Income Distribution, Lorenz Curve, and Gini Index: A Complex Calculus Project for the Classroom

Erich Prisner[1]
Franklin College, Switzerland
November 2009

Abstract: Many Calculus books discuss Lorenz curves and Gini indices, but usually they assume that the Lorenz curve is given, and then computing the Gini index is a rather simple exercise in integration. In this paper we will see how the Gini index can be computed by a rather complicated multistep procedure, starting with the income distribution. The items that are touched include integration by parts and by substitution method, improper integrals, parametric curves and area under parametric curves, inverse functions, and solving equations, transformations of functions, and even global optimization. Therefore this example could serve as a rather complex project in Calculus towards the end of the semester, combining many of the topics covered.

Inequality of income or wealth is a very important and also controversial topic. Christians and socialists claim equality as the ideal. Liberals and economists insist that some extend of inequlity is unavoidable, and even necessary for a prospering society. Still all agree that too much inequality may be dangerous for the stability of society. We, as humble mathematicians, concentrate in this paper on the modest task to measure inequality.

Two of the tools that have been proposed are Lorenz curve and Gini index. See [K 2008] and [X ?] for surveys on the vast literature on these topics. In our paper we demonstrate how many classical features of an ordinary Calculus curriculum are touched when these questions are attacked, and we will propose the topic as an extended applied project for the Calculus classroom

1 Income Distributions

In real countries, statisticians count how many persons (or households) fall in different ranges of income, like below $10,000, between $10,000 and $20,000, and so on. Instead of using the frequencies, one might want to use relative frequencies—the ratio of number whose income lies in the corresponding interval and the whole popluation. These numbers are usually presented in a table, but can also be graphed. Such graphs are called bar graphs. The larger the intervals for income, the less precise the picture gets. For this reason, one may be tempted to increase the number of categories (also called classes) considered and decrease their width. However, for real data the graph one obtains will eventually look very coarse and less telling, having most frequencies equal to 0 except some, which are equal to $1/N$, where N is the number of individuals considered. Examples are given in Figures 1 and 2 for a data set of 100 numbers and 10 classes respectively 400 classes.

[1]Franklin College, Via Ponte Tresa 29, 6924 Sorengo-Lugano, Switzerland, eprisner@fc.edu

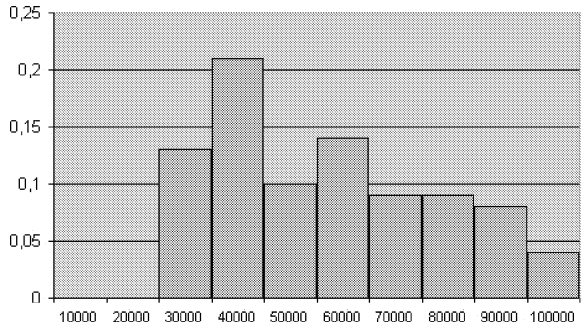

Figure 1: a bar graph with 10 classes

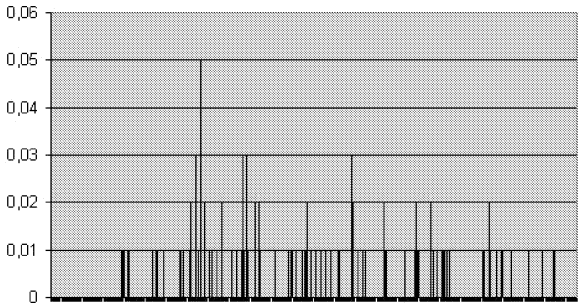

Figure 2: a bar graph with 400 classes

But if we assume very large populations of millions of persons, and don't overdo this process of making the intervals smaller, the histogram may look close to a smooth curve. These curves are considered here.

An **income distribution** $f(t)$, either defined on a closed interval $0 \leq t \leq b$ or defined for all $t \geq 0$, obeys $f(t) \geq 0$ for all t in the domain, and the additional property that the area under the curve equals 1, $\int_{t=0}^{b} f(t)dt = 1$ respectively $\int_{t=0}^{\infty} f(t)dt = 1$.

In the first part of the paper we will demonstrate the concepts with the rather simple income distribution $f(t) = \frac{1}{40} - \frac{1}{3200}t$, defined for $0 \leq t \leq 80$, whose graph is shown in Figure 3. As we will see later, starting with too complicated income distribution functions may make the computation of Lorenz curve, Gini index, or some other features difficult.

