Measuring agreement between raters

by

Stian Lydersen

Presentation at NTNU
24 January 2014

Slides Updated 22 January 2014

Measures of agreement:
- Categorical data:
  - Cohen’s kappa, alternatives and generalizations.
- Continuous data:
  - Intraclass correlation coefficient (ICC), different versions

Examples
- X-rays rated by radiologists
- Claims for compensation after alleged birth trauma judged by medical experts.
- Video recordings of parent – child interaction. Emotional attachment scored by psychologists.

Categorical data: Cohen’s Kappa

<table>
<thead>
<tr>
<th></th>
<th>Normal</th>
<th>Benign</th>
<th>Suspected cancer</th>
<th>Cancer</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 1</td>
<td>21</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>33</td>
</tr>
<tr>
<td>Rater 2</td>
<td>4</td>
<td>17</td>
<td>1</td>
<td>0</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>9</td>
<td>15</td>
<td>2</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>28</td>
<td>30</td>
<td>21</td>
<td>3</td>
<td>85</td>
</tr>
</tbody>
</table>

Gisev et al (2013), Table 2:
Examples of interrater indices suitable for use with various types of data (not exhaustive)

<table>
<thead>
<tr>
<th>Level of measurement</th>
<th>Nominal / categorical</th>
<th>Ordinal</th>
<th>Interval and ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 raters</td>
<td>Cohen’s kappa</td>
<td>ICC</td>
<td>Bland-Altman plots</td>
</tr>
<tr>
<td></td>
<td>Weighted kappa</td>
<td>ICC</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>ICC</td>
<td></td>
</tr>
<tr>
<td>&gt;2 raters</td>
<td>Fleiss’ kappa</td>
<td>Kendall’s coefficient of concordance</td>
<td>ICC</td>
</tr>
<tr>
<td></td>
<td>ICC</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Now, consider a situation where two raters each classify subjects in \( c \) categories, numbered from 1 to \( c \). Let \( p_{ij} \) denote the probability that a subject is classified in category \( i \) by rater 1 and category \( j \) by rater 2, respectively. An intuitive measure of agreement is the probability that the raters agree, which is
\[
P_{agreement} = \sum_{i=1}^{c} p_{ii}.
\]

But part of this agreement is due to chance. Suppose that rater 1 assigns to category \( i \) with probability \( p_i = \sum_{j=1}^{c} p_{ij} \) and rater 2 assigns to category \( j \) with probability \( p_j = \sum_{i=1}^{c} p_{ij} \) independently of rater 1. Then, Cohen’s probability of agreement by chance is given by
\[
P_{chance} = \sum_{i=1}^{c} p_i p_j.
\]

Cohen’s kappa is defined as the relative proportion of agreements exceeding that by chance, which is
\[
kappa = \frac{P_{agreement} - P_{chance}}{1 - P_{chance}}.
\]

If only two categories:

Table 14.9: Assessments of 85 mammograms by two radiologists, dichotomized in two categories based on Table 14.8.

<table>
<thead>
<tr>
<th>Rater 1</th>
<th>Normal or benign</th>
<th>Suspected cancer or cancer</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal or benign</td>
<td>34</td>
<td>1</td>
<td>35</td>
</tr>
<tr>
<td>Suspected cancer or cancer</td>
<td>12</td>
<td>18</td>
<td>30</td>
</tr>
<tr>
<td>Total</td>
<td>46</td>
<td>19</td>
<td>85</td>
</tr>
</tbody>
</table>

\( \kappa = 0.63 \) for two categories

\( \kappa = 0.47 \) when using all four categories.

A weighted kappa, described later, may be more appropriate for ordered categories.

