The Economic Journal of Nepal, Vol. 16, No.4, Issue 64, October-December 1993 © DOE-TU

Lorenz Curve and Gini Coefficient: Conceptual Considerations

Nav R. Kanel*

INTRODUCTION

The data that we deal with are usually unequal in nature. Much of the problems in statistics would be simplified if the data under investigation show equality within the given dataset and across similar other datasets. Therefore, because of inequality in the values of any variable, the study of inequality has a very important role in statistics as well as in macroeconomic and development studies.

Inequality refers to the situation in which a particular variable under inquiry does not show equality in its values. Many economic variables such as income, assets, land, educational level, to mention a few, are not distributed equally (or proportionately). This has not only created a conflict among various socioeconomic classes, but it has also alarmed the policy makers about how to reduce, let alone eliminate, such inequalities. Societies have been divided into *rich* and *poor* because of the maldistribution of such economic variables. This has been a problem not only in less developed countries but also in developed countries. The people of developed countries are also suffering from economic inequalities among social classes, and between races (in the United States, for example). The problem arising from unequal distribution of a variable under investigation has always been a matter of apprehension everywhere.

An income distribution is usefully described by a measure of central tendency as well as a measure of dispersion. It is relatively easier to find out a measure of central tendency, but how to measure such inequalities in the distribution of the variable under inquiry ? Various measures of inequality (dispersion) have been suggested in the literature and are in use - range, relative mean deviation, variance, coefficient of variation, standard deviation of logarithms, quartile points, Lorenz Curve, and Gini Coefficient. From the standpoint of welfare economics, it is customary to study the concentration of income, which is just another way of looking at the income dispersion.

OBJECTIVE

The objective of this paper is to show a method of deriving the formula for calculating Gini Coefficient from definition, the Lorenz Curve. Many books (Anand, 1983; Kakwani, 1980, for example) merely describe that the Gini Coefficient is derived from the Lorenz Curve, and furnish only the formula for the calculation of Gini Coefficient. None of them, however, show the proof of the formula clearly and in a simplified manner. Because of this problem many people keep wondering about the derivation of the formula. This paper attempts to solve the problem of those people who might be wishing to derive the formula for Gini Coefficient. In this paper,

* *Dr. Kanel is a Lecturer at Central Department of Economics, T.U., Kirtipur, Kathmandu, Nepal.*

therefore, the concepts of the Lorenz Curve and the Gini Coefficient are briefly examined, and the formulas for the computation of the Gini Coefficient are derived. Other measures of dispersion (inequality) are not discussed here as they are described elsewhere (Sen, 1973).

LORENZ CURVE AND GINI COEFFICIENT

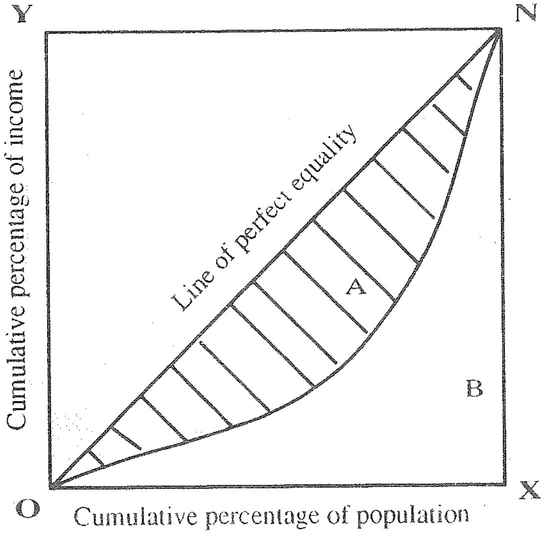
A measure that has been very widely used to represent the extent of inequality is the Gini Coefficient, also known as Gini Concentration Ratio, attributed to Gini (1912, 1921). Gini Coefficient is calculated from the help of a Lorenz Curve, due to Lorenz (1905). These measures, Gini Coefficient and Lorenz Curve, were originally developed to measure the concentration of income, and now also they are being extensively used to measure income disparity. Nowadays, these measures are also used in macroeconomic analysis of such variables as income-, assets-, and land distributions in various countries and among various socioeconomic classes within a country. The United Nations usually employs these measures to compare the distribution of wealth and land across various countries (UNDP, 1993). Besides, Gini Coefficients are also used in the field of public economics to examine the change in the income- and wealth distributions before and after certain tax schedules, and similar applications of these measures are also used in the domain of monetary economics to examine the effects of inflation on similar variables.

