Assessing inter-rater agreement in Stata

Daniel Klein
klein.daniel.81@gmail.com
klein@incher.uni-kassel.de

University of Kassel
INCHER-Kassel

15th German Stata Users Group meeting
Berlin
June 23, 2017
Interrater agreement and Cohen’s Kappa: A brief review

Generalizing the Kappa coefficient

More agreement coefficients

Statistical inference and benchmarking agreement coefficients

Implementation in Stata

Examples
An imperfect working definition

Define interrater agreement as the propensity for two or more raters (coders, judges, . . . ) to, independently from each other, classify a given subject (unit of analysis) into the same predefined category.
Interrater agreement

How to measure it?

Consider

- $r = 2$ raters
- $n$ subjects
- $q = 2$ categories

<table>
<thead>
<tr>
<th>Rater A</th>
<th>Rater B</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$n_{11}$</td>
<td>$n_{12}$</td>
</tr>
<tr>
<td>2</td>
<td>$n_{21}$</td>
<td>$n_{22}$</td>
</tr>
<tr>
<td>Total</td>
<td>$n.1$</td>
<td>$n.2$</td>
</tr>
</tbody>
</table>

The observed proportion of agreement is

$$p_o = \frac{n_{11} + n_{22}}{n}$$
Cohen’s Kappa
The problem of chance agreement

The problem

▶ Observed agreement may be due to . . .
  ▶ subject properties
  ▶ chance

Cohen’s (1960) solution

▶ Define the proportion of agreement expected by chance as

\[ p_e = \frac{n_1}{n} \times \frac{n_1}{n} + \frac{n_2}{n} \times \frac{n_2}{n} \]

▶ Then define Kappa as

\[ \kappa = \frac{p_o - p_e}{1 - p_e} \]
Cohen’s Kappa
Partial agreement and weighted Kappa

The Problem

- For $q > 2$ (ordered) categories raters might partially agree
- The Kappa coefficient cannot reflect this

Cohen’s (1968) solution

- Assign a set of weights to the cells of the contingency table
  - Define linear weights
    \[
    w_{kl} = 1 - \frac{|k - l|}{|q_{max} - q_{min}|}
    \]
  - Define quadratic weights
    \[
    w_{kl} = 1 - \frac{(k - l)^2}{(q_{max} - q_{min})^2}
    \]
Cohen’s Kappa
Quadratic weights (Example)

- Weighting matrix for $q = 4$ categories
- Quadratic weights

<table>
<thead>
<tr>
<th>Rater A</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.89</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.56</td>
<td>0.89</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.00</td>
<td>0.56</td>
<td>0.89</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Generalizing Kappa

Missing ratings

The problem

- Some subjects classified by only one rater
- Excluding these subjects reduces accuracy

Gwet’s (2014) solution

(Also see Krippendorff 1970, 2004, 2013)

- Add a dummy category, $X$, for missing ratings
- Base $p_o$ on subjects classified by both raters
- Base $p_e$ on subjects classified by one or both raters

- Potential problem: no explicit assumption about type of missing data (MCAR, MAR, MNAR)
# Missing ratings

## Calculation of $p_o$ and $p_e$

<table>
<thead>
<tr>
<th>Rater A</th>
<th>1</th>
<th>2</th>
<th>...</th>
<th>$q$</th>
<th>$X$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$n_{11}$</td>
<td>$n_{12}$</td>
<td>...</td>
<td>$n_{1q}$</td>
<td>$n_{1X}$</td>
<td>$n_{1.}$</td>
</tr>
<tr>
<td>2</td>
<td>$n_{21}$</td>
<td>$n_{22}$</td>
<td>...</td>
<td>$n_{2q}$</td>
<td>$n_{2X}$</td>
<td>$n_{2.}$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$q$</td>
<td>$n_{q1}$</td>
<td>$n_{q2}$</td>
<td>...</td>
<td>$n_{qq}$</td>
<td>$n_{qX}$</td>
<td>$n_{q.}$</td>
</tr>
<tr>
<td>$X$</td>
<td>$n_{X1}$</td>
<td>$n_{X2}$</td>
<td>...</td>
<td>$n_{Xq}$</td>
<td>0</td>
<td>$n_{X.}$</td>
</tr>
<tr>
<td>Total</td>
<td>$n_{.1}$</td>
<td>$n_{.2}$</td>
<td>...</td>
<td>$n_{.q}$</td>
<td>$n_{.X}$</td>
<td>$n$</td>
</tr>
</tbody>
</table>

- Calculate $p_o$ and $p_e$ as

$$p_o = \sum_{k=1}^{q} \sum_{l=1}^{q} \frac{w_{kl}n_{kl}}{n - (n_{.X} + n_{X.})}$$

and

$$p_e = \sum_{k=1}^{q} \sum_{l=1}^{q} w_{kl} \frac{n_{k.}}{n - n_{.X}} \times \frac{n_{.l}}{n - n_{X.}}$$
Consider three pairs of raters \{A, B\}, \{A, C\}, \{B, C\}

Agreement might be observed for . . .

