<table>
<thead>
<tr>
<th>Weighted $\kappa$ coefficient</th>
<th>A</th>
<th>0.75A</th>
<th>0.50A</th>
<th>0.25A</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.75A</td>
<td>A</td>
<td>0.75A</td>
<td>0.50A</td>
<td>0.25A</td>
</tr>
<tr>
<td>0.50A</td>
<td>0.75A</td>
<td>A</td>
<td>0.75A</td>
<td>0.50A</td>
</tr>
<tr>
<td>0.25A</td>
<td>0.50A</td>
<td>0.75A</td>
<td>A</td>
<td>0.75A</td>
</tr>
</tbody>
</table>

Arbitrary weights (two standards implemented in SAS)
Marginal homogeneity

Beyond agreement we would want

\[ \Pr(X_1 = k) = \Pr(X_2 = k) \quad \text{for all } k = 0, 1, \ldots, K \]

Bowkers test of Symmetry tests this hypothesis.

For \( K = 1 \) this is McNemars test.

\[
\frac{\Pr(X_1 = 1, X_2 = 0)}{\Pr(X_1 = 1, X_2 = 0) + \Pr(X_1 = 0, X_2 = 1)}
\]

Continuous data: regression model

\[ X_{rs} = \delta_r + \gamma_s + \epsilon_{rs} \quad \epsilon_{rs} \sim N(0, \omega^2) \]

Limits of agreement / Bland-Altman plot

\[ X_{1s} - X_{2s} = \delta_1 - \delta_2 + (\epsilon_{1s} - \epsilon_{2s}) \]

95% reference interval \( V(\epsilon_{1s} - \epsilon_{2s}) \sim N(0, 2\omega^2) \)
Ordinal data: regression models / IRT

divide-by-total models

\[
\Pr(X_{rs} = x | \theta) = \begin{cases} 
\frac{\exp(x\theta_s - \sum_{k=1}^{x} \beta_{rk})}{\sum_{l} \exp(l\theta_s - \sum_{k=1}^{l} \beta_{rk})}, \\
\frac{\exp(\alpha_r(x\theta_s - \sum_{k=1}^{x} \beta_{rk}))}{\sum_{l} \alpha_r(l\theta_s - \sum_{k=1}^{l} \beta_{rk})}
\end{cases}
\]

\((K = 1: \text{logistic regression})\)

threshold models

\[
\Pr(X_{rs} = x | \theta) = \begin{cases} 
\Phi(\ldots) - \Phi(\ldots), \\
\expit(\ldots) - \expit(\ldots)
\end{cases}
\]

\( \theta = \theta_s \) latent location of subject \( s \)
$\theta = \theta_s$ latent location of subject $s$
$\theta = \theta_s$ latent location of subject $s$
$\theta = \theta_s$ latent location of subject $s$
\[ \theta = \theta_s \text{ latent location of subject } s, \quad (\beta_{rk})_{k=1,...,K} \text{ rater parameters} \]
Marginal homogeneity

Rater parameters \( \beta_r = (\beta_{r1}, \ldots, \beta_{rK}) \). Test

\[ H_0 : \beta_r = \beta \text{ for all } r = 1, \ldots, R \]

using likelihood ratio test based on (2) or (3)

Example: 150 subjects, two raters

<table>
<thead>
<tr>
<th></th>
<th>( X_1 = 0 )</th>
<th>( X_1 = 1 )</th>
<th>( X_1 = 2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X_2 = 0 )</td>
<td>9</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>( X_2 = 1 )</td>
<td>22</td>
<td>59</td>
<td>14</td>
</tr>
<tr>
<td>( X_2 = 2 )</td>
<td>3</td>
<td>25</td>
<td>7</td>
</tr>
</tbody>
</table>

Bowker’s test \( S = 8.6, \ df = 3, \ p = 0.0351 \).
LRT based on (2) \(-2 \log Q = 12.2, \ df = 2, \ p = 0.0022\).
Quantify agreement

Randomly chosen person $s$ with location $\theta_s = \theta$.

Compute reference interval for $|X_{1s} - X_{2s}|$, $\Pr(X_{1s} = X_{2s})$ or $\Pr(|X_{1s} - X_{2s}| > 1)$ if $\theta \sim N(0, \omega^2)$: computations for 'typical' person $\theta = 0$.

Compare to population distribution of $\theta$. 

Example $X \in \{0, 1, 2\}$

marginal homogeneity $H_0 : \beta_{r1} = \beta_{r2}$ accepted.

Common estimate $(\beta_1, \beta_2) = (-0.82, -0.75)$

Table

$$
\begin{bmatrix}
\Pr(X_1 = 0, X_2 = 0|\theta) & \Pr(X_1 = 0, X_2 = 1|\theta) & \Pr(X_1 = 0, X_2 = 2|\theta) \\
\Pr(X_1 = 1, X_2 = 0|\theta) & \Pr(X_1 = 1, X_2 = 1|\theta) & \Pr(X_1 = 1, X_2 = 2|\theta) \\
\Pr(X_1 = 2, X_2 = 0|\theta) & \Pr(X_1 = 2, X_2 = 1|\theta) & \Pr(X_1 = 2, X_2 = 2|\theta)
\end{bmatrix}
$$

(normal latent distribution: Typical patient $\sim \theta = 0$)
Example $X \in \{0, 1, 2\}$

Agreement $\Pr((X_1, X_2) \in \{(1, 1), (2, 2), (3, 3)\} | \theta)$
Examples of clinical applications

Tönnis grade 0,1,2,3: rating of x-rays in hip surgery population.

