A note on the relationship between top income shares and the Gini coefficient

Facundo Alvaredo

Department of Economics, Oxford University, Manor Road Building, Manor Road, OX1 3UQ, Oxford, UK and CONICET-UTDT, Miñones 2177, C1428ATG, Buenos Aires, Argentina

A R T I C L E   I N F O

Article history:
Received 15 January 2010
Received in revised form 1 September 2010
Accepted 5 October 2010
Available online 13 October 2010

JEL classification:
D31
H2

Keywords:
Gini coefficient
Top income shares
Pareto distribution

A B S T R A C T

When a very top group of the income distribution, infinitesimal in numbers, owns a finite share \( S \) of total income, the Gini coefficient \( G \) can be approximated by \( G^*(1-S) + S \), where \( G^* \) is the Gini coefficient for the rest of the population. We provide a simple formal proof for this expression, give a general formula of the relationship when the top group is not infinitesimal, and offer two applications as illustrations.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

In a typical income distribution, the rich may appear insignificant. The most commonly used measure of inequality, the Gini coefficient, is more sensitive to transfers at the center of the distribution than at the tails. In a textbook-sized Lorenz curve, the top 0.1% or even the top 1% are scarcely distinguishable on the horizontal axis from the vertical endpoint. However, changes in top income shares are capable of impacting on changes in overall inequality significantly, as advanced by Atkinson (2007): “If we treat the very top group as infinitesimal in numbers, but with a finite share \( S \) of total income, then the Gini coefficient \( G \) can be approximated by \( G^*(1-S) + S \), where \( G^* \) is the Gini coefficient for the rest of the population” (p. 19). The relevance of the last expression has increased with the recent developments of the literature on top incomes (Atkinson and Piketty, 2007, 2010) and the comparison of inequality statistics from survey data and tax records (Burkhauser et al., forthcoming; Leigh, 2007).

The purposes of this note are (i) to provide a simple formal proof of the last statement about the connection between top income shares and the Gini coefficient when the top group is infinitesimal (not given in Atkinson, 2007), (ii) to give a general formula of the relationship when the top group is not infinitesimal, and (iii) to offer two illustrative examples of their application: survey data are usually affected by severe under-reporting at the top, and it is possible to improve the survey-based Gini coefficients by incorporating top income shares estimates coming from other sources (typically tax data).

From a graphical perspective, Atkinson’s result is rather intuitive: when the very top group owns a large share of total income \( S \), the Lorenz curve \( L(p) \) almost touches the right \( y \)-axis at \( 1-S \). Let us call \( L^*(p) \) the Lorenz curve for the non-top group (the bottom 99%, the bottom 99.9%, etc.). Given that \( L(p) = L^*(p)(1-S) \), and that the Gini coefficient \( G \) (in continuous space) is \( 1 - 2 \int L(p)dp \), then it is straightforward to note that \( G = 1 - 2 \int L^*(p)(1-S)dp = G^*(1-S) + S \). More formally, we start from the decomposition of the Gini coefficient in discrete space proposed by Dagum (1997).

2. The decomposition of the Gini coefficient

Let us consider a population of \( N \) individuals with mean income \( \mu \), partitioned in \( j = 1,2,...,k \) non-overlapping subpopulations of \( N_j \) individuals with mean income \( \mu_j \). Each individual \( i \) in group \( j \) has income \( y_{ij} \). The Gini coefficient of the whole population is

\[
G = \frac{\sum_{j=1}^{k} \sum_{h=1}^{j} \sum_{i=1}^{N_j} \sum_{r=1}^{N_h} |y_{ij} - y_{hj}|}{2N^2\mu}
\]

The Gini coefficient within the \( j \)-th group (simply the Gini of the \( j \)-th group) is

\[
G_{jj} = \frac{\sum_{i=1}^{N_j} \sum_{r=1}^{N_j} |y_{ij} - y_{jr}|}{2N_j^2\mu_j}
\]
The Gini coefficient between the \( j \)-th and the \( h \)-th groups is (Dagum, 1987)

\[
G_{jh} = \sum_{i=1}^{N_j} \sum_{h=1}^{N_h} \frac{|y_{ij} - y_{ih}|}{N_j N_h (\mu_j + \mu_h)}
\]

from which it is straightforward to note that \( G_{jh} = G_{hj} \).

