Working Paper Series

A discreet approach to study the distribution-free downward biases of Gini coefficient and the methods of correction in cases of small observations

Amlan Majumder
Takayoshi Kusago

ECINEQ WP 2013 - 298

ECINEQ 2013 - 298
August 2013

www.ecineq.org

A discreet approach to study the distribution-free downward biases of Gini coefficient and the methods of correction in cases of small observations*

Amlan Majumder†
Dinhata College, West Bengal, India

Takayoshi Kusago
Kansai University, Osaka, Japan

Abstract

It is well-known that Gini coefficient is influenced by granularity of measurements. When there are few observations only or when they get reduced due to grouping, standard measures exhibit a non-negligible downward bias. At times, bias may be positive when there is an apparent reduction in sample size. Although authors agreed on distribution-free and distribution-specific parts of it, there is no consensus in regard to types of bias, their magnitude and the methods of correction in the former. This paper deals with the distribution-free downward biases only, which arise in two forms. One is related to scale and occurs in all the cases stated above, when number of observations is small. Both occur together if initial number of observations is not sufficiently large and further they get reduced due to grouping. Underestimations associated with the former is demonstrated and addressed, for discontinuous case, through alternative formulation with simplicity following the principle of mean difference without repetition. Equivalences of it are also derived under the geometric and covariance approaches. However, when it arises with the other, a straightforward claim of it in its full magnitude may be unwarranted and quite paradoxical. Some exercises are done consequently to make Gini coefficient standardized and comparable for a fixed number of observations. Corrections in case of the latter are done accordingly with a newly proposed operational pursuit synchronizing the relevant previous and present concerns. The paper concludes after addressing some definitional issues in regard to convention and adjustments in cases of small observations.

Keywords: Gini coefficient, maximum inequality Lorenz curve, mean difference approach, small observations, underestimation.
JEL Classification: D310, D630.

---

*This theoretical research has been in commemoration of 100 years of Gini's mean difference: 1912-2012. It was supported by 'FY 2010 JSPS Postdoctoral Fellowship for Foreign Researcher (P 10316)' and 'KAKENHI - Grant-in-Aid for Scientific Research (C) (2200316)', Japan Society for the Promotion of Science. An earlier version of it was presented at the Fifth Meeting of the ECINEQ, University of Bari, Italy during 22-24 July 2013. We are thankful to participants for comments and suggestions. Responsibility of errors rests with authors.

†Contact details of A. Majumder: Dinhata College, Dinhata, Cooch Behar, W. B. 736135, India, E-mail: amlan@amlan.co.in.

1. INTRODUCTION

It is well-known that Gini coefficient is influenced by granularity of measurements. For example, five quintiles (low granularity) usually yield significantly a lower Gini coefficient than twenty ventiles (high granularity) taken from the same distribution. This is commonly perceived as an often encountered problem with measurements. However, there is no consensus on the issue in theoretical and empirical literature. In his recent study, Milanovic (2010) computes Gini coefficients from micro data for some countries with large number of observations and then he squeezes micro data for each of the countries into twenty ventiles and computes Gini coefficients again. He observes that for each of such cases loss in Gini coefficient is negligible. For example, in table 3 of his paper (considering the case of Belarus only, p. 10), when number of observations is changed from 5227 to 20, Gini coefficient is reduced from 28.67 to 28.5 meaning 0.6 per cent underestimation. On the contrary, according to Deltas (2003, p. 227), "… a reduction in the sample size leads to a reduction in apparent inequality as measured by the Gini coefficient". He examines the extent of bias using Monte Carlo simulations and claims that Gini coefficient may exhibit a bias of up to 7.5 per cent for a sample size of 20 in contrast to that of the whole population with a finite size. According to him (p. 227), "… the bias varies substantially across distributions". In line with these methodological notes on 'small-sample bias' he also cites example of 'small-size bias', which occurs for availability of few observations only (p. 227). At times, positive bias may arise as well in adjusted Gini when incomes are generated by a uniform distribution[1] (Deltas, 2003, p. 230). Considering its distribution-free and the distribution-specific parts, he terms those as first-order and second-order biases respectively. The underestimation due to first-order bias is caused by the fact that in cases of small observations, standard measures produce results in truncated scales, not in 0-100 point scale. In regard to its correction, he suggests one multiplicative factor [n/(n-1), where n = number of observations] with geometric intuition[2], which adjusts results by putting them in 0-100 point scale. Interestingly, this correction factor corresponds to the contribution of Deaton[3] (1997, p. 139), who derives one formula of computing Gini coefficient more directly without relating it to the Lorenz curve framework and without mentioning any inconsistency under discussion. Deaton (1997, p. 139) claims

[1] An example is cited in section 5.3 showing positive bias even before adjustments.

[2] Does not eliminate bias associated with stochastic processes.

[3] Some other formulae also correspond to this, such as Majumder (2007, pp. 4-5) and World Bank (2003, p. 22).

that: " … the Gini coefficient is often defined from the Lorentz curve, but can also be defined directly". It is shown later that his formula is clearly related to (maximum inequality) Lorenz curve framework. Qurti and Clarke (2010) contribute significantly on the issue of underestimation or bias associated with Gini coefficient due to grouping. They report that the grouping of income into relatively small number of categories imparts a non-negligible downward bias, as found in a study by Lerman and Yitzhaki (1989), where biases from using grouped data with ten and five income categories are about 2.5 per cent and 7 per cent respectively as compared to the Gini calculated from micro data. As above, although authors go similarly in views of the approaches and on the subject of concern, they differ significantly in terms of reasoning. Also there is no consensus in regard to types of bias, which are distribution-free, their magnitude and the methods of correction. It is clear when Qurti and Clarke (2010, p. 983) comment that the correction factor proposed by Deltas (2003) in regard to the first-order bias "… neglects that the small-sample bias of the Gini is distribution specific". They also classify bias similarly into two types: first-order and second-order considering the distribution-free and distribution-specific parts of it respectively and consequently derive a first-order correction factor [$k^2/(k^2-1)$, where k = number of equally sized groups] from a measurement error framework. As the correction factors proposed by them differ, the concept of so-called 'first-order' bias will also do. We demonstrate later that the 'first-order' bias as highlighted by Deltas (2003) is not distribution-specific. Further, we reveal that the distribution-free bias, in such cases[4], arises in two forms. They are to be corrected first by the 'first-order correction factor' proposed by Qurti and Clarke (2010) and then by that proposed by Deltas (2003). Although the issue has been addressed methodologically by Qurti and Clarke (2010) as well as used by Milanovic (2010), there is no reflection of such an operational pursuit in their works. However, in order to proceed further, we term the first-order bias classified by Deltas (2003) as the bias of 'Type I', the first-order bias perceived by Qurti and Clarke (2010) as that of 'Type II' and the remaining second order distribution-specific one as bias of 'Type III'. Underestimations associated with the first two are the main concern of this paper. Nevertheless, as above, the present paper plans to synchronise the concerns and approaches on the issue of underestimation with more emphasis to use them in alternative theoretical derivations relating them with the standard sets of

[4] Of course when initial number of observations is not sufficiently large and further they reduced due to grouping.

measures. For example, it may be observed that under the Gini's mean difference approach, the inconsistency with small observations occurs when the computation is based on the principle of mean difference with repetition. Consequently, one formula is suggested where computation is done following the principle of mean difference without repetition. Under the geometric approach, such inconsistencies occur when computation of Gini coefficient includes the area beyond the maximum inequality Lorenz curve. An alternative formulation is proposed accordingly within the framework of maximum inequality Lorenz curve. Under the covariance approach, the alternative derivation follows that of the geometric approach. The standard formula considers all the income ranks corresponding to the all cumulative proportions of population. However, as the correction under the geometric approach abandons one cumulative proportion of population (beyond the maximum-inequality Lorenz curve), the new formula under the covariance approach follows that. With such precise objectives of deriving alternative sets of measures (including the narrations presented above while making arguments), the paper also maintains that computation of Gini coefficient should be more user-friendly in regard to the use of simple spreadsheet programmes[5]. As literature on Gini coefficient is vast and wide and as it seems fairly impossible to cover all, the tradition of discrete approach is followed, which is popularised, among other, by Anand (1983), Milanovic (1994, 1997), Subramanian (1997) and Xu (2004).

