The econometrics of inequality and poverty

Chapter 4: Lorenz curves, the Gini coefficient and parametric distributions

Michel Lubrano

October 2017

Contents

1 Introduction

Some authors like Sen (1976) prefer to use a discrete representation of income, which is based on the assumption that the population is finite. Atkinson (1970) and his followers prefer to suppose that income is a continuous variable. It implies that the population is implicitly infinite, but the sample can be finite. Discrete variables and finite population are at first easy notions to understand while continuous variables and infinite population are more difficult to accept. But as far as computations and derivations are concerned, continuous variables lead to integral calculus which is an easy topic once we know some elementary theorems. Considering a continuous random variable opens the way for considering special parametric densities such as the Pareto or the lognormal which have played an important role in studying income distribution. Discrete mathematics are quite complicated.

In this chapter, we have marked with an asterisk the sections that can be skipped at a first reading, because they inolved a more specialized material.

2 General notions

We are interesting in the income distribution. Income is supposed to be a continuous random variable X with cumulative distribution $F(.)$.

2.1 Distributions

Definition 1 *The distribution function $F(x)$ gives the proportion of individuals of the population having a standard of living below or equal to x.*

F is a non-decreasing function of its argument x. We suppose that $F(0) = 0$ while $F(\infty) = 1$. $F(x)$ gives the percentage of individual with an income below x. We usually call p that proportion.

A natural estimator is obtained for $F(.)$ by considering:

$$\hat{F}(x) = \frac{1}{n}\sum_{1=1}^{n} \mathbb{I}(x_i \leq x).$$

where $\mathbb{I}(.)$ is the indicator function. This estimator is easy to implement. The resulting graph of the density might seem discontinuous for very small sample sizes, but get rapidly smoother as soon as $n > 30$. So there is in general no need for non-parametric smoothing.

Let us now order the observations by increasing order from the smallest to the largest and call $x_{[j]}$ the observation which has rank j. We can write the natural estimator of F as

$$\hat{F}(x_{[j]}) = j/n.$$

It is common to call $x_{[j]}$ an order statistics. They will play an important role for estimation. It will be used for qunatiles, for Lorenz curves and so on.

We can give a short numerical example written in R to illustrate the performance of the natural estimator of a distribution. We draw two samples from a normal distribution of size $n = 100$ and then of size $n = 1000$.

```
n = 100
xr = rnorm(n)
x = sort(xr)
y = seq(0,1,length=n)
plot(x,y,type="l",xlim=c(-3,3),ylab="Cumulative",xlab="X")

n = 1000
xr = rnorm(n)
x = sort(xr)
y = seq(0,1,length=n)
lines(x,y,col=2)
```

```
lines(x,pnorm(x),col=3)
text(-2.8,0.8,"n=1000",col=2,pos=4)
text(-2.8,0.75,"n=100",col=1,pos=4)
text(-2.8,0.70,"True",col=3,pos=4)
```

The abscises of the cumulative are obtained by ordering the draws, while the ordinates are simply an ordered index between 0 and 1. The curve in black corresponds to the sample of size 100. It

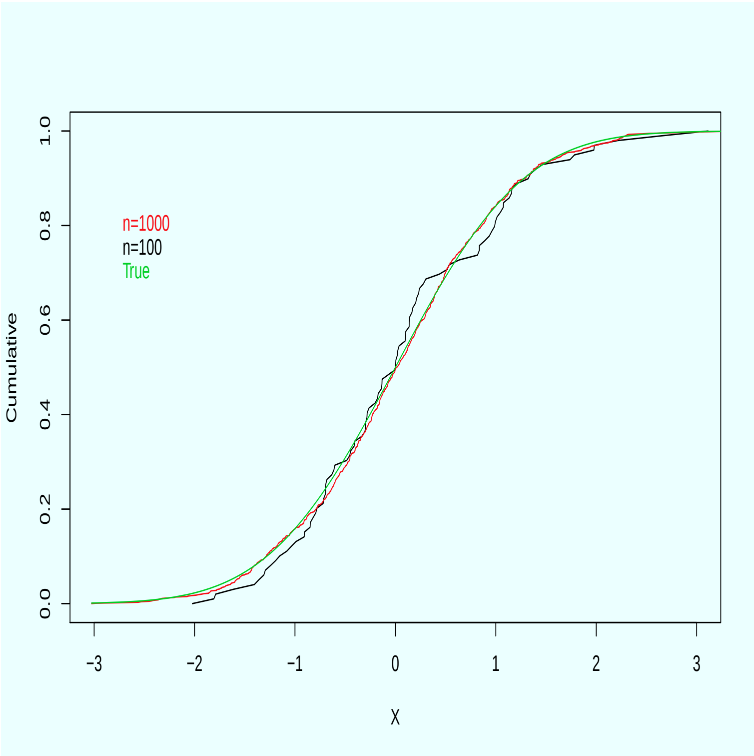

Figure 1: Natural estimator of the cumulative distribution for a Gaussian random variable

is rather rough. But the curve in red, corresponding to a sample of 1000 is perfectly smooth. It is roughly the same as the true cumulative in green.

2.2 Densities

We shall suppose that F is continuously differentiable so that there exist a density defined by

$$f(x) = F'(x).$$

So, for a given x, the value of p such that $X < x$ can be defined alternatively as

$$p = \int_0^x f(t)\,dt = F(x).$$

Densities are much more complicated to estimate. There exist no natural estimator as for distributions, simply because $\hat{F}(.)$ is not differentiable. Some kind of non-parametric smoothing is needed. Non-parametric density estimation will be detailed in a next Chapter.

If $f(x)$ is the density, then the probability that the random variable X belongs to the interval $[x_{k-1}, x_k]$ is given by

$$p(x_{k-1} < x < x_k) \simeq f(x_k)\Delta x_k,$$

where $\Delta x_k = x_k - x_{k-1}$. If the interval $[a, b]$ is sliced in m smaller slices, then:

$$p(a < x < b) \simeq \sum_{k=2}^{m} f(x_k)\Delta x_k.$$

If we take the limit for $m \to \infty$, we have

$$p(a < x < b) \simeq \lim_{m\to\infty} \sum_{k=2}^{m} f(x_k)\Delta x_k = \int_a^b f(x)dx.$$

Of course this limit exists only if $F(.)$ is sufficiently smooth, i.e. it has no jumps or kinks.

2.3 Quantiles

Once a distribution is given, it is always possible to compute its quantiles (this is not the case for moments that exists only under specific conditions). Deciles are a convenient way of slicing a distribution in intervals of equal probability, each interval being of probability 1/10. More generally, a quantile is a function $x = q(p)$ that gives the value of x such that $F(x) = p$. Quantiles are implicitly defined by the relation:

$$x = q(p) = F^{-1}(p).$$

$q(p)$ is thus the living standard level below which we find a proportion p of the population. The median of a population is the value of x such that half of the population is below x and half of the population is above $x = q(0.50)$. Using quantiles is also a way to normalize the characteristics of a population between 0 and 1. This facilitates comparisons between two populations, ignoring thus scale problems.

Quantiles are rather easy to estimate once we know the order statistics. Suppose that we have an ordered sample of size n. The estimator of a quantile comes directly from the natural estimator of the distribution. The p quantile is simply the observation that has rank $[p \times n]$. Quantiles are directly estimated in R using the instruction

```
quantile(x,p),
```

where x is a vector containing the sample and p the level of the quantile.

Piketty (2000) in his book on the history of high incomes in France makes an extensive use of quantiles to study the French income distribution and particularly its right tail concerning high incomes. High incomes concern the last decile, which means $q_{0.90}$. That decile however cover a variety of situations where wages, mixed incomes and capital incomes have a varying importance. The interval $q_{0.90} - q_{0.95}$ concerns what he calls the middle class, formed mainly by salaried executives, which thus in fact corresponds more to the higher middle class. The interval $q_{0.95} - q_{0.99}$ is the upper middle class, formed mainly by holder of intermediate incomes like layers, doctors. The really rich persons correspond to the last centile $q_{0.99}$ and over. It corresponds to holders of capital income.

Defining social classes using quantiles is a very hazardous project. The poverty line can be defined as 50% of the median. However, this does not correspond to a precise quantile. In their paper about inequality in China, Piketty et al. (2017) prefer to speak about the 50% bottom which they implicitly consider as being the poor class, the top 10% which correspond to the rich class while the remaining 40% represents the middle incomes.

2.4 Some useful math results

Three main rules are important to understand the next coming calculations:

1. *Integration by parts*. It comes from the rule giving the derivative of a product of two functions of x, $u(x)$ and $v(x)$:
$$(uv)' = u'v + uv'.$$
Let us take the integral of this expression.
$$uv = \int u'v\,du + \int uv'\,dv.$$
We deduce the integration by part formula by simply rearranging the terms:
$$\int u'v\,du = uv - \int uv'\,dv.$$

2. *Compound derivatives*.
$$\frac{\partial f(u(x))}{\partial x} = f'(u(x))u'(x).$$

3. *Change of variable and densities*. Let $x \sim f(x)$ and a transformation $y = h(x)$ with inverse $x = h^{-1}(y)g(y)$. Then the density of y is given by:
$$\phi(y) = |J(x \to y)| f(h^{-1}(y)),$$
where J is the absolute value of the Jacobian of the transformation
$$J(x \to y) = |\partial x_i / \partial y_i|.$$

4. *Change of variable and integrals*. Consider the integral
$$\int_a^b f(x)\,dx$$
and the change of variable $x = h(u)$ with reciprocal $u = h^{-1}(x)$. Then the original integral can be expressed as
$$\int_a^b f(x)\,dx = \int_{h^{-1}(a)}^{h^{-1}(b)} f[h(u)]\,h'(u)\,du.$$

2.5 Means and truncated means

We are now going to investigate the properties of partial means. As a by-product, we shall obtain useful formulae to define the Lorenz curve in the next subsection. We thus start from an income distribution with continuous density $f(x)$. The average standard of living in the total population is given by the total mean:

$$\mu = \int_0^\infty x dF(x) = \int_0^\infty x f(x) dx.$$

We now consider a threshold z and the population which is below that threshold, sometimes the population over that threshold. We can compute the average standard of living of the first group, the one which is below z. This is especially interesting for computing certain poverty indices. This is equivalent to the expectation of a truncated distribution, defined as:

$$\mu_1 = \frac{\int_0^z x f(x) dx}{F(z)}.$$

For $z \to \infty$, we recover the mean income of the population as $F(\infty) = 1$. Using integration by parts with $u = x$ and $v' = f(x)$, we can rewrite the integral in the numerator as:

$$\begin{aligned} \int_0^z x f(x) dx &= [xF(x)]_0^z - \int_0^z F(x) dx \\ &= zF(z) - \int_0^z F(x) dx. \end{aligned}$$

Noting that $z = \int_0^z dx$, it comes that:

$$\mu_1 = \frac{\int_0^z x f(x) dx}{F(z)} = \int_0^z \left[1 - \frac{F(x)}{F(z)} \right] dx.$$

Incidently, if we now let z tend to infinity, we arrive at an alternative expression for the the mean:

$$\mu = \int_0^\infty [1 - F(x)] dx.$$

Note also that another expression of the mean can be obtained as follows, using the quantiles. Let us start from:

$$\mu = \int_0^\infty x f(x) dx.$$

By the change of variable $x = F^{-1}(p)$ and $p = F(x)$, we have $dp = f(x)dx$ and thus:

$$\mu = \int_0^\infty x f(x) dx = \int_0^1 F^{-1}(p) dp = \int_0^1 q(p) dp.$$

This expression will be used for explaining the Lorenz curve.

3 Lorenz curves

The Lorenz curve is a graphical representation of the cumulative income distribution. It shows for the bottom $p_1\%$ of households, what percentage $p_2\%$ of the total income they have. The percentage of households is plotted on the $x-$axis, the percentage of income on the $y-$axis. It was developed by Max O. Lorenz in 1905 for representing inequality in the wealth distribution. As a matter of fact, if $p_1 = p_2$, the Lorenz curve is a straight line which says for instance that 50% of the households have 50% of the total income. Thus the straight line represents perfect equality. And any departure from this 45° line represents inequality.