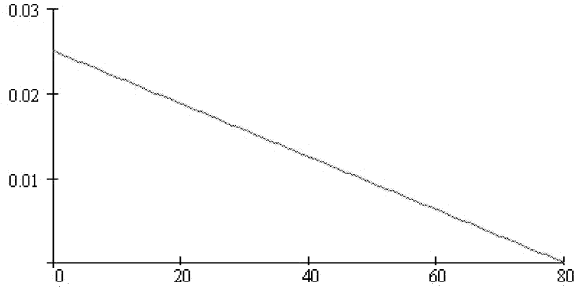

Figure 3: A simple income distribution

It is not straightforward to tell what $f(t)$ expresses. It does **not** express the number or percentage of persons with income t. Not $f(t)$ itself has meaning, but only integrals

of the form $\int_{t=a}^{c} f(t)$, which is the percentage of the population whose income lies between a and c. But since the rectangle with height $f(t)$ and width 1 is about the area under the curve between $t - 1/2$ and $t + 1/2$, the number $f(t)$ is about the percentage of the population with income between $t - 1/2$ and $t + 1/2$.

2 Density Function F and Relative Cumulative Income Function H_N

We will derive two functions from f. The first one is just a special antiderivative. The second one is expressed in terms of the first one and a second antiderivative of f. Therefore, in order to compute both these functions F and H_N, f should rather be integrable twice.

The antiderivative of f obeying $F(0) = 0$, i.e. $F(t) = \int_{s=0}^{t} f(s)ds$ is called *densitiy function* of f. Since f is positive, F is increasing. By the norming of the area under f, $F(b) = 1$. $F(x)$ expresses the percentage of the population whose income is x or less.

Let M be the number of persons considered in the income distribution. How many people make between $a - 1/2$ and $a + 1/2$? Well, since $\int_{t=a-1/2}^{a+1/2} f(t)dt$ expresses the percentage of these people, their *number* equals $M \int_{t=a-1/2}^{a+1/2} f(t)dt = M(F(a - 1/2) - F(a + 1/2))$, which is about $Mf(a)$. And how much money do these people make? Well, it is the product of the number of people, which is about $Mf(a)$, and the amount each gets, which is about a. Therefore this total amount is about $Mf(a)a$. When we ask about the total amount of money all people making less than a fixed value t obtain together, we would add all these values for integers $a < t$ to get an approximation, but for the precise number we would need to integrate. Therefore this amount is preciely

$$M \int_{s=0}^{t} sf(s)ds.$$

Let H be the antiderivative of the function $h(t) = tf(t)$ with $H(0) = 0$, that is

$$H(t) = \int_{s=0}^{t} sf(s)ds.$$

Therefore $M \cdot H(t)$ denotes the total amount of money that all persons with income at most t make.

However, more than in the absolute value of that money, we are interested in the relative value of it, in the ratio of the total money $M \cdot H(t)$ of all persons with income at most t divided by the total amount of money $M \cdot H(b)$ of all persons. This ratio is expressed by the *relative cumulative income function* $H_N(t) = \frac{H(t)}{H(b)}$, and gives the percentage of the total income that is made by those whose income is limited by the value t.

Using integration by parts we get

$$\int tf(t)dt = tF(t) - \int F(t)dt = tF(t) - \Phi(t) + C$$

, where Φ is an antiderivative of F, a second antiderivative of f. If we choose Φ with $\Phi(0) = 0$, then we get

$$H(t) = tF(t) - \Phi(t).$$

In our example $f(t) = \frac{1}{40} - \frac{1}{3200}t$, we get $F(t) = \frac{1}{40}t - \frac{1}{6400}t^2$ and $\Phi(t) = \frac{1}{80}t^2 - \frac{1}{19200}t^3$. We also get

$$H(t) = tF(t) - \Phi(t) = \frac{1}{40}t^2 - \frac{1}{6400}t^3 - \frac{1}{80}t^2 + \frac{1}{19200}t^3 = \frac{1}{80}t^2 - \frac{1}{9600}t^3,$$

and, since $H(80) = \frac{80}{3}$, the normed function H_N equals

$$H_N(t) = \frac{3}{6400}t^2 - \frac{1}{256000}t^3.$$

These two special functions F and H_N are shown in Figure 4.

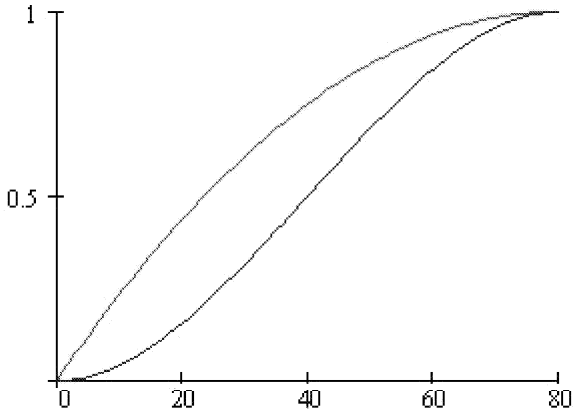

Figure 4: The functions F and H_N

Although this figure displays the functions associated with a concrete function f, some of the features shown in this graph are typical. Both functions start at a value of 0 at $t = 0$, are increasing, and reach a value of 1 at $x = b$. Moreover, $F(t)$ is always larger or equal to $H_N(t)$—the percentage of the persons with income less than a given boundary exceeds the precentage of the income of these persons of the whole income. After all, these are the poor persons.

3 Mean, Median, Mode

Mean, median, and mode are three important features of any distribution.

