# Income and Allied Size Distributions

M Pal

# Syllabus

- **Analysis of income and allied size distributions:**

  Pareto and log-normal distributions, genesis, specification and estimation, Lorenz curve, Gini coefficient.

  Income and consumer expenditure distribution, poverty.

# Introduction

- The size distributions of certain economic and socioeconomic variables—incomes, wealth, firms, plants, cities, etc.—display remarkably regular patterns. These patterns, or distribution laws, are usually skew, the most important being the Pareto and the lognormal.
- Some disagreement about the patterns observed still exists.
- The study of size distributions is concerned with explaining why the observed patterns exist and persist.
- Various particle size distributions follow Lognormal distribution.

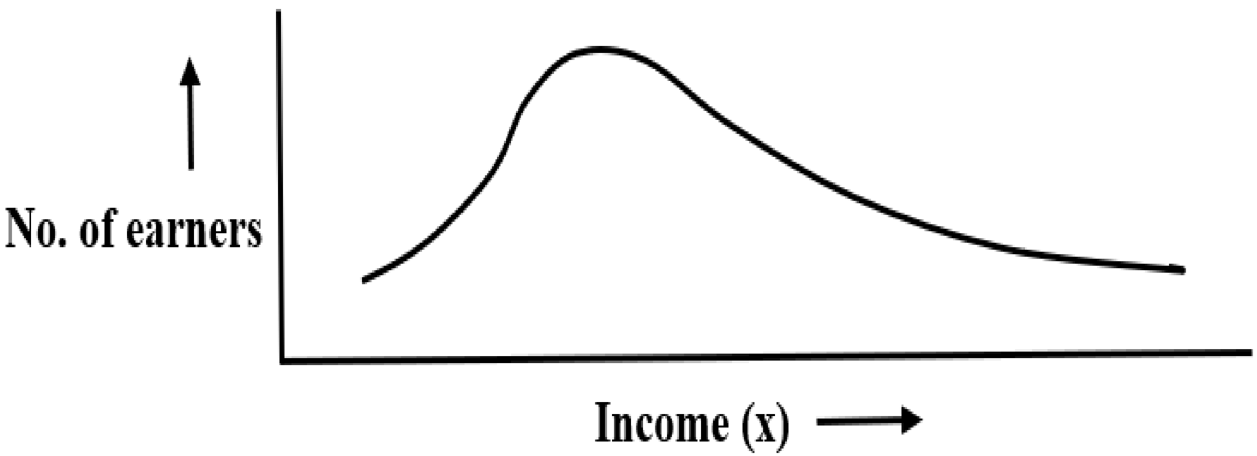
# Size Distribution

- By size distribution we usually mean a nonnegative valued continuous random variable with finite expectation.
- We may also want the shares of different groups of individuals to be meaningful in the sense that we can conceive of redistributing the aggregate amount in different ways among the members of the community. E.g., if $x_1, x_2, \dots x_n$ are incomes of n persons then we should be able to talk meaningfully the share $\frac{\sum_1^m x_i}{\sum_1^n x_i}$, provided $x_1 < x_2 < \cdots < x_n$ and m < n.
- Also, we should be able to talk about transfer of some amount of money from one person to another person.
- To make it clearer let us take the distribution of heights, it obeys all the properties of a size distribution except the shares and redistribution/ transfer.
- However, all these constraints of a size distribution are not strictly adhered to in many cases.

# Examples of Size Distributions

- **Examples of Size Distributions:**

1. Distribution of earners by size of income




2. Distribution of all persons by per capita household consumer expenditure

3. Distribution of households by land owned/possessed/operated or in other words distribution of land holdings

# Examples of Size Distributions

4. Distribution of households by wealth owned or in other words wealth distribution

5. Distribution of firms by size

- Giant firms control the market. We study concentration in business and industry using such size distributions. By analysing the size distribution of firm, we study the concentration of economic power and prevalence of monopolistic/oligopolistic practices.

