On the Gini coefficient, Cohen's kappa, weighted kappa, ICCs and the SB formula

Matthijs J Warrens

November 6, 2017

Outline

Four topics

- Negative bias of the Gini coefficient when data are grouped
- Cohen's kappa can always be increased and decreased by combining categories
- Pearson correlation is a special case of weighted kappa
- Transforming ICCs with the SB formula

Gini coefficient

Measure of statistical dispersion (Gini 1912)

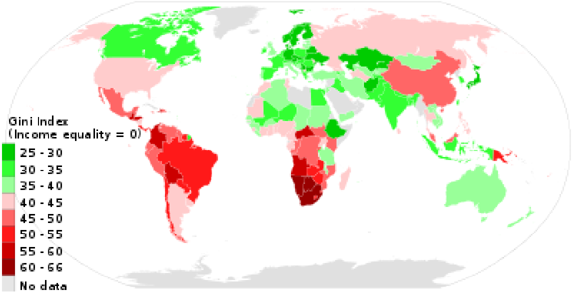
- inequality of income, wealth or opportunity
- widely used in economics (sociology, health science)
- range [0,1], 0 = uniform distribution (equality)

World Gini coefficient (income)

1988	.80
1993	.76
1998	.74
2003	.72
2008	.70
2013	.65

Country Gini coefficients (income)

World Bank (2014)
Source: Wikipedia



Relative mean difference

Gini coefficient of real numbers $x_1, x_2, \ldots, x_n$

$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2n \sum_{i=1}^{n} x_i}$$

Examples

$G = .17$ for $x = \{1, 1, 2, 2\}$

$G = .00$ for $x = \{2, 2, 2\}$

Negative bias

Gini coefficient usually not calculated on microdata

- microdata combined into households
- other grouped data with 5 to 30 categories
- income or tax statistics are grouped for confidentiality reasons

Literature: smaller Gini-values observed when data are grouped

Interpretation Gini coefficient, take into account

- type of households
- demographic structure of a country or region

Grouped data

Income USA 2010 ($G = .47$)

income category	% of pop.
under 15,000	13.7%
15,000–24,999	12.0%
25,000–34,999	10.9%
35,000–49,999	13.9%
50,000–74,999	17.7%
75,000–99,999	11.4%
100,000–149,999	12.1%
150,000–199,999	4.5%
200,000 and over	3.9%

Gini coefficient does not always decrease when data are grouped

$x = \{1, 1, 2, 2\}$
$G = .17$

$x' = \{2, 2, 2\}$
$G = .00$

$x' = \{1, 1, 4\}$
$G = .33$

A theorem

Theoretical gap

- specific grouping conditions for downward bias have not been formulated

A theorem

- G strictly decreases if values are partitioned into equal sized groups
- values may be 0 or negative (Warrens 2018)

Limitation

- groups must have equal size

Example

Individuals		Small households		Large households	
No.	Income	No.	Income	No.	Income
1	12,000	1 – 2	33,000	1 – 6	121,000
2	21,000				
3	35,000	3 – 4	53,000		
4	18,000				
5	24,000	5 – 6	35,000		
6	11,000				
7	47,000	7 – 8	70,000	7 – 12	260,000
8	23,000				
9	57,000	9 – 10	100,000		
10	43,000				
11	39,000	11 – 12	90,000		
12	51,000				
G	0.27		0.23		0.18

Cohen's kappa

Coefficient of agreement (Cohen 1960)

- agreement between two nominal classifications
- quality of ratings and rating instrument
- widely used in social and behavioral sciences, medical sciences
- $1 =$ perfect agreement, 0 if classifications independent

Examples

- Personality type
- Personality disorder (Schizotypal, Depression, ...)

Two nominal classifications

Objects	Rater 1	Rater 2
1	A	A
2	B	A
3	A	B
4	C	C
5	C	A
6	B	B
⋮	⋮	⋮

Ratings done independently

3 categories

- nominal
- defined in advance

Summarize data in table

R1/R2	A	B	C	total
A	p_{11}	p_{12}	p_{13}	p_{1+}
B	p_{21}	p_{22}	p_{23}	p_{2+}
C	p_{31}	p_{32}	p_{33}	p_{3+}
total	p_{+1}	p_{+2}	p_{+3}	1

Definition

Cohen (1960)

$$\kappa := \frac{\sum_{i=1}^{m} (p_{ii} - p_{i+}p_{+i})}{1 - \sum_{i=1}^{m} p_{i+}p_{+i}}$$

R1/R2	A	B	C	total
A	.10	.05		.15
B		.25	.20	.45
C		.20	.20	.40
total	.10	.50	.40	1

