Applications of Mathematics

Petr Zizler

Gini indices and the moments of the share density function

*Applications of Mathematics*, Vol. 59 (2014), No. 2, 167–175

Persistent URL: `http://dml.cz/dmlcz/143627`

# GINI INDICES AND THE MOMENTS OF THE SHARE DENSITY FUNCTION

Petr Zizler, Calgary

(Received August 13, 2012)

*Abstract.* The expected value of the share density of the income distribution can be expressed in terms of the Gini index. The variance of the share density of the income distribution is interesting because it gives a relationship between the first and the second order Gini indices. We find an expression for this variance and, as a result, we obtain some nontrivial bounds on these Gini indices. We propose new statistics on the income distribution based on the higher moments of the share density function. These new statistics are easily computable from the higher order Gini indices. Relating these moments to higher order Ginis suggests new estimates on these quantities.

*Keywords*: Gini index; income distribution; share density function

*MSC 2010*: 62P20, 91B15

## 1. Introduction

One of the summary measures of how income is distributed in a society is the Gini index. There is a vast literature on the subject but for a recent mathematical treatment of the Gini index we refer the reader to [2]. This single number statistic measures how equitably the income is spread in a population. It is defined as

$$G = 2\int_0^1 (p - L(p))\,\mathrm{d}p,$$

where $L(p)$ is the Lorenz curve. Here the quantity $p$, referred to as the percentile variable, is the fraction of the population, ranging from zero to one, that holds $L(p)$ proportion of the whole income. Note that we must have $L(0) = 0$ and $L(1) = 1$. It is not difficult to see that the graph of the function $L$ must be convex, see [2]. For example, if 60 % of the households in a society hold 20 % of the income, we then have

$L(0.6) = 0.2$. Geometrically speaking, the Gini index $G$ is twice the area between the graph of $y = p$ and the graph of $y = L(p)$ for $p \in [0, 1]$. The reason for the factor of 2 is to have the Gini index range from zero to one as opposed to a range from zero to one half.

We assume the ideal scenario where the function $L$ is a real-valued function on the interval $[0, 1]$ which is twice differentiable. At the perfectly equitable income distribution we have $L(p) = p$ and therefore $G = 0$. The maximal value for $G$ is the value of 1 which occurs in the extreme case in which all income is concentrated at a point (one household). Technically speaking, the maximum value of $G = 1$ is attained when $L(p) = 0$ for all $p \in [0, 1)$ and $L(p) = 1$ for $p = 1$, which is clearly not a diffentiable function. We can get as close to $G = 1$ as we wish with our function $L$ and we will treat the extreme case as thus attainable for simplicity of our exposition.

There are also Gini indices of higher order defined, see [2] for example, as

$$G_k = k(k+1) \int_0^1 (p - L(p))(1-p)^{k-1} \, \mathrm{d}p,$$

where the extra weighting factor of $(1-p)^{k-1}$ is added to weight the extreme poverty more than the original Gini index $G$ does. Note the values of all $G_k$ also range from zero to one with $G_k = 0$ in the case of the perfectly equitable income distribution. The value $G_k = 1$ is attained in the extreme case of all the income being concentrated at a point (one household).

If we consider the first derivative of $L$ with respect to $p$

$$\frac{\mathrm{d}L}{\mathrm{d}p} = s(p)$$

we obtain the *share density* function that measures the share of the whole that is owned by the portion of the population that falls in the given percentile range. In the case of a perfectly equitable income distribution we have the constant share density function $s(p) = 1$ for all $p \in [0, 1]$.

The expected value of the share density $s(p)$ can be thought of as the percentile level of a household which earns the average dollar. This concept was introduced in [2] and is given by

$$\bar{p} = \int_0^1 ps(p) \, \mathrm{d}p$$

and it was shown in [2] that there is a nice relationship between $\bar{p}$ and $G$, namely $G = 2\bar{p} - 1$. In our paper we give an expression for the variance of the share density function which in turn will yield a set of inequalities between the first and the second order Gini indices.

