On the Mathematics of Income Inequality: Splitting the Gini Index in Two

Robert T. Jantzen (robert.jantzen@villanova.edu) and Klaus Volpert (klaus.volpert@villanova.edu)
Dept of Mathematics and Statistics
Villanova University, Villanova, PA 19085
November 29, 2011

Abstract
A two-parameter model for the Lorenz curve describing income distribution interpolating between self-similar behavior at the low and high ends of the income spectrum naturally leads to two separate Gini indices describing the low and high ends individually. These new indices accurately capture realistic data on income distribution and give a better picture of how income data is shifting over time.

Introduction
In early 2011, Forbes Magazine [1] reported that hedge fund manager John Paulson’s income for 2010 was $5 billion (give or take $100 million). That’s a staggering amount. Just consider that it takes no fewer than 50,000 professors with an average income of $100,000 (we are being generous) to equal that income. That’s more than all the math professors in this country combined. Looking in the other direction of the ladder of wealth, it is not implausible to imagine that there would be 50,000 among the world’s poor whose combined wealth is less than that of an average professor in the US. The world’s economic inequality is stunning.

But is it static? Is it simply a given, something one cannot do anything to change? One of the authors’ high school teachers memorably made clear that inequality is a fact of life: “Even if one could distribute all the money in the world equally for once”, he said, “inequality would be back in an instant. For one man would take his money to the bank, another to the bar.”

So is inequality just the story of the wise and the foolish?

A closer look at income data reveals that the degree of inequality is not at all a constant, but rather varies greatly from country to country, and even within
one country over time. Consider, for example, the following data on the share of national income (including capital gains) in the US received by the richest 1% over the period 1913–2008, obtained from the IRS by Emmanuel Saez of UC Berkeley and Thomas Piketty of the Paris School of Economics [2].

By this measure, we can see a peak of inequality was reached at the height of the stock market frenzy of the 1920’s, when the top 1% took a share of 23.5% of the national income and again in 2007 right before the latest financial crisis. The low was reached in the 1970’s, when this share was a ‘mere’ 9%. This indicates a remarkable variation in inequality over a 100-year period, and provides grist for sociological inquiries: what causes the rising and falling of inequality? Is it change in the political leadership? Is it mainly economic factors? Or is it reflective of a fundamental change in what our society values?

To help figure this out, we first turn our attention to a more mathematical problem: as vivid an illustration this 1%-index yields, it is not sufficiently representative of income inequality. By concentrating on the top 1% this measure says little about the welfare of the middle class or the poor. So how could we more accurately portray income inequality so that it remains simple, yet comprehensive? The well-known Gini-Index is a first answer. But in this paper we shall go beyond that index and introduce and promote the use of a two-parameter index, that is derived from a two-parameter family of functions used to model the Lorenz curve of incomes. This family was first proposed in [3], but our derivation, interpretation and analysis is new.
The Lorenz Curve and the Gini-Index

A long-known index that is more comprehensive than the 1%-measure is the Gini-Index, named after Italian statistician Corrado Gini (1884-1965) who defined it in 1912 [4]. It is based on the Lorenz curve, introduced by American economist Max Lorenz in 1905 [5]. This curve plots the percentage $L(x)$ of the total income of a population that is cumulatively earned by the bottom $x\%$ of the population in terms of income levels. In 2007 in the US, for example, the value was $L(99) = .765$, since the top 1% took 23.5%, and the bottom 99% took the remaining 76.5%. Instead of using percentages, we use the corresponding proper fractions, so that the graph of $L(x)$ becomes a curve from the origin $(0,0)$ to the point $(1,1)$ in the unit square representing the two anchoring points: none of the people have none of the wealth, while all of the people have all of the wealth.

In a (Utopian) society, where everyone has the same income, the Lorenz curve would approach the $L(x) = x$ line. The other extreme is the case where one person receives all the income, in which case $L(x) = 0$ for all $x < 1$, and $L(1) = 1$. A realistic Lorenz curve lies between these extremes.

The more equal the distribution of income, the closer the curve is to the diagonal line. It is therefore natural that one defines as a measure of inequality the area between the diagonal and the Lorenz curve, as a percentage of the entire area between the two extremes (which is $1/2$). This number is the Gini-Index, and its formula is therefore
\[ G = 2 \int_0^1 x - L(x) \, dx. \]

As an example, if \( L(x) = x^p \) with \( p \geq 1 \), one quickly calculates that \( G = \frac{p-1}{p+1} \). So this index is a summary measure, giving weight to the income shares of the poor and the wealthy. Here are the latest numbers for different countries, compiled by the CIA [6].