# Sources of Income Data

- **Sources of Income Data**

**1.** ***Income Tax Statistics***

- Income Tax Statistics give only truncated data. It covers only the upper tail of the distribution. It is also highly unreliable especially for those who are self-employed.

**2.** ***Specialized Sample Surveys***

- National Sample Survey Organization (NSSO) regularly (every year) collects data on Household Consumer Expenditure. Besides, NSSO collects data on Household Consumer Expenditure every five-year interval on a higher scale taking larger sample size. These are called quinquennial surveys.

- NSSO usually collects data on monthly consumer expenditures of household because these data are much more reliable than income data. Usually, expenditures are overstated, and incomes are understated.

# Gini Concentration Curve or Lorenz Curve

- **Gini Concentration Curve or Lorenz Curve**
- Lorenz curve is the most well-known technique of studying inequality.
- Suppose X is a size distribution with pdf f(x), x ≥ 0 and cumulative distribution function F(x). We have

$$F(x) = \int_0^x f(t)dt. \qquad (I)$$

- F(x) is empirically the proportion of earners with income ≤ x.
- We define

$$F_1(x) = \frac{\int_0^x tf(t)dt}{\int_0^\infty tf(t)dt}. \qquad (II)$$

- $F_1(x)$ is the proportion of aggregate income enjoyed by earners with income ≤ x.

# Lorenz Curve (Continued)

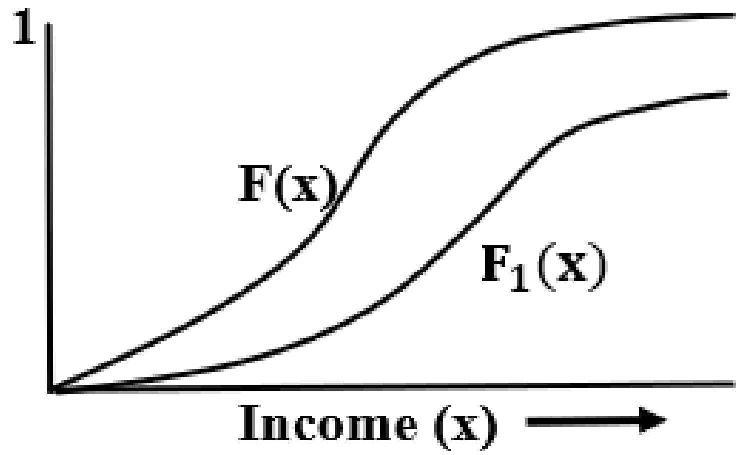
- To visualize why it is so let us write $F_1(x)$ as

$$F_1(x) = \frac{\int_0^x Ntf(t)dt}{\int_0^\infty Ntf(t)dt}.$$

- Nf(t)dt is the number of earners with income in between t and t+dt and the total income enjoyed by these earners is Ntf(t)dt.

# Lorenz Curve (Continued)

- The outline of the graph of F(x) and $F_1(x)$ is given below.




- $F_1$ has all the property of a distribution function
- $F_1(0) = 0$, $F_1(\infty) = 1$ and it is an increasing function of x. Thus, $F_1$ is called the first moment distribution of X.

# Lorenz Curve (Continued)

- Verify that

$$\frac{dF_1(x)}{dx} = \frac{xf(x)}{\mu} > 0,$$

- where μ is the expectation of X.

$$\frac{dF(x)}{dx} = f(x) > 0.$$

- Again,

$$\frac{dF_1(x)}{dF(x)} = \frac{dF_1(x)/dx}{dF(x)/dx} = \frac{x}{\mu} > 0.$$

- It implies that $F_1$ is an increasing function of F.
- Equations (I) and (II) define the concentration curve – the graph of $F_1$ against F in the parametric form, x being the parameter. Often, we can eliminate x between them to get the explicit form $F_1$ = $F_1$(F). Graph of $F_1$ = $F_1$(F) is called the concentration curve or Lorenz Curve.

