On the negative bias of the Gini coefficient due to grouping

Warrens, Matthijs J.

Published in:
Journal of Classification

DOI:
10.1007/s00357-018-9267-9

Publication date:
2018

Journal of Classification 35:580-586 (2018)
DOI: 10.1007/s00357-018-9267-9

On the Negative Bias of the Gini Coefficient due to Grouping

Matthijs J. Warrens

University of Groningen, The Netherlands

Abstract: The Gini coefficient is a measure of statistical dispersion that is commonly used as a measure of inequality of income, wealth or opportunity. Empirical research has shown that the coefficient may have a nonnegligible downward bias when data are grouped. It is unknown under which grouping conditions the downward bias occurs. In this note it is shown that the Gini coefficient strictly decreases if the data are partitioned into equal sized groups.

Keywords: Statistical dispersion; Measure of inequality; Inequality of income; Inequality of wealth; Grouping data.

1. Introduction

The Gini coefficient (Gini, 1912) is a measure of statistical dispersion that is commonly used in various scientific disciplines, including economics, sociology, health science and engineering. It is commonly used to quantify inequality of wealth, income and opportunity, and inequality in education between countries (Sen, 1977). The coefficient can be defined in various different ways (Jasso, 1979; Yitzhaki, 1998; Ceriani and Verme, 2012). Here, we define the coefficient as the relative mean difference of the values of a frequency distribution (Damgaard and Weiner, 2000). The Gini coefficient of the real numbers $x_1, x_2, \ldots, x_n$ is

The author thanks Professor Doug Steinley and an anonymous referee for their helpful comments and valuable suggestions on earlier versions of this note.

Corresponding Author's Address: M.J. Warrens, University of Groningen, Grote Rozenstraat 3, 9712 TG Groningen, The Netherlands, phone: +31 50 36 36691, email: m.j.warrens@rug.nl.

Published online: 3 October 2018

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

Formula (1) is equal to the mean of the difference between every possible pair of values, divided by the mean value size.

The Gini coefficient measures the dispersion among the values of the frequency distribution. If all values are positive the coefficient produces values between 0 and 1. The coefficient has value 0 if all the values are equal. In this case there is perfect equality. Values near 1 express high inequality among the values. Furthermore, the coefficient is invariant under multiplication of a positive constant. Moreover, different frequency distributions may have the same value of the Gini coefficient.

In many applications the Gini coefficient is estimated from grouped data with 5 to 30 categories instead of the microdata (Gastwirth, 1972; Abounoori and McCloughan, 2003). For example, income or tax statistics are often grouped for confidentiality reasons (Van Ourti and Clarke, 2011). Empirical research has shown that the Gini coefficient may have a nonnegligible downward bias when data are grouped (Lerman and Yitzhaki, 1989; Davies and Shorrocks, 1989; Kwok, 2010). Vice versa, Kwok (2010) noted that the Gini coefficient increases if a combined household is split into several smaller households or people living alone. Thus, the Gini coefficient may produce different results for income when the units of analysis are individuals instead of households (Deininger and Squire, 1996). Therefore, in interpreting the Gini coefficient the demographic structure of a country or region should be taken into account.

The Gini coefficient may decrease if the data are grouped, but it may also increase. For example, for $x = \{1, 1, 2, 2\}$ we have $G = .167$. Indeed, adding the first and second value of x we obtain $x' = \{2, 2, 2\}$ and $G = 0$. However, if we add the second and third value of x we obtain $x' = \{1, 3, 2\}$ and $G = .222$, whereas if we add the third and fourth value of x we get $x' = \{1, 1, 4\}$ and $G = .333$. The latter two cases show that the Gini-value may also increase if the data are grouped. Specific grouping conditions under which the downward bias of the Gini coefficient occurs have not been formulated. New insights into the properties of the coefficient with respect to data grouping are therefore welcomed.

In this research note, it is proved that the Gini coefficient strictly decreases if the values of a frequency distribution are partitioned into equal sized groups and the combined values are analyzed. An immediate consequence is that, vice versa, the Gini coefficient increases if the units are split into equally sized parts. A theorem that formalizes these statements is presented in the next section. An example and discussion are presented in the last section.

2. A Theorem

In this section, we show (Theorem 1) that the Gini coefficient strictly decreases if the data are partitioned into equal sized groups.