Let \( P_j \) be the \( j \)-th group share in total population

\[
P_j = \frac{N_j}{N}
\]

and \( S_j \) the \( j \)-th group income share

\[
S_j = \frac{N_j \mu_j}{N \mu}
\]

Dagum (1997) has shown that the Gini coefficient for the whole population can be decomposed as follows:

\[
G = \sum_{j=1}^{k} G_{j} P_j S_j + \sum_{j=1}^{k-1} \sum_{h=1}^{j} G_{h} (P_j S_h + P_h S_j) = G_w + G_b
\]

\( G_w \) measures the contribution of inequality within groups, and \( G_b \) measures the contribution of inequality between groups.

3. Top income shares

We consider a population partitioned in two (\( k = 2 \)). In group \( j = 1 \) we have individuals at the top of the distribution (e.g. the top 0.01%, the top 0.1%, etc.), with income share \( S \) and population share \( P \). The rest of the population is in group \( j = 2 \), with income share \( 1 - S \) and population share \( 1 - P \). Then Eq. (1) can be expressed as

\[
G = G_{11}PS + G_{22}(1-P)(1-S) + G_{12}(1-S) + G_{12}(1-P)S
\]

\[
= G_{11}PS + G_{22}(1-P)(1-S) + G_{12}(1-P)S + G_{12}(1-P+S)
\]

In this case (with only 2 subpopulations and with higher-income individuals in \( j = 1 \)), \( G_b \) can be further simplified:

\[
G_b = G_{12}(1-S) + (1-P)S
\]

\[
= \sum_{i=1}^{N_1} \sum_{h=1}^{N_2} \frac{|y_{ij} - y_{ih}|}{N_1 N_2 (\mu_1 + \mu_2)} (1-P)(1-S)
\]

\[
= \frac{1}{\mu_1 + \mu_2} (1-P)(1-S) + (1-P)S
\]

\[= (1-P)S-P(1-S)\]

\[= S-P\]

This is equivalent to the result described graphically in Atkinson and Bourguignon (2000), pp. 7–8, for the two-class case. Incorporating Eq. (3) in Eq. (2) and relabeling the Gini coefficients for the non-top and the top groups as \( G^* \) and \( G^{**} \), we get the general formula

\[
G = G^{**}PS + G^*(1-P)(1-S) + S-P
\]

For a very top group, infinitesimal in numbers \((P \to 0)\), but with a finite share \( S \) of total income, we have

\[
\lim_{P \to 0} G^{**}PS + G^*(1-P)(1-S) + S-P = G^*(1-S) + S
\]

When the top group is small but not infinitesimal, the general formula given in Eq. (4) can be transformed in a useful way under the assumption that the distribution at the top takes the Pareto form, with Pareto coefficient \( \alpha \), or inverted-Pareto coefficient \( \beta = \frac{1}{\alpha} \). In this case \( G^{**} \) can be easily expressed as a decreasing function of \( \alpha \), or an increasing function of \( \beta \)

\[
G^{**} = \frac{1}{2\alpha - 1} = \frac{\beta - 1}{\beta + 1}
\]

and then (4) becomes

\[
G = \frac{\beta - 1}{\beta + 1} PS + G^*(1-P)(1-S) + S-P.
\]

Expressions (5) and (6) can be useful empirically: when working with survey data, generally affected by severe under-reporting not only for the top 1%, but also for groups as large as the top 5% or top 10%, it is possible to improve the survey-based Gini coefficients by incorporating top income shares estimates coming from other sources (typically tax data). In the next section we show how both formulas differ in practical cases.

4. Applications

4.1. Case 1: United States

Burkhauser et al. (forthcoming) have tried to reconcile Piketty and Saez’s (2003) tax-based top income share series with top income shares from the United States internal CPS. The internal CPS is less affected by top code than the public CPS. They find that their CPS-based top income share series closely match the Piketty and Saez’s series for the top 10–1% (the top decile excluding the top percentile). However, even if the top-code effect is less pervasive, the top 1% measured by the internal CPS is consistently lower than the top 1% measured with tax data.

According to the results in Burkhauser et al. (forthcoming), the internal CPS Gini in the United States increased from 50.3 in 1976 to 58.8 in 2006, the change between those 2 years (net of measurement adjustments in 1992–1993) being 6.2 percentage points. With the formulas in (5) we “corrected” G using G* and the top 1% share from tax data (Table 1). G increased from 52.8 to 64.5 (top share including capital gains) and from 52.3 to 62.3 (top share excluding capital gains) over the same period. If the series including capital gains are taken as benchmark, then the rise in G, 11.7 percentage points, is almost twice as large as the 6.2 percentage points increase recorded by the CPS series. As Atkinson, Piketty and Saez (2009) state, “the top percentile plays a major role in the increase in the Gini over the last three decades and CPS data which do not measure top incomes fail to capture about half of this increase in overall inequality”.