3.1 A partial moment function

The standard definition of the Lorenz curve is in term of two equations. First, one has to determine a particular quantile, which means solving for z the equation:

$$p = F(z) = \int_0^z f(t)dt,$$

and then write:

$$L(p) = \frac{1}{\mu}\int_0^z t\, f(t)\, dt.$$

So the Lorenz curve is an unscaled partial moment function. Unscaled, because it is not divided by $F(z)$.

A notation popularized by Gastwirth (1971) used the fact that $z = F^{-1}(p)$ to write the Lorenz curve in a direct way, using a change of variable:

$$L(p) = \frac{1}{\mu}\int_0^p q(t)\, dt = \frac{1}{\mu}\int_0^p F^{-1}(t)\, dt.$$

Alternatively, using the relation $\mu = \int_0^1 q(t)\, dt$, we can have another writing:

$$L(p) = \frac{\int_0^p q(t)\, dt}{\int_0^1 q(t)\, dt}.$$

The numerator sums the incomes of the bottom p proportion of the population. The denominator sums the incomes of all the population. $L(p)$ thus indicates the cumulative percentage of total income held by a cumulative proportion p of the population, when individuals are ordered in increasing income values.

3.2 Properties

The Lorenz curve has several interesting mathematical properties.

1. It is entirely contained into a square, because p is defined over [0,1] and $L(p)$ is at value also in [0,1]. Both the $x-$axis and the $y-$axis are percentages.

2. The Lorenz curve is not defined is μ is either 0 or ∞.

3. If the underlying variable is positive and has a density, the Lorenz curve is a continuous function. It is always below the 45° line or equal to it.

4. $L(p)$ is an increasing convex function of p. Its first derivative:

$$\frac{dL(p)}{dp} = \frac{q(p)}{\mu} = \frac{x}{\mu} \text{ with } x = F^{-1}(p)$$

is always positive as incomes are positive. And so is its second order derivative (convexity). The Lorenz curve is convex in p, since as p increases, the new incomes that are being added up are greater than those that have already been counted. (Mathematically, a curve is convex when its second derivative is positive).

5. The Lorenz curve is invariant with positive scaling. X and cX have the same Lorenz curve.

6. The mean income in the population is found at that percentile at which the slope of $L(p)$ equals 1, that is, where $q(p) = \mu$ and thus at percentile $F(\mu)$ (as shown on Figure 2). This can be shown easily because the first derivative of the Lorenz curve is equal to x/μ.

7. The median as a percentage of the mean is given by the slope of the Lorenz curve at $p = 0.5$. Since many distributions of incomes are skewed to the right, the mean often exceeds the median and $q(p = 0.5)/\mu$ will typically be less than one.

The convexity of the Lorenz curve is revealing of the density of incomes at various percentiles. The larger the density of income $f(q(p))$ at a quantile $q(p)$, the less convex the Lorenz curve at $L(p)$. On Figure 2, the density is thus visibly larger for lower values of p since this is where the slope of the $L(p)$ changes less rapidly as p increases.

By observing the slope of the Lorenz curve at a particular value of p, we know the $p-$quantile relative to the mean, or, in other words, the income of an individual at rank p as a proportion of the mean income. An example of this can be seen on Figure 2 for $p = 0.5$. The slope of $L(p)$ at that point is $q(0.5)/\mu$, the ratio of the median to the mean. The slope of $L(p)$ thus portrays the whole distribution of mean-normalized incomes.

3.3 A mathematical characterization*

Lorenz curves were defined by reference to a given distribution function $F(.)$. Is it possible to characterize a Lorenz curve directly, without making reference to a particular distribution? Let us consider directly the expression of function that we consider to be a potential Lorenz curve. In this case, this curve has to verify some properties in order to be a true Lorenz curve. From Sarabia (2008), we have a first theorem:

Theorem 1 *Suppose $L(p)$ is defined and continuous on [0,1] with second derivative $L''(p)$. The function $L(p)$ is a Lorenz curve if and only if $L(0) = 0$, $L(1) = 1$, $L'(0+) \geq 0$, $L''(p) \geq 0$ in (0,1).*

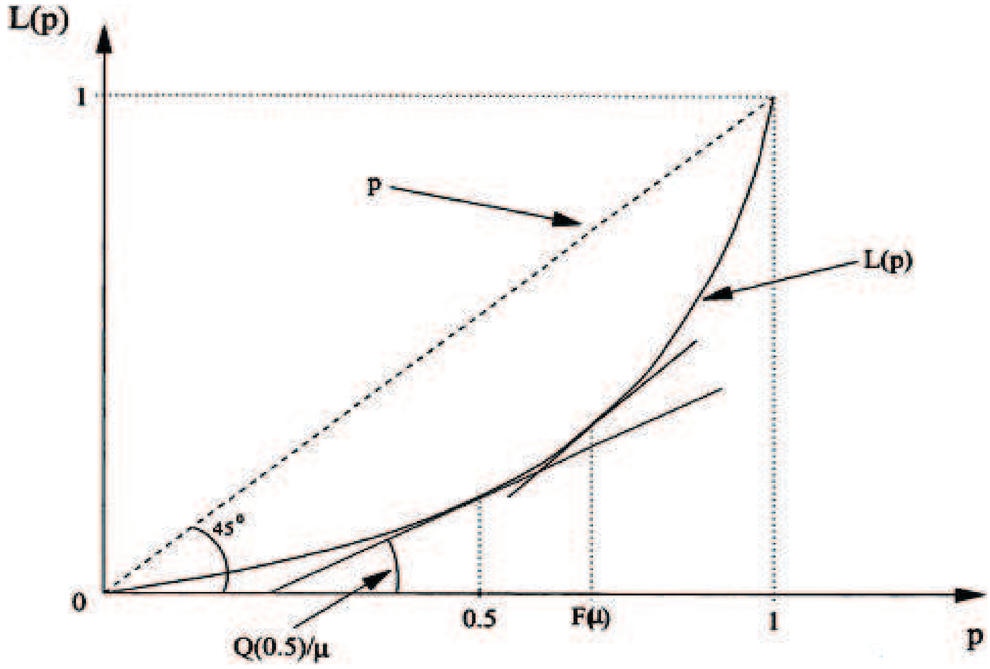

Figure 2: Lorenz curve (source Duclos and Araar 2006)

If a curve is a Lorenz curve, it determines the distribution of X up to a scale factor which is the mean μ. How could we find it? Let us take the definition of the Lorenz curve:

$$L_X(p) = \frac{1}{\mu_X} \int_0^p F_X^{-1}(t)\, dt$$

and express it as:

$$\mu L(F(x)) = \int_0^x y dF(y)).$$

Let us differentiate it using the derivative of a compound function:

$$\mu L'(F(x)) f(x) = x f(x).$$

We simplify by $f(x)$ and take the derivative it a second time so that

$$\mu L''(F(x)) f(x) = 1.$$

We get the following theorem from Sarabia (2008):

Theorem 2 *If $L''(p)$ exists and is positive everywhere in an interval (x_1, x_2), then F_X has a finite positive density in the interval $(\mu L'(x_1^+), \mu L'(x_2^-))$ which is given by*

$$f_X(x) = \frac{1}{\mu L''(F_X(x))}.$$

4 The Gini coefficient revisited

The Gini coefficient can be written in many different forms. In this section, we shall see how to pass from the standard definition of the Gini as a surface to its various expressions (covariance, mean of absolute difference). We shall use the surveys of Yitzhaki (1998) and of Xu (2003), using however a simplification. We shall suppose that the mean of F exists. As a consequence:

$$\lim_{t\to 0} tF(t) = \lim_{t\to\infty} t(1 - F(t)) = 0,$$

which means that both limits exists, which simplifies greatly the computation of some integrals when considering an infinite bound.

4.1 Gini coefficient as a surface

If everybody had the same income, the cumulative percentage of total income held by any bottom proportion p of the population would also be p. The Lorenz curve would then be $L(p) = p$: population shares and shares of total income would be identical. A useful informational content of a Lorenz curve is thus its distance, $p - L(p)$, from the line of perfect equality in income. Compared to perfect equality, inequality removes a proportion $p - L(p)$ of total income from the bottom $100 \cdot p\%$ of the population. The larger that “deficit”, the larger the inequality of income. There is thus an interest in computing the average distance between these two curves or the surface between the diagonal p and the Lorenz curve $L(p)$. We know that the Lorenz curve is contained in the unit square having a normalized surface of 1. The surface of the lower triangle is 1/2. If we want to obtain a coefficient at values between 0 and 1, we must take twice the integral of $p - L(p)$, i.e.:

$$G = 2\int_0^1 (p - L(p))\, dp = 1 - 2\int_0^1 L(p)dp,$$

which is nothing but the usual Gini coefficient. Xu (2003) gives a good account of the algebra of the Gini index. We have given above an interpretation of the Gini index as a surface. The initial definition we gave was in term of a mean of absolute differences in the previous chapter. There are other formula too. All of these formula are equivalent. We have to prove this. A large survey of the literature can also be found in the article *Gini coefficient* of Wikipedia.

4.2 Gini as a covariance

Let us us start from the above definition of the Gini coefficient and use integration by parts with $u' = 1$ and $v = L(p)$. Then

$$\begin{aligned} G &= 1 - 2\int_0^1 L(p)dp \\ &= 1 - 2\left[pL(p)\right]_0^1 + 2\int_0^1 pL'(p)\, dp \\ &= -1 + 2\int_0^1 pL'(p)\, dp. \end{aligned}$$

We are then going to apply a change of variable $p = F(y)$ and use the fact proved above that $L'(p) = y/\mu$. We have

$$G = \frac{2}{\mu}\int_0^\infty yF(y)f(y)dy - 1 = \frac{2}{\mu}\left[\int_0^\infty yF(y)f(y)dy - \frac{\mu}{2}\right].$$

This formula opens the way to an interpretation of the Gini coefficient in term of covariance as

$$\text{Cov}(y, F(y)) = \text{E}(yF(y)) - \text{E}(y)\text{E}(F(y)).$$

Using this definition, we have immediately that

$$G = \frac{2}{\mu}\text{Cov}(y, F(y)),$$

which means that *the Gini coefficient is proportional to the covariance between a variable and its rank*. The covariance interpretation of the Gini coefficient open the way to numerical evaluation using a regression.

Meanwhile, noting that $\text{Cov}(y, F(y)) = \int y(F(y) - 1/2)dF(y)$, using integration by parts, we get

$$\text{Cov}(y, F(y)) = \frac{1}{2}\int F(x)[1 - F(x)]dx,$$

so that we arrive at the integral form

$$G = \frac{1}{\mu}\int F(x)[1 - F(x)]dx.$$

We can remark that $F(x)(1 - F(x))$ is largest at $F(x) = 0.5$, which explains why the Gini index is often said to be most sensitive to changes in incomes occurring around the median income.

The above integral form can also be written as

$$G = 1 - \frac{1}{\mu}\int [1 - F(x)]^2 dx.$$

We shall prove this equivalence by considering the last interpretation of the Gini which is the scaled mean of absolute differences.