Theorem 1. *Suppose d divides n with $1 < d < n$. The Gini coefficient of $x_1, \ldots, x_n$ decreases if the values are partitioned into groups of size d and the combined values are analyzed.*

Proof. Let $m = n/d$. Let x_{ik} denote value $i \in \{1, 2, \ldots, d\}$ in group $k \in \{1, 2, \ldots, m\}$. Furthermore, define for each group k the sum

$$s_k = \sum_{i=1}^{d} x_{ik}. \tag{2}$$

The Gini coefficient of the sums is

$$G_s = \frac{\sum_{k=1}^{m}\sum_{\ell=1}^{m} |s_k - s_\ell|}{2m \sum_{k=1}^{m} s_k}. \tag{3}$$

Repeated application of the triangle inequality to the sum

$$\sum_{i=1}^{d} |x_{ik} - x_{i\ell}|, \tag{4}$$

yields the inequality

$$\sum_{i=1}^{d} |x_{ik} - x_{i\ell}| \geq \left| \sum_{i=1}^{d} x_{ik} - \sum_{i=1}^{d} x_{i\ell} \right| = |s_k - s_\ell|. \tag{5}$$

In the absolute difference $|x_{ik} - x_{i\ell}|$, the value x_{ik} of group k is compared to the value $x_{i\ell}$ from group ℓ. The value x_{ik} can also be compared to one of the $d - 1$ other values $x_{j\ell}$ in group ℓ. Thus, we have d variants of inequality (5) such that in each variant a value of group k is compared to a different value in group ℓ. If we sum these d variants of (5) we obtain

$$\sum_{i=1}^{d}\sum_{j=1}^{d} |x_{ik} - x_{j\ell}| \geq d|s_k - s_\ell|. \tag{6}$$

Summing the right-hand side of (6) over all combinations of k and ℓ with $k \neq \ell$, we obtain the identity

$$2d \sum_{k=1}^{m-1} \sum_{\ell=k+1}^{m} |s_k - s_\ell| = d \sum_{k=1}^{m}\sum_{\ell=1}^{m} |s_k - s_\ell|, \tag{7}$$

since $s_k - s_\ell = 0$ if $k = \ell$. However, summing the left-hand side of (6) over all combinations of k and ℓ with $k \neq \ell$ yields

$$2\sum_{k=1}^{m-1}\sum_{\ell=k+1}^{m}\sum_{i=1}^{d}\sum_{j=1}^{d}|x_{ik}-x_{j\ell}| = \sum_{k=1}^{m}\sum_{\ell=1}^{m}\sum_{i=1}^{d}\sum_{j=1}^{d}|x_{ik}-x_{j\ell}| - \sum_{k=1}^{m}\sum_{i=1}^{d}\sum_{j=1}^{d}|x_{ik}-x_{jk}|. \quad (8)$$

The triple summation in (8) is only equal to zero in the rare case that, in each group, all values are equal to one another. If we exclude this very particular case, equality (8) implies the inequality

$$\sum_{k=1}^{m}\sum_{\ell=1}^{m}\sum_{i=1}^{d}\sum_{j=1}^{d}|x_{ik}-x_{j\ell}| > 2\sum_{k=1}^{m-1}\sum_{\ell=k+1}^{m}\sum_{i=1}^{d}\sum_{j=1}^{d}|x_{ik}-x_{j\ell}|. \quad (9)$$

Combining the left-hand side of (9) with the identity

$$\sum_{k=1}^{m}\sum_{\ell=1}^{m}\sum_{i=1}^{d}\sum_{j=1}^{d}|x_{ik}-x_{j\ell}| = \sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|, \quad (10)$$

and the right-hand side of (9) with identity (7), it follows that, summing (6) over all combinations of k and ℓ, yields

$$\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j| > d\sum_{k=1}^{m}\sum_{\ell=1}^{m}|s_k - s_\ell|. \quad (11)$$

Dividing both sides of (11) by $2n\sum_{i=1}^{n} x_i$, and using the identities $n = dm$ and $\sum_{i=1}^{n} x_i = \sum_{k=1}^{m} s_k$ on the right-hand side of the result, we obtain the strict inequality $G > G_s$, which completes the proof.
■

3. An Example and Discussion

Table 1 presents the income in Australian dollars from 2013 of twenty-four individuals. The numbers were made freely available by the Australian Government (http://data.gov.au). The particular numbers in Table 1 are the twenty-four top numbers on income of the 2012-13 Individual sample file. Six people did not have an income in 2013. The maximum income is 192669, the average income is 45066, and the total income for the twenty-four individuals is 1081585. For Table 1 we have $G = .540$.

Table 1. 2013 Income of twenty-four persons in Australian dollars.