# Diagram of Lorenz Curve

- The curve goes through (0, 0) corresponding to x = 0 and through (1, 1) corresponding to x = ∞. It is customary to draw the diagonal ($F_1$ = F) and the four sides of it to complete the square.

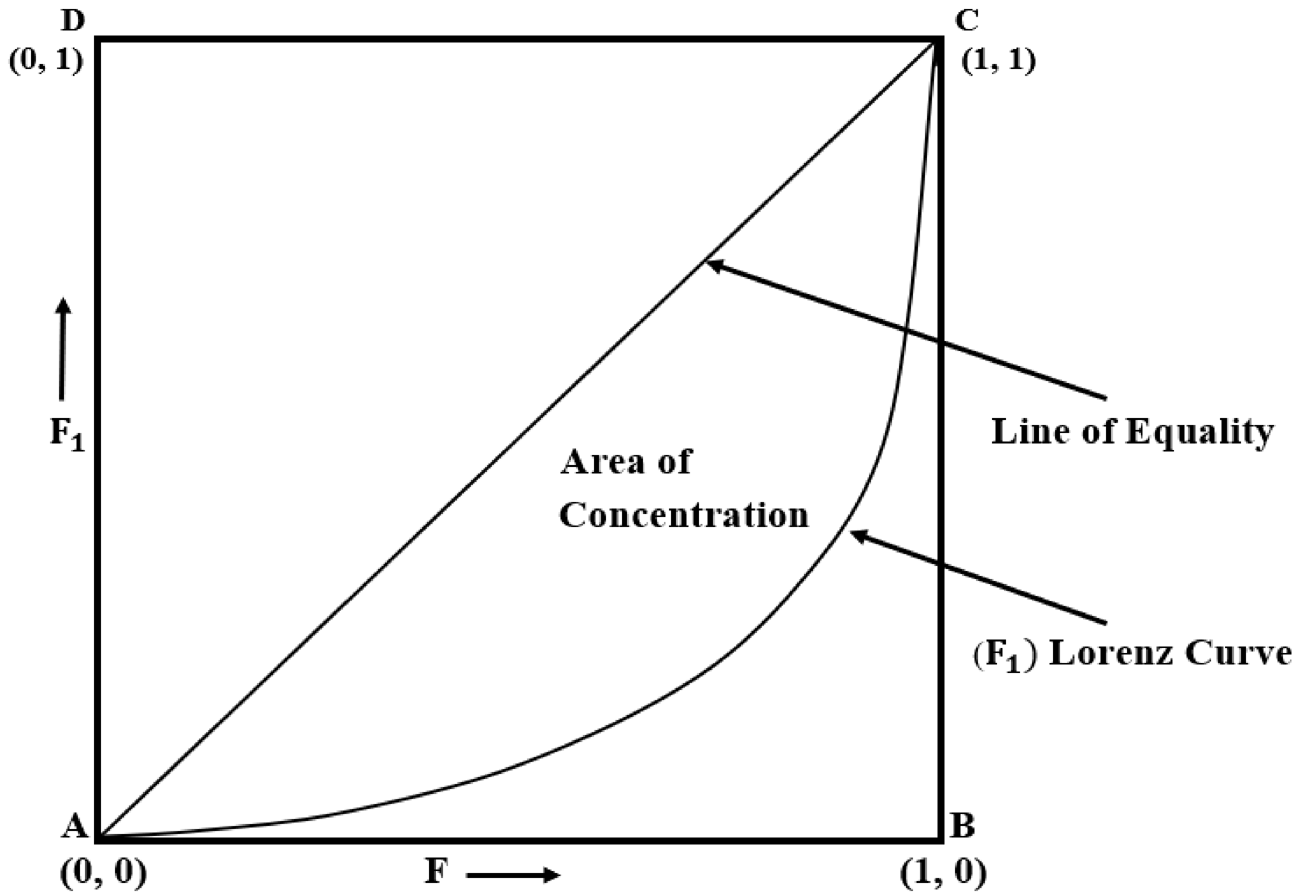
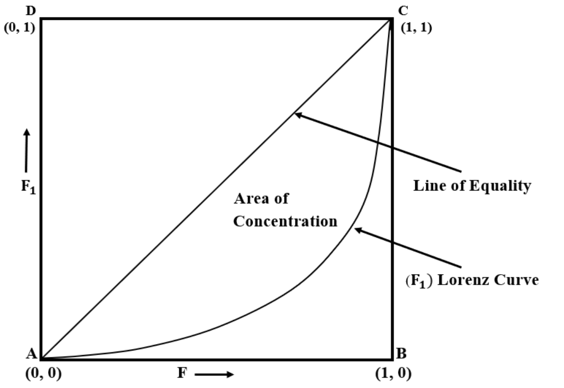