Taking the top 1% group as infinitesimal is a rough approximation that can be improved by applying the formula given in Eq. (6). This is done in columns 9 and 10 of Table 1. The Gini coefficients thus obtained are lower, but the increase in G is very similar: 11.8

1 The average income above a given threshold is \( \beta \) times that threshold. A higher \( \beta \) (lower \( \alpha \)) coefficient generally means larger top income shares and higher income inequality.

2 For a formal proof, see Atkinson and Brown (1954), p. 101.

3 These values, taken from Burkhauser et al. (forthcoming), Table C1, correspond to the income distribution of tax units (not households), and were chosen for comparability with the unit of analysis of Piketty and Saez (2003).

4 The results in columns 5–7 of Table 1 are numerically different from those in Atkinson et al. (2009) for two reasons: (i) these authors pinned down G* for the bottom 99% of the population from Expression (5), while we take it from the direct computations on CPS data from Burkhauser et al. (forthcoming); and (ii) the Gini coefficients in Atkinson et al. (2009) correspond to the household distribution, while we use the tax unit distribution from Burkhauser et al. that is more comparable to the tax-based top income share estimates. The qualitative results are of course the same.
Discrepancies are due to rounding.

Following Expression (6), and using the inverted-Pareto coefficient in column (8), the Gini coefficient in column (7) is computed as 100\(^*\)((1.69−1)/(1.69 + 1))\(^*\)0.01\(*\)0.089+0.469\(*\)0.99\(*\)1/(1−0.089)+0.089−0.01)=51.4. Discrepancies are due to rounding.

### Table 1

<table>
<thead>
<tr>
<th>Year</th>
<th>Top 1% income share from CPS data (%)</th>
<th>Top 1% income share from tax data excluding K gains (%)</th>
<th>Top 1% income share from tax data including K gains (%)</th>
<th>Gini Coeff. G based on Pareto Coefficient (α)</th>
<th>Gini Coeff. G corrected with tax-based top 1% excluding K gains (%)</th>
<th>Gini Coeff. G corrected with tax-based top 1% including K gains (%)</th>
<th>Case A: top group considered infinitesimal</th>
<th>Case B: top group considered small but not infinitesimal</th>
</tr>
</thead>
<tbody>
<tr>
<td>1976</td>
<td>6.7</td>
<td>7.9</td>
<td>8.9</td>
<td>50.3</td>
<td>48.2</td>
<td>52.3</td>
<td>52.8</td>
<td>1.69</td>
</tr>
<tr>
<td>2006</td>
<td>13.7</td>
<td>18.0</td>
<td>22.8</td>
<td>58.8</td>
<td>54.0</td>
<td>62.3</td>
<td>64.5</td>
<td>2.83</td>
</tr>
</tbody>
</table>

Sources: Columns (2)-(3) and (8): taken from Atkinson, Piketty and Saez (2009); columns (1) and (4)-(5): reproduced from Burkhauser et al. (forthcoming); and columns (6)-(7) and (9)-(10): author's calculations.

Notes: Following expression (5), and using the Gini coefficient \(G^*\) for the bottom 99% for 1976 (column 5) and the tax-based top 1% income share including capital gains (column 3), the corrected Gini coefficient in column (7) is computed as 100\(^*\)((1.69−1)/(1.69 + 1))\(^*\)0.01\(*\)0.089+0.469\(*\)0.99\(*\)1/(1−0.089)+0.089−0.01)=51.4. Discrepancies are due to rounding.

![Table 2](image)

