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

Rater A	Rater B 1	2	Total
1	n_{11}	n_{12}	$n_{1.}$
2	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	n

- The observed proportion of agreement is

$$p_o = \frac{n_{11} + n_{22}}{n}$$

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

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

Rater A	Rater B 1	2	3	4
1	1.00			
2	0.89	1.00		
3	0.56	0.89	1.00	
4	0.00	0.56	0.89	1.00

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

Rater A	Rater B 1	2	$\dots$	q	X	Total
1	n_{11}	n_{12}	$\dots$	n_{1q}	n_{1X}	$n_{1.}$
2	n_{21}	n_{22}	$\dots$	n_{2q}	n_{2X}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
q	n_{q1}	n_{q2}	$\dots$	n_{qq}	n_{qX}	$n_{q.}$
X	n_{X1}	n_{X2}	$\dots$	n_{Xq}	0	$n_{X.}$
Total	$n_{.1}$	$n_{.2}$	$\dots$	$n_{.q}$	$n_{.X}$	n

- 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

$$p_e = \sum_{k=1}^{q}\sum_{l=1}^{q} w_{kl} \frac{n_{k.}}{n - n_{.X}} \times \frac{n_{.l}}{n - n_{X.}}$$

Generalizing Kappa

Three or more raters

- 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

Observed agreement

- Organize the data as $n \times q$ matrix

Subject	Category 1	$\dots$	k	$\dots$	q	Total
1	r_{11}	$\dots$	r_{1k}	$\dots$	r_{1q}	r_1
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
i	r_{i1}	$\dots$	r_{ik}	$\dots$	r_{iq}	r_i
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
n	r_{n1}	$\dots$	r_{nk}	$\dots$	r_{nq}	r_n
Average	$\bar{r}_{1.}$	$\dots$	$\bar{r}_{k.}$	$\dots$	$\bar{r}_{q.}$	$\bar{r}$

- Average observed agreement over all pairs of raters

$$p_o = \frac{1}{n'} \sum_{i=1}^{n'} \sum_{k=1}^{q} \sum_{l=1}^{q} \frac{r_{ik}\left(w_{kl} r_{il} - 1\right)}{r_i \left(r_i - 1\right)}$$

Three or more raters

Chance agreement

- Fleiss (1971) expected proportion of agreement

$$p_e = \sum_{k=1}^{q} \sum_{l=1}^{q} w_{kl} \pi_k \pi_l$$

  with

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

- Fleiss' Kappa does not reduce to Cohen's Kappa
  - It instead reduces to Scott's π
  - 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

Data metric	Weights
nominal/categorical	none (identity)
ordinal	ordinal
interval	linear, quadratic, radical
ratio	any

- Rating categories must be predefined

More agreement coefficients

A general form

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

$$\kappa_{\cdot} = \frac{p_o - p_e}{1 - p_e}$$

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

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

  - Gwet's (2008, 2014) AC (κ_G)

$$p_e = \frac{\sum_{k=1}^{q} \sum_{l=1}^{q} w_{kl}}{q(q-1)} \sum_{k=1}^{q} \pi_k \left(1 - \pi_k\right)$$

More agreement coefficients

Krippendorff's alpha

- 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

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

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

$$V(\kappa) = \frac{1-f}{n(n-1)} \sum_{i=1}^{n} \left(\kappa_i^{\star} - \kappa\right)^2$$

where

$$\kappa_i^{\star} = \kappa_i - 2\left(1-\kappa\right) \frac{p_{e_i} - p_e}{1-p_e}$$

with

$$\kappa_i = \frac{n}{n'} \times \frac{p_{o_i} - p_e}{1-p_e}$$

p_{e_i} and p_{o_i} are the subject-level expected and observed agreement

Benchmarking agreement coefficients

Benchmark scales

- How do we interpret the extent of agreement?
- Landis and Koch (1977) suggest

Coefficient			Interpretation
	$<$	0.00	Poor
0.00	to	0.20	Slight
0.21	to	0.40	Fair
0.41	to	0.60	Moderate
0.61	to	0.80	Substantial
0.81	to	1.00	Almost Perfect

- Similar scales proposed (e.g., Fleiss 1981, Altman 1991)

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
  - Caseweise 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

- `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

Rater A	Rater B 1	Rater B 2	Total
1	45	15	60
2	25	15	40
Total	70	30	100

Rater A	Rater B 1	Rater B 2	Total
1	25	35	60
2	5	35	40
Total	30	70	100

$$
\begin{aligned}
p_o &= 0.60 \\
\kappa_n &= 0.20 \\
\kappa &= 0.13 \\
\kappa_F &= 0.12 \\
\kappa_G &= 0.27 \\
\kappa_\alpha &= 0.13
\end{aligned}
$$

$$
\begin{aligned}
p_o &= 0.60 \\
\kappa_n &= 0.20 \\
\kappa &= 0.26 \\
\kappa_F &= 0.19 \\
\kappa_G &= 0.21 \\
\kappa_\alpha &= 0.20
\end{aligned}
$$

Tables from Feinstein and Cicchetti 1990

Kappa paradoxes

High agreement, low Kappa

Rater A	Rater B 1	Rater B 2	Total
1	118	5	123
2	2	0	2
Total	120	5	125

$$
\begin{aligned}
p_o &= 0.94 \\
\kappa_n &= 0.89 \\
\kappa &= -0.02 \\
\kappa_F &= -0.03 \\
\kappa_G &= 0.94 \\
\kappa_\alpha &= -0.02
\end{aligned}
$$

Table from Gwet 2008

Kappa paradoxes

Independence of center cells, row and columns with quadratic weights

Rater A	Rater B 1	2	3	Total
1	1	15	1	17
2	3	0	3	6
3	2	3	2	7
Total	6	18	6	30

$$
\begin{aligned}
p_o &= 0.10 \\
p_{o_{w2}} &= 0.70 \\
\kappa_{n_{w2}} &= 0.10 \\
\kappa_{w2} &= 0.00 \\
\kappa_{F_{w2}} &= -0.05 \\
\kappa_{G_{w2}} &= 0.15 \\
\kappa_{\alpha_{w2}} &= -0.03
\end{aligned}
$$

Rater A	Rater B 1	2	3	Total
1	1	1	1	3
2	3	17	3	23
3	2	0	2	4
Total	6	18	6	30

$$
\begin{aligned}
p_o &= 0.67 \\
p_{o_{w2}} &= 0.84 \\
\kappa_{n_{w2}} &= 0.53 \\
\kappa_{w2} &= 0.00 \\
\kappa_{F_{w2}} &= 0.00 \\
\kappa_{G_{w2}} &= 0.69 \\
\kappa_{\alpha_{w2}} &= 0.02
\end{aligned}
$$

Tables from Warrens 2012

Benchmarking

Set up from Gwet (2014)

```
. 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
```

	raterb			
ratera	1	2	3	Total
1	75	1	4	80
2	5	4	1	10
3	0	0	10	10
Total	80	5	15	100

```
. kappaetc ratera raterb

Interrater agreement                             Number of subjects =       100
                                                Ratings per subject =         2
                                        Number of rating categories =         3
```

	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´ 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

Benchmarking

Probabilistic method

```
. kappaetc , benchmark showscale

Interrater agreement                             Number of subjects =     100
                                                Ratings per subject =       2
                                        Number of rating categories =       3
```

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

```
  Benchmark scale

        <0.0000      Poor
  0.0000-0.2000      Slight
  0.2000-0.4000      Fair
  0.4000-0.6000      Moderate
  0.6000-0.8000      Subtantial
  0.8000-1.0000      Almost Perfect
```