The *median* of the distribution f is that value t for which $\int_{s=0}^{t} f(t)dt = 1/2$, i.e. $F(t) = 1/2$. If the inverse F^{-1} of the function F can be computed, and we will later see that this should better be the case, then the median is the value $F^{-1}(1/2)$.

The **mean** or **average** is the quotient of total income $MH(b)$ and number M of persons considered, thus it equals value $H(b) = \int_{t=0}^{b} tf(t)dt$. If there is no upper boundary b for the distribution considered, then the mean equals $\lim_{t\to\infty} H(t)$.

The **mode** is the value t maximizing $f(t)$. Therefore computing the mode requires differentiation and methods to find the global maximum of a function.

In our example $f(t) = \frac{1}{40} - \frac{1}{3200}t$, the inverse of F can be found by solving the equation $x = F(t) = \frac{1}{40}t - \frac{1}{6400}t^2$ for t. This is done using the quadratic formula as $t = 80 - 80\sqrt{1-x}$. The other theoretically possible solution $t = 80 + 80\sqrt{1-x}$ is

not admissible since these t values are larger than the upper boundary $b = 80$. Thus $F^{-1}(x) = 80 + 80\sqrt{1-x}$, and therefore the median equals $F^{-1}(1/2) = 80 + 80\sqrt{1/2} \approx 23.4$. The mean equals $H(80) = \frac{80}{3} \approx 26.7$, and the mode is the value $t = 0$ in this example.

If we consider discrete descriptions of income using bar graphs, and compare mean, median, and mode to that obtianed from a income distribution fucntion that approximates the data, then the means are supposed to be close in both concepts. Actually the smaller the intervals defining the categories are, the closer the values will be. The same holds for the median, but not for the mode, which depends very much on the class width, and may get rather random for small class width.

4 Lorenz Curve

For every number $0 \leq x \leq 1$, let $L(x)$ denote the proportion of the total income that the poorest x of the population generates together. If both functions $F(t)$—which expresses the percentage of the total population that all persons with income at most t make— and $H_N(t)$—which is the percentage of the total income generated by those persons with income at most t— are known, then we can graph the curve L easily: All we need to do is compute the values $F(t)$ and $H_N(t)$ for various values of t, and plot all pairs $(F(t), H_N(t))$. The resulting **Lorenz curve** connects these points. It is a parametric curve, which means that both x- and y-coordinates of the curve are defined as expressions $x = \alpha(t)$ and $y = \beta(t)$ of a third variable t. Varying t one gets different values for x and y, and the curve consists of all pairs $(\alpha(t), \beta(t))$.

In our example, let's choose $t = 0, 10, \ldots, 80$. We get the following values for $x = F(t)$ and $y = H_N(t)$:

t	$x = F(t)$	$y = H_N(t)$
0	0	0
10	0.234	0.043
20	0.438	0.156
30	0.609	0.316
40	0.75	0.5
50	0.859	0.684
60	0.938	0.844
70	0.984	0.957
80	1	1

This function can be formualted as an ordinary function $y = L(x)$, with y equal to an expression $L(x)$ in x, provided the equation $x = F(t)$ can be solved for t, which means that the inverse function F^{-1} of F can be computed. Thus $L(x) = H_N(F^{-1}(x))$. Note that F^{-1} is defined for $0 \leq x \leq 1$ and is increasing.

The Lorenz curve has a few interesting properties:

- L is increasing, since $L'(x) = H_N'(F^{-1}(x)) \cdot F^{-1\prime}(x)$ and both H_N and F^{-1} are increasing.
- L is concave up.

Having seen already the graph of the Lorenz curve for the simple linear example above, let us also compute the expression defining it and the corresponding Gini index.

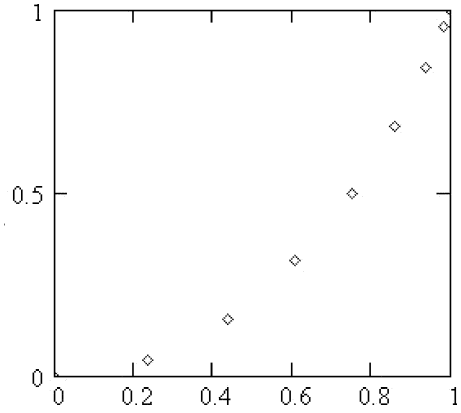

Figure 5: Graphing the six points

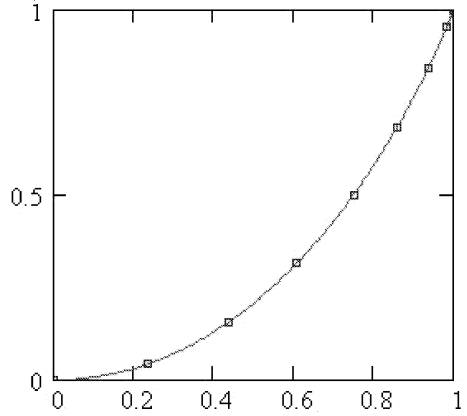

Figure 6: The Lorenz curve

As we have already seen in Section 3, we obtain $F^{-1}(x) = 80 - 80\sqrt{1-x} = 80(1 - \sqrt{1-x})$. Then