<table>
<thead>
<tr>
<th>Country</th>
<th>Gini-Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sweden</td>
<td>.230</td>
</tr>
<tr>
<td>Germany</td>
<td>.270</td>
</tr>
<tr>
<td>Egypt</td>
<td>.344</td>
</tr>
<tr>
<td>China</td>
<td>.415</td>
</tr>
<tr>
<td>Russia</td>
<td>.422</td>
</tr>
<tr>
<td>Brazil</td>
<td>.539</td>
</tr>
<tr>
<td>South Africa</td>
<td>.650</td>
</tr>
</tbody>
</table>

One needs to take these numbers with a grain of salt, as the precision of the economic data is not everywhere equally good. Also, other factors, such as the size of the basic economic units (e.g., household, family or individuals) and what counts as income (e.g., capital gains, government support) are not always strictly comparable. Finally these data are from different years, and, as we have mentioned, the values do change with time.

How does the US stack up?

It turns out that the US Census Bureau [7] has compiled and published data on the household income distribution here, covering the years 1967–2010, allowing us to calculate the Gini-Index ourselves. Our only problem is, that the data are published in quintiles, i.e., we only know the values of \( L(x) \) for \( x = 0, .2, .4, .6, .8 \) and 1. For example, for 2009, the values given in [7] are

\[
\begin{array}{cccccc}
 x & 0 & .2 & .4 & .6 & .8 & 1 \\
 L(x) & 0 & .034 & .120 & .266 & .498 & 1 \\
\end{array}
\]

The data are plotted as
How shall we calculate the Gini-Index from this? Numerical integration using the trapezoidal rule (connecting the dots by straight line segments) would substantially underestimate the Gini-Index. A better approach is to approximate the data with a smooth curve, but what type of curve should we use? The first idea might be to use the lowest power polynomial that fits well. If we use a fifth-degree polynomial, the fit looks very good:

Because of the number of data points, and the number of coefficients in the
polynomial, the curve goes exactly through each data point. The curve is

\[ L(x) = 0.17083 x - 0.70208 x^2 + 4.70833 x^3 - 6.82292 x^4 + 3.64583 x^5 \]

and the Gini-Index with this curve calculates to \( G = 0.457 \).

However, there are two problems with this:

1. The Gini-Index calculated and published for 2009 by the US-Census-Bureau, presumably using their complete data, is \( G = 0.468 \), a substantial discrepancy. So while the model curve meets every data point, it does not do the right things in between the points.

2. The coefficients of the polynomial yield no information, because the model lacks any economic meaning. In other words, when we compare modeling equations over the years, the coefficients change substantially without giving us a clue what is really happening in the different segments of society.

The Phenomenon of (Right-Sided) Self-Similarity

A hint what to do comes from more data by Piketty and Saez: consider that in 2008 the top 10% received 48.23% of the total income (including capital gains). The top 1% received 20.95%, the top .1% had 10.40%, the top .01% had 5.03%. There is a pattern: The degree of inequality repeats even as we restrict ourselves to richer and richer slivers of the population! Informally expressed, the top 10% of the whole population got roughly half of the total pie. The top 1% got again (almost) half of that pie. The top .1% again half that pie, and so on. One could say that ‘money begets money’. We shall call this phenomenon a (right-sided) self-similarity. It is also known as the Pareto-principle. Strikingly, with small deviations, it is present in the Piketty-Saez data in all years from 1913–2007, even as the degree of inequality varies. For example, in this graph we plot the ratio of income shares of the top .1% to the income shares of the top 1% (with asterisks), and the ratio of income shares of the top 1% to the income shares of the top 10% (with squares). They are remarkably similar over nearly 100 years.
What equation for the Lorenz function would model such a phenomenon? Let us assume that the richest segment $R$ of the population receives a portion $P$ of the total income. Assume furthermore that this ratio repeats itself within that segment, and so on. Then the Lorenz-function must have the following property: $L(1 - R^N) = 1 - P^N$ for any $N$. Solve the input $x = 1 - R^N$ of this unknown function $L(x) = 1 - P^N$ for $N$ in terms of $x$ and then substitute the result into the right hand side of $L(x)$, also using the fact that $a^\ln(b) = b^{\ln(a)}$, to find that $L(x) = 1 - (1 - x)^q$. We’ll abbreviate this to $L(x) = 1 - (1 - x)^q$ where $0 < q \leq 1$ for this to be a Lorenz curve. This model is known as the Pareto Distribution. Here is a plot of a family of functions of this form, with $q$ ranging over the reciprocals of 1 through 10. Note that $q = 1$ corresponds to complete equality and $q \to 0$ to complete inequality.
For this income distribution, the inequality repeats also in the sense that the Gini-Index is right-invariant. In other words, for any cutoff point \( r \) between 0 and 1, the Gini-Index among the richest \( r \% \) is the same as in the whole population:

To make this statement more precise, let us define for any Lorenz curve \( L(x) \) the right-sided Gini-function as

\[
Gr(r) = \frac{\int_r^1 \left( \frac{1 - L(r)}{1 - r} (x - 1) + 1 \right) - L(x) \, dx}{\frac{1}{2}(1 - r)(1 - L(r))},
\]

that is, as the area between the Lorenz curve from a point \( x = r \) to \( x = 1 \) and the idealized line of equal distribution among that segment, again as a percentage of the maximum area of a completely unequal distribution among that segment.
In the case that \( L(x) = 1 - (1 - x)^q \) we calculate that

\[
Gr(r) = \frac{\int_r^1 \left( \frac{1 - (1 - (1 - r)^q)}{1 - r} (x - 1) + 1 \right) - (1 - (1 - r)^q) \, dx}{\frac{1 - r}{1 - (1 - r)^q}} = \frac{1 - q}{1 + q}
\]

which is independent of \( r \) and equal to the Gini-index for the whole population. So with the Pareto distribution the Gini-Index is right-invariant.

Sociologically speaking, this self-similarity phenomenon explains why rich people might not ‘feel’ so rich. Making $1 million a year places a person easily in the top 1% of earners in the US, yet this income pales in comparison to the neighbor who makes $1 billion.

How good is this function in modeling a realistic Lorenz curve? The 2008 data of Saez and Pickety for the top US-incomes are well modeled with \( R = 0.01 \) and \( P = 0.213 \) (hence \( q = 0.336 \)), especially for the top 5%:

However, when we try to extend this model to the Census Bureau data for the entire population we see that it does not really work:
The right-sided self-similarity holds only in the upper echelons of earners. The economical dynamics in the lower echelons is different.

**Left-Sided Self-Similarity**

Let’s return to the Census data. A loose reading suggests a left-sided self-similarity: the bottom 40% receives only slightly more than one fourth of what the bottom 80% receives. The bottom 20% receives only slightly more than one fourth of what the bottom 40% receives, and so on. Even among the poor there appears to be a gradient of inequality comparable to the inequality in a larger segment of the society.

Mirroring the argument for right-sided similarity, the Lorenz curve $L(x)$ with strict left-sided self-similarity would obey $L(R^N) = P^N$, where the poorest segment $R$ of the population receives the proportion $P$ of the total income. Solving $x = R^N$ for $N$, substituting on the right side we find that such a Lorenz curve would have to have the form $L(x) = x^{\frac{\ln(P)}{\ln(R)}} = x^p$, a simple power function, where the power must satisfy $1 < p$ for it to be below the line of strict equality. However, strict self-similarity is not warranted by the data. But a modified model $L(x) = cx^p$ works well among the bottom 60%: for example for 2008, we have the least-square-fit curve $L(x) = 0.71135 x^{1.9288}$. (This curve cannot be used on the entire interval since it falls short of the value 1 at $x = 1$.)
This power law equation still exhibits the property of the left-invariance of the Gini-Index—for any segment between 0 and \( R < 1 \), the Gini-Index over that segment is the same (the figure illustrates the case \( c = 1 \)):

![Graph illustrating the case c = 1](image)

The proof is again easy. We define the left-sided Gini-function as

\[
G_l(s) = \frac{\int_0^s L(s) - L(x) \, dx}{\frac{1}{2} s L(s)}.
\]

If \( L(x) = cx^p \), we calculate that

\[
G_l(s) = \frac{\int_0^s c s^p x - cx^p \, dx}{\frac{1}{2} (s)(c)(s^p)} = \frac{p - 1}{p + 1} = G
\]

which is independent of \( s \), and equal to the Gini-index \( G \) of the entire population. So if the Lorenz function is a power law function, the Gini-index is
left-invariant. Note that setting $p = 1/q$ transforms this Gini index into that of the Pareto distribution, corresponding to the fact that reflection across the line $y = 1 - x$ transforms the power function $y = x^{1/q}$ into the corresponding Pareto distribution with exponent $q$.