4.3 Gini as mean of absolute differences*

The initial definition of the Gini coefficient is the mean of the absolute differences divided by twice the mean. If y and x are two random variables of the same distribution F, this definition implies

$$I_G = \frac{1}{2\mu}\int_0^\infty \int_0^\infty |x - y|dF(x)dF(y).$$

As $F(x)$ and $1 - F(x)$ are simply the proportions of individuals with incomes below and above x, integrating the product of these proportions across all possible values of x gives again the Gini coefficient, in its form $\frac{1}{\mu}\int F(x)[1-F(x)]dx$. If we decide to proceed step by step, we first note that $|x-y| = (x+y) - 2\min(x,y)$, so that the expectation of this absolute difference is

$$\Delta = \text{E}|x-y| = 2\mu - 2\text{E}(\min(x,y)).$$

To compute the last expectation, we need the distribution of the Min of two random variables having the same distribution. We know or we can show that it is equal to $1-(1-F(y))^2$, while its derivative is $-d(1-F(y)$. So that

$$\Delta = 2\mu + 2\int_0^\infty y\, d(1-F(y))^2.$$

The last integral can be transformed using integration by parts with $u = y$ and $v = (1-F(y))^2$:

$$\int_0^\infty y\, d(1-F(y))^2 = \left[y(1-F(y))^2\right]_0^\infty - \int [1-F(y)]^2 dy.$$

So that we get the integral form of the Gini

$$I_G = \frac{\Delta}{2\mu} = 1 - \frac{1}{\mu}\int [1-F(x)]^2 dx,$$

because the first right hand term is zero.

4.4 S-Gini*

We underlined that the Gini coefficient was very sensitive to changes in the middle of the income distribution. A generalization of the Gini coefficient, obtained by adding a aversion for inequality parameter as in the Atkinson index, was proposed in the literature by Donaldson and Weymark (1980) and other papers following this contribution. Starting from

$$G = -2\text{Cov}(\frac{y}{\mu}, 1-F(y)),$$

the S-Gini is found by introducing α so as to modify the shape of the income distribution

$$G = -\alpha\text{Cov}(\frac{y}{\mu}, (1-F(y))^{\alpha-1}).$$

For $\alpha = 2$, of course, we recover the usual Gini index. With a value of α greater than 2, a greater weight is attached to low incomes.

We can run a small experiment, generating $n = 1000$ observations of a lognormal distribution and then computing the Gini according to the above formula, with various values of α. We then compare the result to the Gini computed using the usual formula corresponding to $\alpha = 2$.

```
n = 10000
x = sort(rlnorm(n))
y = seq(0,1,length=n)
for (alpha in c(1.2,2,3,4)){
  g = -alpha*cov(x/mean(x),(1-y)^(alpha-1))
  cat("Gini = ",g," alpha = ",alpha,"\n")}
Gini(x)
```

Table 1: Computing the α-Gini
using the empirical cumulative distribution

α	α-Gini	Usual Gini
1.2	0.2077537	-
2.0	0.5288477	0.5277905
3.0	0.6692843	-
4.0	0.7362263	-

For $\alpha = 1$, the modified Gini is equal to zero. For $\alpha = 2$, this method based on the empirical covariance is only approximate. In small samples, the difference can be substantial. For $n = 100$, the covariance method gives $G = 0.5413686$, while the correct methods gives $G = 0.5305954$.

5 Estimation of the Gini coefficient

5.1 Numerical evaluation

The definition of the Gini coefficient in term of the mean of absolute differences yield several ways of estimating it, without any assumption on the shape of F. The direct approach using a double summation is not feasible. We have first to order the observations to compute the order statistics $x_{[i]}$. Several methods were proposed in the literature:

- Deaton (1997) in his book orders the observations and proposes to use

$$G = \frac{n+1}{n-1} - \frac{2}{n(n-1)\mu} \sum (n+1-i) x_{[i]}.$$

  Note that this formula points out that there are $n(n-1)$ distinct pairs.

- Sen (1973) uses a slight simplification of this with

$$G = \frac{n+1}{n} - \frac{2}{n^2 \mu} \sum (n+1-i) x_{[i]}.$$

- The interpretation of the Gini coefficient in term of covariance between the variable and its rank implies that a simple routine can be used

$$G = \frac{2}{n\mu} \text{Cov}(y_{[i]}, i).$$

For the covariance approach, we note that the mean of the ranks is given by

$$\bar{i} = \frac{1}{n} \sum i = \frac{n+1}{2}.$$

So the covariance is estimated by

$$\text{Cov}(i, y_{[i]}) = \frac{1}{n}\sum(i - \bar{i})y_{[i]} = \frac{1}{n}\sum i\, y_{[i]} - \frac{n+1}{2}\mu,$$

and the Gini coefficient is obtained as:

$$G = \frac{2}{n^2\mu}\sum i\, y_{[i]} - \frac{n+1}{n},$$

which is the formula of Sen (1973).

5.2 Inference for the Gini coefficient

The main question is to find a standard deviation for the Gini coefficient. This is not an easy task because the observations are ordered and thus are not independent. We can find essentially two methods in the recent literature.

Giles (2004) found that the Gini can be estimated as

$$I_G = \frac{2\hat{\theta}}{n} - \frac{n+1}{n}, \tag{1}$$

where $\hat{\theta}$ is the OLS estimate of θ in the weighted regression

$$i\sqrt{x_{[i]}} = \theta\sqrt{x_{[i]}} + u_i\sqrt{x_[i]}. \tag{2}$$

where $x_{[i]}$ is an order statistics and i its rank. An appropriate standard error for the Gini coefficient is then

$$SE(I_G) = \frac{2\sqrt{\text{Var}(\hat{\theta})}}{n}. \tag{3}$$

This estimation is biased because the usual regression assumptions are not verified in the above regression. For instance the residuals are dependent.

Davidson (2009) gives an alternative expression for the variance of the Gini which is not based on a regression, but simply on the properties of the empirical estimate of $F(x)$. If we note $\hat{I}_G$ the numerical evaluation of the sample Gini, we have:

$$V\hat{a}r(\hat{I}_G) = \frac{1}{(n\hat{\mu})^2}\sum(\hat{Z}_i - \bar{Z})^2, \tag{4}$$

where $\bar{Z} = (1/n)\sum_{i=1}^n \hat{Z}_i$ is an estimate of $E(Z_i)$ and

$$\hat{Z}_i = -(\hat{I}_G + 1)x_{[i]} + \frac{2i-1}{n}x_{[i]} - \frac{2}{n}\sum_{j=1}^{i} x_{[j]}.$$

This is however an asymptotic result which is general gives lower values than those obtained with the regression method of Giles. Small sample results can be obtained if we adjust a parametric density for y and use a Bayesian approach.

6 Lorenz curve and other inequality measures*

Simple summary measures of inequality can readily be obtained from the graph of a Lorenz curve. The share in total income of the bottom p proportion of the population is given by $L(p)$; the greater that share, the more equal is the distribution of income. Analogously, the share in total income of the richest p proportion of the population is given by $1 - L(p)$; the greater that share, the more unequal is the distribution of income.

6.1 Schutz or Pietra index

An interesting but less well-known index of inequality is given by the Pietra index. What is the proportion of total income that would be needed to be reallocated across the population in order to achieve perfect equality. This proportion is given by the maximum value of $p - L(p)$, which is attained where the slope of $L(p)$ of the Lorenz curve is 1 (i.e., at $L(p = F(\mu))$). It is therefore equal to

$$F(\mu) - L(F(\mu)).$$

This index is called the *Schutz* coefficient in Duclos and Araar (2006), but is also known under the name of the Pietra index. In a stricter mathematical framework and following Sarabia (2008), the Pietra index is defined as the maximal deviation between the Lorenz curve and the egalitarian line

$$P_X = \max_{0 \leq p \leq 1} \{p - L_X(p)\}.$$

If we assume that F is strictly increasing on its support, the function $p - L_X(p)$ will be differentiable everywhere on $(0, 1)$ and its maximum will be reached when its first derivative in p

$$1 - F^{-1}(x)/\mu$$

is zero, that is, when $x = F(\mu)$. The value of $p - L_X(p)$ at this point is given by

$$P_X = F(\mu) - \frac{1}{\mu} \int_0^{F(\mu)} [\mu - F^{-1}(t)] dt = \frac{1}{2\mu} \int_0^{\infty} |t - \mu| dF(t).$$

Consequently

$$P_X = \frac{\mathrm{E}|X - \mu|}{2\mu},$$

which is an alternative formula for the Pietra index.

6.2 Other inequality measures

It is possible also to give a formulation of the Atkinson index and of the Entropy index as transformations of the Lorenz curve. We first give the expression of these two indices when X is a continuous random variable.

The Atkinson inequality indices are defined as

$$I_A(\varepsilon) = 1 - \left[\int_0^\infty (x/\mu)^{1-\varepsilon}\, dF(x)\right]^{1/(1-\varepsilon)}, \varepsilon > 0,$$

where ε is the parameter that controls inequality aversion. The limiting case $\varepsilon \to 1$ is

$$I_A(1) = 1 - \frac{1}{\mu} \exp\left\{\int_0^\infty \log(x) dF(x)\right\}.$$

The family of generalized entropy indices is

$$I_G(c) = \frac{1}{c(c-1)} \int_0^\infty [(x/\mu)^c - 1] dF(x), \quad c \neq 0, 1$$

The two particular cases obtained for $c = 0$ and $c = 1$ are

$$I_G(0) = \int_0^\infty \log(\mu/x)\, dF(x),$$

and

$$I_G(1) = \int_0^\infty (x/\mu)\; log(x/\mu)\, dF(x).$$

These two indices can be written in terms of the Lorenz Curve. We have for the Atkinson index

$$I_A(\varepsilon) = 1 - \left\{\int_0^1 [L'_X(p)]^{1-\varepsilon}\, dp\right\}^{1/(1-\varepsilon}, \varepsilon > 0.$$

For the generalized entropy index:

$$I_G(c) = \frac{1}{c(c-1)} \int_0^1 \{[L'_X(p)]^c - 1\} dp, c \neq 0, 1.$$

These formulas allow these indices to be obtained directly from the Lorenz curve without the necessity of knowing the underlying cumulative distribution function.

7 Main parametric distributions and their properties

Several densities have been proposed in the literature to model the income distribution. Of course all these densities are defined for a positive support. The most simple distributions, and consequently the widely used ones are the Pareto and the log-normal. These distributions have two parameters. The gamma and the Weibull are also two parameter distributions. In order to fit better the tails, three parameters distributions were proposed. We shall examine the mainly the Singh-Maddala distribution. We must note that all these densities are uni-modal. Four parameter densities were proposed in the literature, without solving the question of multi-modality. At this stage, mixture of simple distributions offer more flexibility without having an overwhelming cost in term of parsimony.

7.1 The Pareto distribution

Pareto (1897) observed that in many populations the income distribution was one in which the number of individuals whose income exceeded a given level x could be approximated by Cx^{α} for some choice of C and α. More specifically, he observed that such an approximation seemed to be appropriate for large incomes, i.e. for x above a certain threshold. If one, for various values of x, plots the logarithm of the income level against the number of individuals whose income exceeds that level, Pareto's intuition suggests that an approximately linear plot will be encountered.

The important role of the Pareto laws in the study of income and other size distributions is somewhat comparable to the central role played by the normal distribution in many experimental sciences. In both settings, plausible stochastic arguments can be advanced in favour of the models, but probably the deciding factor is that the models are analytically tractable and do seem to adequately fit observed data in many cases.