By definition, Gini Coefficient is the proportion of the total area (of the triangle) under the diagonal that lies in the area between the diagonal and the Lorenz Curve. The Gini Coefficient also satisfies the Pigou-Dalton Condition, a desirable property of any measure of inequality (The desirable property of any measure of inequality is that any transfer of income from a poorer person to a richer person, other things being the same, must always increase the value of an inequality measure. This condition is referred to as the Pigou-Dalton Condition). A transfer of income from a richer to a poorer person raises the entire Lorenz Curve between the corresponding percentiles; hence it reduces the Gini Coefficient. As the definition of Gini Coefficient involves the concept of Lorenz Curve as well, it will be relevant to first define and describe what a Lorenz Curve is.

The Lorenz Curve, first expounded in 1905, has long been used to measure inequalities in the distribution of wealth and income. It is a graphical depiction of the concentration of income and wealth. It has also been used to show the state (as opposed to the process) of concentration of population and of other demographic aggregates. To plot the curve, the units are first either arrayed individually or grouped in class intervals according to the appropriate independent variable. If we assume that we are interested to find out the measurement of income distribution, the percentage of population is arranged systematically from the poorest to the richest and is arranged on the horizontal axis, and the percentage of income enjoyed by the bottom x percent of the population is shown on the vertical axis (Proportions of population and income can also be used in stead of percentages of these variables). Then the cumulative percentage of the population (X_i) is plotted against the cumulative percentage of the total income (Y_i). The curve derived thus from these two variables is called a Lorenz Curve. For comparison, a diagonal line is drawn at 45° to show the condition of perfect equality.

The following figure might be helpful for a clearer exposition of the concepts involved in the discussion.

Figure 1


Y
N
Cumulative percentage of income
Line of perfect equality
A
B
O
X
Cumulative percentage of population

Because 0 percent of the population enjoys 0 percent of the income, a Lorenz Curve must originate from O (as shown in Figure 1). Also, 100 percent of the population enjoys 100 percent of the income, a Lorenz Curve must terminate at N. If the income is distributed equally x percent of the total population will receive exactly x percent of the total income. In this case, the graph of such income distribution will be the 45^0 line, which is the diagonal ON, showing perfect equality between the values of the horizontal axis and the vertical axis. On the other extreme, if there is perfect inequality in the distribution (i.e., only one person gets all the income) the rest of the population gets nothing (i.e., they get 0 percent of the income). In such situation, the curve will run from O along OX until we reach to X where the curve will jump to N, because the last person gets 100 percent of the total income. In this case, the graph of an such income distribution will be OXN. Therefore, the total area included in between these two extreme cases is given by the area of the triangle OXN.

Perfect equality and perfect inequality are two extreme situations, which we usually do not observe in the real world. What we usually observe more often is that the lower (higher) income groups will enjoy a proportionately lower (higher) share of income. Plotting all such values, we will obtain a curve which runs from O to N and is in between these two extremes as shown in Figure 1. Obviously, a Lorenz Curve in this case must lie below the diagonal, and its slope will increasingly rise -- at any rate will not fall -- as we move to richer and richer sections of the population (If we instead arrange the population from the richest to the poorest (in non-ascending order) the Lorenz Curve will lie above the diagonal (in the triangular area OYN), and its slope will increasingly fall -- at any rate will not rise -- as we move to poorer and poorer sections of the population). From the figure it can also be shown that for more unequal

(equal) income distribution, the Lorenz Curve will be farther away from (nearer to) the diagonal line. The converse of this rule is also true: the greater the departure of the Lorenz Curve from the diagonal, the larger is the inequality in income distribution.