$$\begin{aligned} L(x) &= H_N(F^{-1}(x)) = \frac{3}{6400}(80(1-\sqrt{1-x}))^2 - \frac{1}{256000}(80(1-\sqrt{1-x}))^3 = \\ &= 3(1-\sqrt{1-x})^2 - 2(1-\sqrt{1-x})^3 = \\ &= 3 - 6\sqrt{1-x} + 3 - 3x - 2 + 6\sqrt{1-x} - 6 + 6x + 2\sqrt{1-x}^3 = \\ &= 3x - 2 + 2\sqrt{1-x}^3. \end{aligned}$$

5 Gini Index

The **Gini index** is defined as the ratio of area between Lorenz curve and diagonal and the area of the triangle under the diagonal. It is therefore twice the area between Lorenz curve and diagonal $y = x$.

$$G = 2\int_{x=0}^{1}(x - L(x))dx$$

It can attain every number between 0 and 1, where a value close to 0 means that the Lorenz curve is close to the diagonal and we have equality, whereas a value close to 1 means inequality.

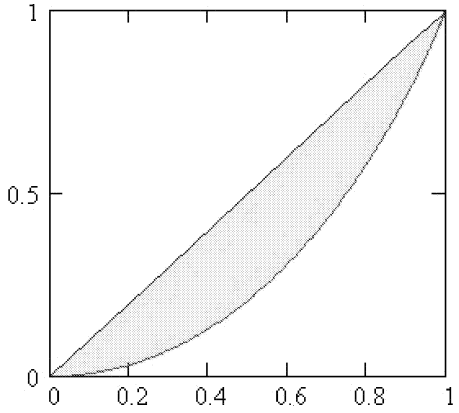

Figure 7: Twice the shaded area equals the Gini index

Note that different Lorenz curves may have the same Gini index.
For our linear income distribution example above, we get

$$\begin{aligned} G &= 2\int_{x=0}^{1}(x - L(x))dx = 2\int_{x=0}^{1}(2 - 2x - 2(1-x)^{3/2})dx = \\ &= 2(2x - x^2 - \frac{4}{5}(1-x)^{5/2}\Big|_0^1) = 2(2 - 1 - \frac{4}{5}) = \frac{2}{5} = 0.4. \end{aligned}$$

6 Gini Index without Lorenz Curve

Remember that if F^{-1} can not be found easily, then the expression defining the Lorenz curve can not be found. What about the Gini index in this case? Actually there is still a method to compute the Gini index, which uses the definition of the Lorenz curve as a parametric curve defined by the two functions F and H_N.

Calculus for parametric curves is not always covered in beginning Calculus courses, but if it is, here is a nice application. For the area under the Lorenz curve $\int_{x=0}^{1} L(x)dx$, using $x = F(t), \frac{dx}{dt} = F'(t) = f(t)$ and the heuristic to replace dx by $f(t)dt$ and $y = L(x)$ by $y = H_N(t)$, we get we get $\int_0^b H_N(t)f(t)dt$. Of course this is not intended as a serious derivation of the formula—such a derivation has to be looked up in the textbooks. Therefore

$$G = 1 - 2\int_0^b H_N(t)f(t)dt.$$

An example for such a distribution is $f(t) = \frac{1}{t+19} - \frac{1}{120}$, defined for $0 \leq t \leq 101$. Actually the area under the curve is not exactly equal to 1, we would have to multiply the function by a constant close to 1, but this is not relevant for our purposes. We get $F(t) = \ln(t+19) - \frac{1}{120}t - \ln(19)$, a function which can not be inverted easily, and whose inverse can not be expressed by an algebraic expression. Still the formula above works, although it needs some effort, which we will not invest here.

7 Another example: A Quadratic Function

Let us now consider the slightly more complicated function

$$f(t) = \frac{3}{100}(\frac{t-100}{100})^2$$

defined for $0 \leq t \leq 100$. Figure 8 displays the graph, together with the graph of the previous linear distribution.

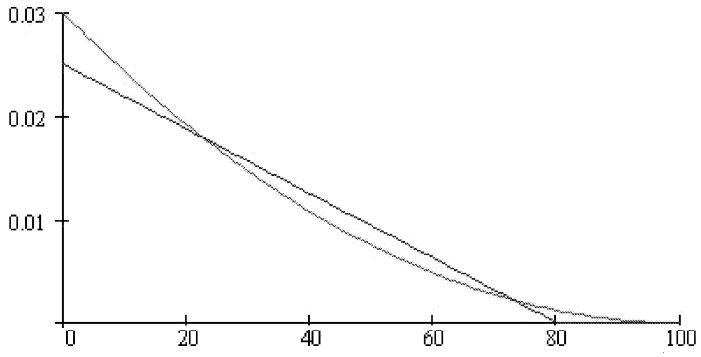

Figure 8: A simple income distribution

Some high incomes larger than 80 are added, also there are more incomes of value less than 25, but the number of incomes between 25 and 75 has decreased. Obviously inequality has increased from the previous distribution—this should also be reflected in Lorenz curve and Gini index.