A random variable X follows a Pareto distribution if its survival function is

$$\bar{F}(x) = P(X > x) = \left(\frac{x}{x_m}\right)^{-\alpha}, \quad x > x_m.$$

The use of the survival function comes from the intuitive characterization of the Pareto. The cumulative function is simply $1 - \bar{F}$ which implies

$$F(x) = P(X < x) = 1 - \left(\frac{x}{x_m}\right)^{-\alpha}.$$

The density is obtained by differentiation

$$f(x) = \alpha x_m^{\alpha} x^{-\alpha-1}, \quad x > x_m.$$

Moments are given in Table 7. We can already see that this density has a special shape. It is

Table 2: Moments of the Pareto distribution

parameters	value	domain
scale	x_m	$x_m > 0$
shape	α	$\alpha > 0$
support	$x \in [x_m; +\infty)$	
median	$x_m \sqrt[\alpha]{2}$	
mode	x_m	
mean	$x_m \frac{\alpha}{\alpha - 1}$	$\alpha > 1$
variance	$x_m^2 \frac{\alpha}{(\alpha-1)^2(\alpha-2)}$	$\alpha > 2$

always decreasing. So it is valuable only to model high or medium incomes. Its moments are restricted to exist only for certain values of α. This is the price to pay for its long tails. In Figure 3, we give the graph of the density for $x_m = 1$ and various plausible values of α. The Gini index (see Table 4 for its expression) is very sensitive to the value of α. Table 3 shows that the most

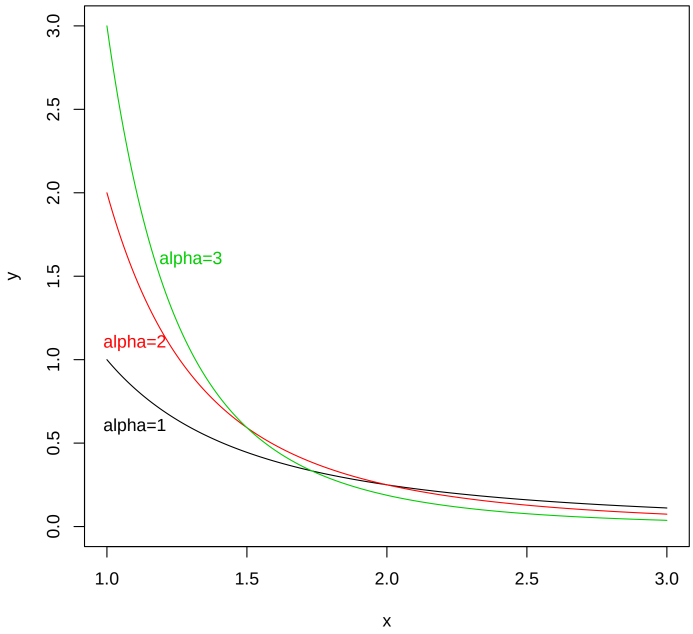

Figure 3: Pareto density

Table 3: Gini and Pietra indices for the Pareto

α	1.2	1.5	2.0	2.5	3.0	3.5
Gini	0.71	0.50	0.33	0.25	0.20	0.17
Pietra	0.58	0.39	0.25	0.19	0.15	0.12

plausible values of the Gini correspond to the very small range $\alpha \in [2, 2.5]$.

The tails of the Pareto distribution have an interesting property which is nice for an empirical test. On a log-log graph, the tail of the Pareto distribution is a straight line as

$$\log(\Pr(X \geq x)) = \alpha \log(x_m) - \alpha \log(x).$$

Because the distribution is available analytically, many interesting characteristics for inequality analysis are directly available and given in Table 4. These expressions are particularly simple. In particular the Lorenz curve of two Pareto distributions can never intersect as soon as the α are different. This is a strong restriction. In Figure 4, we have displayed Lorenz curves associated to the Pareto densities for various values of α. The Pareto density is very unequal for low values of α. It is particularly able to give a good place to rich people in the income distribution. These Lorenz curve are totally different from those that will be obtained for the log-normal density.

Many variants of the Pareto distribution were proposed in the literature, see for instance Arnold (2008). Usual generalizations are Pareto II-IV which introduce more parameters. Those variants can be interesting to model top incomes as in Jenkins (2017).

Table 4: Various coefficients for the Pareto distribution

Coefficient	expression	domain
Coefficient of variation	$(\alpha^2 - 2\alpha)^{-1/2}$	$\alpha > 2$
Lorenz curve	$L(p) = 1 - (1-p)^{(\alpha-1)/\alpha}$	$\alpha > 1$
Pietra index	$\frac{(\alpha-1)^{\alpha-1}}{\alpha^\alpha}$	$\alpha > 1$
Gini index	$(2\alpha - 1)^{-1}$	$\alpha > 1/2$
Atkinson	$1 - \frac{\alpha-1}{\alpha}\left[\frac{\alpha}{\alpha+\varepsilon-1}\right]^{1/(1-\varepsilon)}$	$\alpha > 1$
Generalized entropy	$\frac{1}{\theta^2-\theta}\left[\left[\frac{\alpha-1}{\alpha}\right]^\theta \frac{\alpha}{\alpha-\theta} - 1\right]$	$\alpha > 1$

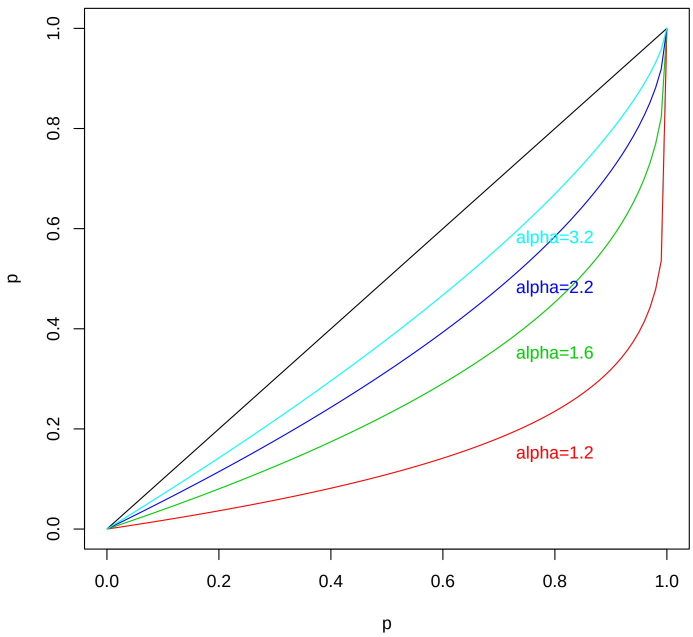

Figure 4: Lorenz curves for the Pareto density

Pareto and more generally power function distributions can appear in a variety of context that are nicely summarized in Mitzenmacher (2004). For instance Champernowne (1953) considers a minimum income x_m and then breaks income and small intervals with bounds defined as $x_m\gamma^j$ with $\gamma > 1$. Over each time step, An individual can move from class i to class j with a probability p_{ij} that depends only on the value of $j - i$. Champernowne (1953) shows that the equilibrium distribution is a Pareto. In fact, a Pareto is obtained in a multiplicative process with a minimum bound.

```
Lp = function(p,alpha) {1-(1-p)^((alpha-1)/alpha)}
p = seq(0,1,0.01)
plot(p,p,type="l")
lines(p,Lp(p,1.2),col=2)
lines(p,Lp(p,1.6),col=3)
lines(p,Lp(p,2.2),col=4)
lines(p,Lp(p,3.2),col=5)
text(0.8,0.15,"alpha=1.2",col=2)
text(0.8,0.35,"alpha=1.6",col=3)
text(0.8,0.48,"alpha=2.2",col=4)
text(0.8,0.58,"alpha=3.2",col=5)
```

7.2 LogNormal distribution

The log-normal density is convenient for modelling small to medium range incomes. A random variable X has a log normal distribution if its logarithm $\log X$ has a normal distribution. If Y is a random variable with a normal distribution, then $X = \exp(Y)$ has a log-normal distribution; likewise, if X is log-normally distributed, then $Y = \log X$ is normally distributed.

Let us suppose that y is N(μ, σ^2) and let us consider the change of variable $x = \exp y$. The Jacobian of the transformation from y to x is given by:

$$J(y \to x) = \frac{\partial y}{\partial x} = \frac{\partial \log x}{\partial x} = \frac{1}{x}.$$

So, the probability density function of a log-normal distribution is:

$$f_X(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp -\frac{(\ln x - \mu)^2}{2\sigma^2}, \quad x > 0.$$

The cumulative distribution function has no analytical form and requires an integral evaluation:

$$F_X(x; \mu, \sigma) = \frac{1}{2}\text{erfc}\left[-\frac{\ln x - \mu}{\sigma\sqrt{2}}\right] = \Phi\left(\frac{\ln x - \mu}{\sigma}\right),$$

where erfc is the complementary error function, and Φ is the standard normal cdf. However, these integrals are easy to evaluate on a computer and built-in functions are standard.

The moment are easily obtained as functions of μ and σ. If X is a log-normally distributed variable, its expected value, variance, and standard deviation are

$$\begin{aligned}\mathrm{E}[X] &= e^{\mu+\frac{1}{2}\sigma^2}, \\ \mathrm{Var}[X] &= (e^{\sigma^2}-1)e^{2\mu+\sigma^2}, \\ \mathrm{s.d}[X] &= \sqrt{\mathrm{Var}[X]} = e^{\mu+\frac{1}{2}\sigma^2}\sqrt{e^{\sigma^2}-1}.\end{aligned}$$

Equivalently, the parameters μ and σ can be obtained if the values of the mean and the variance are known:

$$\begin{aligned}\mu &= \ln(\mathrm{E}[X]) - \tfrac{1}{2}\ln\left(1+\frac{\mathrm{Var}[X]}{\mathrm{E}[X]^2}\right), \\ \sigma^2 &= \ln\left(1+\frac{\mathrm{Var}[X]}{\mathrm{E}[X]^2}\right).\end{aligned}$$

The mode is:

$$\mathrm{Mode}[X] = e^{\mu-\sigma^2}.$$

The median is:

$$\mathrm{Med}[X] = e^{\mu}.$$

The above graph was made for $\mu = 0$. The two densities have the same median, but of course

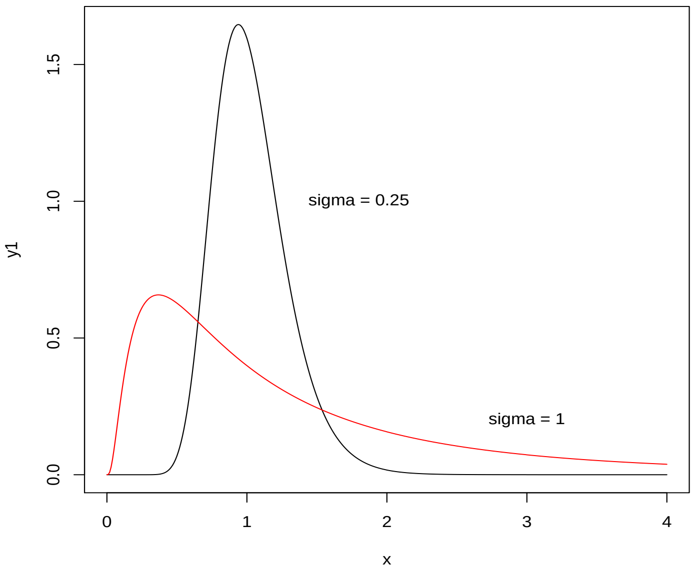

Figure 5: Log-normal density

not the same mean.

```
library(ineq)
x = seq(0,4,0.01)
y1 = dlnorm(x,meanlog=0,sdlog=0.25)
y2 = dlnorm(x,meanlog=0,sdlog=1.0)
plot(x,y1,type="l")
lines(x,y2,type="l",col="red")
text(1.8,1,"sigma = 0.25")
text(3,0.20,"sigma = 1")
```

The log-normal has some nice properties.

1. Suppose that all incomes are changed proportionally by a random multiplicative factor, which is different for everybody and that follows a gaussian process. Then the distribution of the population income will converge to a log-normal, if the process is active for a long enough period.

2. The log normal fits well to many data sets

3. Lorenz curves associated to the log-normal are symmetric around a line which is given by the points corresponding to the mean of x. This is a good visual test to see if the log-normal fits well to a data set.

4. Inequality depends on a single parameter σ which uniquely determines the shape of the Lorenz curves. The latter do not intersect. The Gini coefficient also depends uniquely on this parameter.

5. Close form under certain transformations

We know that if $X \sim N(\mu, \sigma^2)$, then $Y = a + bX$ is also normal with $Y \sim N(a + b\mu, b^2\sigma^2)$. Let us now consider a log-normal random variable $Y \sim \Lambda(\mu, \sigma^2)$ and the transformation $Y = aX^b$. Then $Y \sim \Lambda(\log(a) + b\mu, b^2\sigma^2)$. There is a nice application for this property. It has been observed in many countries that the tax scheduled can be approximated by

$$t = x - ax^b.$$

The disposable income is given by

$$y = ax^b.$$

So if the pre-tax income follows a log-normal, the disposable income will also follow a log-normal.