The Gini Coefficient, G, is defined as the area between the Lorenz Curve and the line of equality (i.e., the diagonal) divided by the area of the triangle OXN (It can also be revealed, though tedious to work it out, that the Gini Coefficient is exactly one-half of the relative mean difference, which is defined as the arithmetic average of the absolute values of differences between all pairs of incomes). If the area between the diagonal and the Lorenz Curve, as shown in Figure 1, is denoted by area A and the area of the triangle OXN below the Lorenz Cure by B, the Gini Coefficient can be specified in algebraic terms as follows:

$$G = \frac{A}{A + B} \quad \ldots (1)$$

The greater the departure of the Lorenz Curve from the diagonal, the larger will be the value of the Gini Coefficient.

When there is a perfect equality in the income distribution, then the diagonal line will be the Lorenz Curve implying that the area A will be zero. In this case, G also will be equal to zero. On the other hand, if there is perfect inequality (i.e., only one person gets all the income), then the lines OXN will be the Lorenz Curve, in which case the area B will be equal to zero. In this case, G will be equal to one.

Therefore, $O \leq G \leq 1$. ... (2)

As G is the ratio of two areas it is always positive, which is also shown by equation (2). But while calculating the value of G one sometimes may encounter with a negative value of G. In such a case, the negative sign should be ignored and only the absolute value of the result should be taken.

As mentioned earlier, to plot a Lorenz Curve the units are first either arranged individually or grouped in class intervals according to the appropriate independent variable, which is income in our case. The most common forms of formulas for computing the Gini Coefficient are as follows:

Grouped Data

$$G = \sum_i X_i Y_{i+l} \; - \; \sum_i X_{i+l} Y_i \quad \ldots (3)$$

Where, X_i denotes the cumulative *proportion* of the population in the *i*th class interval, and

Y_i denotes the cumulative *proportion* of the population in the *i*th class interval.

It should be noted that X_i and Y_i have been defined as the *percentages* (in stead of proportions) of respective variables in the discussion of Lorenz Curve. This should not make any difference in our results. If the variables are measured as percentages, then both of them have to be divided by 100 to change them into proportions. In this case, equation (3) will reduce to:

$$\left.\begin{aligned} G &= \frac{1}{100}\left[\sum_i X_i Y_{i+1} - \sum_i X_{i+1} Y_i\right] \text{ percent} \\ \text{or } G &= \frac{1}{100^2}\left[\sum_i X_i Y_{i+1} - \sum_i X_{i+1} Y_i\right]. \end{aligned}\right] \quad \ldots (4)$$

Ungrouped Data

$$G = (1 + 1/n) - \frac{2}{n^2\mu}\left[y_n + 2y_{n-1} + \ldots + ny_1\right]$$

for $y_1 \leq y_2 \leq \ldots \leq y_n$,

where, n = number of observations, and
μ = mean value of y.

Both these formulas are proved in the following section.

PROOF OF FORMULAS

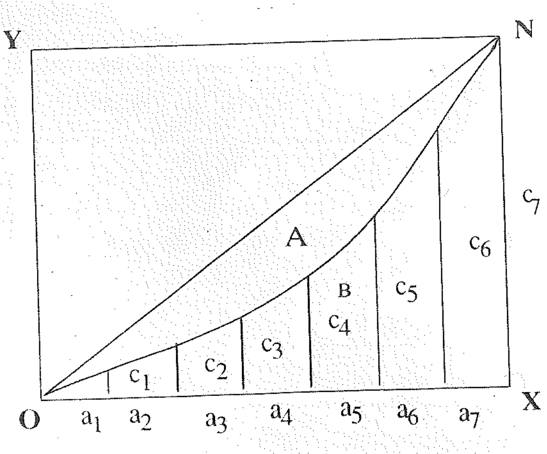
For Grouped Data

For simplicity, I will assume that there are only seven groups (class intervals) in this case. Any number of groups could have been taken. Nevertheless, I will generalize this case to n groups.