To obtain the density function we integrate using substitution method and the substitution $u = (\frac{t-100}{100})$. Then we obtain

$$\int f(t)dt = \int \frac{3}{100}(\frac{t-100}{100})^2 dt = \int \frac{3}{100}u^2 100du = \ = u^3 + C = (\frac{t-100}{100})^3 + C.$$

Since $F(0) = 0$, the get the equation $0 = (\frac{-100}{100})^3 + C$, hence the integration constant C must be equal to 1, and we get

$$F(t) = (\frac{t-100}{100})^3 + 1.$$

The function F can be inverted easily. We get x = F(t)

$$F^{-1}(x) = 100\sqrt[3]{x-1} + 100$$

In order to derive the other function $H_N(t)$, we need the second derivative $\Phi(t)$. We use the same substitution and obtain $\Phi(t) = 25(\frac{t-100}{100})^4 + t + C$. Since $\Phi(0) = 0$, we obtain $C = -25$, and

$$\Phi(t) = 25((\frac{t-100}{100})^4 - 1) + t.$$

and

$$H(t) = tF(t) - \Phi(t) = t(\frac{t-100}{100})^3 + t - 25((\frac{t-100}{100})^4 - 1) - t = t(\frac{t-100}{100})^3 - 25((\frac{t-100}{100})^4 - 1).$$

Since $H(100) = 25$, we get

$$H_N(t) = \frac{t}{25}(\frac{t-100}{100})^3 - ((\frac{t-100}{100})^4 - 1)$$

Next let's compute mode, median, and mean. The mode is obviously again $t = 0$. The mean equals $H(100) = 25$. For the median, we have to solve the equation $F(t) = 1/2$, i.e. $(\frac{t-100}{100})^3 = -1/2$ or $t = 100 \cdot \sqrt[3]{-\frac{1}{2}} + 100 = 20.63$.

For the Lorenz curve function we get

$$\begin{aligned} L(x) = H_N(F^{-1}(x)) &= (4\sqrt[3]{x-1} + 4)(\sqrt[3]{x-1})^3 - ((\sqrt[3]{x-1})^4 - 1) = \\ &= (4\sqrt[3]{x-1} + 4)(x-1) - (\sqrt[3]{x-1})^4 + 1 = 3\sqrt[3]{x-1})^4 + 4x - 3 \end{aligned}$$

and a Gini index of

$$G = 2\int_{x=0}^{1}(x - L(x))dx = 2\int_{x=0}^{1}(3 - 3x - 1)^{4/3} - 3x)dx = 2(3 - \frac{3}{2} + \frac{9}{7}(-1))) = 0.429.$$

That the Gini index for the quadratic function is higher than for the linear example confirms our initial observation of the section. The Lorenz curve lies also totally below that for the linear example, as can be seen in Figure 12 where both Lorenz curves are shown.

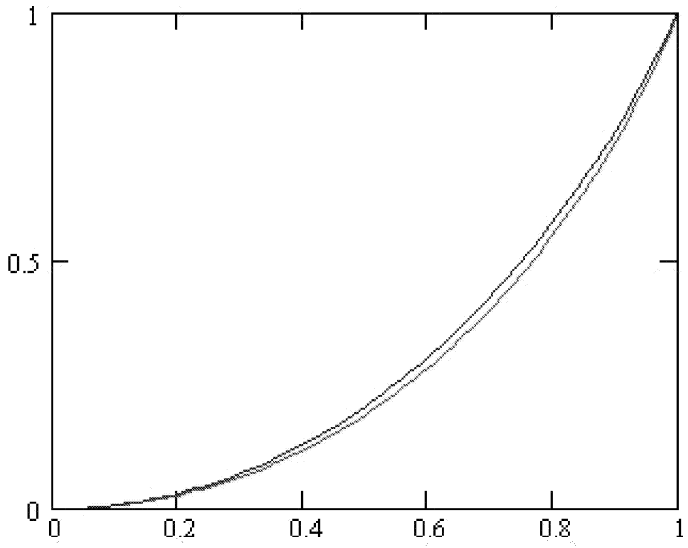

Figure 9: The Lorenz curves for the linear and the quadratic example

8 Another example: Decreasing Exponential Functions

The two examples discussed are mathematically simple, and for this reason well suited for treatment by students in a Calculus class. But how relevant are they? What kind of functions are used to model real-world distributions? Among the distributions that have been used to model real-world data are Pareto distributions, that are discussed

in the appendix, so-called log-normal distributions and gamma distributions, but also decreasing exponential functions

$$f(t) = ke^{-kt}.$$

These functions have been proposed in [DY 2000] as a reslut of a statistical model of money exchange. The density function is $F(t) = 1 - e^{-kt}$, and since $lim_{b\to\infty}F(b) = 1$, the function f is an income distribution with no maximum possible income. Thus in this example improper integrals come into play.