The right tail of the lognormal density behaves very differently from the Pareto tail, just because the log normal has got all its moment when the Pareto in general has no finite moment when α is too small. However, for large values of σ, the two distributions might have quite similar tails. This can be seen on a log-log graph. Let us take the log of the density

$$\begin{aligned}\log f(x) &= -\log x - \log\sqrt{2\pi}\sigma - \frac{(\log x-\mu)^2}{2\sigma^2} \\ &= -\frac{\log^2 x}{2\sigma^2} + (\frac{\mu}{\sigma^2}-1)\log x - \log\sqrt{2\pi}\sigma - \frac{\mu^2}{2\sigma^2} \\ &\simeq (\frac{\mu}{\sigma^2}-1)\log x - \log\sqrt{2\pi}\sigma - \frac{\mu^2}{2\sigma^2} \qquad \text{for large } \sigma\end{aligned}$$

The left tail of the log density behaves like a straight line for a large range of x when σ is large enough.

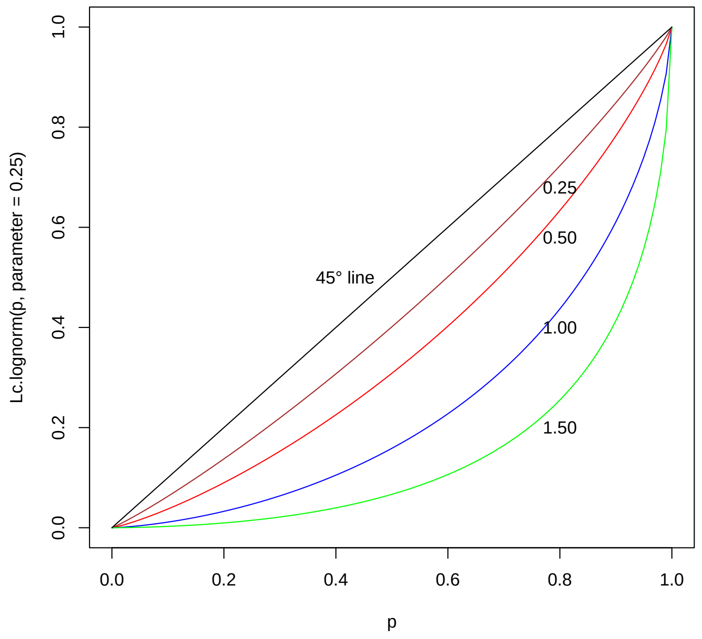

Figure 6: Log-normal Lorenz curves

```
library(ineq)
p = seq(0,1,0.01)
plot(p,Lc.lognorm(p, parameter=0.25),type="l",col="brown")
lines(p,Lc.lognorm(p, parameter=0.5),col="red")
lines(p,Lc.lognorm(p, parameter=1.0),col="blue")
lines(p,Lc.lognorm(p, parameter=1.5),col="green")
lines(p,p)
text(0.42,0.5,"45° line")
text(0.8,0.68,"0.25")
```

```
text(0.8,0.58,"0.50")
text(0.8,0.40,"1.00")
text(0.8,0.20,"1.50")
```

We can give some more details on this distribution, concerning Gini coefficient and the Lorenz curve. Let us call $\Phi(x)$ the standard normal distribution with $\Phi(x) = Prob(X < x)$. From Cowell (1995), we have Table 5. The Pietra index was found in Moothathua (1989).

Table 5: Various coefficients for the Log-Normal distribution

Coefficient of variation	$\sqrt{\exp(\sigma^2) - 1}$
Lorenz curve	$\Phi(\Phi^{-1}(p) - \sigma)$
Pietra index	$2\Phi(\sigma^2/2) - 1$
Gini index	$2\Phi(\sigma/\sqrt{2}) - 1$
Atkinson	$1 - \exp(-1/2\varepsilon\sigma^2)$
Generalized entropy	$\frac{\exp((\theta^2 - \theta)\sigma^2/2) - 1}{\theta^2 - \theta}$

The lognormal has an interesting poverty for poverty analysis. The mean is given by $\exp(\mu + \sigma^2/2)$ while the mode is $\exp(\mu)$. A usual practice for defining a poverty line is the take either $z_1 = 0.5\times$ the mean or $z_2 = 0.6\times$ the mode. Using the properties of the lognormal, we can show that these choices are not equivalent and can give rather different results. The two poverty lines are the same when $\sigma^2 = 2 \times \log(0.6/0.5) = 0.37$, which corresponds to a Gini index of 0.33. So, if we adopt a lognormal distribution for the French income, then $z_1 < z_2$ because the Gini index is lower than 0.30 while for China, we shall have just the contrary because the Gini index is greater than 0.50.

Lognormal distributions are usually generated by multiplicative models. The first explanation of this type was proposed by Gibrat (1930). We start with an initial value for income X_0. In the next period, this income can grow or diminish according to a multiplicative and positive random variable F_t

$$X_t = F_t X_{t-1}.$$

Taking the logs and using a recurrence, we have

$$\log X_t = \log(X_0) + \sum_k \log(F_k).$$

By the central limit theorem, we get a log normal distribution. Note that the mechanism designed by Champernowne (1953) was very similar. We got a Pareto distribution only because a minimum value was imposed.

7.3 Singh-Maddala distribution*

Singh and Maddala (1976) propose a justification of the old Burr XII distribution by considering the log survival function as a richer function of x than what the Pareto does. With the Pareto we

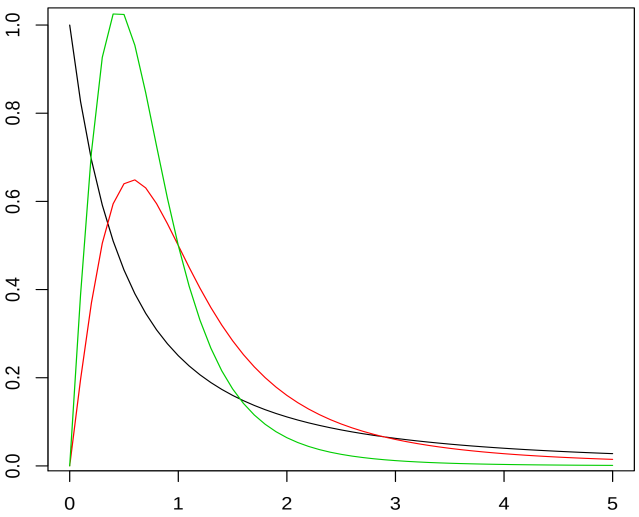

Figure 7: Singh-Maddala income distribution
The line in black corresponds to $a_2 = a_3 = 1$. Then, for the curve in red, $a_2 = 2$, while for the curve in green $a_3 = 3$.

had $\log(1-F) = \alpha\log(x_m) - \alpha\log(x)$. Here, the relation is no longer linear with:

$$\log(1-F) = -a_3\log(1+a_1x^{a_2}),$$

following the notations of Singh and Maddala (1976). Consequently, the cumulative distribution is

$$F_{SM}(x) = 1 - \frac{1}{(1+a_1x^{a_2})^{a_3}}.$$

The corresponding density is obtained by differentiation

$$f_{SM}(x|a,b,q) = a_1a_2a_3\frac{x^{a_2-1}}{[1+a_1x^{a_2}]^{a_3+1}}.$$

Let us plot this density for various values of the parameters. First of all, a_1 is just a scale parameter and we set it equal to 1. Then we use the following code in R:

```
x = seq(0,5,0.1)
f_SM = function(x,a_2,a3){
  f = a_2*a_3*(x^(a_2-1))/(1+x^(a_2))^(a_3+1)
  }
a_2 = 1
```

```
a_3 = 1
plot(x,f_SM(x,a_2,a3),type="l",ylab="",xlab="")
a_2=2
lines(x,f_SM(x,a_2,a3),col=2)
a_3 = 2
lines(x,f_SM(x,a_2,a3),col=3)
```

The parameter of the Pareto distribution could easily be estimated using a linear regression of $\log(1-\hat{F})$ over $\log(x)$ where $\hat{F}$ is the natural estimator of the cumulative distribution. Here a non linear regression can be applied which minimized:

$$\sum[\log(1-\hat{F}(x)) + a_3 \log(1+a_1 x^{a_2})]^2.$$

The uncentered moments of order h and the Gini coefficient are expressed in term of the Gamma function and can be found in McDonald and Ranson (1979) and McDonald (1984):

$$\mathrm{E}(X^h) = b^h \frac{\Gamma(1+h/a_2)\Gamma(a_3-h/a_2)}{\Gamma(a_3)}$$

with $b=(1/a_1)^{1/a_2}$ as well as the Gini index:

$$G = 1 - \frac{\Gamma(a_3)\Gamma(2a_3-1/a_2)}{\Gamma(a_3-1/a_2)\Gamma(2a_3)}.$$

All the moment do not exist in this distribution. For a moment of order h, we must have

$$a_3 > \frac{h}{a_2}.$$

If $a_3 > 1/a_2$, we can derive the Lorenz curve as

$$\begin{aligned} LC(p) &= \frac{1}{\mu}\int_0^p b[(1-y)^{-1/a_3}-1]^{1/a_2} dy \\ &= \frac{a_3}{\mu}\int_0^z t^{1/a_2}(1-t)^{a_3-1/a_2-1} dt \\ &= I_z(1+1/a_2, a_3-1/a_2) \end{aligned}$$

where $z = 1-(1-a_3)^{1/a_3}$ and $I_z(a,b)$ denotes the incomplete beta function ratio defined by:

$$IB_z(a,b) = \frac{\int_0^z t^{a-1}(1-t)^{b-1} dt}{\int_0^1 t^{a-1}(1-t)^{b-1} dt}.$$

The Singh-Maddala distribution admit two limiting distributions, depending on the value of a_3. For $a_3 = 1$, we have the Fisk (1961) distribution. For $a_3 \to \infty$, we have the Weibull distribution, to be detailed later on. So, depending on the value of a_3, the associated Lorenz curves are supposed to cover a wide range of shapes. In the left panel, we kept $a_2 = 2$ and let a_3 vary between 0.7 and 2. In the right panel, we kept $a_3 = 0.7$ and let a_2 vary between 2 and 3.5. The two black curves are identical. In one case the modification is more in the right part and in the other case more in the left part. However, we note that the flexibility is not very strong.