This paper consists of sections, which: (i) simplifies the standard measure under the Gini's mean difference approach in regard to use of simple spreadsheet programmes, (ii) highlights the source of Type I bias under the Gini's mean difference approach, geometric approach and the covariance approach and proposes alternative formulations, (iii) discusses about Type II bias and its possible solution, (iv) demonstrates the method of correction when both Type I and Type II biases occur together with numerical examples, and (v) discusses the case when Type I bias cannot be claimed in its full magnitude and consequently does some (numerical) exercises to make Gini coefficient standardised and comparable for a fixed number of observations. The paper concludes after addressing some definitional issues in regard to convention and adjustments in cases of small observations.

[5] For example, by presenting a simple measure, Milanovic (1997) claims that since all the components of the formula are easy to calculate, the Gini coefficient can be obtained using a simple hand calculator.

2. THE GINI'S MEAN DIFFERENCE APPROACH

2.1. The standard measure

Although there are various ways of expressing and calculating Gini coefficient, we begin with the following (see Anand, 1983, p. 313):

$$G_1 = (1/2n^2\mu)\sum_{i=1}^{n}\sum_{j=1}^{n}|y_i - y_j|, \qquad (1)$$

where, G stands for Gini coefficient, the subscript denotes its sequence in the paper and y_i is the income of person or group i, y_j is that of person or group j, μ is the average income, i = 1, 2, 3, …, n, j = 1, 2, 3, …, n and $y_1 \leq y_2 \leq \ldots \leq y_n$. The formula tells that Gini coefficient is one-half the average value of absolute differences between all pairs of incomes divided by the mean income. For operational advantage, we plan to work with distribution of income. In that case we comprehend y_i and y_j as proportions or shares of income corresponding to person or group i and j respectively (such that $\mu = 1/n$, as summation over all proportions of income is equal to one).

The above formulation is due to Kendall (1948, p. 42), who, for discontinuous case, puts forward the principles of computing 'mean difference with repetition' and 'mean difference without repetition'[6]. The case of repetition is for all i = j = 1, 2, 3, …, n; and that of without repetition is for all i = 1, 2, 3, …, n and $i \neq j$. Formula (1) corresponds to the first case of 'with repetition'. However, while working out Gini's mean difference (with repetition), Kendall (1948, p. 45) considers one set of terms of type $(y_i - y_j)$ only in the numerator, sum of which is half the sum of all pairs of differences. While demonstrating equivalence of formula (1) with the ones under different approaches, Anand (1983, pp. 313-314) also restricts number of elements to the lower triangular matrix of the following symmetric matrix for i = 1, 2, 3, …, n and $j \leq i$:

$$\begin{bmatrix} |y_1 - y_1| & |y_1 - y_2| & \ldots & |y_1 - y_n| \\ |y_2 - y_1| & |y_2 - y_2| & \ldots & |y_2 - y_n| \\ \ldots & \ldots & \ldots & \ldots \\ |y_n - y_1| & |y_n - y_2| & \ldots & |y_n - y_n| \end{bmatrix}.$$

Gini index, in response to the above restriction, is expressed as following:

[6] This is, according to him, unimportant when n is large.

$$G_2 = (1/n^2\mu)\sum_{i=1}^{n}\sum_{j\le i}(y_i - y_j),$$

as sum of all (absolute) terms is twice the sum of terms in the lower triangular matrix:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\left|y_i - y_j\right| = 2\sum_{i=1}^{n}\sum_{j\le i}(y_i - y_j).$$

As,

$$\sum_{i=1}^{n}\sum_{j\le i}\left|y_i - y_j\right| = \sum_{i=1}^{n}\{(i-1)y_i - (n-i)y_i\},$$

$$G_2 = \frac{1}{n^2\mu}(\sum_{i=1}^{n} iy_i - \sum_{i=1}^{n} y_i - n\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} iy_i),$$

$$= \frac{1}{n^2\mu}(2\sum_{i=1}^{n} iy_i - n - 1). \qquad (2)$$

As the composite term inside the parenthesis in the above expression (formula 2) can be computed easily using simple spreadsheet programmes, it is one of the simplest formulae of Gini coefficient. It says that Gini coefficient is a function of weighted sum of share of income[7], where the weight is nothing but the rank of individuals or groups in the distribution, when arranged in ascending order.

2.2. Bias of Type I with the standard measures

If we imagine the extreme situation, where all resources are given to one individual or group, share of that person or group will be 1. For n = 20 (and when G is multiplied by 100 and μ is replaced by 1/n, henceforth we will not mention these):

$$G_2 = 100 * \frac{1}{20}(2 * 20 * 1 - 20 - 1)$$

$$= 95.$$

In ideal case, when all resources are equally divided among all, share of each individual or group will be 0.05. For n = 20,

$$G_2 = 100 * \frac{1}{20}\{2(1 * 0.05 + 2 * 0.05 + ... + 20 * 0.05) - 20 - 1\},$$

$$= 0.$$

[7] Income only (rather than share of income), if work is done with absolute income levels.

So, in case of n = 20, the standard measure produce results in 0-95 point scale. It may be checked that, in extreme case, for n = 10, $G_2 = 90$; for n = 5, $G_2 = 80$. The lower limit (in ideal case) is however, always equal to zero.

Alternatively, this inconsistency may also be checked from formula (1). For example, when n = 5, in the extreme case (when all resources are given to one group or individual) we have the following pairs of difference (considering the absolute values only):

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix}.$$

Sum of all pairs of difference is 8.

$$G_1 = 100 \times \frac{1}{2n^2} \times \frac{1}{1/n} \sum_{i=1}^{n} \sum_{j=1}^{n} \left| y_i - y_j \right|,$$

$$= 100 \times \frac{1}{2 \times 5} \times 8,$$

$$= 80.$$

In ideal condition (when all resources are equally divided among all), the pairs of difference are:

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

In that case:

$$G_1 = 0.$$

So, when n = 5, standard measure produces results in 0-80 point scale. So, it is clear that for small observations the standard measure of Gini coefficient does not produce result in 0-100 point scale. A careful observation reveals that in such cases, maximum value depends up on the factor: (n-1)/n. As above, under the standard measure, Gini coefficient is underestimated by 5 per cent when n = 20, by 10 per cent when n = 10, and by 20 per cent when n = 5, and so on. This is an example of underestimation or downward bias of Type I. It is distribution-free, as we understand that the concept of distribution breaks in the extreme case when all

resources are given to one individual. It is rather related to scale. Results can be corrected by the correction factor proposed by Deltas (2003): n/(n-1).

2.3. The alternative formulation to check bias of Type I

According to Kendall (1948, p. 42), in case of obtaining Gini's mean difference without repetition, in the above symmetric matrix, we are confined to elements for all i = 1, 2, 3, …, n and $j \neq i$. In other words, we are confined to the elements in the lower (for $j < i$) and upper (for $j > i$) triangular portions of it except the diagonal ones. The divisor, according to him, will be as many as elements under consideration, exactly n(n-1) for the situation stated above. Now, following Kendall (1948, p. 45), if we consider one set of terms of type $(y_i - y_j)$ only from the lower triangular portion of the matrix, number of terms reduces to: n(n-1)/2. Also, as

$$\sum_{i=1}^{n} \sum_{j \neq i} \left| y_i - y_j \right| = 2 \sum_{i=1}^{n} \sum_{j<i} (y_i - y_j),$$

Gini coefficient appears to be:

$$G_3 = \frac{1}{n(n-1)\mu} \sum_{i=1}^{n} \sum_{j<i} (y_i - y_j). \tag{3}$$

The above formula says that (when we work with absolute income levels) Gini coefficient is one-half the average value of differences between the least possible combinations of pairs of incomes divided by the mean income.