Let a_i denote the *proportion* of the x-axis variable in the *i*th class interval,
c_i denote the *cumulative proportion* of the y-axis variable in the *i*th class interval, and
A and B denote the areas as defined earlier.

Then the Lorenz Curve for these seven groups can be shown as follows:

Figure 2



We know that Gini Coefficient, $G = \dfrac{A}{A + B}$.

Here we have to find out the areas A and B.
A + B denotes the total area of the triangle OXN.

Therefore, $A + B = \frac{1}{2}(a_1 + a_2 + a_3 + a_4 + a_5 + a_6 + a_7).\ c_7.$ ... (5)

Area B consists of seven smaller areas: one triangle, and six trapeziums, as shown in Figure 2. Let these seven little areas be denoted by $B_1, B_2, B_3 \ldots, B_7$.

Then,

$$B_1 = \frac{1}{2}c_1\ a_1,$$
$$B_2 = \frac{1}{2}(c_1 + c_2)\ a_2,$$
$$B_3 = \frac{1}{2}(c_2 + c_3)\ a_3,$$
$$B_4 = \frac{1}{2}(c_3 + c_4)\ a_4,$$
$$B_5 = \frac{1}{2}(c_4 + c_5)\ a_5,$$
$$B_6 = \frac{1}{2}(c_5 + c_6)\ a_6, \text{ and}$$
$$B_7 = \frac{1}{2}(c_6 + c_7)\ a_7.$$

Therefore, Area $B = B_1 + B_2 + B_3 + B_4 + B_5 + B_6 + B_7$

$$= \frac{1}{2}\left[c_1a_1+(c_1+c_2)a_2+(c_2+c_3)a_3+(c_3+c_4)a_4+(c_4+c_5)a_5+(c_5+c_6)a_6+(c_6+c_7)a_7\right]$$

$$= \frac{1}{2}\left[(a_1+a_2)c_1+(a_2+a_3)c_2+(a_3+a_4)c_3+(a_4+a_5)c_4+(a_5+a_6)c_5+(a_6+a_7)c_6+a_7c_7\right]$$

$$= \frac{1}{2}[(a_1+a_2)c_1 + (a_1+a_2+a_3)c_2 + (a_1+a_2+a_3+a_4)c_3 + (a_1+a_2+\ldots+a_5)c_4 + (a_1+a_2+\ldots+a_6)c_5 + (a_1+a_2+\ldots+a_6+a_7)c_6 + (a_1+a_2+\ldots+a_7)\ c_7]$$

$$- \frac{1}{2}[a_1c_2 + (a_1+a_2)c_3 + (a_1+a_2+a_3)c_4 + (a_1+a_2+a_3+a_4)c_5 + (a_1+a_2+\ldots+a_5)c_6 + (a_1+a_2+\ldots+a_6)c_7].$$

Let $a_1+a_2+ \ldots + a_i = x_i$ (cumulative proportion of the x-axis variable), and
$c_i = y_i$ (cumulative proportion of the y-axis variable).

$$\text{Then, } B = \frac{1}{2}\left[X_2Y_1 + X_3Y_2 + X_4Y_3 + X_5Y_4 + X_6Y_5 + X_7Y_6 + X_7Y_7\right]$$
$$- \frac{1}{2}\left[X_1Y_2 + X_2Y_3 + X_3Y_4 + X_4Y_5 + X_5Y_6 + X_6Y_7\right]$$
$$= \frac{1}{2}\left[X_2Y_1 + X_3Y_2 + X_4Y_3 + X_5Y_4 + X_6Y_5 + X_7Y_6\right] + \frac{1}{2}X_7Y_7$$

$$-\frac{1}{2}\left[X_1Y_2 + X_2Y_3 + X_3Y_4 + X_4Y_5 + X_5Y_6 + X_6Y_7\right]. \qquad \ldots (6)$$

From equation (5), $A + B = \frac{1}{2}\ (a_1 + a_2 + \ldots + a_7)\ c_7$

$$= \frac{1}{2} X_7Y_7 \text{ (in our new notations)}$$

$$= \frac{1}{2}.1.1\ [\text{Because, } X_7 = 100\% = 1,\quad Y_7 = 100\% = 1]$$

$$= \frac{1}{2}. \qquad \ldots (7)$$

For generalization, use summation notations for equation (6).