**Table 2**

<table>
<thead>
<tr>
<th>Year</th>
<th>Top 1% income share from tax data (%)</th>
<th>Top 0.1% income share from tax data (%)</th>
<th>Gini Coeff. G*</th>
<th>Gini Coeff. G corrected with the top 0.1% income share</th>
<th>Gini Coeff. G corrected with the top 1% income share</th>
<th>Inverted-Pareto coefficient (\beta)</th>
<th>Gini Coeff. G corrected with tax-based top 1% income share</th>
<th>Gini Coeff. G corrected with tax-based top 0.1% income share</th>
</tr>
</thead>
<tbody>
<tr>
<td>1997</td>
<td>12.39</td>
<td>4.27</td>
<td>46.9</td>
<td>53.5</td>
<td>49.2</td>
<td>2.16</td>
<td>52.1</td>
<td>49.0</td>
</tr>
<tr>
<td>1998</td>
<td>12.57</td>
<td>4.37</td>
<td>48.5</td>
<td>55.0</td>
<td>50.8</td>
<td>2.18</td>
<td>53.6</td>
<td>50.6</td>
</tr>
<tr>
<td>1999</td>
<td>13.53</td>
<td>5.22</td>
<td>47.0</td>
<td>54.1</td>
<td>49.7</td>
<td>2.42</td>
<td>52.8</td>
<td>49.6</td>
</tr>
<tr>
<td>2000</td>
<td>14.34</td>
<td>5.68</td>
<td>48.6</td>
<td>56.0</td>
<td>51.6</td>
<td>2.49</td>
<td>54.7</td>
<td>51.4</td>
</tr>
<tr>
<td>2001</td>
<td>12.91</td>
<td>5.22</td>
<td>51.1</td>
<td>57.4</td>
<td>53.6</td>
<td>2.54</td>
<td>56.0</td>
<td>53.5</td>
</tr>
<tr>
<td>2002</td>
<td>15.53</td>
<td>6.92</td>
<td>51.9</td>
<td>59.4</td>
<td>55.2</td>
<td>2.85</td>
<td>58.0</td>
<td>55.1</td>
</tr>
<tr>
<td>2003</td>
<td>16.85</td>
<td>7.40</td>
<td>50.9</td>
<td>59.2</td>
<td>54.6</td>
<td>2.80</td>
<td>57.8</td>
<td>54.4</td>
</tr>
<tr>
<td>2004</td>
<td>16.75</td>
<td>7.02</td>
<td>48.8</td>
<td>57.3</td>
<td>52.4</td>
<td>2.65</td>
<td>56.0</td>
<td>52.2</td>
</tr>
</tbody>
</table>

Source: Top shares in columns (1) and (2) are taken from Alvaredo (2010).

Notes: \(G^*\) denotes the Gini coefficient of individual income based on the Greater Buenos Aires households’ survey. All results correspond to October surveys, except for 2003 (May). Only income recipients with positive income were considered and no further adjustments were applied. The Greater Buenos Aires households survey is taken as representative of Argentina. The \(\beta\) coefficients reported in column (6) are computed by using the top income shares series in \(\beta = \alpha/\log(10)\) where \(\alpha\) is the income share of the top 1% and \(\log(10)\) is the Gini coefficient \(G^*\) in column (4) is computed as \(100\times(0.469\times(1−0.089)+0.089−0.01)=51.4\). Discrepancies are due to rounding.

percentage points for the Gini coefficient corrected with the tax-base 1% income share estimate including capital gains, and 10.1% percentage points for the series excluding capital gains.

### 4.2. Case 2: Argentina

Szákeley and Hilgert (1999) have analyzed a large number of Latin American surveys to confirm that surveys’ top incomes generally correspond to the prototype of highly educated relatively rich families or capital owners. They find that the income of the ten richest households is generally similar to the average wage of a manager of a medium to large size firm (and, in many cases, even below that level). We take the case of Argentina as the second example. Table 2 displays the tax-based top 1% and top 0.1% income shares from Alvaredo (2010) and the survey-based Gini coefficient \(G^*\) between 1997 and 2004. We computed G in two hypothetical cases, namely that the top 1% and the top 0.1% are not represented in the surveys, both considering formula (5) and formula (6). Three unsurprising facts are readily noticeable. Firstly, G can be several percentage points above \(G^*\). Secondly, not only can levels be different, but also the trends of G and \(G^*\) can diverge. According to the survey’s results, \(G^*\) displays virtually no change when 2001 and 2003 are compared, going from 51.1 to 50.9. However, G “corrected” with the top 1% income share (column 4) was 57.4 in 2001 and 59.2 in 2003 (almost a two percentage points increase). Finally, the discrepancy between the two formulas is larger, the larger the top group considered.

If top incomes ignored by surveys experience a large enough relative increase, then the true dynamics of overall inequality may display a rising trend even when survey-based estimates show very few differences. Discrepancies are due to rounding.

5 Slightly different from the strategy followed in the previous subsection, in this case we assume that top individuals are completely ignored by the survey, and we directly consider \(G^*\) the result arising from the whole survey sample.
opposite results. As long as surveys do not record what is happening with the true distribution at the top, survey-based estimates showing a decline in inequality can at most indicate that those reductions are happening within non-top individuals.

References