**Diagram: Lorenz Curve Shown in the Lorenz Box**

# Lorenz Curve

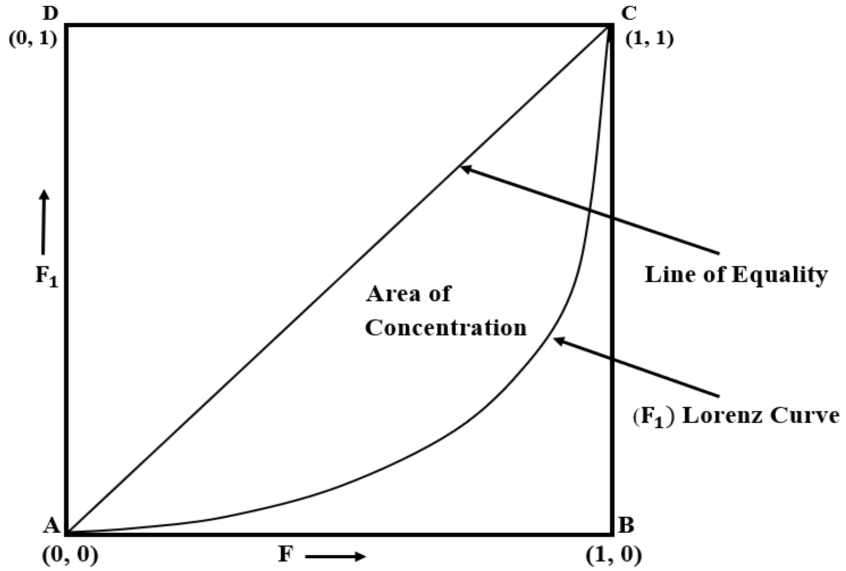



- It is easy to see that

$$\frac{d^2F_1}{dF^2} = \frac{1}{\mu f(x)} > 0.$$

- Thus, the curve is increasing and convex from below. I.e., the curve goes through somewhere below the diagonal line. The diagonal line is called the egalitarian line or the line of equality, because in the limit when all incomes become equal, the LC approaches this line.
- Lorenz Ratio (or Gini's Concentration Coefficient) = 2×(Area of Concentration)
- $F_1 = F_1$(F) is the equation of the Lorenz Curve (LC)
- $F_1$ = F is the egalitarian line or line of equal distribution.
- Area of Concentration = Area enclosed between the egalitarian line and the Lorenz Curve.

