Instituto Tecnológico Autónomo de México
Maestría en Finanzas
Economía Financiera (Eco-44105), 2015
The Discrete Lorenz Curve

Ricard Torres

Contents

1 Definition for distributions with finite support

Let us take the viewpoint of a tax agency that collects data about the income of individuals in its jurisdiction. Suppose there are m distinct income levels, which we number in increasing order:

$$0 < y_1 < y_2 < \cdots < y_{m-1} < y_m.$$

(For the sake of simplification, we assume we are dealing with precise income levels, instead of ranges, which would be more common in practice.) Define also the *income vector* as $y = (y_1, y_2, \ldots, y_m)$. For each level k ($1 \leq k \leq m$), let n_k be the number of individuals who have income y_k, so the *population vector* is $n = (n_1, n_2, \ldots, n_m)$. The *aggregate income* of this society is:

$$Y := \sum_{k=1}^{m} n_k\, y_k = n_1\, y_1 + n_2\, y_2 + \cdots + n_{m-1}\, y_{m-1} + n_m\, y_m.$$

The total number of individuals is:

$$N := \sum_{k=1}^{m} n_k = n_1 + n_2 + \cdots + n_{m-1} + n_m.$$

In order to be able to compare the income distributions of different jurisdictions, or of the same jurisdiction in different periods, we are going to *normalize* the relevant variables and express them in terms of proportions.

As regards the population, define the *frequency* of individuals in each category: $f_k = n_k/N$, for $1 \leq k \leq m$. Note that frequencies have properties similar to probabilities: for all k, $f_k \geq 0$, and $\sum_{k=1}^{m} f_k = 1$. The *frequency vector* is $f = (f_1, f_2, \ldots, f_m)$. Cumulative frequencies (the distribution function) can be defined recursively:

$$F_o \equiv 0, \quad F_k = F_{k-1} + f_k, \text{ for } 1 \leq k \leq m.$$

Thus, $F_k = f_1 + f_2 + \cdots + f_{k-1} + f_k$. We have that, for all k, $0 \leq F_k \leq 1$, $F_k \leq F_{k+1}$ whenever $k < m$, and $F_m = 1$. The *cumulative frequency vector* is $F = (F_0, F_1, \ldots, F_m)$.

As regards income, consider the *share* of total income that corresponds to the aggregate income of all individuals in each category: $s_k = (n_k\, y_k)/Y$, for $1 \leq k \leq m$. Note that shares have properties similar to probabilities: for all k, $s_k \geq 0$, and $\sum_{k=1}^{m} s_k = 1$. The *shares vector* is $s = (s_1, s_2, \ldots, s_m)$. We also define for this case cumulative shares:

$$S_o \equiv 0, \quad S_k = S_{k-1} + s_k, \text{ for } 1 \leq k \leq m.$$

Thus, $S_k = s_1 + s_2 + \cdots + s_{k-1} + s_k$. We have that, for all k, $0 \leq S_k \leq 1$, $S_k \leq S_{k+1}$ whenever $k < m$, and $S_m = 1$. The *cumulative shares vector* is $S = (S_0, S_1, \ldots, S_m)$.

Let us represent in a Cartesian diagram the cumulative distribution of the population, F, on the abscissa axis, and the cumulative distribution of shares, S, on the ordinate one. Our data consist of a number of points: (F_0, S_0), (F_1, S_1), (F_2, S_2), ..., (F_m, S_m). The ***Lorenz Curve*** is the result of joining consecutive points by means of line segments, that is, of *linearly interpolating* the successive sample points. By construction, the curve always begins at $(F_0, S_o) = (0, 0)$ and ends at $(F_m, S_m) = (1, 1)$.

Consider a simple example of a given income distribution A. Let $m^A = 5$, $y^A = (10, 20, 40, 50, 80)$, and $n^A = (40, 40, 80, 20, 20)$. Let us represent the data in a table:

y^A	0	10	20	40	50	80
n^A	0	40	40	80	20	20
f^A	0	4/20	4/20	8/20	2/20	2/20
s^A	0	4/70	8/70	32/70	10/70	16/70
F^A	0	4/20	8/20	16/20	18/20	20/20
S^A	0	4/70	12/70	44/70	54/70	70/70

Thus, the Lorenz curve is the result of joining the points of the last two rows. The result for the present distribution is shown in Figure 1.