Example: Table 14.7:

Estimated agreement proportion:
\[
\hat{p} = (21 + 17 + 15 + 1)/85 = 54/85 = 0.64
\]

Cohen’s probability of agreement by chance:
\[
\hat{p}_c = (28 \times 33 + 38 \times 22 + 16 \times 29 + 3 \times 1)/85^2 = 0.31
\]

Cohen’s kappa:
\[
k = \frac{0.64 - 0.31}{1 - 0.31} = 0.47
\]

Interpretation of kappa values


<table>
<thead>
<tr>
<th>Value of ( k )</th>
<th>Strength of agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.20 – 0.40</td>
<td>Fair</td>
</tr>
<tr>
<td>0.41 – 0.60</td>
<td>Moderate</td>
</tr>
<tr>
<td>0.61 – 0.80</td>
<td>Good</td>
</tr>
<tr>
<td>0.81 – 1.00</td>
<td>Very good</td>
</tr>
</tbody>
</table>

Recommendation:
Show the original table data, not only the measure of agreement.
Confidence intervals for Cohen’s kappa

The approximate standard error of kappa for dichotomous or nominal categories is given by Altman et al. (2000) as

\[ SE(k) = \sqrt{\frac{\hat{p}(1-\hat{p})}{N(1-\hat{p})}} \]

An approximate 1 - α confidence interval is given by \( \hat{k} \pm z_{\alpha/2} SE(k) \).

A 95% CI based on the data in Table 14.9 is (0.45, 0.82). Some software uses other formulae, see Lydersen (2012) and references therein.

Cohen’s kappa: Unexpected results or paradoxes.

- Depends on the number of categories, especially for nominal categories
- Depends on the marginal distribution (prevalence) of the categories
- Raters who disagree more on the marginal distribution may produce higher kappa values

Kappa depends on the marginal distribution:

Table 14.9

<table>
<thead>
<tr>
<th>Rater 2</th>
<th>Normal</th>
<th>Cancer</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>54</td>
<td>1</td>
<td>55</td>
</tr>
<tr>
<td>Cancer</td>
<td>12</td>
<td>18</td>
<td>30</td>
</tr>
<tr>
<td>Total</td>
<td>66</td>
<td>19</td>
<td>85</td>
</tr>
</tbody>
</table>

\( \hat{k} = 0.63 \)

Table 14.10

<table>
<thead>
<tr>
<th>Rater 2</th>
<th>Normal</th>
<th>Cancer</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>68</td>
<td>1</td>
<td>69</td>
</tr>
<tr>
<td>Cancer</td>
<td>12</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>Total</td>
<td>80</td>
<td>5</td>
<td>85</td>
</tr>
</tbody>
</table>

\( \hat{k} = 0.32 \)

Cohen’s weighted kappa:

Weights the degree of agreement (distance from the diagonal)

Linear weighted kappa:

\[ w_j = 1 - \frac{|x - j|}{c-1} \]

With 4 categories, the weights are 1 on the diagonal, and 2/3, 1/3 and 0 off the diagonal:

Quadratic weighted kappa:

\[ w_j = 1 - \frac{((x - j)^2)}{(c-1)} \]

With 4 categories, the weights are 1 on the diagonal, and 8/9, 5/9 and 0 off the diagonal.

Unweighted kappa:

The weights are 1 on the diagonal, and always 0 off the diagonal.
Linear versus quadratic weighted kappa?

- No clear advice in the literature
- For the case of equal marginal distributions, that is, \( n_i = n_j \) for all \( i \), then the quadratic weighted \( \hat{\kappa}_i \) is equal to the intraclass correlation coefficient \( \hat{\text{ICC}}_i \) described in Section 14.8, except for a term involving the factor \( 1/N \).

Table 14.8: Assessments of 85 xeromammograms by two radiologists

<table>
<thead>
<tr>
<th>Rater 1</th>
<th>Normal</th>
<th>Benign</th>
<th>Suspect cancer</th>
<th>Cancer</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>21</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>33</td>
</tr>
<tr>
<td>Benign</td>
<td>4</td>
<td>17</td>
<td>1</td>
<td>0</td>
<td>22</td>
</tr>
<tr>
<td>Suspected cancer</td>
<td>3</td>
<td>9</td>
<td>15</td>
<td>2</td>
<td>29</td>
</tr>
<tr>
<td>Cancer</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>28</td>
<td>38</td>
<td>16</td>
<td>3</td>
<td>85</td>
</tr>
</tbody>
</table>

Unweighted kappa: 0.47
Linear weighted kappa: 0.57
Quadratic weighted kappa: 0.67
User-defined (example) 0.59