1.	67848	7.	27439	13.	81586	19.	83468
2.	50335	8.	0	14.	36241	20.	0
3.	14537	9.	62941	15.	48822	21.	14723
4.	37495	10.	30822	16.	13419	22.	78472
5.	0	11.	20992	17.	0	23.	192669
6.	0	12.	79069	18.	140707	24.	0

Table 2. Gini-values corresponding to groupings of the data from Table 1.

Number of groups	G
24	.540
12	.303
8	.253
6	.226
4	.161
3	.193
2	.138

Table 2 presents the Gini-values that are obtained by grouping the data in Table 1. The first line corresponds to the case of no grouping (or twenty-four groups). The second line with twelve groups corresponds to the case in which the first two incomes are grouped ($67848 + 50335$), the second two values are grouped ($14537 + 37495$), and so on. The third line with eight groups corresponds to the case where the first three values are grouped ($67848 + 50335 + 14537$), and so on. The bottom line with two groups corresponds to the case where the first and last twelve incomes are grouped.

Table 2 shows that, for the data in Table 1, the Gini coefficient tends to decrease when the number of groups becomes smaller. Theorem 1 applies to two partitions that are nested. For two numbers from the first column of Table 2, if the bottom number is a divisor of the top number, then the two partitions are nested, and the partition associated with the top number is finer than the partition associated with the bottom number. Consider, for example, the sequence starting with 24 to 12 to 6 to 3 groups. In each step the values are partitioned into groups of equal size. Furthermore, the associated G-values strictly decrease (from .540 to .303 to .226 to .193). Another illustration of Theorem 1 is the sequence from 24 to 8 to 4 to 2 groups. Again, the associated G-values strictly decrease (from .540 to .253 to .161 to .138).

Theorem 1 is also applicable to income and wealth distribution tables (see, e.g. Kerbo, 2000). These tables typically summarize the income and

wealth frequency distributions using a number of quantiles. Quantiles divide a frequency distribution into equal groups, each containing the same fraction of the total population. If two sets of quantiles are nested, Theorem 1 tells us that the set with higher granularity (higher number of quantiles) will have a higher Gini coefficient. For example, five 20% quantiles (low granularity) will yield a lower Gini coefficient than twenty 5% quantiles taken from the same distribution. Hence, it is important in applications that the Gini-value is reported together with the proportions of the quantiles used for measurement.

Finally, a limitation of Theorem 1 is that the units of analysis must be partitioned into equal sized groups. As demonstrated in the introduction, if the units are partitioned into groups of different sizes, it depends on the data at hand whether the Gini coefficient increases or decreases. On the other hand, the theorem puts no restrictions on which values are combined. Furthermore, some of the values are allowed to be negative or zero.

References

ABOUNOORI, E., and MCCLOUGHAN, P. (2003), "A Simple Way to Calculate the Gini Coefficient for Grouped as well as Ungrouped Data," *Applied Economics Letters, 10,* 505–509.

CERIANI, L., and VERME, P. (2012), "The Origins of the Gini Index: Extracts from Variabilità e Mutuabilità (1912) by Corrado Gini," *Journal of Economic Inequality, 10,* 421–443.

DAMGAARD, C., and WEINER, J. (2000), "Describing Inequality in Plant Size and Fecundity," *Ecology, 81,* 1139–1142.

DEININGER, K., and SQUIRE, L. (1996), "A New Data Set Measuring Income Inequality," *World Bank Economic Review, 10,* 565–591.

GASTWIRTH, J. (1972), "Robust Estimation of the Lorenz Curve and Gini Index," *The Review of Economics and Statistics, 54,* 306–316.

GINI, C. (1912), *Variabilità e Mutuabilità*, Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche, Bologna: C. Cuppini.

JASSO, G. (1979), "On Gini's Mean Difference and Gini's Index of Concentration," *American Sociological Review, 44,* 867–870.

KERBO, H.R. (2000). *Social Stratification and Inequality: Class Conflict in Historical, Comparative, and Global Perspective* (4th ed.), Boston: McGraw-Hill.

KWOK, K.C. (2010), *Income Distribution of Hong Kong and the Gini Coefficient*, China: The Government of Hong Kong.

SEN, A. (1977), *On Economic Inequality* (2nd ed.), Oxford: Oxford University Press.

VAN OURTI, T., and CLARKE, P. (2011), "A Simple Correction to Remove the Bias of the Gini Coefficient due to Grouping," *The Review of Economics and Statistics, 93,* 982–994.

YITZHAKI, S. (1998), "More Than a Dozen Alternative Ways of Spelling Gini," *Economic Inequality, 8,* 13–30.