By looking at the diagram we can understand the idea of constructing the curve by means of linear interpolation. Suppose we wish to know what is the share of income that corresponds to the 30% fraction of individuals with less income. The points of the cumulative curves show that 20% of the population with less income (40 individuals) have income $y_1 = 10$, which corresponds to a 4/70 fraction of aggregate income, and that the lower 40% has a 12/70 fraction of the aggregate. If we want to consider the 30% fraction of the population with less income, to the 40 individuals with income $y_1 = 10$, that account for 20% of the total, we have to add half of the 40 individuals that belong to the income category $y_2 = 20$, which account for another 20%. Hence, this 30% of the population will have total income $40 \times 10 + 20 \times 20 = 800$, which amounts to $800/7000 = 8/70$ of the total. Since $F = 0.3 = 0.5 * 0.2 + 0.5 * 0.4$, the linear interpolation gives to the Lorenz curve at 0.3 the value $L(0.3) = 0.5 * (4/70) + 0.5 * (12/70) = 8/70$.

In general, if we let F satisfy $F_{k-1} \leq F \leq F_k$; then F is a convex combination of the two extremes, ie, there is $0 \leq \lambda \leq 1$ such that $F = (1 - \lambda)\, F_{k-1} + \lambda\, F_k$. The linear interpolation assigns to F the value $L(F) = (1 - \lambda)\, S_{k-1} + \lambda\, S_k$. By substituting λ we obtain:

$$L(F) = \frac{F_k - F}{F_k - F_{k-1}}\, S_{k-1} + \frac{F - F_{k-1}}{F_k - F_{k-1}}\, S_k = \frac{F_k S_{k-1} - F_{k-1} S_k}{F_k - F_{k-1}} + \frac{S_k - S_{k-1}}{F_k - F_{k-1}}\, F = \frac{F_k S_{k-1} - F_{k-1} S_k}{f_k} + \frac{s_k}{f_k}\, F.$$

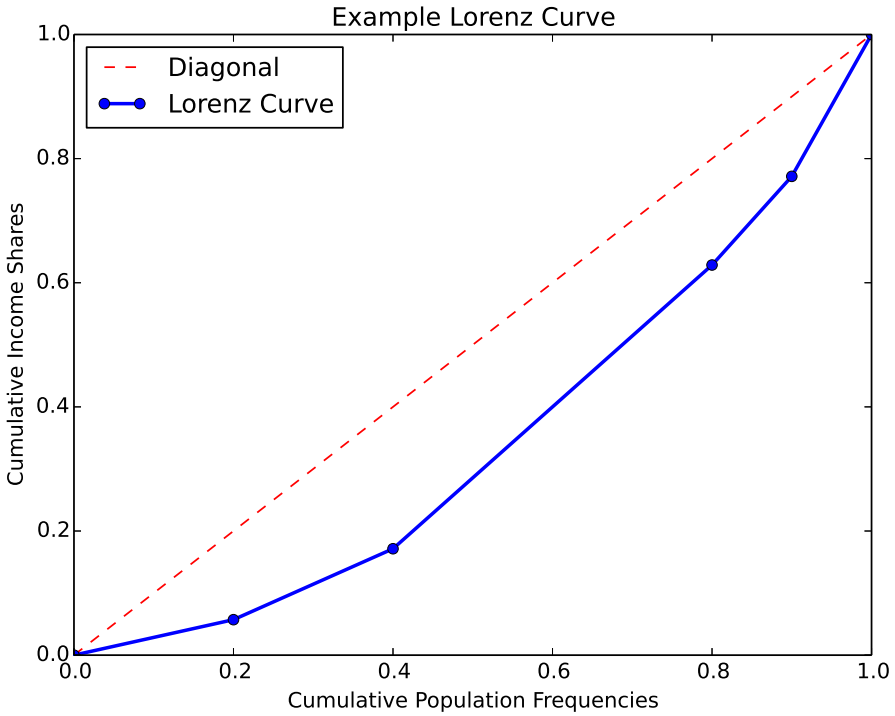

Figure 1: Lorenz curve of a discrete distribution.

So the Lorenz curve consists of m line segments indexed by k, for $1 \leq k \leq m$, in which the kth segment is given by the above expression. In particular, the slope of this line segment is s_k/f_k.

The Lorenz curve of a discrete distribution has the following properties:

1. If $0 < F_k < 1$, then $S_k < F_k$. Geometrically, the curve lies below the diagonal, and coincides with it if, and only if, $m = 1$.