**Lemma 2.1.** *If $s(p)$ is the share density function, then its second moment is given by*

$$\int_0^1 p^2 s(p)\,\mathrm{d}p = \frac{1}{3} - \frac{1}{3}G_2 + G.$$

Proof. From the definition of the Gini index $G$ we observe

$$\int_0^1 L(p)\,\mathrm{d}p = \frac{1-G}{2}.$$

Recall

$$\begin{aligned} G_2 &= 6\int_0^1 (p - L(p))(1-p)\,\mathrm{d}p \\ &= 6\int_0^1 p - L(p)\,\mathrm{d}p - 6\int_0^1 p(p-L(p))\,\mathrm{d}p \\ &= 3G - 6\int_0^1 p^2\,\mathrm{d}p + 6\int_0^1 pL(p)\,\mathrm{d}p \\ &= 3G - 2 + 6\int_0^1 pL(p)\,\mathrm{d}p \end{aligned}$$

and thus we have

$$\int_0^1 pL(p)\,\mathrm{d}p = \frac{1}{3} + \frac{1}{6}G_2 - \frac{1}{2}G.$$

Now we obtain the desired result

$$\begin{aligned} \int_0^1 p^2 s(p)\,\mathrm{d}p &= p^2 L(p)]_0^1 - 2\int_0^1 pL(p)\,\mathrm{d}p \\ &= \frac{1}{3} - \frac{1}{3}G_2 + G. \end{aligned}$$

□

**Theorem 2.1.** *Let $s(p)$ be the share density function. Then its variance is given by*

$$\int_0^1 (p-\bar{p})^2 s(p)\,\mathrm{d}p = \frac{1}{12} - \frac{1}{3}G_2 + \frac{1}{2}G - \frac{1}{4}G^2.$$

Proof. Using Lemma 2.1 we have, recalling $G = 2\bar{p} - 1$,

$$\begin{aligned} \int_0^1 (p-\bar{p})^2 s(p)\,\mathrm{d}p &= \int_0^1 p^2 s(p)\,\mathrm{d}p - 2\bar{p}\int_0^1 ps(p)\,\mathrm{d}p + \bar{p}^2\int_0^1 s(p)\,\mathrm{d}p \\ &= \int_0^1 p^2 s(p)\,\mathrm{d}p - \bar{p}^2 \\ &= \frac{1}{12} - \frac{1}{3}G_2 + \frac{1}{2}G - \frac{1}{4}G^2. \end{aligned}$$

□

**Corollary 2.1.** *If $G$ is the Gini index and $G_2$ is the Gini index of the second order, then*

$$\frac{3}{2}G - \frac{3}{4}G^2 \leqslant G_2 \leqslant \frac{1}{4} + \frac{3}{2}G - \frac{3}{4}G^2.$$

P r o o f. Observe the function $s(p)$ is continuous and nondecreasing. Moreover, it is not difficult to see that among all nondecreasing continuous density functions on $[0,1]$ the one with the highest variance is the constant function $s(p) = 1$ for all $p \in [0,1]$ and the variance is equal to $1/12$.

Using Theorem 2.1, we must have

$$\frac{1}{12} - \frac{1}{3}G_2 + \frac{1}{2}G - \frac{1}{4}G^2 \leqslant \frac{1}{12}$$

and hence

$$G_2 \geqslant \frac{3}{2}G - \frac{3}{4}G^2.$$

The inequality $G_2 \leqslant 1/4 + (3/2)G - (3/4)G^2$ follows directly from the fact that variance must be nonnegative, in particular $1/12 - (1/3)G_2 + (1/2)G - (1/4)G^2 \geqslant 0$.

□

One of the consequences of the above inequalities is the following. For a fixed Gini index $G$, we observe the higher the second order Gini index $G_2$ is, the lower the variation of the share density becomes. Thus we can claim that the second order Gini index carries important information about the variation of the share density.

Various data in regards to income distributions or the Gini index are readily available. Pikkety and Saez report [1] that the world Gini index increased from 0.61 in 1910 to 0.657 in 1992. On the same note, the US Gini index increased from 0.398 in 1976 to 0.470 in 2006. The 2006 second order US Gini index can be obtained from [2], where Farris estimated it from the 2006 US income data using numerical techniques. It is reported that $G_2 = 0.61$.

For the world wide data, in the year 1910 the percentile level of the household which earned the average dollar was $0.5(1 + G) = 0.8050$ (80.5 %). In the year 1992 it increased to 0.8285 (82.9 %). Note the increase of almost 3 % over this time period of this statistic. Using the 2006 US data, the percentile level of the household which earned the average dollar was $0.5(1 + G) = 0.7350$ (73.5 %). Since we know the $G_2$ value as well for this distribution we can go further with our analysis. The variance of the share density function (which needs the $G_2$ value) was given by 0.0598, which corresponds to the standard deviation of about 24.5 %. Note that the maximal variance would occur at the perfectly equitable income distribution and its value would be $1/12 \approx 0.083$, with the standard deviation of about 28.9 %.

Another interpretation of the summary Gini indices is as follows, see the work of Farris [2] and the work of Leiber and Kotz [3]. Consider the following experiment where you pick at random $k$ household incomes and record their lowest value $Y_k^{\min}$. Let $\overline{X}$ denote the mean household income. Farris showed [2] that

$$\frac{\overline{Y_k^{\min}}}{\overline{X}} = 1 - G_k,$$

where $\overline{Y_k^{\min}}$ is the expected value for $Y_k^{\min}$. Ideas presented in [2] can be generalized.