As,

$$\sum_{i=1}^{n} \sum_{j<i} \left| y_i - y_j \right| = \sum_{i=1}^{n} \{(i-1)y_i - (n-i)y_i\},$$

following the similar steps as we do in case of G_2, we have:

$$G_4 = \frac{1}{n(n-1)\mu} (2 \sum_{i=1}^{n} i y_i - n - 1). \tag{4}$$

G_3 and G_4 are same, but expressed somewhat differently. The expression in G_3 is similar to the formula of Gini coefficient proposed by Deaton (1997, p. 139):

$$\gamma = \frac{1}{\mu N(N-1)} \sum_{i>j} \sum_{j} \left| x_i - x_j \right|. \tag{5}$$

They are free from the issue of underestimation, as they produce results in 0-100 point scale as demonstrated below.

In the extreme case, for n = 20:

$$G_4 = 100 * \frac{1}{(20-1)}(2*20*1-20-1)$$

$$= 100.$$

It may be checked that in ideal case, $G_4 = 0$ irrespective of number of observations.

3. THE GEOMETRIC APPROACH

3.1. The standard geometric formula

Equivalence of formula (2) with the ones under geometric approach is demonstrated by Anand (1983, pp. 311-313). However, we derive the geometric formula following Majumder (2007) in order to move with the spirit of the paper[8]. Figure 1 shows a simple illustration of Lorenz curve framework with five quintiles. If the area between the line of equality and the Lorenz curve be A and the total area below the line of absolute equality be (A + B), by definition, Gini coefficient = A / (A + B). As, the total area under the line of absolute equality is 1/2, Gini coefficient appears to be 2A.

If we look at the figure 1, we see that B consists of five (n in general) triangles and four [(n-1) in general] rectangles. Area of each triangle is:

$$\frac{1}{2} \times \text{base} \times \text{height} = \frac{1}{2} \times \frac{1}{n} \times y_i .$$

Sum of all triangles is:

$$\frac{1}{2n}, \text{ as} \sum_{i-1}^{n} y_i = 1.$$

Area of each rectangle is:

$$\text{base} \times \text{height} .$$

Sum area of all (n-1) rectangles is:

$$\frac{1}{n} \times y_1 + \frac{1}{n} \times (y_1 + y_2) + ... + \frac{1}{n} \times (y_1 + y_2 + ... + y_{n-1})$$

[8] As we need to demonstrate the role of the area beyond the maximum inequality Lorenz curve.

or, $\frac{1}{n}[(n-1)y_1 + (n-2)y_2 + ... + \{n-(n-1)\}y_{n-1}]$

or, $\frac{1}{n}\sum_{i=1}^{n-1}(n-i)y_i$

or, $\frac{1}{n}\sum_{i=1}^{n-1}(ny_i - iy_i)$

or, $\sum_{i=1}^{n-1}y_i - \frac{1}{n}\sum_{i=1}^{n-1}iy_i$. (6)

We understand that:

$y_n + \sum_{i=1}^{n-1}y_i = 1$ or,

$\sum_{i=1}^{n-1}y_i = 1 - y_n$.

Also,

$\sum_{i=1}^{n-1}iy_i = \sum_{i=1}^{n}iy_i - ny_n$.

If we replace these results in the above expression (6), the sum area of (n-1) rectangles is:

$(1-y_n) - \frac{1}{n}(\sum_{i=1}^{n}iy_i - ny_n)$, or,

$1 - \frac{1}{n}\sum_{i=1}^{n}iy_i$.

B = Sum area of n triangles + Sum area of (n-1) rectangles,

$$= \left(\frac{1}{2n}\right) + \left(1 - \frac{1}{n}\sum_{i=1}^{n}iy_i\right). \quad (7)$$

A = 1/2 - B, as (A + B) = 1/2.

So, $A = \left(\frac{1}{2} - \left(\frac{1}{2n}\right) - \left(1 - \frac{1}{n}\sum_{i=1}^{n}iy_i\right)\right)$,

$= \left(\left(\frac{n-1}{2n}\right) - \left(1 - \frac{1}{n}\sum_{i=1}^{n}iy_i\right)\right)$,

$$= \frac{1}{2n}(2\sum_{i=1}^{n} iy_i - n - 1) \tag{8}$$

As, G = 2A,

$$G_5 = \frac{1}{n}(2\sum_{i=1}^{n} iy_i - n - 1).$$

Multiplying the numerator and denominator of the above expression by n we have:

$$G_5 = \frac{1}{n^2\mu}(2\sum_{i=1}^{n} iy_i - n - 1), \quad \text{as} \quad \mu = 1/n. \tag{9}$$

$$= G_2.$$

G_5 is the standard formula of computing Gini coefficient under the geometric approach, which corresponds to the case of Gini's mean difference with repetition. It is needless to say that G_5 is subject to the bias of Type I as discussed in case of G_2 in the previous section.

However, formula (9) can be given a geometric look as per the Lorenz curve specification:

$$G_6 = (2\sum_{i=1}^{n} p_i y_i - p - 1), \tag{10}$$

where, p_i = i/n or cumulative proportion of population, and p = 1/n or proportion of population in each group.

3.2. The alternative geometric formulation to check bias of Type I

In figure 2, we spotted one area: C, which is taken away from the area previously denoted by B in figure 1. Now, the effective area beyond the Lorenz curve is $\hat{B}$. A light reasoning will reveal the rationale behind taking away the area denoted by C. When number of observations is sufficiently large, C does not arise and results are expressed in 0-100 point scale. When number of observations is small, C arises and consequently we get underestimated results (not expressed in 0-100 point scale) as it is included in the computation procedure. It is not a part of the maximum inequality Lorenz curve[9] when number of observations is small and its presence in the denominator in the process of computation unnecessarily truncates the scale.

[9] Please see figure 2, as presented by Deltas (2003, p. 229).

Geometrically, when number of observations is small and when all resources are given to one particular group or individual (the n^{th} or the richest one after rearrangement in ordering), Lorenz curve will never coincide with the right-hand side boundary of the triangle below the line of absolute equality. Although it starts from the point (0, 0) and ends at the point (1, 1), it never passes through point (1, 0) in case of small observations. Rather, before reaching the end of the base, it takes a turn to the right-hand side upper corner of the triangle from a point where proportion of population is: (n-1)/n. As we made this illustration with five quintiles, the turn takes place at a point where proportion of population is: (5-1)/5 = 0.8. Also, as n is small, the area C is prominent. If n is sufficiently large, it disappears. However, in principle, such an area should not be taken into account in computation of Gini coefficient, when number of observations is small. In new case the maximum possible area below the Lorenz curve is:

$$\hat{B} = B - C.$$

$$C = \frac{1}{2} \times \text{base} \times \text{height}$$

$$= \frac{1}{2} \times \frac{1}{n} \times 1$$

$$= \frac{1}{2n}. \qquad (11)$$

So,

$\hat{B} = \left(\frac{1}{2n}\right) + \left(1 - \frac{1}{n}\sum_{i=1}^{n} iy_i\right) - \frac{1}{2n}$, substituting the values of B from expression (7).

$$\hat{B} = \left(1 - \frac{1}{n}\sum_{i=1}^{n} iy_i\right).$$

Also, the total effective area below the line of equality is not 1/2, it is (1/2 - C) or,

$$A + \hat{B} = \left(\frac{1}{2} - \frac{1}{2n}\right),$$

$$= \left(\frac{n-1}{2n}\right). \qquad (12)$$

The modified geometric Gini coefficient is:

$$G_7 = \frac{A}{A + \hat{B}}$$

$$= \left(\frac{n}{n-1}\right)2A \text{, putting the results of } (A + \hat{B}) \text{ from expression (12) above.}$$

$$= \left(\frac{n}{n-1}\right)2\left(\frac{1}{2n}(2\sum_{i=1}^{n} iy_i - n - 1)\right) \text{, putting the result of A from expression (8) from the}$$

previous section.

$$= \frac{1}{(n-1)}(2\sum_{i=1}^{n} iy_i - n - 1).$$

Multiplying the numerator and denominator of the above expression by n we have:

$$G_7 = \frac{1}{n(n-1)\mu}(2\sum_{i=1}^{n} iy_i - n - 1), \text{ as } \quad \mu = 1/n. \tag{13}$$

$$= G_3 = G_4.$$

When the geometric formula is based on the maximum possible area below the maximum inequality Lorenz curve, it corresponds to that under the Gini's mean difference without repetition or to the one proposed by Deaton (1997, p. 139) as shown in formula (5). Although Deaton (1997, p. 139) claims that his formula is independent of Lorenz curve framework, we see that it is also related to the geometric one based on the concept of maximum inequality Lorenz curve.