2. The slopes of the successive linear segments are strictly increasing. As a result, the Lorenz curve is a convex function.

3. Two Lorenz curves, L^A and L^B, are equal if, and only if, there exist constants $p > 0$ and $q > 0$ such that $y^B = p\, y^A$ and $n^B = q\, n^A$ (provided we let the ns have components that are not integers).

Note also that, by construction, the Lorenz ordering is *anonymous*, in the sense that it treats the individuals symmetrically, because it only depends on income categories and the number of people in them, not on the identity of those individuals.

Property 3 is sometimes broken into two categories. First, let $p = 1$ and $q > 0$ vary: that is, there is a proportional increase or decrease in the population (if q is integer, people refer to this property as *replication invariance*). Second, let $q = 1$ and $p > 0$ vary: this is referred to as *scale invariance*.

Property 1: $0 < F_k < 1$ implies $S_k < F_k$.

It will be convenient to create new notation. Suppose initially that $m \geq 2$. For $1 \leq k \leq m-1$, define the *partial sums*:

$$N_k = \sum_{i=1}^{k} n_i, \quad Y_k = \sum_{i=1}^{k} n_i\, y_i.$$

Define also the *residual sums*:

$$N_{-k} = \sum_{i=k+1}^{m} n_i, \quad Y_{-k} = \sum_{i=k+1}^{m} n_i\, y_i.$$

Thus, we have that, for each k, $1 \leq k \leq m-1$: $N = N_k + N_{-k}$, and $Y = Y_k + Y_{-k}$. Substituting:

$$S_k = \sum_{i=1}^{k} \frac{n_i\, y_i}{Y} = \frac{Y_k}{Y} = \frac{Y_k}{Y_k + Y_{-k}}, \quad F_k = \sum_{i=1}^{k} \frac{n_i}{N} = \frac{N_k}{N} = \frac{N_k}{N_k + N_{-k}}.$$

Hence:

$$S_k \leq F_k \Leftrightarrow \frac{Y_k}{Y_k + Y_{-k}} \leq \frac{N_k}{N_k + N_{-k}} \Leftrightarrow Y_k\, N_{-k} \leq N_k\, Y_{-k} \Leftrightarrow \frac{Y_k}{N_k} \leq \frac{Y_{-k}}{N_{-k}}.$$

Now:

$$\frac{Y_k}{N_k} = \sum_{i=1}^{k} \frac{n_i}{N_k}\, y_i, \text{ with } \sum_{i=1}^{k} \frac{n_i}{N_k} = 1, \text{ and } \frac{Y_{-k}}{N_{-k}} = \sum_{i=k+1}^{m} \frac{n_i}{N_{-k}}\, y_i, \text{ with } \sum_{i=k+1}^{m} \frac{n_i}{N_{-k}} = 1.$$

This means that Y_k/N_k is a convex combination of $\{y_1, y_2, \ldots, y_k\}$, which implies $y_1 \leq Y_k/N_k \leq y_k$. Analogously, Y_{-k}/N_{-k} is a convex combination of $\{y_{k+1}, y_{k+2}, \ldots, y_m\}$, which implies $y_{k+1} \leq Y_{-k}/N_{-k} \leq y_m$. In particular: $Y_k/N_k \leq y_k < y_{k+1} \leq Y_{-k}/N_{-k}$. Therefore, we will have $S_k < F_k$ whenever $k \leq m-1$. On the other hand, we always have $S_0 = 0 = F_0$ and $S_m = 1 = F_m$. The Lorenz curve will coincide with the diagonal if, and only if, there is a unique income category (ie, there is perfect income equality). If there is more than one income category, the Lorenz curve will lie *strictly* below the diagonal at all interior points.

Property 2: The slopes of the successive line segments are strictly increasing.

As we have seen above, the slope of the linear segment between (F_{k-1}, S_{k-1}) and (F_k, S_k) is:

$$\frac{s_k}{f_k} = \frac{n_k\, y_k/Y}{n_k/N} = \frac{N\, y_k}{Y}.$$

Therefore, the slopes of successive segments are proportional to the y_k, which are strictly increasing.

Property 3: Let $p > 0$ and $q > 0$ be arbitrary numbers. Let $y^B = p\, y^A$ and $n^B = q\, n^A$. Then it is easy to see that, for each i, $f_i^A = f_i^B$ and $s_i^A = s_i^B$, so the two Lorenz curves satisfy $L^A(F) = L^B(F)$ for each F.