$\kappa = .25$

Combining categories

R1/R2	A	B	C	total
A	.10	.05		.15
B		.25	.20	.45
C		.20	.20	.40
total	.10	.50	.40	1

$\kappa = .25$

Some confusion between categories B and C

$\rightarrow$ combine categories B and C

R1/R2	A	B+C	total
A	.10	.05	.15
B+C		.85	.85
total	.10	.90	1

$\kappa = .77$

Kappa increases from .25 to .77

A theorem

Common believe: higher kappa-value when fewer categories

Theorem (existence): kappa can always be increased or decreased by combining categories (Warrens 2010)

Outline of proof (Warrens 2011)

- consider all kappas corresponding to a partition type of the categories
- overall kappa is a weighted average of these kappas
- smallest kappa is smaller than overall kappa
- highest kappa is higher than overall kappa

Three partitions

R1/R2	A	B	C	total
A	.10	.05		.15
B		.25	.20	.45
C		.20	.20	.40
total	.10	.50	.40	1

$\kappa = .25$

R1/R2	A+B	C	total
A+B	.40	.20	.60
C	.20	.20	.40
total	.60	.40	1

$\kappa = .17$

R1/R2	A	B+C	total
A	.10	.05	.15
B+C		.85	.85
total	.10	.90	1

$\kappa = .77$

R1/R2	A+C	B	total
A+C	.30	.20	.50
B	.25	.25	.50
total	.55	.45	1

$\kappa = .10$

Increasing kappa

R1/R2	A	B	C	D	E	total
A	.02	.01		.01		.04
B	.01	.05	.02	.04	.02	.14
C		.02	.03	.06	.03	.15
D		.04	.05	.20	.13	.43
E		.01	.02	.09	.12	.24
total	.03	.14	.13	.41	.29	1

Agresti (1990)

$\kappa = .18$ {A} {B} {C} {D} {E}

$\kappa = .26$ {A} {B} {C} {D,E}

$\kappa = .33$ {A} {B} {C,D,E}

$\kappa = .41$ {A} {B,C,D,E}

Decreasing kappa

R1/R2	A	B	C	D	E	total
A	.02	.01		.01		.04
B	.01	.05	.02	.04	.02	.14
C		.02	.03	.06	.03	.15
D		.04	.05	.20	.13	.43
E		.01	.02	.09	.12	.24
total	.03	.14	.13	.41	.29	1

Agresti (1990)

$\kappa = .18$ {A} {B} {C} {D} {E}

$\kappa = .17$ {A,C} {B} {D} {E}

$\kappa = .16$ {A,C} {B,D} {E}

$\kappa = -.28$ {A,C,E} {B,D}

Weighted kappa

Coefficient of agreement (Cohen 1968)

- agreement between two ordinal classifications
- quality of ratings and rating instrument
- widely used in social and behavioral sciences, medical sciences
- 1 = perfect agreement, 0 if classifications independent

Examples

- Severity of lesions or fractions
- Degree of anxiety
- Degree of disability (severe, moderate, good recovery)

Two ordinal classifications

Objects	Rater 1	Rater 2
1	A	A
2	B	A
3	A	B
4	C	C
5	C	A
6	B	B
⋮	⋮	⋮

Ratings done independently

3 categories

- ordinal (A<B<C)
- defined in advance

Summarize data in table

R1/R2	A	B	C	total
A	p_{11}	p_{12}	p_{13}	p_{1+}
B	p_{21}	p_{22}	p_{23}	p_{2+}
C	p_{31}	p_{32}	p_{33}	p_{3+}
total	p_{+1}	p_{+2}	p_{+3}	1

Definition

Cohen (1968)

Weights $w_{jk} \geq 0$

$$\kappa_w := 1 - \frac{\sum_{i=1}^{m}\sum_{j=1}^{m} p_{ij} w_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{m} p_{i+} p_{+j} w_{ij}}$$

All weights essentially arbitrary

Identity weights

- $w_{ij} = 1_{i \neq j}$ (Cohen 1960)

Linear weights

- $w_{ij} = |i - j|$

Quadratic weights

- $w_{ij} = (i - j)^2$
- most widely used
  Fleiss and Cohen (1973):
  proportion of variance

Some more notation

	Rater 1	Rater 2
Objects	X	Y
1	x_1	y_1
2	x_2	y_2
3	x_3	y_3
4	x_4	y_4
5	x_5	y_5
6	x_6	y_6
$\vdots$	$\vdots$	$\vdots$