Then, $$B = \frac{1}{2}\left[\sum_i X_{i+1}\ Y_i\right] + (A{+}B) - \frac{1}{2}\left[\sum_i X_i\ Y_{i+1}\right].$$

or $$B = \frac{1}{2}\left[\Sigma X_{i+1}\ Y_i - \Sigma X_i\ Y_{i+1}\right] + (A{+}B).$$

Therefore, $$A = \frac{1}{2}\left[\Sigma X_i\ Y_{i+1} - \Sigma X_{i+1}\ Y_i\right]$$

As $$G = \frac{A}{A+B}, \text{ and } A{+}B = \frac{1}{2} \text{ from equation (7).}$$

Then $$G = \frac{\frac{1}{2}\left[\Sigma X_i\ Y_{i+1} - \Sigma X_{i+1}\ Y_i\right]}{\frac{1}{2}}$$

i.e., $G = \Sigma X_i\ Y_{i+1} - \Sigma X_{i+1}\ Y_i$.

For Ungrouped Data |

Most of the literature on Gini Coefficient employ the method of absolute mean difference to find out the Coefficient as used by its originator Gini (Kakwani, 1980; Sen, 1973). However, the method that I am employing here to find out the formula for Gini Coefficient for ungrouped data is somewhat different from the conventional one. I am utilizing the same formula that was derived earlier to find out the Gini Coefficient for grouped data.

Let y_i denote the values of the variable under investigation. Let us also assume that there are *n* observations in the given distribution, for which we are interested in calculating the Gini Coefficient. First of all, we arrange the data in a non-descending order as:

$y_1, y_2, ..., y_n$; where $y_1 \leq y_2 \leq ... \leq y_n$.

Let μ be the mean value of y, so that $\mu = \frac{\Sigma y_i}{n}$.

Therefore, $\Sigma y_i = n\mu$.

Then the proportion of the total value of y for the *ith* observation is $\frac{y_i}{\Sigma y_i} = \frac{y_i}{n\mu}$.

Similarly, as there is only one observation for each value of the variable, the proportion of total observations to any observation is $\frac{1}{n}$.

Suppose q_i denotes the *proportion* of the total value of the variable and pi denotes the corresponding *proportion* of total frequencies for the ith observation. Then the distribution can be written as:

i	q_i	p_i	X_i	Y_i
1	$\frac{y_1}{n\mu}$	$\frac{1}{n}$	$\frac{1}{n}$	$\frac{y_1}{n\mu}$
2	$\frac{y_2}{n\mu}$	$\frac{1}{n}$	$\frac{2}{n}$	$\frac{(y_1+y_2)}{n\mu}$
3	$\frac{y_3}{n\mu}$	$\frac{1}{n}$	$\frac{3}{n}$	$\frac{(y_1+y_2+y_3)}{n\mu}$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
n-1	$\frac{y_{n-1}}{n\mu}$	$\frac{1}{n}$	$\frac{n-1}{n}$	$\frac{(y_1+y_2+ ... +y_{n-1})}{n\mu}$
n	$\frac{y_n}{n\mu}$	$\frac{1}{n}$	$\frac{n}{n}$	$\frac{(y_1+y_2+ ... +y_n)}{n\mu}$

Where X_i and Y_i denote the cumulative values of p_i and q_i respectively. Note that both X_i and Y_i are equal to one for i = n.