On the other hand, suppose that $L^A(F) = L^B(F)$ for each F. In particular, Property 2 implies that $F^A = F^B$ (as vectors). This in turn implies, on the one hand, that $f^A = f^B$, and on the other that $S^A = S^B$, so that $s^A = s^B$. Given these equalities, define

$$p = \frac{\sum_{i=1}^{m} f_i^B y_i^B}{\sum_{i=1}^{m} f_i^A y_i^A}, \quad q = \frac{\sum_{i=1}^{m} n_i^B}{\sum_{i=1}^{m} n_i^A}.$$

Then we can check that $y^B = p\, y^A$ and $n^B = q\, n^A$.

2 The Lorenz ordering

Property 1 above shows that, when you depart from a perfectly equitable distribution, the Lorenz curve moves away from the diagonal. The ***Lorenz ordering*** among income distributions is defined

as follows: we say that income distribution A is *more unequal* than income distribution B (or that B *Lorenz-dominates* A) if, for all $F \in [0,1]$, $L_A(F) \leq L_B(F)$.

In order to understand how to check for Lorenz domination in the case of distributions with finite support, we just need to take into account an elementary fact about linear interpolation: given four points in $\mathbb{R}^2$, (x_1, y_1), (x_1, z_1), (x_2, y_2), and (x_2, z_2), satisfying $x_1 < x_2$, let f and g denote the respective linear interpolations, $f : [x_1, x_2] \to [y_1, y_2]$ and $g : [x_1, x_2] \to [z_1, z_2]$, then we have that, for all x such that $x_1 < x < x_2$, $f(x) \leq g(x)$ if, and only if, $y_1 \leq z_1$ and $y_2 \leq z_2$; moreover, the inequality is strict if, and only if, $\max\{z_1 - y_1, z_2 - y_2\} > 0$. The proof can be easily derived from our above explanation of linear interpolation, so we skip it here.

Let now A and B be two income distributions, that is, there are $m^A \geq 1$ and $m^B \geq 1$, and collections of points in $\mathbb{R}^2$, $\{(F_1^A, S_1^A), (F_2^A, S_2^A), \ldots, (F_{m^A}^A, S_{m^A}^A)\}$ and $\{(F_1^B, S_1^B), (F_2^B, S_2^B), \ldots, (F_{m^B}^B, S_{m^B}^B)\}$ such that the corresponding Lorenz curves L^A and L^B are the result of linearly interpolating between consecutive points in each case. Let $\mathscr{F}$ be the union of the abscissa coordinates:

$$\mathscr{F} = \{(F_1^A, F_2^A, \ldots, F_{m^A}^A, F_1^B, F_2^B, \ldots, F_{m^B}^B\}$$

Then, the fact about linear interpolation mentioned above implies that $L^A(F) \leq L^B(F)$ for all $F \in [0,1]$ if, and only if, $L^A(F) \leq L^B(F)$ for all $F \in \mathscr{F}$. Additionally, there is $F \in (0,1)$ such that $L^A(F) < L^B(F)$ if, and only if, there is $F \in \mathscr{F}$ such that $L^A(F) < L^B(F)$. In other words, in order to check for Lorenz domination we can restrict ourselves to the points of $\mathscr{F}$ and their images under the two Lorenz curves. This comparison, as well as the computation of Lorenz curves, is easily implementable in programs like `Matlab`, or the freely available `Octave`, or `Julia`.

2.1 Mean preserving spreads

We have seen that the point of departure of the Lorenz curve construction is to work in terms of proportions so as to be able to *abstract from scale* and compare different distributions. Unfortunately, just working with proportions is not enough in order to get rid of all scale effects. It is very easy to illustrate why with a few variations of our previous example.

Let us generate a new distribution from the old one that, at least intuitively, should unambiguously lead to less inequality. With this aim, we will transfer half of the people in the lowest income category to the second-lowest one. This will result in $n = (20, 60, 80, 20, 20)$. The lowest category represents now $20/200 = 10\%$ of the individuals. If the total wealth were unchanged, this would lead to exactly the same share of total wealth that the lowest 10% of the population has under the previous distribution. However, there is a *scale effect*: by transferring people to a superior income category, the total wealth has increased, which implies that now the share of the lowest 10% of the population has decreased, so that, at this point, the new Lorenz curve lies *below* the previous one (even though for most values of F the new Lorenz curve is above the original one). The same phenomenon is repeated if we transfer people from the highest to the second-highest category, or between intermediate categories: the scale effects prevent the new Lorenz curve to dominate the previous one at all population shares F.