Means $\bar{x}$, $\bar{y}$

Variances s_x^2, s_y^2

Covariance s_{xy}

$x_i, y_i \in \{z_1, \ldots, z_m\}$

z_i is used for coding category i

A theorem

We have

$$\kappa_r := 1 - \frac{\sum_{i=1}^{m}\sum_{j=1}^{m} p_{ij}\left(\frac{z_i-\bar{x}}{s_x} - \frac{z_j-\bar{y}}{s_y}\right)^2}{\sum_{i=1}^{m}\sum_{j=1}^{m} p_{i+}p_{+j}\left(\frac{z_i-\bar{x}}{s_x} - \frac{z_j-\bar{y}}{s_y}\right)^2} = \frac{s_{xy}}{s_x s_y} = r.$$

Pearson correlation is a special case of weighted kappa

- commonly used in data analysis
- use of correlation is unquestioned

Weighted kappa coefficients are obsolete

Intraclass correlations

Measures of reliability

- Shrout and Fleiss (1979), McGraw and Wong (1996)
- assess reliability of quantitative measurements
- widely used in social and behavioral sciences, medical sciences
- range [0,1]

Models

- random raters, or same raters
- absolute agreement, or consistency
- single measurements, average measurements

Quantitative measurements

Objects	Rater 1	Rater 2
1	5	4
2	1	3
3	7	3
4	3	4
5	4	5
6	9	8
⋮	⋮	⋮

Ratings done independently

- quantitative

Definition ICCs

Variance components

- σ_t^2 for objects of measurement (subjects, students, scans)
- $\sigma_1^2 + \cdots + \sigma_q^2$ for other sources

General formulas are

$$\text{ICC}(1) = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_1^2 + \cdots + \sigma_q^2}$$

for single measurements, and

$$\text{ICC}(k) = \frac{\sigma_t^2}{\sigma_t^2 + (\sigma_1^2 + \cdots + \sigma_q^2)/k}$$

for average measurements with k raters.

SB formula

Spearman-Brown formula

- from classical test theory
- relates reliability to length of a test

Formula is given by

$$\rho^* = \frac{n\rho}{1 + (n-1)\rho}.$$

where

- ρ^* is reliability of extended test
- ρ is reliability of original test
- n is ratio length extended test/length original test

A discovery

De Vet, Mokkink, Mosmuller and Terwee (2017)

- use SB formula to transform ICC(1) into ICC(k) of same form
- only requires ICC(1)-value
- no knowledge of the variance components needed

De Vet et al. presented examples to support discovery

Discovery is remarkable

- ICCs and SB formula come from quite different disciplines

Example

From single measurements to average measurements

Using $\rho = \text{ICC}(1) = .166$ and $n = 4$ in

$$\rho^* = \frac{n\rho}{1 + (n-1)\rho}$$

yields

$\rho^* = .443 = \text{ICC}(4)$

A theorem

A theorem (Warrens 2017)

- if ICC(m) and ICC(u) have same variance components
- ICC(m) can be transformed into ICC(u)
- using $n = m/u$ and ICC(m) in SB formula

Specifically

- ICC(1) can be transformed into corresponding ICC(k)
- ICC(k) can be transformed into corresponding ICC(1)

Another example

From average measurements to single measurements

Using $\rho = \text{ICC}(4) = .443$ and $n = 1/4$ in

$$\rho^* = \frac{n\rho}{1 + (n-1)\rho}$$

yields

$\rho^* = .166 = \text{ICC}(1)$

References

- Cohen J (1960) A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20:37-46
- Cohen J (1968) Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin 70:213-220
- De Vet HCW, Mokkink LB, Mosmuller DGM, Terwee CB (2017) Spearman-Brown prophecy and Cronbachs alpha: different faces of reliability and opportunities for new applications. Journal of Clinical Epidemiology 85:45-49
- McGraw KO, Wong SP (1996) Forming inferences about some intraclass correlation coefficients. Psychological Methods 1:30-46
- Shrout PE, Fleiss JL (1979) Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin 86:420-428

References 2

- Warrens MJ (2018) On the negative bias of the Gini coefficient due to grouping. Journal of Classification
- Warrens MJ (2010) Cohen's kappa can always be increased and decreased by combining categories. Statistical Methodology 7:673-677.
- Warrens MJ (2011) Cohen's kappa is a weighted average. Statistical Methodology 8:473-484
- Warrens MJ (2014) Corrected Zegers-ten Berge coefficients are special cases of Cohen's weighted kappa. Journal of Classification 31:179-193
- Warrens MJ (2017) Transforming intraclass correlation coefficients with the Spearman-Brown formula. Journal of Clinical Epidemiology 85:14-16