Dichotomized table kappa (Table 14.9): 0.63

Categorical data: Alternatives to Cohen’s kappa

- Assuming independence between raters:
  - Cohen’s kappa (1960)
  - Scott’s pi (1955)
  - Bennett’s sigma (1954)

- Assuming some subjects are easy, other difficult to agree on:
  - Gwet’s AC1 (Gwet’s gamma) (2001, 2008)
  - Aickin’s alpha (1990)
  - Martin and Femia’s Delta (2004, 2008) for multiple choice tests

Measures which differ only in terms of calculating chance agreement:

Cohen’s kappa (1960) uses the product of the marginals,
\[
\hat{p}_c = \sum \hat{p}_i \hat{p}_i
\]
where \( \hat{p}_i = n_i / n \), and \( \hat{p}_c = n_c / n \)

Scott’s pi (1955) uses the squared average of the marginals,
\[
\hat{p}_c = \sum \left( \frac{\hat{p}_i + \hat{p}_i}{2} \right)^2
\]

Bennet’s sigma (1954) assumes a uniform marginal:
\[
\hat{p}_c = \frac{1}{c}
\]

Gwet’s gamma (2001, 2008) (Also called Gwet’s AC1):
\[
\hat{p}_c = \frac{1}{c-1} \sum \hat{p}_i (1 - \hat{p}_c)
\]
where \( \hat{p}_c = \frac{n_c}{n} \), and \( \hat{p}_c = \frac{n_c}{n} \)

When \( c = 2 \), the equation reduces to
\[
\hat{p}_c = 2\hat{p}_1 \hat{p}_2
Gwet’s gamma and Aickin’s alpha:

Easy subjects to classify (E) will be classified (deterministic) in the same category by both raters.

Hard subjects to classify (H) will be random classified. Probability \(1/c\) for each of the \(c\) categories.

Aickin assumes each subject is either hard for both raters (HH), or easy for both raters (EE).

Gwet allows also a subject to be hard for Rater 1 and easy for Rater 2 (HE), or vice versa (EH)

Possible outcomes with Aickin’s theory (Gwet, 2012):

<table>
<thead>
<tr>
<th>Rater A</th>
<th>Hard Subjects</th>
<th>Easy Subjects</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hard</td>
<td>(\sum_{i=1}^{c} N_{EH}^{i})</td>
<td>(\sum_{i=1}^{c} N_{EH}^{i})</td>
<td>(N_{HH})</td>
</tr>
<tr>
<td>Subjects</td>
<td>(\sum_{i=1}^{c} N_{HE}^{i})</td>
<td>(\sum_{i=1}^{c} N_{HE}^{i})</td>
<td>(N_{EE})</td>
</tr>
<tr>
<td>Easy</td>
<td>(\sum_{i=1}^{c} N_{HE}^{i})</td>
<td>(\sum_{i=1}^{c} N_{HE}^{i})</td>
<td>(N_{HE})</td>
</tr>
<tr>
<td>Subjects</td>
<td>(\sum_{i=1}^{c} N_{HE}^{i})</td>
<td>(\sum_{i=1}^{c} N_{HE}^{i})</td>
<td>(N_{EE})</td>
</tr>
<tr>
<td>Total</td>
<td>(\sum_{i=1}^{c} N_{EH}^{i})</td>
<td>(\sum_{i=1}^{c} N_{EH}^{i})</td>
<td>(N)</td>
</tr>
</tbody>
</table>

Possible outcomes with Gwet’s theory (Gwet, 2012):

Gwet’s gamma:

\[
\gamma = \frac{\sum_{i=1}^{c} N_{EH}^{i} - \left(\frac{1}{c}\right)\sum_{i=1}^{c} N_{EH}^{i}}{\left(\frac{1}{c}\right)\sum_{i=1}^{c} N_{EH}^{i}}
\]

Aickin’s alpha:

\[
\alpha = \frac{\sum_{i=1}^{c} N_{EH}^{i}}{N}
\]
Multiple choice tests:
Assume the student knows, say, 40% of the answers ($\Delta = 0.4$). He/she will answer 40% correct, and randomly choose the answers for the remaining questions.