Formula (13) can be given a geometric look as per the Lorenz curve specification, as we do in case of formula (10) as follows:

$$G_8 = \frac{n}{(n-1)}(2\sum_{i=1}^{n} p_i y_i - p - 1), \text{ as } \mu = 1/n,$$

where, p_i = i/n or cumulative proportion of population, and p = 1/n or proportion of population in each group. Now, if we move a step further, we have:

$$G_8 = 2\sum_{i=1}^{n} \hat{p}_i y_i - \hat{p} - k, \tag{14}$$

where, $\hat{p}_i = np_i/(n-1)$, $\hat{p} = np/(n-1)$ and $k = n/(n-1)$. Each factor in G_8 is multiplied by another factor: n/(n-1). This is how it differs from G_6 or any standard measure. It is to be noted that

this factor is identical with the correction factor proposed by Deltas (2003). This further confirms that such a derivation corresponds to Deltas's (2003) proposition in regard to the correction for underestimation. G_7 and G_8 are identical and it is obvious that they are free from bias of Type I. They produce results in 0-100 point scale.

4. THE COVARIANCE APPROACH

4.1. The standard formula under the covariance approach

Under the covariance approach, Gini coefficient is defined by Pyatt, Chen and Fei (1980) and Anand (1983, p. 315)[10] independently in a similar fashion:

$$G_9 = \frac{2}{n\mu}\mathrm{cov}(i, y_i).$$

According to Anand (1983, pp. 315-316), up to a multiplicative constant, Gini coefficient can be expressed as the covariance of income and its rank such that:

$$\mathrm{cov}(i, y_i) = \frac{1}{n}\sum_{i=1}^{n} iy_i - \frac{(n+1)}{2}\mu, \text{ and}$$

$$G_9 = \frac{2}{n^2\mu}\sum_{i=1}^{n} iy_i - \frac{(n+1)}{n}.$$

If we move a step further,

$$G_9 = \frac{1}{n^2\mu}(2\sum_{i=1}^{n} iy_i - n - 1), \qquad (15)$$

$$= G_2 = G_5 = G_6.$$

In case of covariance approach, the standard formula is equivalent to the one based on Gini's mean difference with repetition and the one under geometric approach corresponding to figure 1. G_9 is subject to the inconsistency under discussion and the empirical exercise of section 2.2 is equivalently applicable here.

4.2. Alternative formulation under the covariance approach

Xu (2004, p. 20), who threw some more light on Anand's (1983, pp. 315-316) work, explained that the basic formula under the covariance approach is obtained by taking covariance between y_i and i and dividing i by the number of observations, n (as there are i/n numbers of cumulative proportions of population) such that:

[10] Xu (2004) reports that Anand's thesis, which is the basis of Anand (1983), was completed in 1978.

$$\text{cov}(y_i, i/n) = \frac{1}{n}\text{cov}(y_i, i).$$

If we move to the case of maximum inequality Lorenz curve in figure 2, we realise that effectively there are (n-1) cumulative proportions of population. If n is replaced by (n-1) we have:

$$\text{cov}\big(y_i, i/(n-1)\big) = \frac{1}{n-1}\text{cov}(y_i, i), \text{ or}$$

if we multiply the ranks (or cumulative proportions of population) i/n by the factor: n/(n-1) as we do in case of G_8, we have:

$$\text{cov}\left(y_i, \left(\frac{i}{n} * \frac{n}{n-1}\right)\right) = \frac{1}{n-1}\text{cov}(y_i, i).$$

In response to the above, Gini coefficient appears to be:

$$G_{10} = \frac{2}{(n-1)\mu}\text{cov}(i, y_i),$$

$$= \frac{2}{(n-1)\mu}\left(\frac{1}{n}\sum_{i=1}^{n} i y_i - \frac{(n+1)}{2}\mu\right).$$

$$= \frac{1}{n(n-1)\mu}(2\sum_{i=1}^{n} i y_i - n - 1). \qquad (16)$$

$$= G_3 = G_4 = G_7 = G_8.$$

G_{10} shows that the formula under the covariance approach is equivalent to that based on Gini's mean difference without repetition and the geometric one based on maximum inequality Lorenz curve corresponding to figure 2. G_{10} is free from underestimation due to bias of Type I.

5. TYPE II BIAS AND POSSIBLE GEOMETRIC SOLUTION

5.1. Source of Type II bias

We observed that Type II bias occurs with the Type I when number of initial observations is not sufficiently large and further they get reduced due to grouping. In section 2.2, we see that the maximum value of the Gini coefficient in the extreme case is: (n-1)/n. So, underestimation due to bias of Type I is:

$$1 - \frac{n-1}{n} = \frac{1}{n}.$$

For example, when n = 20, underestimation is 1/20 = 0.5 or 5 per cent (when expressed as percentage). In section 3.2, we take away the area C from B to remove the bias of Type I. In

order to check bias of Type II, we take away n^{th} fraction of the area C, (say, $\hat{C} = C/n$) from B. In new case, the possible area beyond the Lorenz curve is denoted by $\dot{B}$ in figure 3. So,

$$A + \dot{B} = (A + B) - \hat{C}.$$

$$= \frac{1}{2} - \hat{C}. \qquad (17)$$

$$\hat{C} = C \times \frac{1}{n}.$$

Substituting the value of C from expression (11), we have:

$$\hat{C} = \frac{1}{2n} \times \frac{1}{n},$$

$$\hat{C} = \frac{1}{2n^2}.$$

Substituting the value of $\hat{C}$ in expression (17), we have:

$$A + \dot{B} = \left(\frac{1}{2} - \frac{1}{2n^2}\right),$$

$$= \frac{1}{2}\left(\frac{n^2 - 1}{n^2}\right). \qquad (18)$$

The manipulated geometric Gini coefficient is:

$$G_{11} = \frac{A}{A + \dot{B}},$$

$$= \left(\frac{n^2}{n^2 - 1}\right) 2A \text{ , putting the results of } (A + \dot{B}) \text{ from expression (18) above.}$$

Substituting the values of A from expression (8), we have:

$$G_{11} = \left(\frac{n^2}{n^2 - 1}\right) \times 2\left(\frac{1}{2n}(2\sum_{i=1}^{n} iy_i - n - 1)\right),$$

$$= \frac{n^2}{n^2 - 1} \times \frac{1}{n}(2\sum_{i=1}^{n} iy_i - n - 1).$$

Multiplying the numerator and denominator of the above expression by n we have:

$$(\text{Adjusted})G_{11} = \frac{n^2}{n^2 - 1} \times \frac{1}{n^2 \mu}(2\sum_{i=1}^{n} iy_i - n - 1) \text{ , as } \mu = 1/n. \qquad (19)$$

$$= \frac{n^2}{n^2 - 1} \times G_2 .$$

G_{11} is a manipulated formula of Gini coefficient and it checks bias of Type II. The correction factor is the same as derived by Qurti and Clarke (2010). The source of bias is the n^{th} fraction of the area C, as shown in figure 3. As a result, when we take out C/n from computation, the bias is checked.