In order to avoid this type of problem and obtain a new distribution which dominates the original one according to the Lorenz order, we may shift individuals from extreme categories towards less extreme ones in such a way that *the total wealth remains unchanged*. In our example, this happens if we move to the distribution $n^B = (30, 40, 90, 30, 10)$, because moving 10 individuals from y_1 to y_3 increases the wealth by $10 \times 30 = 300$, and moving another 10 individuals from y_5 to y_4 decreases the wealth by the same amount. This results in unambiguous Lorenz domination of the original curve by the new one. One can check that this distribution gives rise to the original one via a *mean preserving spread* by comparing the areas under the respective distribution functions. We can also

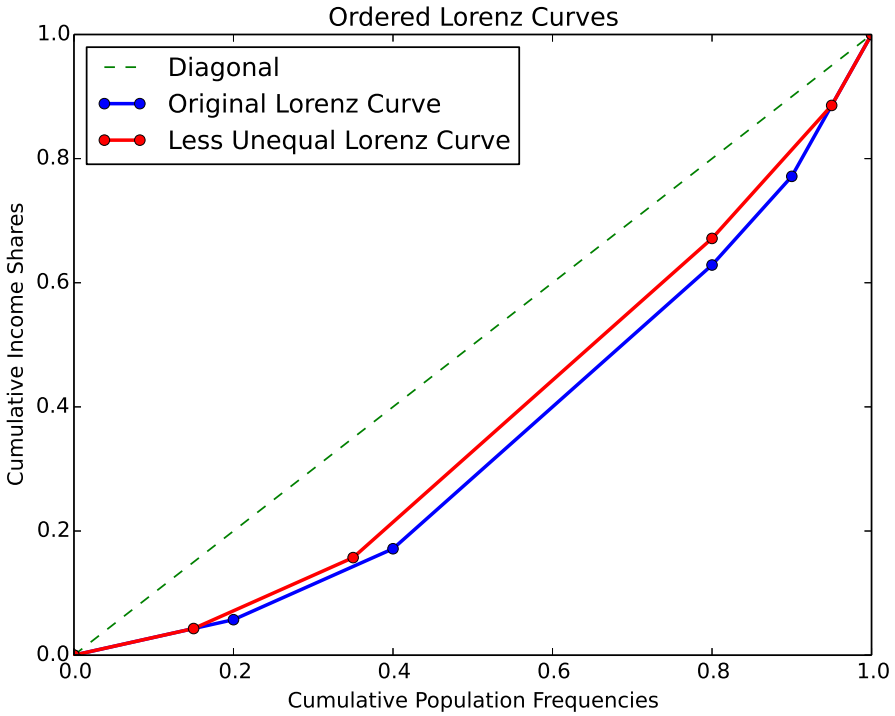

Figure 2: Lorenz domination between discrete distributions.

directly verify that A can be obtained from B via the following mean-preserving spread: out of the $n_3^B = 90$ individuals in category $y_3 = 40$, transfer 10 to $y_1 = 10$, leave 65 in $y_3 = 40$, transfer 10 to $y_4 = 50$, and transfer 5 to $y_5 = 80$; out of the $n_4^B = 30$ individuals in category $y_4 = 50$, transfer 15 to $y_3 = 40$, leave 10 in $y_4 = 50$, and transfer 5 to $y_5 = 80$. In general, a mean-preserving spread will always result in Lorenz domination. Figure 2 shows the resulting ordered Lorenz curves.

Property 3 above shows that Lorenz domination can occur without equal means. Just take one distribution that is a mean preserving spread of another one, and then rescale the income and/or population vectors.

2.2 Pigou-Dalton transfers

One step-by-step way to implement mean preserving spreads is by using *Pigou-Dalton transfers*.

A *Pigou-Dalton progressive transfer* is one in which money is transferred from a higher-income individual to a lower-income one without altering the overall ranking. Analogously, a *Pigou-Dalton regressive transfer* is one in which money is transferred from a lower-income individual to a higher-income one without altering the overall ranking. The idea is that a sequence of progressive transfers should result in a less unequal distribution, and a sequence of regressive transfers in a more unequal one.