Martin and Femia (2004) suggested this estimator:

$$\hat{\Delta} = \hat{p}_1 + \hat{p}_2 - 2\sqrt{\hat{p}_1 \hat{p}_2}$$

Table 14.9

<table>
<thead>
<tr>
<th>(Rater 1)</th>
<th>(Rater 2)</th>
<th>Normal</th>
<th>Cancer</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>54</td>
<td>1</td>
<td>18</td>
<td>30</td>
</tr>
<tr>
<td>Cancer</td>
<td>12</td>
<td>1</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>Total</td>
<td>66</td>
<td>19</td>
<td>22</td>
<td>85</td>
</tr>
</tbody>
</table>

$$\kappa = 0.635, \tilde{\kappa} = 0.627, \sigma = 0.694, \gamma = 0.741, \Delta = 0.766$$

Table 14.10

<table>
<thead>
<tr>
<th>(Rater 1)</th>
<th>(Rater 2)</th>
<th>Normal</th>
<th>Cancer</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>68</td>
<td>1</td>
<td>69</td>
<td></td>
</tr>
<tr>
<td>Cancer</td>
<td>12</td>
<td>4</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>80</td>
<td>5</td>
<td>85</td>
<td></td>
</tr>
</tbody>
</table>

$$\kappa = 0.320, \tilde{\kappa} = 0.294, \sigma = 0.694, \gamma = 0.805, \Delta = 0.766$$

Table 14.11: Symmetrical imbalance

<table>
<thead>
<tr>
<th>(\text{Rater 1})</th>
<th>(\text{Rater 2})</th>
<th>(\text{disease})</th>
<th>(\text{healthy})</th>
<th>(\text{Total})</th>
</tr>
</thead>
<tbody>
<tr>
<td>disease</td>
<td>50</td>
<td>10</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>healthy</td>
<td>20</td>
<td>20</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>70</td>
<td>30</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

$$\kappa = 0.348, \tilde{\kappa} = 0.341, \sigma = 0.400, \gamma = 0.450, \Delta = 0.417$$

Table 14.12: Asymmetrical imbalance

<table>
<thead>
<tr>
<th>(\text{Rater 1})</th>
<th>(\text{Rater 2})</th>
<th>(\text{disease})</th>
<th>(\text{healthy})</th>
<th>(\text{Total})</th>
</tr>
</thead>
<tbody>
<tr>
<td>disease</td>
<td>30</td>
<td>30</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>healthy</td>
<td>0</td>
<td>40</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>30</td>
<td>70</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

$$\kappa = 0.444, \tilde{\kappa} = 0.394, \sigma = 0.400, \gamma = 0.406, \Delta = 0.700 (or 0.585)$$

Comparisons of measures for 2 raters:

(Ato, Lopez, & Benavente 2011) compare measures in terms of their ability to estimate the systematic agreement proportion. Hence, the construct (estimand) is $\Delta$ (7).

Recommend Bennett’s sigma, and Martin and Femia’s Delta (of course), since these have least bias.

(Wongpakaran, Wongpakaran, Wedding, & Gwet 2013) compare Cohen’s kappa and Gwet’s gamma. “Our results favored Gwet’s method over Cohen’s kappa with regard to prevalence or marginal probability problem.”

But ref [18] only illustrates that AC1 is resistant to the prevalence paradox.

Gwet’s gamma is paradox-resistant (Gwet, 2012)

Wongpakaran, Wongpakaran, Wedding and Gwet (2013):

“It is interesting to note that although Gwet proved that the AC1 is better than Cohen’s Kappa in 2001, a finding subsequently confirmed by biostatisticians [18], few researchers have used AC1 as a statistical tool, or are even aware of it, especially in the medical field.”

Categorical data: Generalizations to more than two raters
More than two raters

- No unique way to generalize
- Fleiss’ kappa (1971) is a generalization of Scott’s pi
- Conger’s chance agreement probability (1980) is a generalization of Cohen’s kappa. Computations are time-consuming if more than three raters.
- Gwet (2012, page 31) recommends using Fleiss’ kappa

Continuous data: Intraclass correlation coefficient (ICC)

The intraclass correlation coefficient (ICC)

- Measures the correlation between one measurement on a subject and another measurement on the same subject (Shrout and Fleiss, 1979).
- Several ICC versions exist for different study designs and study aims
- The term ICC is also used in other settings, such as replicated measurements per subject, or patients within clinics.