We get $\Phi(t) = t + e^{-kt}/k - 1/k$ and

$$H(t) = tF(t) - \Phi(t) = t - te^{-kt} - t - e^{-kt}/k + 1/k = 1/k - (t + 1/k)e^{-kt}.$$

Since the limit of this function as t goes to infinity is 1/k, the mean is $1/k$, and

$$H_N(t) = kH(t) = 1 - (kt + 1)e^{-kt}.$$

The median is the solution of the equation $e^{-kt} = 1/2$, i.e. the value $t = \ln 2/k$.

To get an explicit formula for the Lorenz curve finction, we need to solve the equation $x = F(t) = 1 - e^{-kt}$ for t, and get $t = -\ln(1 - x)/k$. Substituting t in the formula for H_N by this expression, we get the Lorenz function $L(x) = 1 - (1 - \ln(1 - x))e^{\ln(1-x)} = 1 - (1 - \ln(1 - x))(1 - x)$, or simplified

$$L(x) = x + \ln(1 - x) - x\ln(1 - x).$$

Note that the Lorenz curve is independent of the parameter k. Of course, $L(1)$ is not defined, but $\lim_{x\to 1} F(x) = 1$.

For the Gini index we get the improper integral

$$\begin{aligned} G &= 2\int_{x=0}^{1}(x - L(x))dx = \\ &= 2\int_{x=0}^{1}(x - 1)\ln(1 - x)dx = -2\int_{u=1}^{0} u\ln(u)du \end{aligned}$$

using the substitution $u = 1 - x$. Proceeding with integration by parts, we obtain

$$\begin{aligned} G &= -2(\frac{1}{2}u^2\ln(u))\Big|_1^0 - \int_{u=1}^{0}\frac{1}{2}u^2\frac{1}{u}du = \\ &= -u^2\ln(u)\Big|_1^0 + \int_{u=1}^{0} udu = (-u^2\ln(u) + \frac{1}{2}u^2)\Big|_1^0 = \frac{1}{2} \end{aligned}$$

9 Family Income

Since some years, statistics on individual income has been replaced by family income in most countries. Under the assumption that all persons are married, all men and women work, and that the most unrealistic assumption that the matching of men and women is independent from their income (love is stronger than money), the family income distribution can be expressed in terms of the income distribution $a(x)$ of men and $b(x)$ of females. A family income of s can be achieved by s for the man and $t - s$

for the woman, and the probability for that (rounded) is about $a(s)b(t-s)$. Therefore the distribution for the family income is the so-called 'convolution'

$$f(t) = \int_0^t a(s)b(t-s)ds.$$

Even if the income distributions are decreasing, the family income distribution is typically increasing for some time, until it decreases. If we take our initial straight line example for both men and women, $a(t) = b(t) = \frac{1}{40} - \frac{1}{3200}t$, we obtain a family density function of

$$f(t) = \begin{cases} 1/61440000 \cdot t^3 - 1/128000 \cdot t^2 + 1/1600 \cdot t & \text{for } 0 < t \leq 80 \\ -1/61440000 \cdot t^3 + 1/128000 \cdot t^2 - 1/800 \cdot t + 1/15 & \text{for } 80 < t \leq 160 \end{cases}$$

Fugure 10 shows the linear individual income distribution together with the two earners family income distribution.

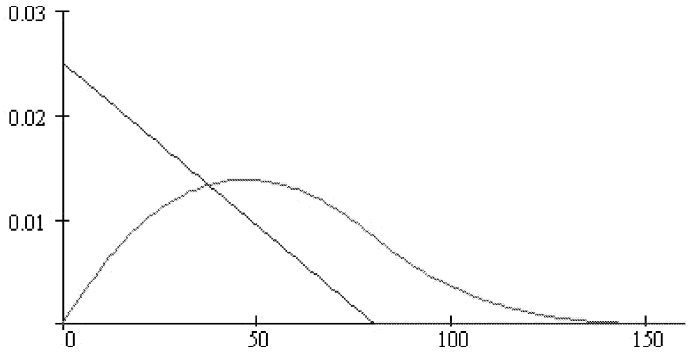

Figure 10: some family income distribution.

Both F and H_N are piecewise-defined functions, with different expressions for the parts $0 \leq t < 80$ and $80 \leq t \leq 160$. Already in this example we get difficulties inverting F, but using the second formula for the Gini index, we can derive a Gini index of 0.284, substantially lower than the Gini index of 0.4 for the linear individual income distribution functions.