# Lorenz Curve

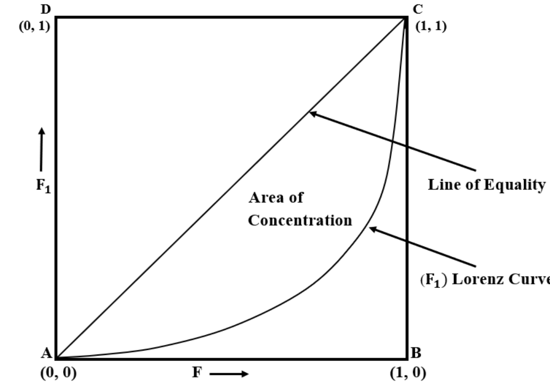



- Any point (F, $F_1$) on the curve tells us that the bottom 100F% people enjoy a share of 100$F_1$% in the aggregate income. [Hence, top 100(1-F) % people enjoy a share of 100(1-$F_1$)% in the aggregate income.]

$$F = 0.4 \text{ and } F_1 = 0.22 \Longrightarrow \text{Bottom 40\% gets 22\%.}$$

- The LC enables us to make a series of such statements about the inequality aspect of the distribution. In fact, ordinates of the curve corresponding to F = 0.1, 0.5 and 0.9 are popular measures of inequality.

# Properties of LC




- **Properties of Lorenz Curve**

1. Every LC goes through (0, 0) corresponding to x = 0 and (1, 1) corresponding to x = ∞.

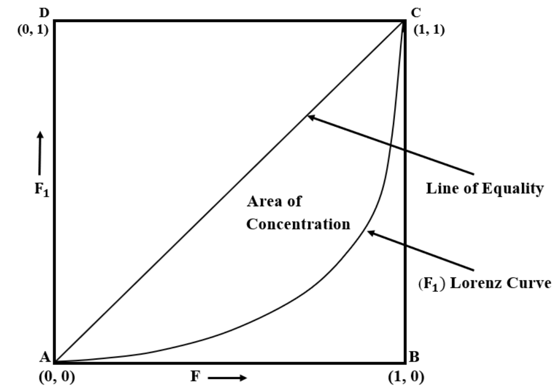
2. $F(x) = \int_0^x f(t)dt$, $F_1(x) = \int_0^x tf(t)dt$ /μ, dF(x)/dx = f(x), d$F_1$(x)/dx = xf(x)/μ and d$F_1$/dF = x/μ >0 ⟹ LC is monotonically increasing.

3. $d^2F_1/dF^2$ = 1/(μf(x)) > 0 ⟹ LC is convex (from below), i.e., LC lies below the egalitarian line. So, except when F = 0 or F = 1, $F_1 \leq$ F.

**4. Case of Perfect Equality:** $F_1$ = F, the egalitarian line, is the case of perfect equality. The LC coincides with the diagonal. In this case everybody's income is same.

**5. Case of Perfect Inequality:** In this case, all incomes are zero except the income of one person. If there are 100 persons in the population then $x_1 = x_2 = \cdots = x_{99} = 0$, and $x_{100} =$ M. This implies that $F_1$ = 0 for all F < 1 and when F = 1, $F_1$ = 1. Geometrically, the curve coincides with the sides AB and BC of the Lorenz Diagram.

# Properties of LC




6. In general, the lower the Lorenz Curve the greater the degree of inequality. For perfect inequality, the area of concentration is 1/2.

- The Lorenz Ratio or Gini's Concentration Curve is defined as
- LR = 2×(Area of Concentration) so that for perfect equality area = 0, and for perfect inequality, area = 1. Thus,

$$\mathrm{LR} = \frac{\text{Area of Concentration}}{\text{Area of } \Delta \mathrm{ABC}} = \frac{\text{Area of Concentration}}{1/2}$$

= 2×(Area of Concentration) = 2(1/2 – Area under the LC)