5.2. Numerical examples of biases of Type I and Type II and methods of correction

One hypothetical example is cited to understand Type I and Type II bias more clearly. We assume an arrangement of income for n = 5200, where the first individual has 1 unit of income, the second individual has 2 units of income and so on such that the 5200^{th} individual has 5200 units of income. We compute total income and make a distribution of income by dividing each individual income by the total income. We have taken n = 5200, to compare our results with that of Milanovic (2010) as cited above. We use simple spreadsheet programme and do the exercise for n = 5200 initially and then squeeze the distribution for n = 20, 10 and 5 respectively. First, we follow a standard formula (say, G_2 or equivalently G_5 or G_9) and then one alternative one (say, G_4 or equivalently G_7 or G_{10}) and summarise the results in table 1.

We see that when n=5200, G = 33.327 in standard measures and G= 33.333 in alternative measures. It is also to be noted that when $n \to \infty$ for a uniform distribution[11], as cited above, G = 33.333. Now, if we move down along column 2 in table 1, we realise presence of bias of Type II. When n is changed from 5200 to 20, it is of 0.25 per cent and is negligible as reported by Milanovic (2010). This continues to increase with the fall in number of observations as shown in column 5. For n = 5, underestimation of Type II is 4 per cent. We need to correct this first:

$$\underset{\text{Standard}}{^{n=5200}}G = \underset{\text{Standard}}{^{n=5}}G \times \frac{n^2}{n^2 - 1},$$

$$= 31.994 \times \frac{5^2}{5^2 - 1},$$

$$= 33.327.$$

As bias of Type I exists with the above result, it has also to be corrected.

[11] Please see table 1 of Deltas (2003, p. 229).

$$\underset{\text{Alternative}}{^{n=5200}}G = \underset{\text{Standard}}{^{n=5200}}G \times \frac{n}{n-1},$$

$$= 33.327 \times \frac{5200}{5200-1},$$

$$= 33.333.$$

Bias of Type III (which is distribution-specific) is absent in a uniform distribution.

We began with the example of Belarus from Milanovic (2010), where G = 28.67 for n = 5227 and G = 28.5 for n = 20. If corrected for Type II bias:

$$\underset{\text{Standard}}{^{n=5227}}G = \underset{\text{Standard}}{^{n=20}}G \times \frac{n^2}{n^2-1},$$

$$= 28.5 \times \frac{20^2}{20^2-1},$$

$$= 28.57.$$

At the second step, it may be corrected for bias of Type I:

$$\underset{\text{Alternative}}{^{n=5227}}G = \underset{\text{Standard}}{^{n=5227}}G \times \frac{n}{n-1},$$

$$= 28.57 \times \frac{5227}{5227-1}$$

$$= 28.58.$$

The remaining part of underestimation (28.67-28.58 = 0.09 or 0.3 per cent) seems to be distribution-specific and hence it may be of Type III. Corrections for Type II bias are shown in column 7 in table 1. After correction, results of the standard measures for small observations ($n \leq 20$) are identical with that based on 5200 observations.

5.3. An example of small-sample bias in the positive direction

At times, positive bias may arise as well when there is an apparent reduction in sample size in a uniform distribution. We draw sample of 20 randomly many times from the above-mentioned uniform distribution. We get few cases of positive bias, one of which is shown in table 2 below. We know that for a uniform distribution (when number of observations is sufficiently large) Gini coefficient is 33.33. In the sampled distribution it is 34.21. It is clear that the result is positively biased. However, as n = 20, it is subject to a downward bias of five per cent. It further implies that at times bias of Type III (distribution-specific) may be positive and in the present case, magnitude of this due to change in distribution in the sample exceeds that (Type I) related to scale.

5.4. Type I bias and the paradox

It is said that underestimation of Type I occurs in all cases when number of observations is small. For example, when n = 20, Gini coefficient is underestimated due to bias of Type I by 5 per cent, as in such a case, standard measures produce results in 0-95 point scale. So, the result of Belarus (Milanovic, 2010) as cited above is also underestimated by 5 per cent for n = 20. When corrected:

$$\underset{\text{Alternative}}{^{n=20}G}_{\text{Belarus}} = 28.5 * \frac{20}{20-1},$$

$$= 30.$$

It is clear that the corrected Gini coefficient is larger than that obtained from micro data. So, a correction for Type I bias may overestimate the Gini coefficient to some extent. Nevertheless, in true sense, this is not an overestimation: 28.5 in 95 (0-95 point scale) means 30 in 100 (0-100 point scale)[12]. This is quite paradoxical. However, it has been an example of downward bias that occurred due to grouping of income into smaller number of categories. It is subject to biases of Type I and Type II together. As demonstrated above, we need to correct it for Type II bias first and then for the Type I to get rid of the paradoxical situation.

However, if we have few observations only, we need to correct it for Type I bias only. For example, when we have n = 5 and the following distribution (say, for Project-Y):

$y_1 = 0.070$, $y_2 = 0.120$, $y_3 = 0.190$, $y_4 = 0.270$, $y_5 = 0.350$;

$$\underset{\text{Standard}}{^{n=5}G}_{Y} = 100 \times \frac{1}{5}[2\{(1\times 0.070) + (2\times 0.170) + (3\times 0.190) + (4\times 0.270) + (5\times 0.350)\} - 5 - 1],$$

$$= 28.4.$$

If corrected for bias of Type I,

$$\underset{\text{Alternative}}{^{n=5}G}_{Y} = 28.4 \times \frac{5}{5-1},$$

$$= 35.50.$$

When n = 10 and we have the following distribution (say, for Project-Z):

$y_1 = 0.007$, $y_2 = 0.027$, $y_3 = 0.060$, $y_4 = 0.097$, $y_5 = 0.110$,

$y_6 = 0.112$, $y_7 = 0.121$, $y_8 = 0.145$, $y_9 = 0.157$, $y_{10} = 0.164$;

$$\underset{\text{Standard}}{^{n=10}G}_{Z} = 100 \times \frac{1}{10}[2\{(1\times 0.007) + (2\times 0.027) + ... + (9\times 0.157) + (10\times 0.164)\} - 10 - 1],$$

[12] Since with given rank order of incomes the Gini index is linear (Chakravarty, 1990, p. 83).

$$= 28.3.$$

If corrected for bias of Type I,

$$^{n=10}_{\text{Alternative}}G_Z = 28.3 \times \frac{10}{10-1},$$

$$= 31.39.$$

In line with the above, correction of results from standard measures and results from alternative derivations are presented in columns 3 and 6 respectively (in table 1). After correction, the results obtained from standard measures are identical with those of alternative ones.

From the above analyses we understand that Gini coefficient is not independent of number of observations[13]. There are two issues to be addressed in this regard. First, even after correction for bias of Type I, Gini coefficients are not comparable for different number of observations (for their dependency on them); and second, a correction makes Gini coefficient inflated to some extent[14]. In order to address these issues we do an exercise to make Gini coefficient standardised and comparable for a fixed number of observations. Primarily, we do an exercise with the previously presented hypothetical example. In table 1, we see that when n = 20:

$$^{n=20}_{\text{Alternative}}G = 34.993.$$

Now, we want to make it standardised for N = 5200 assuming that the pattern of income distribution in cases of n = 20 and N = 5200 remains unchanged. In that case:

$$^{n=20}_{\text{Standardised}}G^{N=5200} = {}^{n=20}_{\text{Alternative}}G \times \frac{n-1}{n} \times \frac{n^2}{n^2-1} \times \frac{N}{N-1}, \quad (20)$$

$$= 34.993 \times \frac{20-1}{20} \times \frac{20^2}{20^2-1} \times \frac{5200}{5200-1},$$

$$= 33.33.$$

From the above hypothetical example we understand that it is possible to make Gini coefficient standardised for a fixed number of observations.

[13] As long as they are not sufficiently large.