- 0 pairs
- 1 pair
- all 3 pairs

It is not possible for only two pairs to agree

Define agreement as average agreement over all pairs

- here $0, 0.33$ or $1$

With $r = 3$ raters and $q = 2$ categories, $p_o \geq \frac{1}{3}$ by design
Three or more raters

Observation agreement

Organize the data as \( n \times q \) matrix

<table>
<thead>
<tr>
<th>Subject</th>
<th>Category</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>1</td>
<td>( r_{11} )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>( i )</td>
<td>( r_{i1} )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>( n )</td>
<td>( r_{n1} )</td>
<td>( \ldots )</td>
</tr>
</tbody>
</table>

Average \
\( \bar{r}_{1.} \) | \( \ldots \) | \( \bar{r}_{k.} \) | \( \ldots \) | \( \bar{r}_{q.} \) | \( \bar{r} \)

Average observed agreement over all pairs of raters

\[
p_o = \frac{1}{n'} \sum_{i=1}^{n'} \sum_{k=1}^{q} \sum_{l=1}^{q} \frac{r_{ik}(w_{kl}r_{il} - 1)}{r_i(r_i - 1)}
\]
Three or more raters

Chance agreement

- **Fleiss (1971) expected proportion of agreement**

\[
pe = \sum_{k=1}^{q} \sum_{l=1}^{q} w_{kl} \pi_k \pi_l
\]

with

\[
\pi_k = \frac{1}{n} \sum_{i=1}^{n} \frac{r_{ik}}{r_i}
\]

- **Fleiss’ Kappa does not reduce to Cohen’s Kappa**
  - It instead reduces to Scott’s \(\pi\)
  - Conger (1980) generalizes Cohen’s Kappa
    (formula somewhat complex)
Generalizing Kappa
Any level of measurement

- Krippendorff (1970, 2004, 2013) introduces more weights (calling them difference functions)
  - ordinal
  - ratio
  - circular
  - bipolar
- Gwet (2014) suggests

<table>
<thead>
<tr>
<th>Data metric</th>
<th>Weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>nominal/categorical</td>
<td>none (identity)</td>
</tr>
<tr>
<td>ordinal</td>
<td>ordinal</td>
</tr>
<tr>
<td>interval</td>
<td>linear, quadratic, radical</td>
</tr>
<tr>
<td>ratio</td>
<td>any</td>
</tr>
</tbody>
</table>

- Rating categories must be predefined
More agreement coefficients

A general form

- Gwet (2014) discusses (more) agreement coefficients of the form

\[ \kappa = \frac{p_o - p_e}{1 - p_e} \]

- Differences only in chance agreement \( p_e \)
  - Brennan and Prediger (1981) coefficient (\( \kappa_n \))

\[ p_e = \frac{1}{q^2} \sum_{k=1}^{q} \sum_{l=1}^{q} w_{kl} \]

- Gwet’s (2008, 2014) AC (\( \kappa_G \))

\[ p_e = \frac{\sum_{k=1}^{q} \sum_{l=1}^{q} w_{kl}}{q(q - 1)} \sum_{k=1}^{q} \pi_k \left( 1 - \pi_k \right) \]
Gwet (2014) obtains Krippendorff’s alpha as

\[
\kappa_\alpha = \frac{p_o - p_e}{1 - p_e}
\]

with

\[
p_o = \left(1 - \frac{1}{n'\bar{r}}\right)p'_o + \frac{1}{n'\bar{r}}
\]

where

\[
p'_o = \frac{1}{n'} \sum_{i=1}^{n'} \sum_{k=1}^{q} \sum_{l=1}^{q} \frac{r_{ik} (w_{kl}r_{il} - 1)}{\bar{r} (r_i - 1)}
\]

and

\[
p_e = \sum_{k=1}^{q} \sum_{l=1}^{q} w_{kl} \pi'_k \pi'_l
\]

with

\[
\pi'_k = \frac{1}{n'} \sum_{i=1}^{n'} \frac{r_{ik}}{\bar{r}}
\]
Statistical inference
Approaches

- Model-based (analytic) approach
  - based on theoretical distribution under $H_0$
  - not necessarily valid for confidence interval construction

- Bootstrap
  - valid confidence intervals with few assumptions
  - computationally intensive

- Design-based (finite population)
  - First introduced by Gwet (2014)
  - sample of subjects drawn from subject universe
  - sample of raters drawn from rater population
Inference conditional on the sample of raters

\[ V(\kappa) = \frac{1 - f}{n(n-1)} \sum_{i=1}^{n} (\kappa_i^* - \kappa)^2 \]

where

\[ \kappa_i^* = \kappa_i - 2 (1 - \kappa) \frac{p_{ei} - p_e}{1 - p_e} \]

with

\[ \kappa_i = \frac{n}{n'} \times \frac{p_{oi} - p_e}{1 - p_e} \]

\( p_{ei} \) and \( p_{oi} \) are the subject-level expected and observed agreement
How do we interpret the extent of agreement?