= 1 - 2× Area under the LC

$$= 1 - 2\int_0^1 \mathrm{F}_1 \mathrm{dF}\,.$$

# Construction of LC from Empirical Income Distribution

| Income Class $(x_{i-1} - x_i)$ | Frequency $(f_i)$ | Average income $(\bar{x}_i)$ | Relative frequency $(p_i)$ | Cum. Rel. frequency $(P_i)$ | Income share $(q_i)$ | Cumulative share $(Q_i)$ |
|---|---|---|---|---|---|---|
| $x_0 - x_1$ | $f_1$ | $\bar{x}_1$ | $p_1 = \frac{f_1}{n}$ | $P_1 = p_1$ | $q_1 = \frac{\bar{x}_1 p_1}{\bar{x}}$ | $Q_1 = q_1$ |
| $x_1 - x_2$ | $f_2$ | $\bar{x}_2$ | $p_2 = \frac{f_2}{n}$ | $P_2 = p_1 + p_2$ | $q_2 = \frac{\bar{x}_2 p_2}{\bar{x}}$ | $Q_2 = q_1 + q_2$ |
| --- | --- | --- | --- | --- | --- | --- |
| $x_{i-1} - x_i$ | $f_i$ | $\bar{x}_i$ | $p_i = \frac{f_i}{n}$ | $P_i = \sum_{j=1}^{i} p_j$ | $q_i = \frac{\bar{x}_i p_i}{\bar{x}}$ | $Q_i = \sum_{j=1}^{i} q_j$ |
| --- | --- | --- | --- | --- | --- | --- |
| $x_{K-1} - x_K$ | $f_K$ | $\bar{x}_K$ | $p_K = \frac{f_K}{n}$ | $P_K = \sum_{j=1}^{K} p_j = 1$ | $q_K = \frac{\bar{x}_K p_K}{\bar{x}}$ | $Q_K = \sum_{j=1}^{K} q_j = 1$ |
| Total | $n = \sum f_i$ | $\bar{x} = \frac{\sum \bar{x}_i f_i}{n}$ | $1 = \sum p_i$ | --- | $1 = \sum q_i$ | --- |