The corresponding code using R is:

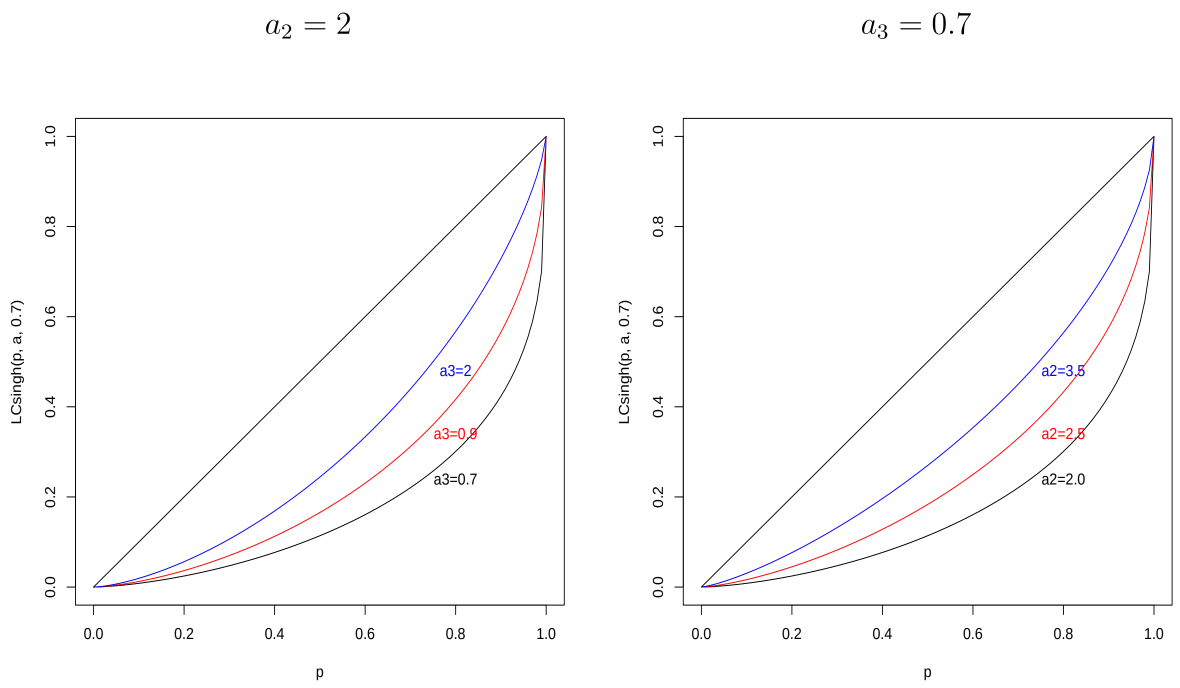

Figure 8: Singh-Maddala Lorenz curves when varying a_2 or a_3

```
LCsingh <- function(p,a,q){
pbeta((1 - (1 - p)^(1/q)), (1 + 1/a), (q-1/a))}
p = seq(0,1,0.01)
a = 2
plot(p,LCsingh(p, a,0.7),type="l")
lines(p,LCsingh(p, a,0.9),type="l",col="red")
lines(p,LCsingh(p, a,2),type="l",col="blue")
lines(p,p)
text(0.8,0.24,"a3=0.7")
text(0.8,0.34,"a3=0.9",col="red")
text(0.8,0.48,"a3=2",col="blue")
```

7.4 Weibull distribution*

The Weibull distribution is a nice two parameter distribution where all moments exists. It is obtained as a special case of the three parameter Singh Maddala distribution, for $a_3 \to \infty$. This relation explains that the cumulative distribution has an analytical form:

$$F(x) = 1 - \exp(-(kx)^\alpha).$$

By differentiation, we get the density

$$f(x) = k\,\alpha\,(k\,x)^{\alpha-1} \exp -(k\,x)^\alpha.$$

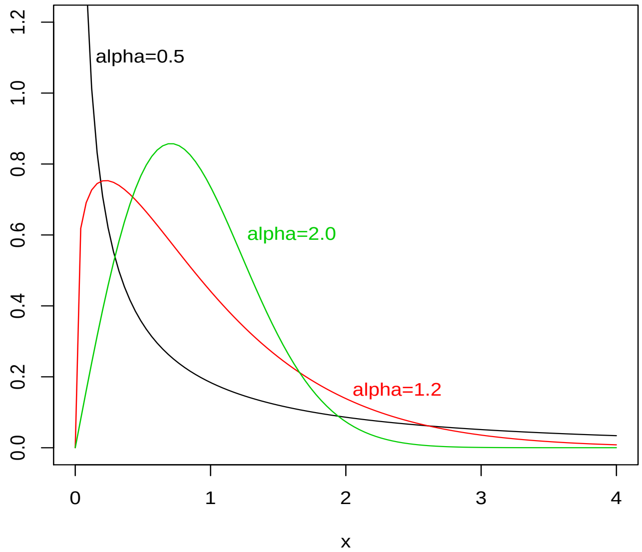

Figure 9: Weibull income distribution

We have a plot of this density in Figure 9. For $\alpha < 1$, the density has the shape of the Pareto density, which means that it has no finite maximum. For $\alpha = 1$, it cuts the y axis. As α grows, there is less and less inequality and the function concentrates around its mean. Plausible values for α corresponding to usual income distributions are $[1.5 - 2.5]$.

The $h - th$ moments around zero are given by

$$\mu_h = \frac{\Gamma(1 + h/\alpha)}{k^h}$$

where $\Gamma(a)$ is the gamma function defined by

$$\Gamma(a) = \int_0^\infty u^a \exp(-u)\, du$$

The coefficient of variation (the ratio between the standard deviation and the mean) is equal to:

$$CV = \frac{\sqrt{\Gamma((\alpha+2)/\alpha) - \Gamma(\alpha+1)/\alpha)^2}}{\Gamma((\alpha+1)/\alpha)}$$

As we have the direct expression of the distribution, the Gini coefficient and the Lorenz curves are directly available. We find the expression of the Lorenz curve and the Gini index for instance in Krause (2014):

$$LC = 1 - \frac{\Gamma(-\log(1-p), 1 + 1/\alpha)}{\Gamma(1 + 1/\alpha)},$$

where $\Gamma(x,\alpha)$ is the incomplete Gamma function.

We regroup in Table 6 some of these results. We did not manage to fully complete this Table, presumably because the Weibull distribution is not very often used for modelling the income distribution.

Table 6: Several indices for the Weibull distribution

Coefficient of variation	$\frac{\sqrt{\Gamma((\alpha+2)/\alpha) - \Gamma(\alpha+1)/\alpha)^2}}{\Gamma((\alpha+1)/\alpha)}$
Lorenz curve	$1 - \frac{\Gamma(-\log(1-p), 1+1/\alpha)}{\Gamma(1+1/\alpha)}$
Pietra index	
Gini index	$1 - 2^{-1/\alpha}$
Atkinson	
Generalized entropy	

Note that there are various ways of writing the density of the Weibull, concerning the scale parameter k. Either $(kx)^\alpha$ or $(x/k)^\alpha$. For inference, it might even be convenient to consider kx^α. So be careful. In R, the density is available as `dweibull(x, shape, scale = 1)` using the parameterizations $(x/k)^\alpha$.

The Weibull distribution shares with the Pareto, the Sing-Maddala distribution a common feature which is to have an analytical cumulative distribution. If we rearrange its expression and take logs, we get:

$$\log(-\log(1-F)) = \alpha \log(kx).$$

So that it is easy to check if a sample has a Weibull distribution. And by the way gives a method to estimate the parameter α.

7.5 Gamma distribution*

The probability density function using the shape-scale parameterizations is

$$f(x;k,\theta) = \frac{x^{k-1}e^{-\frac{x}{\theta}}}{\theta^k\Gamma(k)} \quad \text{for } x>0 \text{ and } k,\theta>0.$$

Here $\Gamma(k)$ is the gamma function evaluated at k. k represent the degrees of freedom. It is also the shape parameter. θ corresponds to the scale parameter in this parameterizations. Using this parameterizations, we can plot this density for $\theta = 1$ and various values of k.

```
n = 1000
x = seq(0,10,length=n)
df = 1.0
s = 1
y = dgamma(x,shape = df, scale = s)
plot(x,y,type="l",ylab="Density")
```

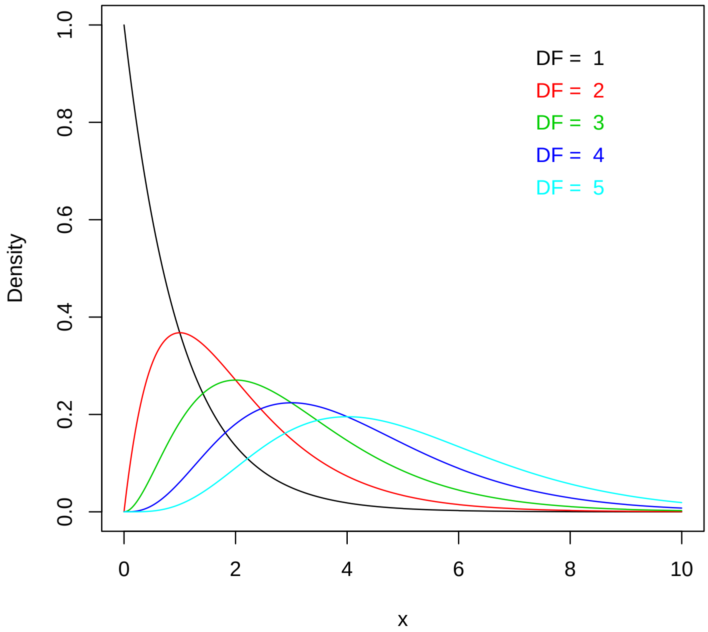

Figure 10: Gamma density

```
text(8,1.0-df/15,paste("DF = ",toString(df)),col=df)
for (df in c(2,3,4,5)){
  y = dgamma(x,shape = df, scale = s)
  lines(x,y,col = df)
  text(8,1.0-df/15,paste("DF = ",toString(df)),col=df)}
```

The cumulative distribution function is the regularized gamma function:

$$F(x; k, \theta) = \int_0^x f(u; k, \theta)\, du = \frac{\gamma\left(k, \frac{x}{\theta}\right)}{\Gamma(k)}$$

where $\gamma(k, x/\theta)$ is the lower incomplete gamma function.

The skewness is equal to $2/\sqrt{k}$, it depends only on the shape parameter k and approaches a normal distribution when k is large (approximately when $k > 10$). The mean is $k\theta$ and the variance $k\theta^2$.

Rather easy to estimate. Bayesian inference. In R, dgamma, pgamma, qgamma, rgamma using the same parametrization.

7.6 Variations around the Pareto distribution*

We have presented the Pareto I distribution. Pareto distribution have a right tail which is a power function. Several variants were proposed in the literature, a good account of which is given in

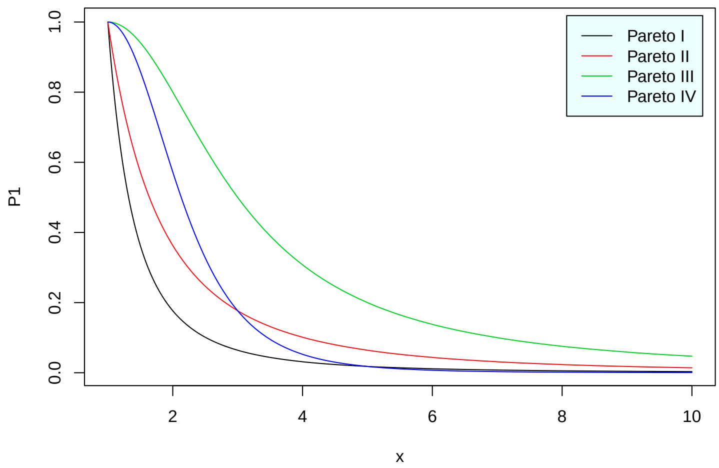

Figure 11: The Pareto family

Arnold (2008). We reproduce the following table 7. These classes provide more flexibility to

Table 7: Pareto distributions

	$1-F(x)$	Support	Parameters
Type I	$(x/x_m)^{-\alpha}$	$x \geq x_m$	$h > 0$, $\alpha > 0$
Type II	$\left[1+\frac{x-\mu}{\sigma}\right]^{-\alpha}$	$x \geq \mu$	$\mu \in R$, $\sigma > 0$, α
Type III	$\left[1+\left(\frac{x-\mu}{\sigma}\right)^{1/\gamma}\right]^{-1}$	$x \geq \mu$	$\mu \in R, \sigma, \gamma > 0$
Type IV	$\left[1+\left(\frac{x-\mu}{\sigma}\right)^{1/\gamma}\right]^{-\alpha}$	$x \geq \mu$	$\mu \in R, \sigma, \gamma > 0, \alpha > 0$

the right tail and in particular in relation with the shape of the left tail for the Pareto IV. Jenkins (2017) has used a lot these variants of the Pareto to model high incomes in the UK. The influence of the different parameters can be seen in Figure 11. Parameters x_m and μ play the same role in defining the support. They were equal to 1. Pareto II introduces $\sigma = 2$ which helps to modify the left bottom of the curve, depending it is greater or lower than 1.0. Pareto II and IV introduce $\gamma = 0.5$. This helps to modify the top part if $\gamma < 1$.

We arrive at four parameters in the last case, but still the mode is at the left limit of the support. A large class of four parameter densities was proposed in McDonald (1984) and the most famous one is the Generalized beta II. The main goal was to provide flexibility for both the left and right tails.