*Fact:* Consider the following experiment where you randomly choose 2 household incomes and record their highest value $Y_2^{\max}$. Let $\overline{X}$ denote the mean household income. Then the expected value for $Y_2^{\max}$, $\overline{Y_2^{\max}}$, divided by $\overline{X}$ is given by

$$\frac{\overline{Y_2^{\max}}}{\overline{X}} = 1 + G.$$

Proof. Let $X$ denote the random variable of the (household) income in a society. Let $f(x)$ denote its density function. Let $x$ be a given (household) income value. Now the probability $P$ that both incomes chosen are lower than $x$ is given by

$$P(Y < x) = P(\text{first income} < x) \cdot P(\text{second income} < x) = F^2(x).$$

The following are key observations:

$$\frac{\mathrm{d}F}{\mathrm{d}x} = f(x);\ F(x) = \int_0^x f(s)\,\mathrm{d}s = p$$

and now we have

$$\frac{\overline{Y_2^{\max}}}{\overline{X}} = 2\int_0^1 ps(p)\,\mathrm{d}p = 1 + G.$$

□

It is interesting to note that no matter how extreme the Gini index value is, $G \approx 1$, the expected value for the maximum income variable (chosen out of two random incomes) can never exceed twice the mean income in the society. The expected value for $Y_2^{\max}$ is given by $1+G$ times the mean income $\overline{X}$. As a proportion of the maximal attainable value of $1 + G = 2$ we have a statistic $(1 + G)/2$ which happens to equal $\bar{p}$ which is the percentile level of the household which earns the average dollar.

**Lemma 2.2.** *Consider the following experiment where you randomly select 3 household incomes and record their highest value $Y_3^{\max}$. Let $\overline{X}$ denote the mean household income. Then the expected value for $Y_3^{\max}$, $\overline{Y_3^{\max}}$, divided by $\overline{X}$ is given by*

$$\frac{\overline{Y_3^{\max}}}{\overline{X}} = 1 - G_2 + 3G.$$

P r o o f. With the notation as above, let $X$ denote the random variable of the (household) income in a society. Let $f(x)$ denote its density function. Let $x$ be a given (household) income value. Now the probability $P$ that all three incomes chosen are less than $x$ is given by

$$\begin{aligned} P(Y < x) &= P(\text{1st income} < x) \cdot P(\text{2nd income} < x) \cdot P(\text{3rd income} < x) \\ &= F^3(x). \end{aligned}$$

Now we have, using Lemma 2.1,

$$\frac{\overline{Y_3^{\max}}}{\overline{X}} = 3 \int_0^1 p^2 s(p)\, \mathrm{d}p = 1 - G_2 + 3G.$$

□

It is interesting to note that the expected value of the minimum income random variable (from $k$ selected) involves the Gini index $G_k$ alone. On the other hand, it can be shown that the maximum income random variable (selected from $k$ random incomes) involves Gini indices up to and including the order $k - 1$.

In general, using the arguments in Lemma 2.2, we have

$$\frac{\overline{Y_k^{\max}}}{\overline{X}} = k \int_0^1 p^{k-1} s(p)\, \mathrm{d}p.$$

For the 2006 US Gini index, $G = 0.47$, $G_2 = 0.61$, we have

$$\frac{\overline{Y_3^{\max}}}{\overline{X}} = 1 - G_2 + 3G = 1.8.$$

So the maximum income variable (chosen out of three random households) has the expected value of 1.8 times the mean household income.

**Corollary 2.2.** *If $G$ is the Gini index and $G_2$ is the Gini index of second order, then*

$$\min\left\{\frac{1}{4}+\frac{3}{2}G-\frac{3}{4}G^2, 2G\right\} \geqslant G_2 \geqslant \max\left\{G, \frac{3}{2}G-\frac{3}{4}G^2\right\}.$$

Proof. Since the expected value for $Y_3^{\max}$ must be greater than or equal to the expected value for $Y_2^{\max}$, we must have

$$1 - G_2 + 3G \geqslant 1 + G$$

and thus $G_2 \leqslant 2G$. Also note that the expected value for $Y_3^{\min}$ must be less than or equal to the expected value for $Y_2^{\min}$ and thus

$$1 - G_2 \leqslant 1 - G$$

which yields $G_2 \geqslant G$. Now the result follows from Corollary 2.1. □

We can use the above inequalities to provide a nontrivial estimate for the second order Gini index $G_2$ based on the $G$ value alone. For example, we deduce that the world second order Gini index in the year 1910 satisfied $0.886 \geqslant G_2 \geqslant 0.636$ and in the year 1992 it satisfied $0.912 \geqslant G_2 \geqslant 0.662$.