**Table 1: Computations towards Lorenz Curve and Lorenz Ratio using Grouped Data**

# Construction of LC and LR

$$\bar{x} = \frac{\sum \bar{x}_i f_i}{n} = \sum \bar{x}_i p_i$$

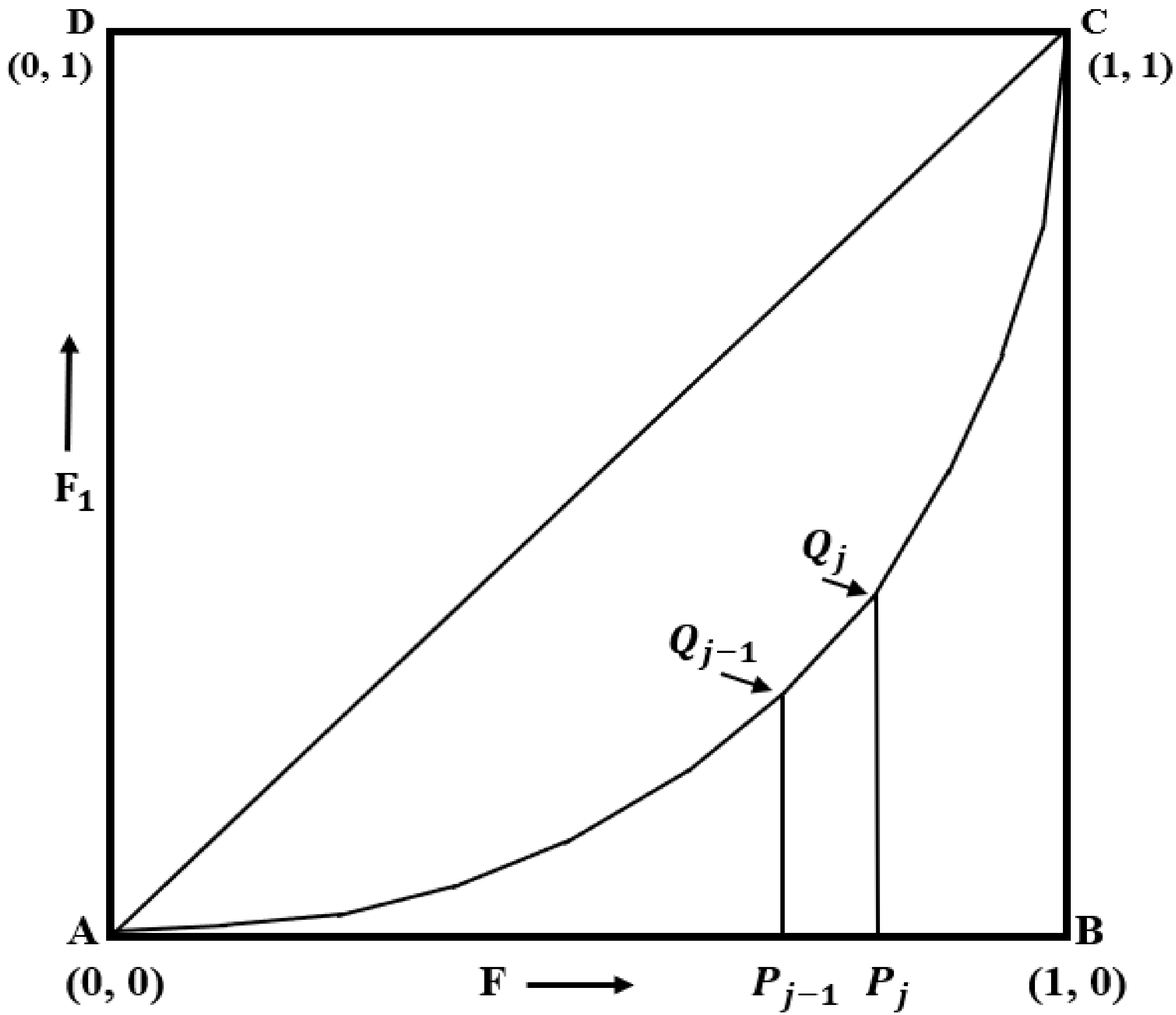
- $P_j$: Cumulative proportion of earners with income ≤ $x_j$ = F$\left(x_j\right)$

$$q_j = \frac{n_j \bar{x}_j}{\sum n_i \bar{x}_i} = \frac{n p_j \bar{x}_j}{\sum n p_i \bar{x}_i} = \frac{p_j \bar{x}_j}{\sum p_i \bar{x}_i} = \frac{p_j \bar{x}_j}{\bar{x}}$$

= share of aggregate income in the jth class.

$Q_j =$ Share (Proportion) of aggregate income of those with income ≤ $x_j$ = $F_1\left(x_j\right)$.

- In the graph paper we plot $Q_j$'s in the Y – axis against the corresponding $P_j$'s in the X – axis for j = 1, 2, …, K. I.e., we plot the pairs $\left(P_j, Q_j\right)$, for i = 0, 1, 2, …, K, taking $(P_0, Q_0)$ = (0,0) and join these points either by straight lines or by smooth curves. Straight line joining is completely objective but ignores the convexity of the curve to some extent and gives an underestimate of the Lorenz Ratio. Smooth curve joining is subjective but gives unbiased estimate of the Lorenz Ratio. The Lorenz Curve is being shown as below.


D
(0, 1)
C
(1, 1)
$F_1$
$Q_j$
$Q_{j-1}$
A
(0, 0)
F
$P_{j-1}$
$P_j$
B
(1, 0)

# Computation of LR

## Numerical Procedure

- In this procedure it is not even necessary to draw the LC. However, the numerical procedure is related to straight line joining procedure of graph.