The Hybrid Model

So we have two models that work well in their respective segments of the population: the Pareto distribution among the highest incomes, the power law function among the lowest incomes. How can we find a model equation that marries the two regimes? Motivated by the fact that the product of any two Lorenz curves is again a Lorenz curve (any nondecreasing nonnegative concave-up function on the interval $[0,1]$ between the endpoints $(0,0)$ and $(1,1)$ has the right properties, and any product of two such curves has the same properties), we can combine the high income fit of the self-similar model with the low income fit by power law functions to get a hybrid model which interpolates between the two endpoint behaviors. The result works exceptionally well over the entire population:

$$L(x) = x^p (1 - (1 - x)^q).$$

(One might ask what happened to the extra parameter $c$. It turns out, perhaps surprisingly, that $q$ plays the role of $c$, so that a third parameter $c$ is unnecessary. As a rule, the fewer parameters, the more natural the model, so we’ll work with just two parameters $p$ and $q$.)

This model behaves correctly asymptotically in the two limiting regimes:

**Proposition:**

- as $x \to 0$, $L(x) \to x^p (qx) = q x^{p+1}$, a power law function. Moreover, the left-sided Gini-function $G_l(s)$ approaches the value of the Gini-index for this power law function, i.e., $\lim_{s \to 0} G_l(s) = \frac{(p+1)-1}{(p+1)+1} = \frac{p}{p+2}$

- as $x \to 1$, $L(x) \to 1 - (1 - x)^q$, the Pareto distribution. Moreover, the right-sided Gini-function $G_r(r)$ approaches the value of the Gini-index for the Pareto distribution, i.e., $\lim_{r \to 1} G_r(r) = \frac{1-q}{1+q}$

**Proof:**

The asymptotics just rephrase that

$$\lim_{x \to 0} \frac{L(x)}{q} x^{p+1} = \lim_{x \to 0} \frac{1 - (1 - x)^q}{q x} = \lim_{x \to 0} \frac{q (1 - x)^{q-1}}{q} = 1$$

and $\lim_{x \to 1} \frac{L(x)}{1 - (1 - x)^q} = \lim_{x \to 1} x^p = 1$. 

12
As for the limit of $Gl(s)$, the calculation requires two successive applications of L’Hospital’s Rule:

$$\lim_{s \to 0} Gl(s) = \lim_{s \to 0} \frac{\int_0^s x^p (1 - (1 - x)^q) dx - x^p (1 - (1 - x)^q)}{s^2 (s^p)(1 - (1 - s)^q)}$$

$$= \lim_{s \to 0} \frac{(p - 1) s^{p-2} (1 - (1 - s)^q) + q (1 - s)^{q-1} s^{p-1}}{s^p (p + 1) (1 - (1 - s)^q) + sq (1 - s)^{q-1}} s^2$$

$$= \frac{(p - 1) q + q}{(p + 1) q + q} = \frac{p}{p + 2}$$

The calculation of $\lim_{r \to 1} Gr(r)$, though a bit lengthy, is similar.

This model works exceedingly well. For example, for 2009, the least-square-fit curve is $L(x) = x^{0.79869} \left(1 - (1 - x)^{0.56077}\right)$.

The Gini-Index with this model is calculated to be 468, which is identical to the value of the index published by the Census Bureau. Recall that the quintic model had estimated it to be 457. Thus the hybrid model is a great fit achieved with only two degrees of freedom (instead of 5). This goodness-of-fit holds for all the Census Bureau data from 1967 to 2010 with the estimated Gini-Index (connected line) consistently within .003 of the published values (diamond points).
Notice the near-steady increase of the index from .386 in 1968 to .469 in 2010. Notice also that as of 2010, the Gini-Index of .469 places the US between Russia and Brazil in terms of income inequality in the above table.

