On the negative bias of the Gini coefficient due to grouping

Matthijs J Warrens

VOC meeting
May 25, 2018

Gini coefficient

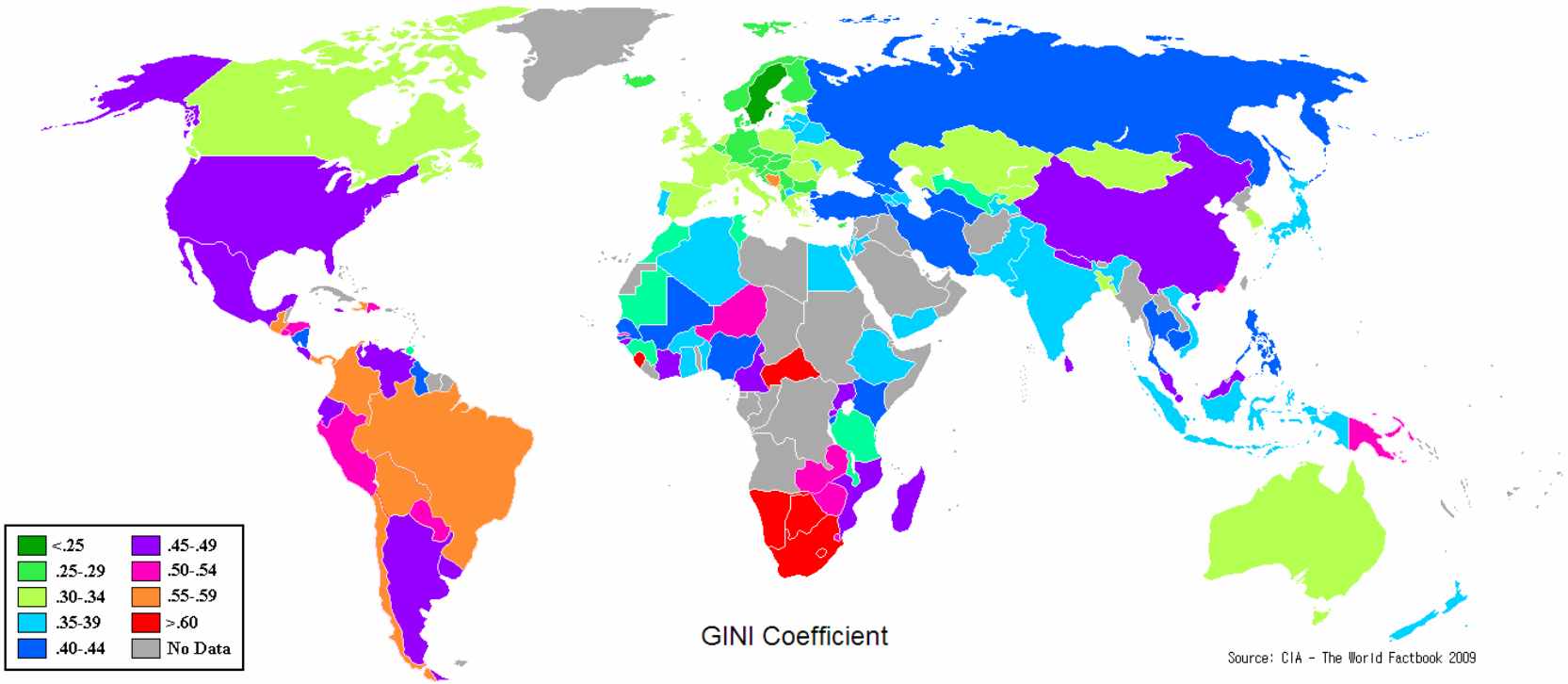
Measure of statistical dispersion (Gini 1912)

- 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)



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)
- no decrease if all values in each group are equal

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

Keys ideas of proof

1) Summation in denominator is constant

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

2) Triangle inequality

$$|x_1 - y_1| + |x_2 - y_2| \geq |x_1 + x_2 - (y_1 + y_2)|$$

3) For values in same group $|x_1 - x_2| \geq 0$

Reference

- Warrens MJ (2018) On the negative bias of the Gini coefficient due to grouping. Journal of Classification