LR = 1 - 2×(Area under the LC) = 1 - 2×(Sum of areas of K trapezia)

$$= 1 - 2\sum_{1}^{K}\left(P_j - P_{j-1}\right)\left(Q_j + Q_{j-1}\right)/2$$

$$= 1 - \sum_{1}^{K} p_j\left(Q_j + Q_{j-1}\right).$$

- Though the above expression is derived for grouped data, it can accommodate ungrouped data also just by taking $f_i = 1$ for all i and K = n. We can simplify this expression to see that it is same as $LR_1$.

# Computation of LR: Numerical Procedure (Continued)

$$\mathrm{LR} = 1 - \sum_{1}^{K} p_j\left(Q_j + Q_{j-1}\right)$$

$$= 1 - \sum_{1}^{K} \frac{1}{n^2}\left(\frac{\left(x_1 + x_2 + \cdots + x_j\right) + \left(x_1 + x_2 + \cdots + x_{j-1}\right)}{\bar{x}}\right)$$

$$= 1 - \frac{1}{n^2\bar{x}}\left(\begin{array}{c}(x_1) + (2x_1 + x_2) + (2x_1 + 2x_2 + x_3) + \\ \ldots + (2x_1 + 2x_2 + \cdots + 2x_{n-1} + x_n)\end{array}\right)$$

$$= 1 - \frac{1}{n^2\bar{x}}\left((2n-1)x_1 + (2n-3)x_2 + \cdots + (2n-2r+1)x_r + \cdots + x_n\right)$$

$$= 1 - \frac{1}{n^2\bar{x}}\sum_{1}^{n}(2n-2r+1)x_r = \mathrm{LR}_1.$$

# Graphical Procedure

- **Graphical Procedure:** Here the successive points may be joined by smooth curves. So, the estimate may be nearly unbiased, but the procedure is subjective. In this method we find the ratio of number squares inside the area of concentration to that of half of the Lorenz box.

$$\mathrm{LR} = \frac{\text{Area of concentration}}{\frac{1}{2}}$$

$$= \frac{\text{Number of squares in the area of concentration}}{1/2(\text{Number of squares in the Lorenz box}}$$

$$= \frac{\mathrm{N}_1 + \mathrm{N}_2 + \mathrm{N}_3/2}{1/2(\text{Number of squares in the Lorenz box})},$$

- where,

$$\mathrm{N}_1 = \text{No. of squares entirely inside the area of concentration}$$

$$\mathrm{N}_2 = \text{No. of squares more than half inside the area of concentration}$$

$$\mathrm{N}_3 = \text{No. of squares exactly half inside the area of concentration}$$

# Definition of LR for Discrete Case

- For discrete case when we have observations $x_1, x_2, \ldots x_n$ such that $x_1 < x_2 < \cdots < x_n$, there are two definitions of LR.

$$LR_1 = \frac{\Delta_1}{2\bar{x}} = \frac{\sum\sum\left|x_i - x_j\right|/n^2}{2\bar{x}} = \frac{\sum(2r-1)x_r}{n^2\bar{x}}$$

$$= \frac{\sum(2r-1)x_r}{n^2\bar{x}} - 1 = \frac{\sum r x_r}{n^2\bar{x}} - \frac{1}{n} - 1.$$

$$LR_2 = \frac{\Delta_2}{2\bar{x}} = \frac{\sum\sum\left|x_i - x_j\right|/(n(n-1))}{2\bar{x}} = \frac{2\sum(r-1)x_r}{n(n-1)\bar{x}} - 1.$$

- In course of time $LR_2$ has been rejected in favour of $LR_1$ by many economists because it was seen that the formula of LR obtained from straight line approximation of the LC for discrete case coincides with the formula of $LR_1$. Observe that $LR_2$ approaches $LR_1$ as n $\rightarrow \infty$.
- The continuous version of both $LR_1$ and $LR_2$ is

$$\Delta = E|X_1 - X_2| = \int_0^\infty \int_0^\infty |x_1 - x_2|\, dF(x_1) dF(x_2),$$

- where $X_1$ and $X_2$ are independent non-negative random variables with common distribution function F.

***Thank You***