Modified Ashworth Scale 0,1,2,3,4,5: Spasticity as complications in spinal cord lesion patients (hospital sample). Sparse tables.

Assessment of exercise-induced laryngeal obstruction 0,1,2,3: sub sampling best and worst cases.
Issues

Conditional inference if persons locations cannot be assumed to be normally distributed.

Reduced rank parametrization if tables are sparse.

Interpretation on original scale.
Ordinal data: regression models / IRT
(Marginal or Conditional inference).

divide-by-total models

\[
Pr(X_{rs} = x|\theta) = \left\{ \begin{array}{l}
\frac{\exp(x\theta_s - \sum_{k=1}^{x} \beta_{rk})}{\sum_{l} \exp(l\theta_s - \sum_{k=1}^{l} \beta_{rk})} \\
\frac{\exp(\alpha_r(x\theta_s - \sum_{k=1}^{x} \beta_{rk}))}{\sum_{l} \alpha_r(l\theta_s - \sum_{k=1}^{l} \beta_{rk})}
\end{array} \right.
\]  

\(K = 1: \) logistic regression

threshold models

\[
Pr(X_{rs} = x|\theta) = \left\{ \begin{array}{l}
\Phi(\cdot) - \Phi(\cdot) \\
\expit(\cdot) - \expit(\cdot)
\end{array} \right.
\]

Ordinal data: regression models / IRT
(Marginal or Conditional inference).

divide-by-total models

\[ \Pr(X_{rs} = x|\theta) = \begin{cases} 
\frac{\exp(x\theta_s - \sum_{k=1}^{x} \beta_{rk})}{\sum_{l} \exp(l\theta_s - \sum_{k=1}^{l} \beta_{rk})} & (C, M) \\
\frac{\exp(\alpha_r(x\theta_s - \sum_{k=1}^{x} \beta_{rk}))}{\sum_{l} \alpha_r(l\theta_s - \sum_{k=1}^{l} \beta_{rk})} & (M) 
\end{cases} \]

\( K = 1: \) logistic regression

threshold models

\[ \Pr(X_{rs} = x|\theta) = \begin{cases} 
\Phi(\ldots) - \Phi(\ldots) & (M) \\
\text{expit}(\ldots) - \text{expit}(\ldots) & (M)
\end{cases} \]

let $X_s = (X_{1s}, \ldots, X_{Rs})$ and $x_s = (x_{1s}, \ldots, x_{Rs})$

**Marginal inference**

$$l_M(\beta) = \sum_s \log \int Pr(X_s = x_s|\theta_s)\varphi(\theta_s)$$  \hspace{1cm} (2)

similar to the model yielding Limits of agreement

**Conditional inference**

$$l_C(\beta) = Pr(X_s = x_s|X_1s + \ldots + X_{Rs} = x_{1s} + \ldots + x_{Rs})$$ \hspace{1cm} (3)

similar to the McNemar test.

Reduced rank parametrization

Interpreting and testing differences in rater parameters

\[ \beta_r = (\beta_{rx})_{x=1,...,K} \text{ and } \beta_{r'} = (\beta_{r'x})_{x=1,...,K} \]

can be difficult for \( K = 4, 5, \ldots \)

Reparametrization using 'location' parameter \( \mu_r \) and 'spread' parameter \( \sigma_r \)

\[ \beta_{rx} = \mu_r + (2x - m - 1)\sigma_r. \]

Reduced rank parametrization

Reparametrize $\left(\frac{\beta_1 + \beta_2}{2}, \frac{\beta_2 - \beta_1}{2}\right)$. Hypotheses:

- Raters differ only wrt. location
- Raters differ only wrt. spread
- Raters do not differ
Interpretation on original scale

Probability of agreement across values of $\theta$ can be compared to modeled distribution: $\varphi(\theta)$.

empirical distribution: $\hat{\theta}_1, \ldots, \hat{\theta}_S$ found by maximizing

$$L(\theta) = \Pr_{\hat{\beta}}(X_s = x_s|\theta).$$

values $E(X_{rs}|\theta_s = \theta)$, same for all $r$ under marginal homogeneity.
Example $X \in \{0, 1, 2\}$

Agreement $\Pr((X_1, X_2) \in \{(1, 1), (2, 2), (3, 3)\} | \theta)$
Example $X \in \{0, 1, 2\}$

Agreement $\Pr((X_1, X_2) \in \{(1, 1), (2, 2), (3, 3)\}|\theta)$
Example $X \in \{0, 1, 2\}$

Agreement $\Pr((X_1, X_2) \in \{(1, 1), (2, 2), (3, 3)\}|\theta)$
Example $X \in \{0, 1, 2\}$

Agreement $\Pr((X_1, X_2) \in \{(1, 1), (2, 2), (3, 3)\}|\theta)$

$E(X|\hat{\theta}) = 0.5, 1.0, 1.5$