A more recent distribution was developed in Reed and Jorgensen (2004), applied for income distributions in Reed (2003) and is known also as the double Pareto. It is closely related to the lognormal and Pareto distributions. A good review of this distribution and its comparison with the Pareto and the lognormal distributions is given in Mitzenmacher (2004). Both the Generalized Beta II and the Double Pareto have four parameters, but are uni-modal.

7.7 Which density should we select?*

In his book, Cowell (1995) is not very optimistic about the more complicated four parameter densities. Their parameters are hard to interpret and they are difficult to estimate. He is more in favour of the Pareto density, which is fact has a single important parameter (x_m defines only the support of the density), the two parameter lognormal and eventually the gamma density. He does not like the more complicated densities like the Singh-Maddala and even more the generalized Beta II. In Lubrano and Protopopescu (2004), we make use of the two parameter Weibull density to estimate generalized Lorenz curves and rank bibliometric distributions. The three parameters Singh-Maddala distribution is quite simple to estimate as the authors propose a method based on a regression. The three parameter generalized gamma density has a very awkward parameterization so that it has the reputation of being not estimable by maximum likelihood on individual data.

The Pareto density and its variants are nice for modelling high incomes, see in particular Jenkins (2017). The gamma density is nice for modelling mid range incomes as well as the log-normal density. Cowell (1995) thus prefers two parameter densities for modelling particular portions of the income distribution. We can conclude that using mixture of two parameter densities might be the best alternative for modelling the complete income distribution.

8 Pigou-Dalton transfers and Lorenz ordering

Pigou-Dalton transfers are mean-preserving equalizing transfers of income. They involve a marginal transfer of 1 from a richer person belonging to percentile p_r to a poorer person belonging to percentile $p_p < p_r$) that keeps total income constant. These equalizing transfers have the consequence of moving the Lorenz curve unambiguously closer to the line of perfect equality. This is because such transfers do not affect the value of $L(p)$ for all p up to p_p and for all p greater than p_r, but they increase $L(p)$ for all p between p_p and p_r.

8.1 Lorenz ordering

Let us consider two income distributions A and B, where distribution B is obtained by applying Pigou-Dalton transfers to A. Hence, the Lorenz curve $L_B(p)$ of distribution B will be everywhere above the Lorenz curve $L_A(p)$ of distribution A. Inequality indices which obey the principle of transfers will unambiguously indicate more inequality in A than in B. We will also say that if

$$L_B(p) - L_A(p) \geq 0 \qquad \forall p$$

then B Lorenz dominates A.

8.2 A numerical example

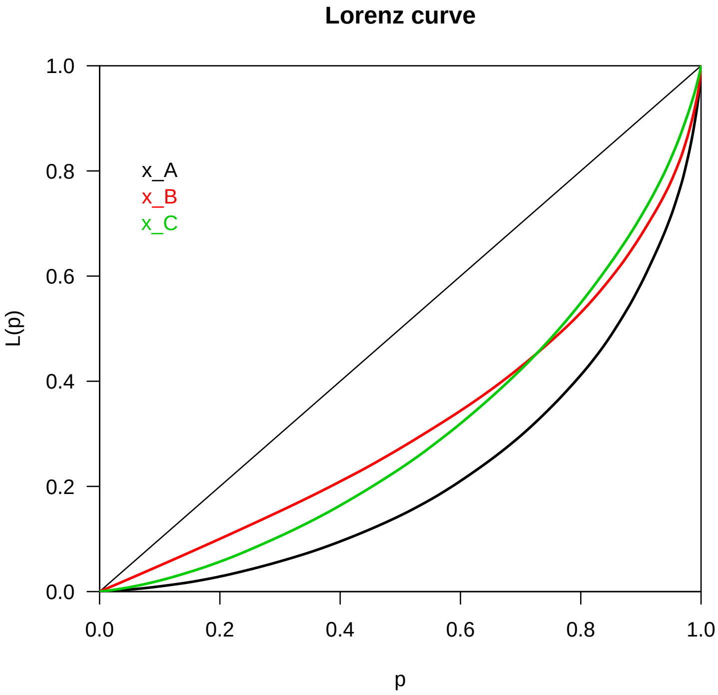

Figure 12: Lorenz dominance for Pigou-Dalton transfers

We are going to illustrate Pigou-Dalton transfers on a simulated example. We first generate an income distribution x_A, using a lognormal distribution with parameters 0 and 1 and $n = 50\ 000$ observations. We then define a flat rate of taxation τ equal to 0.25. A Pigou-Dalton transfer takes money from the rich to redistribute to the poor without changing the mean income and without changing the order of the incomes. We can thus define the transfers as

$$Tr = \tau * sort(x_A, decreasing = T)$$

where $sort(x_A, decreasing = T)$ is the reverse order of x_A, provided x_A is sorter by increasing values. The new income distribution x_B is

$$x_B = (1 - \tau) \times x_A + Tr$$

We finally draw n values of x_C from a lognormal with a σ that should produce the same Gini as in x_B ($\sigma = \sqrt{2}\Phi^{-1}(G_B + 1)/2)$) and a mean close to that of x_B ($\mu = \log(\bar{x}_B) - \sigma^2/2$). In Table 8, we report the mean and the Gini coefficient of each distribution. We illustrate these numbers in Figure 12 where we have drawn the Lorenz curve of x_A in black. It is the farthest away from the diagonal. Inequality is rather large in this income distribution. Pigou-Dalton transfers do not change the mean and the ordering, but reduce greatly the Gini coefficient. The Lorenz curve

Table 8: The effects of Pigou-Dalton transfers

Distribution	Mean	Gini
Before redistribution x_A	1.647	0.521
After redistribution x_B	1.647	0.331
Log normal x_C	1.648	0.331

corresponding to x_B is in red. It does not intersect L_A even if the distribution of x_B cannot be a lognormal.

The last sample x_C should have a Gini coefficient close to that of x_B. However, its Lorenz curve crosses that of x_B because it is obtained in a totally different way, implying different transfers which are not Pigou-Dalton.

The R code is as follows:

```
n = 50000
x_A = sort(rlnorm(n,0,1))
tau = 0.25
Tr = tau*sort(x_A,decreasing = T)
x_B = (1-tau)*x_A + Tr

s = sqrt(2)*qnorm((gini(x_B)+1)/2)
mu = log(mean(x_A))-0.5*s^2
x_C = rlnorm(n,mu,s)


cat(mean(x_A),gini(x_A),"\n")
cat(mean(x_B),gini(x_B),"\n")
cat(mean(x_C),gini(x_C),"\n")

plot(Lc(x_A))
lines(Lc(x_B),col=2)
lines(Lc(x_C),col=3)
```

8.3 Generalized Lorenz Curve

The generalized Lorenz curve (GLC) introduced by Shorrocks (1983) is the most important variation of the Lorenz curve (LC). The LC is scale invariant and is thus only an indicator of relative inequality. However, it does not provide a complete basis for making social welfare comparisons. The Shorrocks proposal is the generalized Lorenz curve defined as:

$$GLC(p) = \mu LC(p) = \int_0^p F^{-1}(y)dy$$

Note that $GLC(0) = 0$ and $GLC(1) = \mu$. A distribution with a dominating GLC provides greater welfare according to all concave increasing social welfare functions defined on individual incomes (Kakwani 1984, and the surveys of Davies et al. 1998 and Sarabia 2008). On the other hand, the GLC is no longer scale-free and in consequence it determines any distribution with finite mean.

The usual Lorenz curve when one focusses his attention on inequality only. The Generalized Lorenz curves mixes concerns for inequality and for the mean, so it is related to welfare comparisons. The order induced by GLC is in fact the second-order stochastic dominance that we shall eventually study in a next chapter. This order is a new partial ordering, and sometimes it allows a bigger percentage of curves to be ordered than in the Lorenz ordering case.

8.4 Lorenz Ordering for usual distributions*

Lorenz curves can be used to define an ordering in the space of the of distributions. If two distribution functions have associated Lorenz curves which do not intersect, they can be ordered without ambiguity in terms of welfare functions which are symmetric, increasing and quasi-concave (see Atkinson 1970. We express this formally with the definition:

Definition 2 *Let A and B be two income distributions. Distribution B is preferred to distribution A in the Lorenz sense iff:*

$$B \succeq_L A \Leftrightarrow L_B(p) \geq L_A(p), \qquad \forall p \in [0, 1].$$

If $B \succeq_L A$, then B exhibits less inequality than A in the Lorenz sense. Note that the Lorenz order is a partial order and is invariant with respect to scale transformation.

It is fairly possible now to characterize Lorenz dominance by restrictions over the parameter space if the two random variables have the same class of distributions. For some parametric families the restrictions will be very simple, and by the way will imply rather simple parametric statistical tests. We have derived Lorenz curves for the most important parametric densities, leaving aside those which were too complex and which are surveyed in Sarabia (2008).

We present first results for the Pareto and the log-normal.

- Pareto: Let $X_i \sim P(\alpha_i, xm_i)$. Then

$$F_{X_1} \succeq_L F_{X_2} \Leftrightarrow \alpha_1 \geq \alpha_2$$

- Log-Normal: Let $X_i \sim LN(\mu_i, \sigma_i^2)$. Then

$$F_{X_1} \succeq_L F_{X_2} \Leftrightarrow \sigma_1 \leq \sigma_2$$

The proof of these results is straightforward because in the two cases, the Lorenz curves never intersect as they depend on a single parameter.

The case of the Singh-Maddala distribution is more difficult to establish. Its Lorenz curve depends on two parameters and may thus intersect. Let us note the normalized distribution as $F = 1 - 1/(1 + x^a)^q$. Then from Sarabia (2008) we get:

Theorem 3 *Let $X_i \sim SM(a_i, q_i), i = 1, 2$ be two Singh-Maddala distributions. Then*

$$X_1 \succeq_L X_2 \Leftrightarrow a_1 q_1 \leq a_2 q_2, \text{ and } a_1 \leq a_2.$$

The proof of this result is more delicate to establish and the statistical test of these restrictions is slightly more difficult to implement.

9 Parametric Lorenz curves*

We first recall in a table the expression of the Lorenz curve for some standard income distribution. We gave a theorem characterizing a Lorenz curve. This means that any function following these

Table 9: Lorenz and Gini indices for classical income distributions

Distribution	Lorenz curve	Gini index
Pareto I	$L(p) = 1 - (1-p)^{1-1/\alpha}$	$\frac{1}{2\alpha-1}$
Lognormal	$L(p) = \Phi(\Phi^{-1}(p) - \sigma)$	$2\Phi(\sigma/\sqrt{2}) - 1$
Weibull	$L(p) = 1 - \dfrac{\Gamma(-\log(1-p), 1+1/\alpha)}{\Gamma(1+1/\alpha)}$	$1 - 2^{-1/\alpha}$
Singh-Maddla	$L(p) = I_z(q + 1/a, q - 1/a)$	$1 - \frac{\Gamma(q)\Gamma(2q-1/a)}{\Gamma(q-1/a)\Gamma(2q)}$

properties is a Lorenz curve. So we can try to investigate this class of functions. We follow Sarabia (2008), but not all the details. The first parametric form which was given in the literature is

$$L(p) = p\alpha \exp(-\beta(1-p)),$$

with $\alpha \geq 1$ and $\beta > 0$.

A family of Lorenz curves which is interesting and easy to understand is build around the Pareto family. We can generalize the Lorenz curve of the Pareto by adding one more parameter, so as to get:

$$L(p) = [1 - (1-p)^{1-1/\alpha}]^{\beta}.$$

If $\beta = 1$, we have the asymmetric Lorenz curve of the Pareto. If $\beta = 1/(1 - 1/\alpha)$, we obtain a symmetric Lorenz curve, thus having a similar property to that of the Lognormal. The underlying density to this Lorenz curve combines properties of the Pareto and of the Lognormal. More general expressions are given in Sarabia (2008).