[14] Although the point is not valid theoretically, but it is quite paradoxical as we see in case of correction of the result of Belarus (Milanovic, 2010) for bias of Type I in the previous section.

Let us now try with the example of Belarus cited above from Milanovic (2010). After correction for bias of Type I, Gini coefficient is inflated from 28.5 to 30. We will now standardise it for N = 5227:

$$
\begin{aligned}
{}^{n=20}_{\text{Standardised}}G^{N=5227}_{\text{Belarus}} &= {}^{n=20}_{\text{Alternative}}G_{\text{Belarus}} \times \frac{n-1}{n} \times \frac{n^2}{n^2-1} \times \frac{N}{N-1}, \\
&= 30 \times \frac{20-1}{20} \times \frac{20^2}{20^2-1} \times \frac{5227}{5227-1}, \\
&= 28.58.
\end{aligned}
$$

The standardised result, as above, exactly matches with the corrected result presented in section 5.2. From these two exercises we realise that the standardisation technique works well and it is independent of income distribution.

We now apply this technique for standardisation for N = 5200 for the results obtained from the two specific income distributions presented in this section above. We have:

${}^{n=5}_{\text{Alternative}}G_Y = 35.50$ and ${}^{n=10}_{\text{Alternative}}G_Z = 31.39$.

Following expression (20), for the first:

$$
\begin{aligned}
{}^{n=5}_{\text{Standardised}}G^{N=5200}_{Y} &= {}^{n=5}_{\text{Alternative}}G_Y \times \frac{n-1}{n} \times \frac{n^2}{n^2-1} \times \frac{N}{N-1}, \\
&= 35.50 \times \frac{5-1}{5} \times \frac{5^2}{5^2-1} \times \frac{5200}{5200-1}, \\
&= 29.59.
\end{aligned}
$$

Similarly, for the second:

$$
\begin{aligned}
{}^{n=20}_{\text{Standardised}}G^{N=5200}_{Z} &= {}^{n=20}_{\text{Alternative}}G_Z \times \frac{n-1}{n} \times \frac{n^2}{n^2-1} \times \frac{N}{N-1}, \\
&= 31.39 \times \frac{20-1}{20} \times \frac{20^2}{20^2-1} \times \frac{5200}{5200-1}, \\
&= 29.90.
\end{aligned}
$$

After standardisation, Gini coefficients are now comparable:

$${}^{n=5}_{\text{Standardised}}G^{N=5200}_{Y} < {}^{n=20}_{\text{Standardised}}G^{N=5200}_{Z}.$$

Although, before standardisation:

${}^{n=5}_{\text{Standard}}G_Y > {}^{n=20}_{\text{Standard}}G_Z$, and

$$\underset{\text{Alternative}}{^{n=5}G_Y} > \underset{\text{Alternative}}{^{n=20}G_Z}.$$

So, ignoring the distribution-specific bias, we may conclude that inequality in the Project Z is higher than that of Project Y.

6. SOME ISSUES WITH DEFINITION

While commenting on Deaton's formula (1997, p. 139), Deltas (2003, p. 232) comments that although the expression incorporates the small-sample adjustment, "… but it is not equal to twice the area above the Lorentz curve". It is to be kept in mind that G = 2A is not a definition, although it appears to be in the context of a Lorenz curve framework as presented in figure 1. The maximum area of the triangle below the diagonal is 1/2, when we take the ratio of A to 1/2, we get the above result. The standard definition is known to all, for example, according to Sen (1997, p. 30), "… it is the ratio of the difference between the line of absolute equality (the diagonal) and the Lorenz curve to the triangular region underneath the diagonal." Further, to keep in mind that such standard definitions are applicable when number of observations is sufficiently large. And the 'triangular region underneath the diagonal' is to mean the maximum possible area in case of total inequality. Obviously, the maximum area is 1/2 when number of observations is sufficiently large; (1/2 - C) otherwise.

According to Yitzhaki (1998, p. 22), "the best known version of the Gini coefficient is as the area between the Lorenz curve and the 45° line divided by the maximum value of the index." This will be an workable definition for us if we replace the term 'value of the index' by 'area underneath the diagonal'. Deltas (2003, p. 229) is right to say that the maximum value of the Gini coefficient equals two times the area beneath the diagonal of the maximum inequality Lorenz curve (figure 2 in his paper, p. 229), that is: (n-1)/n. It follows that the maximum area beneath the diagonal is: (n-1)/2n. If we take the ratio of A to (n-1)/2n:

$$G = \left(\frac{n}{n-1}\right)2A = G_7,$$

not simply 2A. It is 2A multiplied by his own correction factor: n/(n-1), which is proposed from 'geometric intuition' only. The derivation is shown geometrically (taking away the area C from B) in section 3.2 while deriving G_7 above.

However, if we modify the definition of Yitzhaki (1998, p. 22), Gini coefficient is:

$$G = \frac{A}{\text{Max. area underneath the diagonal}}.$$

$$= \frac{A}{(n-1)/2n},$$

$$= \left(\frac{n}{n-1}\right)2A,$$

$$= G_7 = G_4,$$

where G_7 and G_4 are alternative formulations of Gini coefficient related to the maximum inequality Lorenz curve framework and the concept of Gini's mean difference without repetition respectively.

7. CONCLUSION

We begin with a popular concern on granularity of measurements and consequent downward bias in Gini coefficient. We review theoretical and empirical literature and do not find any consensus on the issue in regard to their types, magnitude and the methods of correction. We identify three types of bias, two of which are distribution-free. The remaining one is uncertain and distribution-specific. Of the first two, one occurs due to availability of few observations only, as in such cases the standard measures do not produce results in 0-100 point scale. The other occurs due to grouping of income into relatively smaller number of parts. And the both occur together if initial number of observations is not sufficiently large and further they get reduced due to grouping. At times, bias may be positive as well when there is an apparent reduction in sample size, if sampling is done from a uniform distribution. However, this paper deals with the distribution-free downward biases only associated with the first three cases stated above. Underestimations associated with the first, when there are few observations only, is demonstrated and addressed with simplicity, for discontinuous case, with alternative formulation following the principle of Gini's mean difference without repetition. Equivalences of it are also derived under the geometric and covariance approaches. We see that in such cases downward bias appears to be five per cent for n = 20, ten per cent for n = 10 and twenty per cent for n = 5, and so on. In case of grouping of income into relatively smaller number of parts (with equally sized groups) bias is estimated as a quarter of a per cent for n = 20, one per cent for n = 10 and four per cent for n = 5 and so on. When both the biases arise together, a straightforward claim of the first (which is related to scale) in its full magnitude may be unwarranted and quite paradoxical. Adjustments in this case are done accordingly with a newly proposed operational pursuit, where the bias that arises due to grouping is corrected first followed by correction for the one that is related to

scale. However, when there are few observations only, alternative measures, or standard measures with correction for the bias seem to inflate Gini coefficient to some extent, although theoretically it is not an overestimation. So, unlike the point from where we began, granularity of measurements, in such cases, may work in the positive direction. Consequently, we do some exercises to make Gini coefficient standardised and comparable for a fixed number of observations. We also dealt with issues related to definition and find that the area beyond the maximum inequality Lorenz curve plays the crucial role to eliminate biases in cases of small observations with conformity to the concept of Gini's mean difference without repetition. By addressing the issues of small-size distribution-free downward biases of Gini coefficient methodologically with a discreet approach with simplicity and synchronising the relevant previous and present concerns, this study tries to contribute significantly a gamut of new knowledge to the existing literature.

REFERENCES

Anand, S. (1983), *Inequality and Poverty in Malaysia: Measurement and Decomposition*, Oxford University Press, New York.

Chakravarty, S. R. (1990), *Ethical Social Index Numbers*, Springer-Verlag, New York.

Deaton, A. (1997), *The Analysis of Household Surveys: A Microeconomic Approach to Development Policy*, The Johns Hopkins University Press, Baltimore.