In the decreasing exponential function example $a(t) = b(t) = ke^{-kt}$ we get

$$f(t) = \int_{s=0}^t ke^{-ks}ke^{-k(t-s)}ds = \int_{s=0}^t k^2e^{-kt}ds = k^2te^{-kt}.$$

In [DY 2001] it was shown that these functions fit the data of family income with two earners in the US in 1996 rather well. Again the resulting function F can not be inverted as a closed expression, but the second formula for the Gini index yields $G = 0.381$, again smaller than the 0.5 for the individual income distribution.

Obviously independent marriage increases equality. Some people attribute increasing inequality of family income in recent years to change in marriage pattern, towards less independent marriages, the rich marry the rich and the poor the poor.

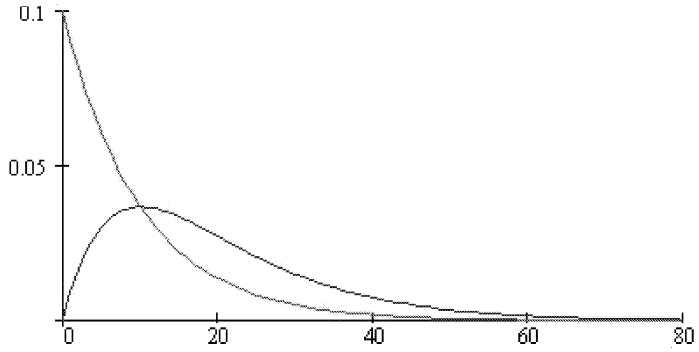

Figure 11: Another family income distribution, together with the individual distribution.

10 Transformations

What happens if over night everyone's income is increased by 10%? Does this make the distribution more equal?

The function obtained from a function f by stretching by a factor k is expressed as $f(\frac{t}{k})$. However, this new function is no longer a distribution—the area between curve and x-axis equals k, since the new upper bound is kb and $\int_0^{kb} f(\frac{t}{k})dt = k\int_0^b f(u)du = k$. Therefore the income distribution function after the rise is $f_1(t) = \frac{1}{k}f(\frac{t}{k})$.

The new density function is $F_1(t) = \int f_1(t)dt = \int \frac{1}{k}f(\frac{t}{k})dt = \int f(u)du = F(u) = F(\frac{t}{k})$. It is obtained from the old one, F, by stretching horizontally with a factor of k. The new function $H_{N,1}$ is also obtained from H_N by a horizontal stretch with factor k, since $H_1(t) = \int tf_1(t)dt = \int \frac{t}{k}f_1(\frac{t}{k})dt = \int kuf(u)du = kH(u) = kH(\frac{t}{k})$.

From these two remarks there follows that the Lorenz curve, and therefore also the Gini index, are the same as before.

References

[DY 2000] A. A. Dragulescu, V. M. Yakovenko, Statistical mechanics of money, *The European Physical Journal* B 17 (2000) 723-729.

[DY 2001] A. Dragulescu, V.M. Yakovenko, Evidence for the exponential distribution of income in the USA, *The European Physical Journal* B 20 (2001) 585-589.

[K 2008] Christian Kleiber, The Lorenz curve in economics and econometrics, in: Gianni Betti und Achille Lemmi (eds.), *Advances on Income Inequality and Concentration Measures*, Collected Papers in Memory of Corrado Gini and Max O. Lorenz, Routledge, London, 2008, 225-242.

[X ?] Kuan Xu, How has the literature on Gini's index evolved in the past 80 years? in *Contemporary Poverty and Welfare: Alleviation Issues*, Frank Columbus (ed.), Nova Science Publishers.

11 Appendix: A few more distributions

11.1 Triangular Distributions

Look at the following two distributions—the quadratic one discussed in a previous section, and another linear one. Which one is more unequal?

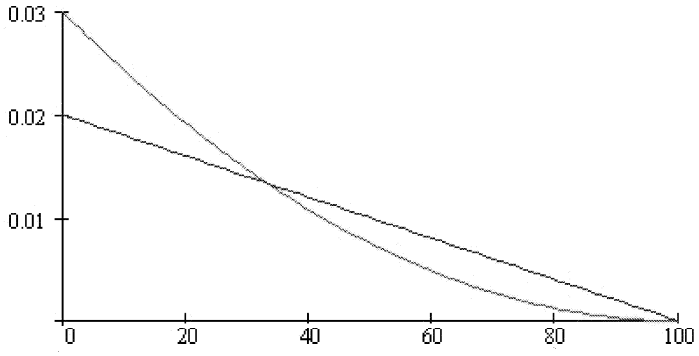

Figure 12: The Lorenz curves for the linear and the quadratic example

We will discuss income distributions whose graph is a straight line between a point on the y-axis and the point $(b, 0)$. According to the area property, these distributions are uniquely determined by the one parameter b. The equation is

$$f(t) = \frac{2}{b} - \frac{2}{b^2}t.$$

Then

$$\begin{aligned} F(t) &= \frac{2}{b}t - \frac{1}{b^2}t^2, \\ \Phi(t) &= \frac{1}{b}t^2 - \frac{1}{3b^2}t^3, \\ H(t) &= tF(t) - \Phi(t) = \frac{2}{b}t^2 - \frac{1}{b^2}t^3 - \frac{1}{b}t^2 + \frac{1}{3b^2}t^3 = \frac{1}{b}t^2 - \frac{2}{3b^2}t^3, \end{aligned}$$