Let us explore these functional forms using R.

```
LCgen <- function(p,alpha,beta){
  smlc <- (1-(1-p)^(1-1/alpha))^beta
  smlc}
p = seq(0,1,0.01)
plot(p,LCgen(p, 1.5,1),type="l")
text(0.93,0.45,"1.5, 1.0")
lines(p,LCgen(p,3,1.5),type="l",col="red")
```

```
text(0.7,0.50,"3.0, 1.5")
lines(p,LCgen(p,4,2),type="l",col="blue")
text(0.5,0.10,"4.0, 2.0")
lines(p,p)
text(0.42,0.5,"45° line")
```

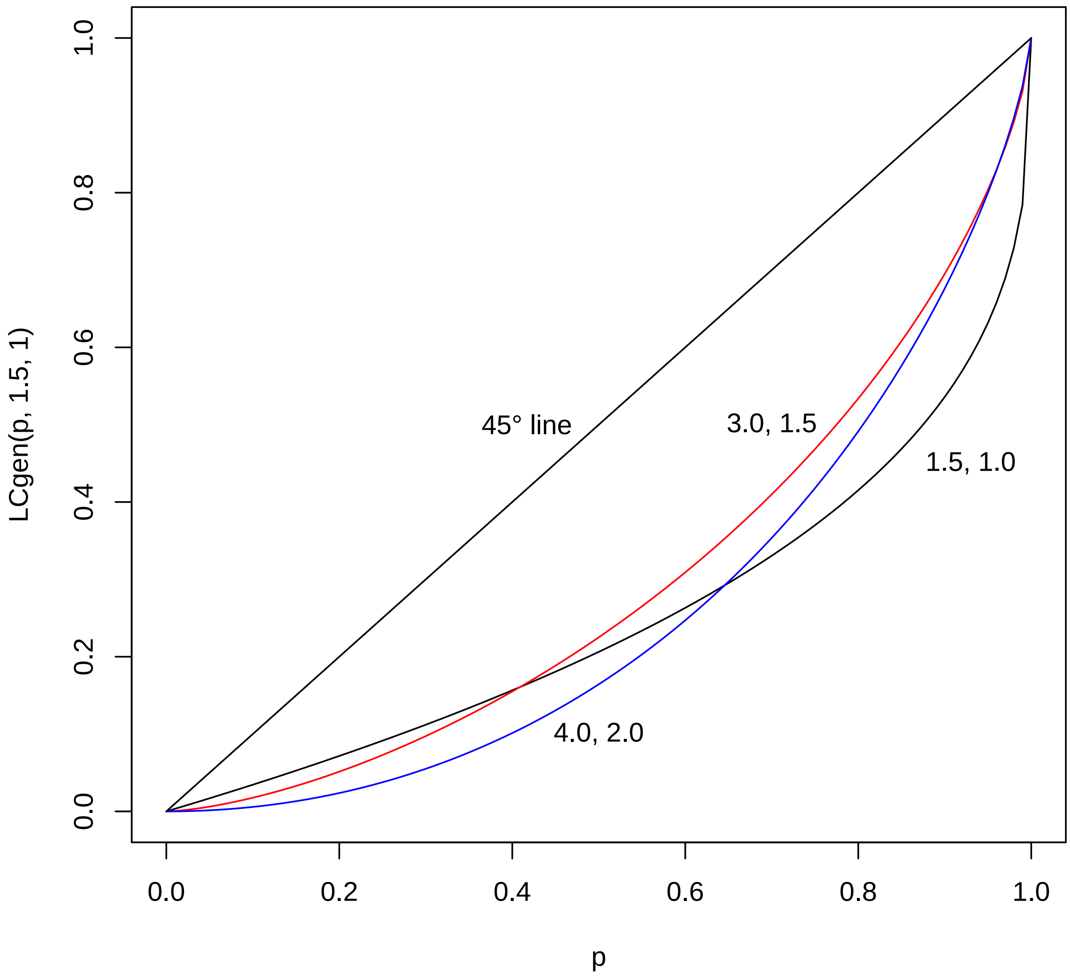

Figure 13: The flexibility of a two parameter Lorenz curve

It is remarkable that play playing with two parameters, we can obtain very different shapes and in particular many points of intersection in a much simpler way than with the Singh-Maddala distribution. The Gini coefficient has a simple expression and is equal to

$$G = 1 - \frac{2}{1 - 1/\alpha} B(1/(1 - 1/\alpha), \beta + 1)$$

where B(.,.) is the incomplete Beta function.

It would be nice to compute the Atkinson and GE indices using the formula given above using the Lorenz curve. Derive the corresponding densities.

10 Exercises

10.1 Empirics

Using the previous FES data set, the software `R` and the package `ineq`, compare the empirical Lorenz curve to those obtained for the Pareto and Log-normal. Say which distribution would fit the best. Redo the same exercise limiting the data to high incomes.

10.2 Gini coefficient

We have seen that the Gini coefficient could be seen as the covariance between a variable and its rank, namely:

$$G = \frac{2}{\mu}\text{Cov}(y, F(y)).$$

As $\text{Cov}(y, F(y)) = \int y(F(y) - 1/2)dF(y)$, use integration by parts to show that

$$\text{Cov}(y, F(y)) = \frac{1}{2}\int F(x)[1 - F(x)]dx,$$

and give the corresponding form of the Gini. Give the value of F for which the Gini is maximum. What can you deduce of this result as a property of the Gini index?

10.3 LogNormal

Compute the value of the Generalized Entropy index for $\theta = 0$ and $\theta = 1$. Comment your result. Does it hold in the general case of a general distribution. Do the same calculation for the Pareto density.

10.4 Uniform

The uniform density between 0 and x_m is sometimes used in theoretical economic paper to describe the income distribution. It writes:

$$f(x) = \frac{1}{x_m}\mathbb{1}(x \leq x_m)$$

This density has strange properties that we shall now explore.

1. Compute the mean and the variance

2. Calculate the expression of the cumulative distribution

3. Using the inverse of this cumulative distribution compute the expression of the Lorenz curve

$$L(p) = \frac{1}{\mu}\int_0^p F^{-1}(t)dt$$

4. Compare $L(p)$ with that of the Pareto distribution
5. Compute the Gini index corresponding to the uniform distribution using

$$G = 1 - 2\int_0^1 L(p)dp$$

6. Verify that you obtain the same result using

$$G = 1 - \frac{1}{\mu}\int_0^{x_m} [1 - F(t)]^2 dt$$

10.5 Singh-Maddala*

Find an example where two Lorenz curves associated to the Singh-Maddala distribution intersect. Use the graphs produced by R for this. Mind that the parametrization adopted in R for the function Lc.singh is awkward. Use the function provided in the text.

10.6 Logistic*

The logistic density is very close to the normal density, but it has nicer properties, such as in particular an analytical cumulative distribution. We have

$$f(x) = \frac{e^{-(x-\mu)/s}}{s(1 + e^{-(x-\mu)/s})^2}$$

$$F(x) = \frac{1}{1 + e^{-(x-\mu)/s}}$$

with mean μ and variance $\pi^2 s^2/3$. Find the log logistic distribution using the adequate transformation. Find the Gini coefficient. This is the Fisk distribution.

10.7 Weibull*

Show that when $a_3 \to \infty$ in the Singh-Maddala distribution, we get the Weibull.

References

Arnold, B. C. (2008). Pareto and generalized Pareto distributions. In Chotikapanich, D., editor, *Modeling Income Distribuions and Lorenz Curves*, volume 5 of *Economic Studies in Equality, Social Exclusion and Well-Being*, chapter 7, pages 119–145. Springer, New-York.

Atkinson, A. (1970). The measurement of inequality. *Journal of Economic Theory*, 2:244–263.

Champernowne, D. (1953). A model of income distribution. *Economic Journal*, 63:318–351.

Cowell, F. (1995). *Measuring Inequality*. LSE Handbooks on Economics Series. Prentice Hall, London.

Davidson, R. (2009). Reliable inference for the gini index. *Journal of Econometrics*, 150:30–40.

Davies, J. B., Green, D. A., and Paarsch, H. J. (1998). Economic statistics and social welfare comparisons: A review. In Ullah, A. and Giles, D. E. A., editors, *Handbook of Applied Economic Statistics*, volume 155 of *Statistics: Textbooks and Monographs*, pages 1–38. Dekker, New York, Basel and Hong-Kong.

Deaton, A. (1997). *The Analysis of Household Surveys*. The John Hopkins University Press, Baltimore and London.

Donaldson, D. and Weymark, J. (1980). A single-parameter generalization of the gini indices of inequality. *Journal of Economic Theory*, 22(1):67–86.

Duclos, J.-Y. and Araar, A. (2006). *Poverty and Equity: Measurement, Policy and Estimation with DAD*. Springer, Newy-York.

Fisk, P. (1961). The graduation of income distributions. *Econometrica*, 29:171–185.

Gastwirth, J. L. (1971). A general definition of the lorenz curve. *Econometrica*, 39(6):1037–1039.

Gibrat, R. (1930). Une loi des réparations économiques: l'effet proportionnel. *Bulletin de Statistique Général, France*, 19:469.

Giles, D. E. A. (2004). Calculating a standard error for the gini coefficient: Some further results. *Oxford Bulletin of Economics and Statistics*, 66(3):425–433.

Jenkins, S. P. (2017). Pareto models, top incomes and recent trends in uk income inequality. *Economica*, 84(334):261–289.

Kakwani, N. (1984). Welfare ranking of income distributions. In Basmann, R. and Rhodes, G., editors, *Advances in Econometrics*, volume 3, pages 191–213. JAI Press.

Krause, M. (2014). Parametric Lorenz curves and the modality of the income density function. *Review of Income and Wealth*, 60(4):905–929.

Lubrano, M. and Protopopescu, C. (2004). Density inference for ranking european research systems in the field of economics. *Journal of Econometrics*, 123(2):345–369.

McDonald, J. (1984). Some generalised functions for the size distribution of income. *Econometrica*, 52(3):647–663.

McDonald, J. B. and Ranson, M. R. (1979). Functional forms, estimation techniques and the distribution of income. *Econometrica*, 47(6):1513–1525.

Mitzenmacher, M. (2004). A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2):226–251.

Moothathua, T. S. K. (1989). On unbiased estimation of Gini index and Yntema-Pietra index of lognormal distribution and their variances. *Communications in Statistics - Theory and Methods*, 18: 2,(2):661–672.

Piketty, T. (2000). *Les hauts revenus en France au 20ième siècle: Inégalités et redistributions, 1901-1998*. Grasset, Paris.

Piketty, T., Yang, L., and Zucman, G. (2017). Capital accumulation, private property and rising inequality in China, 1978-2015. Technical Report Working Paper 23368, NBER.

Reed, W. J. (2003). The Pareto law of incomes: an explanation and an extension. *Physica*, A 319:469–486.

Reed, W. J. and Jorgensen, M. (2004). The double pareto-lognormal distribution: A new parametric model for size distributions. *Communications in Statistics - Theory and Methods*, 33(8):1733–1753.

Sarabia, J. M. (2008). Parametric lorenz curves: Models and applications. In Chotikapanich, D., editor, *Modeling Income Distribuions and Lorenz Curves*, volume 5 of *Economic Studies in Equality, Social Exclusion and Well-Being*, chapter 9, pages 167–190. Springer, New-York.

Sen, A. (1976). Poverty: an ordinal approach to measurement. *Econometrica*, 44(2):219–231.

Sen, A. K. (1973). *On Economic Inequality*. Clarendon Press, Oxford.

Shorrocks, A. F. (1983). Ranking income distributions. *Economica*, 50(197):3–17.

Singh, S. and Maddala, G. (1976). A function for the size distribution of incomes. *Econometrica*, 44:963–970.

Xu, K. (2003). How has the literature on gini's index evolved in the past 80 years? Economics working paper, Dalhousie University. Available at SSRN: http://ssrn.com/abstract=423200 or doi:10.2139/ssrn.423200.

Yitzhaki, S. (1998). More than a dozen alternative ways of spelling Gini. *Research in Economic Inequality*, 8:13–30.