For reader's convenience we have attached the plot (in Figure 1) of the possible $G_2$ region with the above upper and lower bounds. The above inequalities indicate that

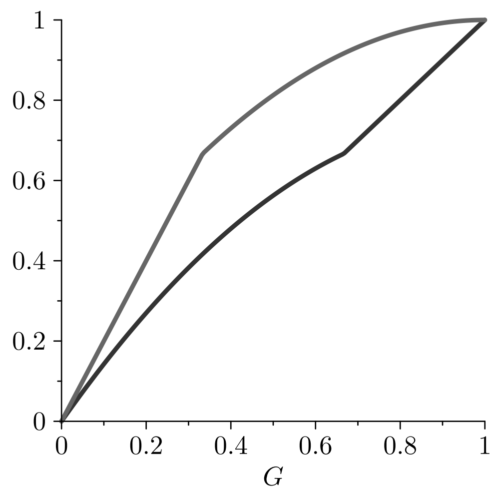


Figure 1

for a given $G$ value providing the $G_2$ value as an extra information on the income distribution carries greater information for the $G$ values near the center value of

0.5 as opposed to the values of $G$ closer to zero or one. The graph below helps to illustrate this observation.

## 3. New indices

We suggest a new sequence of indices $\{q_k\}$, $k \in \{1, 2, \ldots\}$, that measure the income distribution in a society using the higher moments of the share density function $s(p)$. The index $q_k$ is defined as

$$q_k = \int_0^1 p^{k-1} s(p)\,\mathrm{d}p.$$

Recall

$$q_k = \int_0^1 p^{k-1} s(p)\,\mathrm{d}p = \frac{\overline{Y_k^{\max}}}{k\overline{X}}.$$

The quantity $q_k$ involves the Gini indices $G_k$ of order less than or equal to $k-1$. It is also easy to see the sequence $\{q_k\}$ is a non-increasing sequence and $q_1 = 1$. Arguments similar to Corollary 2.1 can show that for each $k$ the range of $q_k$ is the interval $[1/k, 1]$ where the value of $1/k$ is attained for the constant share density function $s(p) = 1$ for all $p \in [0, 1]$ and the value of 1 is attained for the extreme case of all the income being concentrated in a single household. As a result we observe that the highest quantity that the value $\overline{Y_k^{\max}}$ can attain is the value $k\overline{X}$.

Therefore, we can think of the statistic $q_k$ as the ratio between the expected value for the random variable of the maximum income chosen out of $k$ random incomes divided by $\overline{X}$ and its highest attainable value $k$. This gives a summary information about the distribution of income in a society which simultaneously involves several Gini indices.

Let us mention a few examples. Suppose we take the expected value of the maximum income variable chosen out of two random household incomes divided by the mean income $\overline{X}$. The maximum value for this statistic is 2 and we observe

$$q_2 = \frac{1}{2}(1 + G).$$

The range for this statistic is $[1/2, 1]$, and we recall that $q_2 = \bar{p}$, the percentile level of the household which earns the average dollar.

Suppose now we take the expected value of the maximum income variable chosen out of three random household incomes divided by the mean income $\overline{X}$. The maximum value that this statistic can attain is the value 3, and we have

$$q_3 = G + \frac{1 - G_2}{3}.$$

The range for this statistic is $[1/3, 1]$. The lower the $G_2$ value (less extreme poverty), with the same $G$, the higher the $q_3$ value. As an example, for the US data, we recall $G = 0.47$, $G_2 = 0.61$, $q_2 = 0.7350$ and we obtain $q_3 = 0.6$. For a fixed $q_2$ value, the higher the $q_3$ value, the lower the extreme poverty in the US. Therefore, to assess the income distribution in a society, we have to use both $q_2$ and $q_3$ simultaneously just like we should use both $G$ and $G_2$ simultaneously.

Acknowledgement. The author would like to thank the referee for valuable suggestions.

*References*

[1] *A. B. Atkinson, T. Piketty, E. Saez*: Top incomes in the long run of history. Journal of Economic Literature *49* (2011), 3–71, http://www.nber.org/papers/w15408.

[2] *F. A. Farris*: The Gini index and measures of inequality. Am. Math. Mon. *117* (2010), 851–864.

[3] *C. Kleiber, S. Kotz*: A characterization of income distributions in terms of generalized Gini coefficients. Soc. Choice Welfare *19* (2002), 789–794.

*Author's address*: *Petr Zizler*, Mount Royal University, Calgary, Alberta, Canada, e-mail: pzizler@mtroyal.ca.