and, since $H(b) = b - \frac{2}{3}b = \frac{b}{3}$, the normed function H_N equals

$$H_N(t) = \frac{3}{b^2}t^2 - \frac{2}{b^3}t^3.$$

We can compute the Lorenz curve explicitely, since the function F can be inverted easily. When solving $x = F(t)$ for t, we get $t^2 - 2bt + b^2x = 0$, or

$$L(x) = 3(1 - \sqrt{1-x})^2 - 2(1 - \sqrt{1-x})^3,$$

a curve which is independent of the particular parameter b. For the Gini index we get $G = 0.4$.

11.2 Trapezoid Distributions

What about distributions whose graph is a straight line between $(0, a_1)$ and (b, a_2)? Since the area under the curve equals 1 but also $b\frac{a_1+a_2}{2}$, we get $b = \frac{2}{a_1+a_2}$ and

$$f(t) = a_1 + \frac{a_2^2 - a_1^2}{2}t.$$

Then

$$\begin{aligned} F(t) &= a_1 t + \frac{a_2^2 - a_1^2}{4}t^2, \\ \Phi(t) &= \frac{a_1}{2}t^2 + \frac{a_2^2 - a_1^2}{12}t^3, \\ H(t) &= tF(t) - \Phi(t) = a_1 t^2 + \frac{a_2^2 - a_1^2}{4}t^3 - \frac{a_1}{2}t^2 - \frac{a_2^2 - a_1^2}{12}t^3 = \\ &= \frac{a_1}{2}t^2 + \frac{a_2^2 - a_1^2}{6}t^3, \end{aligned}$$

and since

$$H(b) = \frac{2a_1}{(a_1+a_2)^2} + \frac{4(a_2^2 - a_1^2)}{3(a_1+a_2)^3} = \frac{2a_1}{(a_1+a_2)^2} + \frac{4(a_2 - a_1)}{3(a_1+a_2)^2} = \frac{2a_1 + 4a_2}{3(a_1+a_2)^2},$$

the normed function H_N equals

$$H_N(t) = \frac{3a_1(a_1+a_2)^2}{4(a_1+2a_2)}t^2 + \frac{(a_2-a_1)(a_1+a_2)^3}{4(a_1+2a_2)}t^3.$$

The inverse function of F is

$$F^{-1}(x) = \frac{-4a_1 + \sqrt{16a_1^2 + 16(a_2^2 - a_1^2)x}}{2(a_2^2 - a_1^2)} = \frac{-2a_1 + 2\sqrt{a_1^2 + (a_2^2 - a_1^2)x}}{a_2^2 - a_1^2}$$

$$L(x) = ...$$

11.3 Pareto Distributions

Pareto introduced the following distribution in 1897. It has two parameters α and K, and has no upper bound b for the domain. It is piecewise-defined as follows:

$$f(t) = \begin{cases} 0 & \text{for } 0 \le t < K \\ \frac{\alpha K^\alpha}{t^{\alpha+1}} & \text{for } K \le t \end{cases}$$

Figure 13 shows the curve for $K = 20$ and $\alpha = 3/2$.
Integrating these functions is rather easy:

$$F(t) = \begin{cases} 0 & \text{for } 0 \le t < K \\ 1 - \left(\frac{K}{t}\right)^\alpha & \text{for } K \le t \end{cases}$$

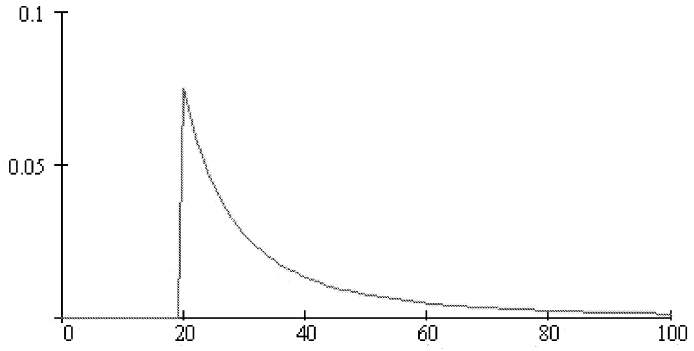

Figure 13: A Pareto distribution, $K = 20, \alpha = 3/2$.

F is not a 1-1 function, but it is 1-1 for $K \leq t$. For this restricted domain, the inverse function is

$$F^{-1}(x) = \frac{K}{(1-x)^{1/\alpha}}$$

$$H(t) = \begin{cases} 0 & \text{for } 0 \leq t < K \\ \frac{\alpha}{\alpha-1}\left(\frac{K^\alpha}{t^{\alpha-1}} - K\right) & \text{for } K \leq t \end{cases},$$

provided $\alpha \neq 1$.