Three study designs:

Case 1:
Each subject is rated by a different set of \( k \) raters, randomly selected from a larger population of raters.

Case 2:
A random sample of \( k \) raters is selected from a larger population of raters. Each subject is rated by each rater.

Case 3:
There are only \( k \) raters of interest. Each subject is rated by each rater.

We shall limit our focus to agreement between single measurements, without interaction, and we use the notation \( ICC(3,k) \) of Barnhart et al. (2007) in Table 14.14.

Alternatively, agreement can be defined for average of \( k \) measurements.

The intraclass correlation \( ICC(3,k) \) in Table 14.14 is equivalent to Cronbach’s alpha, a commonly used measure of the internal consistency of items on a psychometric scale.
Case 1:  
One-way random effect model  
\[ X_{ij} = \mu + b_i + w_{ij} \]

where  
\[ X_{ij} \] is rating number \( j \) on subject number \( i \),  
\[ b_i \sim \mathcal{N}(0, \sigma_b^2) \] is the random effect of subject number \( i \),  
\[ w_{ij} \sim \mathcal{N}(0, \sigma_u^2) \] is a residual term.

In case 1, 2, and 3, all random effects and residual terms are assumed independent.

The correlation between two ratings \( X_{ij} \) and \( X_{ij'} \) on subject number \( i \) is  
\[
\text{ICC}_1 = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_u^2 + \sigma_e^2}
\]

In Case 1, \( w_{ij} \) includes a rater effect and an error term. In cases 2 and 3, the components of \( w_{ij} \) are specified:

\[ X_{ij} = \mu + b_i + c_j + \epsilon_{ij} \]

where  
\[ c_j \] is the effect of rater \( j \),  
\[ \epsilon_{ij} \] is the residual random error.  
Case 2:  \[ c_j \sim \mathcal{N}(0, \sigma_c^2) \]

Case 3, \( c_j \) is a fixed effect with constraint  
\[
\sum_{j=1}^{J} c_j = 0
\]

The correlation between two ratings \( X_{ij} \) and \( X_{ij'} \) on subject number \( i \) is  
\[
\text{ICC}_2 = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_c^2 + \sigma_e^2}
\]

Be aware of:  
– ICC, like Pearson’s correlation coefficient, is highly influenced by the variability in the subjects. The larger variation between subjects, and ICC will be closer to one.  
– ICC combines any systematic difference between the raters and the random measurement variation, in one measure.  
– If the purpose is to compare two measurement methods rather than two raters, Bland and Altman (1986) recommend to not use a correlation coefficient. Rather, they recommend plotting the difference between two measurements as a function of their mean, commonly termed a Bland-Altman plot.

Example: Video recordings of parent–child interaction.  
– An RCT of Marte Meo versus treatment as usual  
– Three time points: Baseline, 2 months, and 8 months  
– Emotional attachment (EA) score based on video recording of parent–child interaction. Rating scored by a psychologist or psychiatrist.
Design of Interrater reliability (IRR) study

- 36 distinct individuals, 12 from each of 3 time points.
- Each was rated by 2 raters, from a pool of 4 raters.
- All 6 combinations of raters rated 2 individuals at each of the 3 time points.

Design ... (continued)

- Three first-raters (A, B, C) at each time point.
- Four second-raters at each time point (A, B, C, D)
- At each time point 12 pairs of raters.

AD  AB
BD  BA
CD  BC
AD  CB
BD  AC
CD  CA

Linear model with crossed random effects of individual and rater
Score on individual i by rater j:

\[ X_{ij} = \beta_i + \beta_{time} + \beta_{rating} + b_j + c_i + \epsilon_{ij} \]

Analyzed in Stata as described by Rabe-Hesketh & Skrondal (2012), page 437-441.
(Show results from Word document)

References

Gwet, K.L. 2012. Handbook of Interrater Reliability, 3 ed. Gaithersburg, Maryland, AdvancedAnalytics, LLC.

Effect of rating 2 versus rating 1 on same individual?

\[ X_{ij} = \beta_i + \beta_{time} + \beta_{rating} + b_j + c_i + \epsilon_{ij} \]