Deltas, G. (2003), The Small-Sample Bias of the Gini Coefficient: Results and Implications for Empirical Research, The Review of Economics and Statistics, Vol. 85, 226-234.

Kendall, M. G. (1948), *The Advanced Theory of Statistics*, Vol. 1, (Fourth edition), Charles Griffin, London .

Lerman, R. and S. Yitzhaki. (1989), Improving the Accuracy of Estimates of Gini Coefficients, Journal of Econometrics, Vol. 42, 43–47.

Majumder, A. (2007), Alternative measures of economic inequality, Artha Beekshan, Vol. 16, 3-20.

Milanovic, B. (1994), The Gini-type functions: an alternative derivation, Bulletin of Economic Research, Vol. 46, 81-90.

Milanovic, B. (1997), A simple way to calculate the Gini coefficient, and some implications, Economics Letters, Vol. 56, 45-49.

Milanovic, B. (2010), Global inequality recalculated and updated: the effect of new PPP estimates on global inequality and 2005 estimates, Journal of Economic Inequality, (Published online: 16 November 2010, DOI 10.1007/s10888-010-9155-y).

Pyatt, D., Chen, C.-N., and Fei, J. (1980), The distribution of Income by Factor Components', Quarterly Journal of Economics, Vol. 95, 451-473.

Qurti, T. V. and P. Clarke. (2010), A simple correction to remove the bias of the Gini coefficient due to grouping, The Review of Economics and Statistics, Vol. 93, 982–994.

Sen, A.K. (1997), *On Economic Inequality*, Clarendon Press, Oxford.

Subramanian, S. (1997), *Measurement of Inequality and Poverty*, Oxford University Press, New Delhi.

Xu, K. (2004), How Has the Literature on Gini's Index Evolved in the Past 80 Years?, available at http://economics.dal.ca/RePEc/dal/wparch/howgini.pdf, accessed: 07 July 2011.

World Bank. (2003), Module 5: Inequality Measures, available at http://info.worldbank.org/etools/docs/library/93518/Hung_0603/Hu_0603/Module5MeasuringInequality.pdf, accessed: 07 July 2011.

Yitzhaki, S. (1998), More Than A Dozen Alternative Ways Of Spelling Gini, in: Slottje, D. J, ed., *Research on Economic Inequality*, JAI Press, London, 13-30.