Landis and Koch (1977) suggest

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 0.00</td>
<td>Poor</td>
</tr>
<tr>
<td>0.00 to 0.20</td>
<td>Slight</td>
</tr>
<tr>
<td>0.21 to 0.40</td>
<td>Fair</td>
</tr>
<tr>
<td>0.41 to 0.60</td>
<td>Moderate</td>
</tr>
<tr>
<td>0.61 to 0.80</td>
<td>Substantial</td>
</tr>
<tr>
<td>0.81 to 1.00</td>
<td>Almost Perfect</td>
</tr>
</tbody>
</table>

Similar scales proposed (e.g., Fleiss 1981, Altman 1991)
Benchmarking agreement coefficients
Probabilistic approach

The Problem

- Precision of estimated agreement coefficients depends on
  - the number of subjects
  - the number of raters
  - the number of categories
- Common practice of benchmarking ignores this uncertainty

Gwet’s (2014) solution

- Probabilistic benchmarking method
  1. Compute the probability for a coefficient to fall into each benchmark interval
  2. Calculate the cumulative probability, starting from the highest level
  3. Choose the benchmark interval associated with a cumulative probability larger than a given threshold
Interrater agreement in Stata

Kappa

- `kap, kappa` (StataCorp.)
  - Cohen’s Kappa, Fleiss Kappa for three or more raters
  - Casewise deletion of missing values
  - Linear, quadratic and user-defined weights (two raters only)
  - No confidence intervals

- `kapci` (SJ)
  - Analytic confidence intervals for two raters and two ratings
  - Bootstrap confidence intervals

- `kappci` (`kaputil`, SSC)
  - Confidence intervals for binomial ratings (uses `ci` for proportions)

- `kappa2` (SSC)
  - Conger’s (weighted) Kappa for three or more raters
  - Uses available cases
  - Jackknife confidence intervals
  - Majority agreement
Interrater agreement in Stata
Krippendorff’s alpha

- krippalpha (SSC)
  - Ordinal, quadratic and ratio weights
  - No confidence intervals
- kalpha (SSC)
  - Ordinal, quadratic, ratio, circular and bipolar weights
  - (Pseudo-) bootstrap confidence intervals (not recommended)
- kanom (SSC)
  - Two raters with nominal ratings only
  - No weights (for disagreement)
  - Confidence intervals (delta method)
  - Supports basic features of complex survey designs
Interrater agreement in Stata
Kappa, etc.

- kappaetc (SSC)
  - Observed agreement, Cohen and Conger’s Kappa, Fleiss’ Kappa, Krippendorff’s alpha, Brennan and Prediger coefficient, Gwet’s AC
  - Uses available cases, optional casewise deletion
  - Ordinal, linear, quadratic, radical, ratio, circular, bipolar, power, and user-defined weights
  - Confidence intervals for all coefficients (design-based)
  - Standard errors conditional on sample of subjects, sample of raters, or unconditional
  - Benchmarking estimated coefficients (probabilistic and deterministic)
  - ...
Kappa paradoxes
Dependence on marginal totals

<table>
<thead>
<tr>
<th>Rater A</th>
<th>Rater B</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>60</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>40</td>
</tr>
<tr>
<td>Total</td>
<td>70</td>
<td>100</td>
</tr>
</tbody>
</table>

\[
p_o = 0.60 \\
k_n = 0.20 \\
k = 0.13 \\
k_F = 0.12 \\
k_G = 0.27 \\
k_\alpha = 0.13
\]

<table>
<thead>
<tr>
<th>Rater A</th>
<th>Rater B</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25</td>
<td>60</td>
</tr>
<tr>
<td>2</td>
<td>35</td>
<td>40</td>
</tr>
<tr>
<td>Total</td>
<td>30</td>
<td>100</td>
</tr>
</tbody>
</table>

\[
p_o = 0.60 \\
k_n = 0.20 \\
k = 0.26 \\
k_F = 0.19 \\
k_G = 0.21 \\
k_\alpha = 0.20
\]

Tables from Feinstein and Cicchetti 1990
Kappa paradoxes

High agreement, low Kappa

<table>
<thead>
<tr>
<th>Rater A</th>
<th>Rater B</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>118</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>120</td>
<td>5</td>
</tr>
</tbody>
</table>