Splitting the Index in Two

This hybrid model has several advantages:

- The fit of the data is excellent across a long range of longitudinal data.
- It can distinguish between intersecting Lorenz curves with identical Gini-Index.
- It requires only two degrees of freedom.
- The parameters have economic meaning that yield interesting information about the development of inequality that is not apparent from the Gini-Index alone.

To better harness this economic information of the two parameters $p$ and $q$, we transform them into Gini-like indices $G_0$ and $G_1$, where $G_0 = \frac{p}{p + 2}$ and $G_1 = \frac{1 - q}{1 + q}$. $G_0$ can be thought of as the ‘low-end’ Gini-index, while $G_1$ is the ‘high-end’ Gini-index. More precisely, the above proposition showed that $G_0$ and $G_1$ are the limits of the left-sided (respectively right-sided) Gini-function as $s$ goes to 0 (respectively as $r$ goes to 1).
Remark.
The Gini index for this hybrid model is an analytic function of the two separate Gini indices.

Proof.
Inverting the two fractional linear relationships gives
\[ p = \frac{2G_0}{1-G_0}, \quad q = \frac{1-G_1}{1+G_1}, \]
while the Gini index integral may be evaluated exactly in terms of the Gamma function:
\[ G = 2 \int_0^1 x - x^p (1 - (1 - x)^q) \, dx = 1 - \frac{2}{(p+1)} + 2 \frac{\Gamma(1+q) \Gamma(p+1)}{\Gamma(2+p+q)}. \]
Composing \( G \) with the expressions for \( p \) and \( q \) yields \( G \) as an explicit function of \( G_0 \) and \( G_1 \).
Note that this allows a reparametrization of the hybrid model Lorenz curve by parameters confined to the unit square with real geometric significance lacked by the exponents \( p \) and \( q \)
\[
L(x) = x^{1-G_0} \left( 1 - (1 - x)^{1-G_1} \right), \quad 0 \leq G_i \leq 1,
\]
where each separate Gini index reduces to the Gini index when the other vanishes, describing the limiting power law model \((G_1 = 0)\) and limiting Pareto model \((G_0 = 0)\), whose intersection gives the line of perfect equality \((G_0 = 0 = G_1)\). The limiting case of perfect inequality corresponds to the boundary \( G_0 \to 1 \) or \( G_1 \to 1 \). The plot of \( G(G_0, G_1) \) is confined to the unit cube. We propose to call this hybrid expression for the Lorenz curve the Pareto-power-law interpolation Lorenz curve.
Let us tabulate the historical development of these two indices: circles indicate $G_0$, asterisks $G_1$.

We see that between 1967 and 2010, income inequality evolved very differently at the two opposite poles of society. This was not apparent in the Gini-Index alone. It sheds interesting light on research by Piketty and Saez.
who have shown that most of the economic gains of the last 30 years have gone to the upper income brackets [2]. In that process the income gradient in this bracket has increased substantially, with the gains ever more skewed to the highest incomes of society. The low-income portion of the population largely did not participate in this process with only a slight increase in the income gradient. Poverty rates in the US have essentially stayed constant from 1975 until 2007 [8]. Median incomes (inflation-adjusted) increased only by an average of .6% per year over this time period [9], while gains in the upper income brackets were the greater the higher the bracket [2].

The dynamics resembles that of a gold rush taking place at the front of a pack of people, pulling apart the fastest runners, while the slow majority hardly changes pace, as if not knowing that there is gold to be had.

What exactly was the gold? There were at least two major speculative bubbles (dot-com, real estate) in this time-period, but it was chiefly the gains of the economy through higher productivity, which kept increasing steadily [10]. There was a rush to grab a large share of those gains producing some exceedingly high incomes. As for what exactly brought on this rush, economic explanations might vary, but factors that certainly played an important role include [11]:

- government policies such as deregulation and privatization;
- the emergence of stock options for compensation of executives;
- financial derivatives (e.g., credit default swaps) that allowed banks and hedge funds to steadily increase risk, leverage and maximum payouts;
- low interest rates that made credit (and leverage) cheap;
- lowered marginal tax rates, that did little to slow down the top earners.

The frequently cited factor of globalization seems less persuasive, as that is not really a new phenomenon and was present even when inequality decreased (as in the 1950’s). Also, some countries have recently decreased their inequality (e.g, South Korea, Brazil) despite the fact that they too face globalization. As for the question of whether fundamental values have changed in our society, and whether they will change again, we’ll leave that to our reader to decide.

References