Formally, given a distribution (y^A, n^A), and given two income levels y_i^A and y_j^A, with $i < j$, let t satisfy $0 < t \leq \min\{y_{i+1}^A - y_i^A, y_j^A - y_{j-1}^A\}$ (except if $j = i + 1$, in which case $2t \leq y_{i+1} - y_i$). Suppose the distribution (y^B, n^B) is obtained from (y^A, n^A) by singling out an individual r (which stands for "receiver") in category y_i^A and another individual d (which stands for "donor") in category y_j^A, and transferring an amount t from d to r, and letting everything else as in the original distribution. Then

we say that the move from (y^A, n^A) to (y^B, n^B) is a *simple Pigou-Dalton progressive transfer*.

Consider again a given distribution (y^A, n^A), and suppose $i < j$, but assume now that $0 < t \leq \min\{y_i^A - y_{i-1}^A, y_{j+1}^A - y_j^A\}$ (with the convention $y_0^A \equiv 0$ and $y_{m^A+1}^A \equiv \infty$). Suppose that the distribution (y^B, n^B) is obtained from (y^A, n^A) by singling out an individual d (which stands for "donor") in category y_i^A and another individual r (which stands for "receiver") in category y_j^A, and transferring an amount t from d to r, and letting everything else as in the original distribution. Then we say that the move from (y^A, n^A) to (y^B, n^B) is a *simple Pigou-Dalton regressive transfer*.

It is clear that (y^B, n^B) can be obtained from (y^A, n^A) by a sequence of progressive transfers if, and only if, the latter can be obtained from the former by a sequence of regressive transfers. The interesting result[1] is that, (y^B, n^B) dominates in the Lorenz-curve sense (y^A, n^A) if, and only if, it can be obtained from it by a sequence of progressive transfers (after normalizing the two series).

By construction, a Pigou-Dalton transfer (and, as a consequence, a sequence of those) does not alter the aggregate income. In a *regressive* transfer, in which income is transferred from a lower income person to a higher income one, the former moves downward and the latter upward in the income distribution, so the result is more spread: it is therefore a mean preserving spread. When the transfer is progressive, the movement goes in the opposite direction, so the original distribution is a mean preserving spread of the resulting one.

3 The Lorenz curve in terms of the generalized inverse for general random variables

In general, let I be the set of all individuals that receive income. This set can be ordered in an arbitrary way (eg, alphabetically) or not at all. If N is the total number of individuals, each individual has weight (ie, probability) $1/N$. Then we can view the income distribution as a random variable defined on this sample space, $X : I \to \mathbb{R}$, so that $X(i)$ denotes the income of individual $i \in I$. Its distribution function is:

$$F(y) = \mathbb{P}\{i \in I : X(i) \leq y\} = \frac{1}{N} \#\{i \in I : X(i) \leq y\}.$$

Here, think of I as the set of names of all individuals, ordered alphabetically. In order to define the Lorenz curve, we would like to order the individuals by income levels, and this is accomplished by considering the generalized inverse $Y(t)$ of $F(y)$. In this case, $Y(t)$ indicates the income level of any individual for whom a fraction t of individuals have a smaller income than him or her. In this sense, by resorting to the generalized inverse we have ranked the individuals by income level.

What follows is equally valid for discrete or continuous distributions (or a mixture of them). Let $\mu = \mathbb{E}X = \mathbb{E}Y = \int_0^1 Y(s)\,ds$ be the average income. The Lorenz curve is the function $L : (0,1) \to [0,1]$ defined as the share of income received by the fraction t of individuals with lower income:

$$L(t) = \frac{1}{\mu} \int_0^t Y(s)\,ds = \frac{\int_0^t Y(s)\,ds}{\int_0^1 Y(s)\,ds}.$$

Additionally, we can (continously) extend it to the extremes of the interval by setting $L(0) = 0$ and $L(1) = 1$. This expression is valid no matter whether the distribution is discrete, continuous, or a combination of the two.

When X_A and X_B are random variables that represent income distributions, one can show that, whenever $\mu_A = \mu_B$, the fact that A Lorenz-dominates B is equivalent to second order stochastic dominance of X_B by X_A. This result was first shown by Atkinson[2] for continuous distributions.

[1] Theorem 2.1 in Fields and Fei, "On Inequality Comparisons," *Econometrica*, vol. 46, n. 2, 1978.

[2] "On the Measurement of Inequality," *Journal of Economic Theory*, 2, pp. 244-263, 1970.