$p_o = 0.94$

$\kappa_n = 0.89$

$\kappa = -0.02$

$\kappa_F = -0.03$

$\kappa_G = 0.94$

$\kappa_\alpha = -0.02$

Table from Gwet 2008
### Kappa paradoxes

Independence of center cells, row and columns with quadratic weights

<table>
<thead>
<tr>
<th>Rater A</th>
<th>Rater B</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>6</td>
<td>18</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rater A</th>
<th>Rater B</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>17</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>6</td>
<td>18</td>
</tr>
</tbody>
</table>

\[
p_0 = 0.10 \quad p_{ow2} = 0.70 \quad \kappa_{n_w2} = 0.10 \\
\kappa_{w2} = 0.00 \quad \kappa_{F_w2} = -0.05 \quad \kappa_{G_w2} = 0.15 \\
\kappa_{\alpha_w2} = -0.03 \quad \kappa_{n_w2} = 0.53 \\
\kappa_{w2} = 0.00 \quad \kappa_{F_w2} = 0.00 \quad \kappa_{G_w2} = 0.69 \\
\kappa_{\alpha_w2} = 0.02
\]

Tables from Warrens 2012
. tabi 75 1 4 \ 5 4 1 \ 0 0 10 , nofreq replace
. expand pop
(2 zero counts ignored; observations not deleted)
(93 observations created)
. drop if !pop
(2 observations deleted)
. rename (row col) (ratera raterb)
. tabulate ratera raterb

<table>
<thead>
<tr>
<th>ratera</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>75</td>
<td>1</td>
<td>4</td>
<td>80</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Total</td>
<td>80</td>
<td>5</td>
<td>15</td>
<td>100</td>
</tr>
</tbody>
</table>
### Benchmarking
### Interrater agreement

#### kappaetc ratera raterb

**Interrater agreement**

|                        | Coef. | Std. Err. | t     | P>|t|  | [95% Conf. Interval] |
|------------------------|-------|-----------|-------|------|---------------------|
| Percent Agreement      | 0.8900 | 0.0314    | 28.30 | 0.000 | 0.8276   0.9524     |
| Brennan and Prediger   | 0.8350 | 0.0472    | 17.70 | 0.000 | 0.7414   0.9286     |
| Cohen/Conger´s Kappa   | 0.6765 | 0.0881    | 7.67  | 0.000 | 0.5016   0.8514     |
| Fleiss´s Kappa         | 0.6753 | 0.0891    | 7.58  | 0.000 | 0.4985   0.8520     |
| Gwet´s AC              | 0.8676 | 0.0394    | 22.00 | 0.000 | 0.7893   0.9458     |
| Krippendorff´s alpha   | 0.6769 | 0.0891    | 7.60  | 0.000 | 0.5002   0.8536     |

**Number of subjects = 100**

**Ratings per subject = 2**

**Number of rating categories = 3**
## Benchmarking Probabilistic method

### Interrater agreement

<table>
<thead>
<tr>
<th></th>
<th>Coef.</th>
<th>Std. Err.</th>
<th>P in.</th>
<th>P cum. &gt; 95%</th>
<th>Probabilistic [Benchmark Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent Agreement</td>
<td>0.8900</td>
<td>0.0314</td>
<td>0.997</td>
<td>0.997</td>
<td>0.8000 - 1.0000</td>
</tr>
<tr>
<td>Brennan and Prediger</td>
<td>0.8350</td>
<td>0.0472</td>
<td>0.230</td>
<td>1.000</td>
<td>0.6000 - 0.8000</td>
</tr>
<tr>
<td>Cohen/Conger´s Kappa</td>
<td>0.6765</td>
<td>0.0881</td>
<td>0.193</td>
<td>0.999</td>
<td>0.4000 - 0.6000</td>
</tr>
<tr>
<td>Fleiss´ Kappa</td>
<td>0.6753</td>
<td>0.0891</td>
<td>0.199</td>
<td>0.998</td>
<td>0.4000 - 0.6000</td>
</tr>
<tr>
<td>Gwet´s AC</td>
<td>0.8676</td>
<td>0.0394</td>
<td>0.955</td>
<td>0.955</td>
<td>0.8000 - 1.0000</td>
</tr>
<tr>
<td>Krippendorff´s alpha</td>
<td>0.6769</td>
<td>0.0891</td>
<td>0.194</td>
<td>0.999</td>
<td>0.4000 - 0.6000</td>
</tr>
</tbody>
</table>

### Benchmark scale

- **<0.0000** Poor
- **0.0000-0.2000** Slight
- **0.2000-0.4000** Fair
- **0.4000-0.6000** Moderate
- **0.6000-0.8000** Substantial
- **0.8000-1.0000** Almost Perfect