Figures

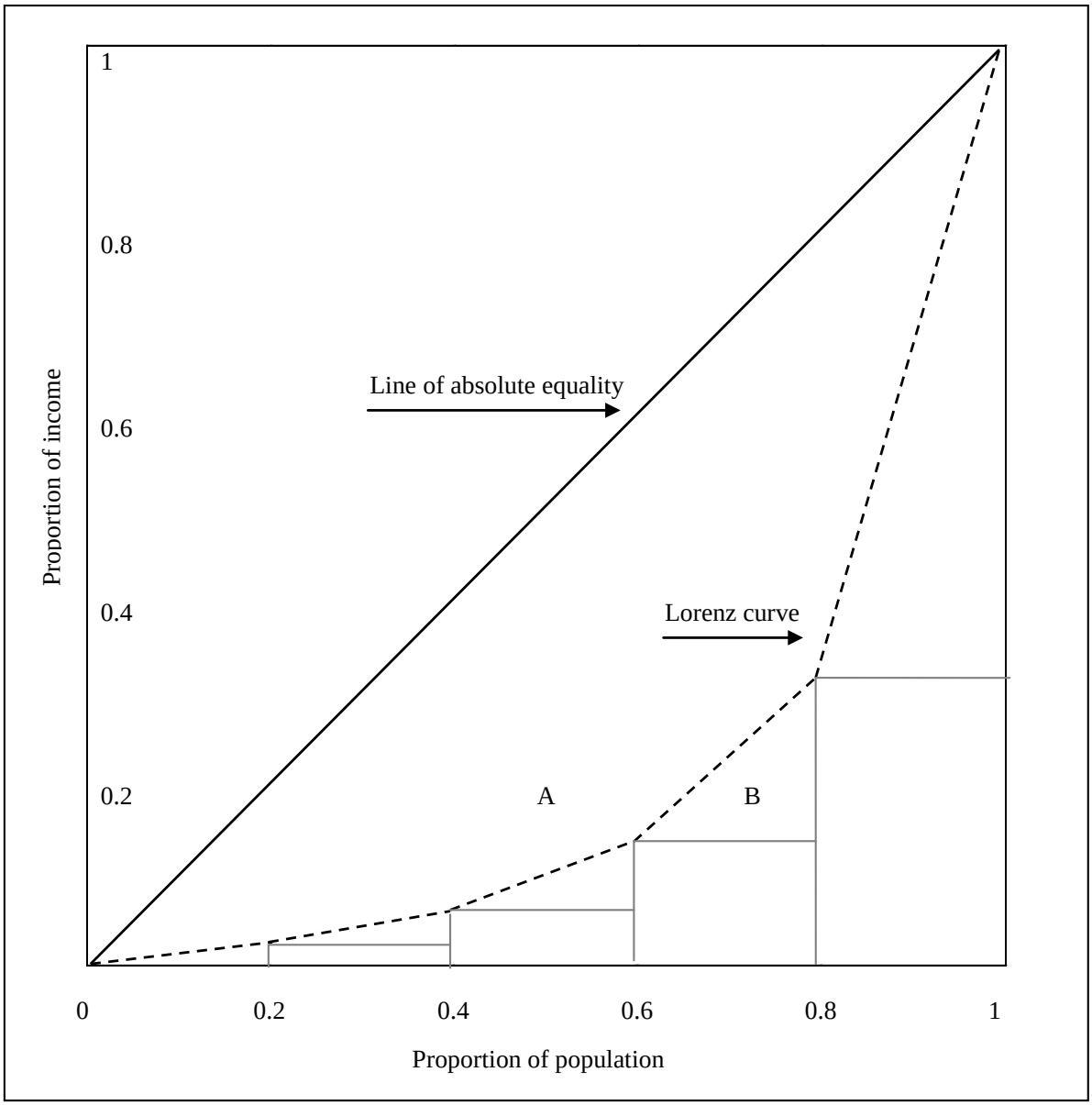

Figure 1

Standard illustration of Lorenz curve framework with five quintiles

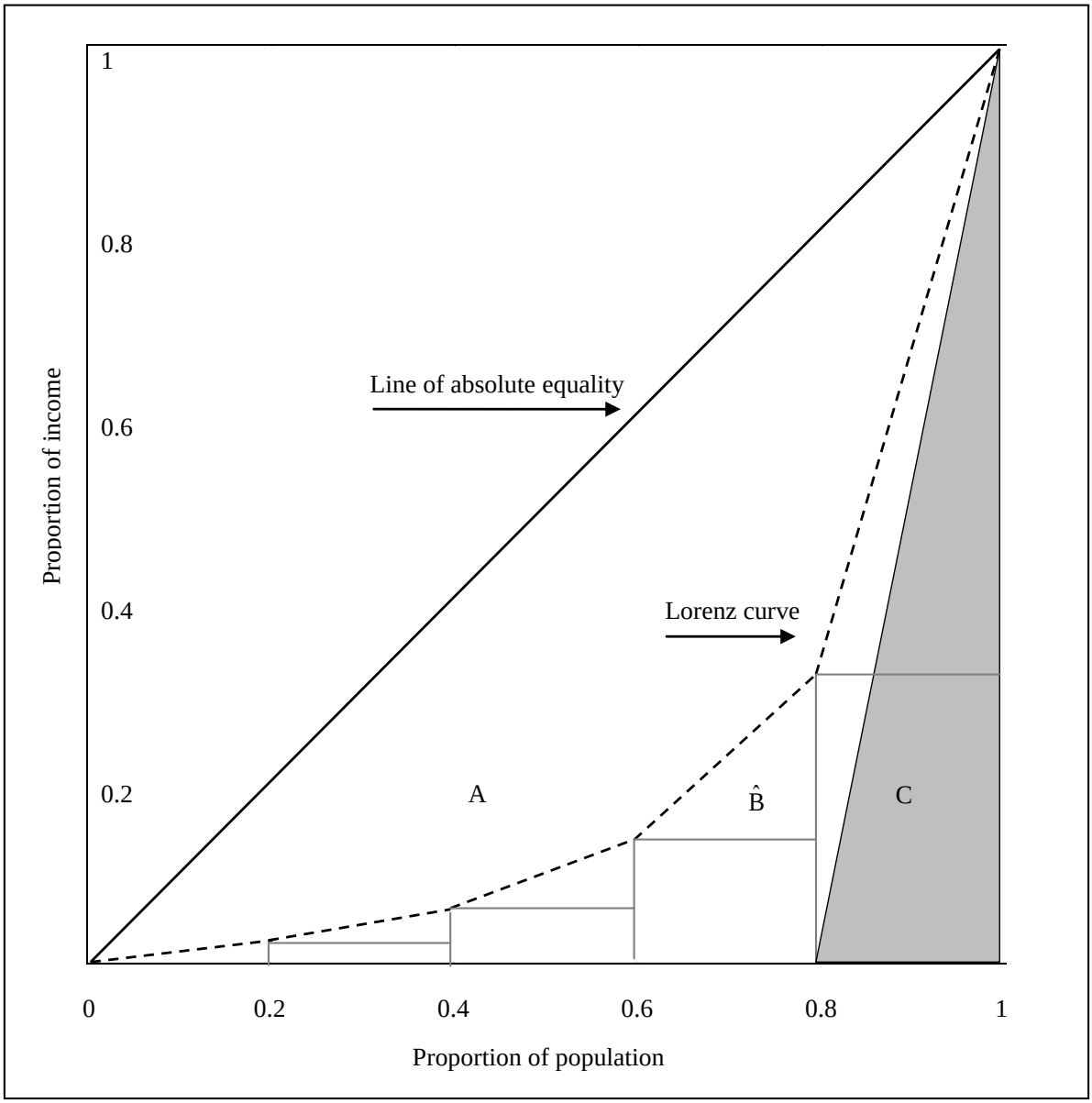

Figure 2

Illustration of maximum inequality Lorenz curve with five quintiles to check Type I bias

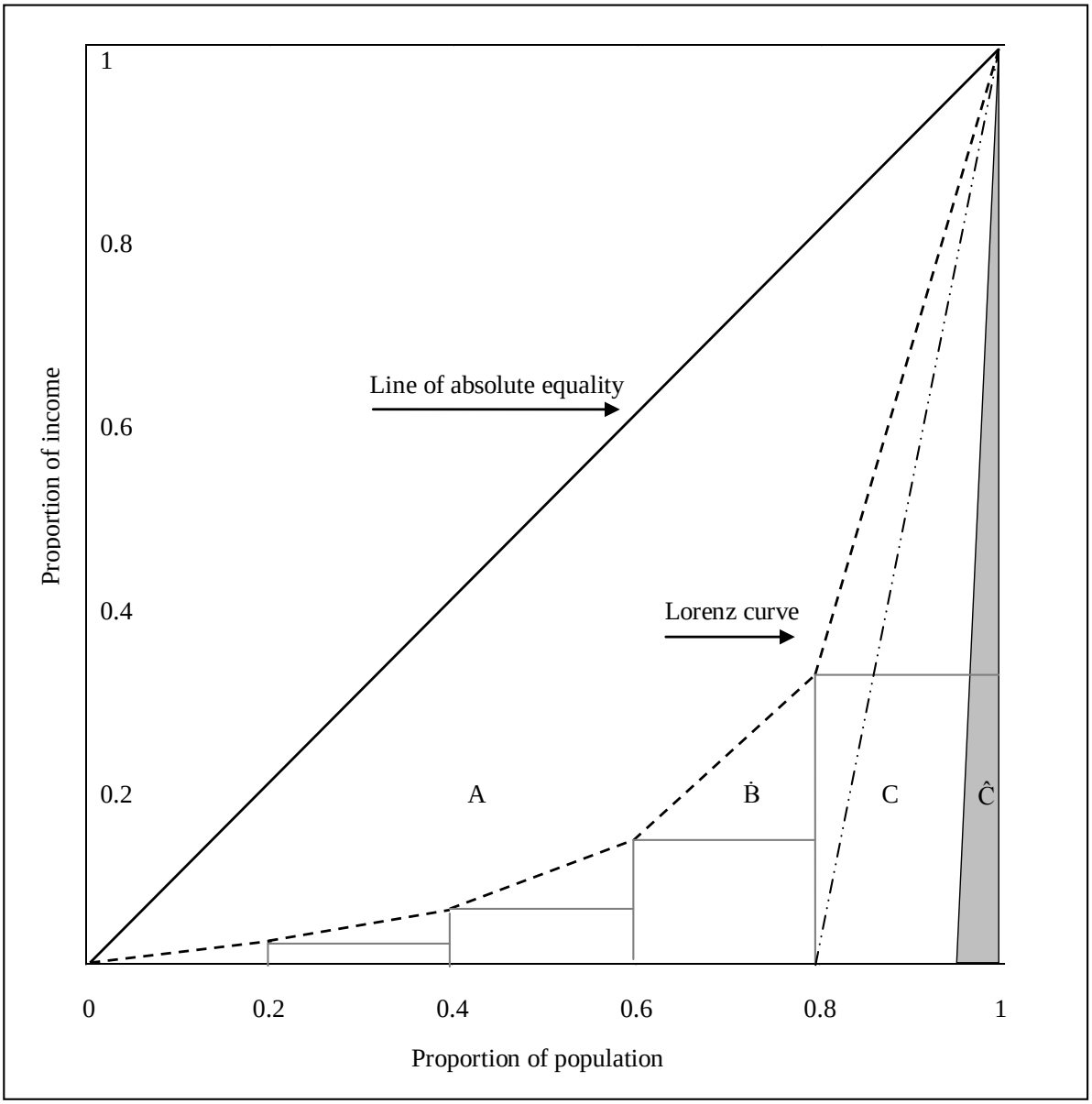

Figure 3

Illustration of Lorenz curve with five quintiles to check Type II bias

Tables

Table 1. Underestimation of Gini coefficient: standard Vs. alternative measures

n	Measures of Gini coefficient (G)		Underestimation (%)		Correction for Type I & II underestimation	
	Standard	Alternative*	Type I	Type II	Type I: Standard G* n/(n-1)	Type II: Standard G* $n^2/(n^2-1)$
(1)	(2)	(3)	(4)	(5)	(6)	(7)
5200	33.327	33.333	0.02	-	33.333	-
20	33.244	34.993	5	0.25	34.993	33.327
10	32.994	36.660	10	1.00	36.660	33.327
5	31.994	39.993	20	4.00	39.992	33.327

* G_{11} is not considered
Source: Self elaboration

Table 2. An example of small-sample bias in the positive direction♣

Rank (i)	Income (Y)	Distribution (y_i)	$i*y_i$
1	181	0.004	0.004
2	337	0.007	0.013
3	431	0.009	0.026
4	499	0.010	0.040
5	878	0.017	0.087
6	1433	0.029	0.171
7	2136	0.042	0.297
8	2302	0.046	0.366
9	2344	0.047	0.419
10	2408	0.048	0.478
11	2423	0.048	0.529
12	2937	0.058	0.700
13	3016	0.060	0.779
14	3104	0.062	0.863
15	3324	0.066	0.990
16	3664	0.073	1.164
17	4025	0.080	1.359
18	4634	0.092	1.657
19	5120	0.102	1.932
20	5158	0.102	2.049
Total	50354	1.000	13.921

♣ A sample of 20 is drawn randomly from the uniform distribution (as cited above) with N = 5200, where Gini coefficient = 33.33.

Gini coefficient (in the above sample) = [100{(2*13.921) – 20 – 1}/20] = 34.21.