The values of x_i Y_{i+1} and x_{i+1} Y_i can be derived as

i	$X_i \, Y_{i+1}$	$X_{i+1} \, Y_i$
1	——	$\frac{1}{n^2\mu} . 2y_1$
2	$\frac{1}{n^2\mu} . (y_1+y_2)$	$\frac{1}{n^2\mu} . 3 (y_1+y_2)$

3	$\frac{1}{n^2\mu} \,.\, 2\,(y_1+y_2+y_3)$	$\frac{1}{n^2\mu} \,.\, 4\,(y_1+y_2+y_3)$
.	.	.
.	.	.
.		.
n-1	$\frac{1}{n^2\mu} \,.\, (n-2)\,(y_1+y_2+ \ldots +y_{n-1})$	$\frac{1}{n^2\mu} \,.\, n\,(y_1+y_2+ \ldots +y_{n-1})$
n	$\frac{1}{n^2\mu} \,.\, (n-1)\,(y_1+y_2+ \ldots +y_n)$	---

Now using the formula derived for grouped data,

$$G = \sum_i X_i Y_{i+1} - \sum_i X_{i+1} Y_i$$

$$= \sum_i (X_i Y_{i+1} - X_{i+1} Y_i)$$

$$= -\frac{1}{n^2\mu}\Big[2y_1+2(y_1+y_2)+2(y_1+y_2+y_3)+\ldots+2(y_1+y_2+\ldots+y_{n-1})-(n-1)(y_1+y_2+\ldots+y_n)\Big]$$

$$= -\frac{1}{n^2\mu}[2y_1+2(y_1+y_2)+2(y_1+y_2+y_3)+\ldots+2(y_1+y_2+\ldots+y_{n-1})+2(y_1+y_2+\ldots+y_n)- (n+1)(y_1+y_2+\ldots+y_n)]$$

$$= -\frac{-2}{n^2\mu}\Big[y_1+(y_1+y_2)+(y_1+y_2+y_3)+\ldots+(y_1+y_2+\ldots+y_n)\Big] + \frac{n+1}{n^2\mu}\,(y_1+y_2+ \ldots +y_n)$$

$$= \frac{-2}{n^2\mu}\,(ny_1 + (n-1)y_2 + \ldots + y_n) + \frac{n+1}{n^2\mu}\,.\,n\mu \quad [\text{Because } y_1 + y_2 + \ldots + y_n = n\mu]$$

$$= \frac{-2}{n^2\mu}\,(ny_1 + (n-1)y_2 + \ldots + y_n) + \frac{n+1}{n}$$

$$= (1+\frac{1}{n}) - \frac{2}{n^2\mu}\,(ny_1 + (n-1)y_2 + \ldots + y_n)$$

Therefore, $G = (1+\frac{1}{n}) - \frac{2}{n^2\mu}\,(y_n + 2y_{n-1} + \ldots + ny_1)$.

As Gini Coefficient is the ratio of two areas, which are never negative, we should take only the absolute value of the result even if we sometimes get a negative result.

In both the above cases, we have as seemed that the data are arranged in a non-descending order. What would happen if we arrange the data in a non-ascending order ? There will be no change in the result, because only the sign of G, which we always ignore, will be different leaving the absolute value of G as the same. Similar will be the case if we interchange the variables Xi and Yi. Therefore, assigning the variable names is also irrelevant so far as the calculation of Gini coefficient is concerned.

Cited References

Anand, Sudhir, (1983), *Inequality and Poverty in Malaysia*, New York: Oxford University Press, (A World Bank Research Publication).

Gini C., (1912), *Variabilita e mutabilita*, Bologna.

-------------- (1921), "Measurement of Inequality of Incomes," *Economic Journal*, Vol. 31, pp. 124-26.

Kakwani, Nanak C., (1980), *Income Inequality and Poverty: Methods of Poverty: Methods of Estimation and Policy Applications*, Washington, DC, Oxford University Press (Published for the World Bank).

Lorenz, M.O., (1905), "Methods of Measuring the Concentration of Wealth," *Quarterly Publications of the American Statistical Association*, 9 (70): 209-219.

Sen, Amartya K. , (1973), *On Economic Inequality*, Oxford: Clarendon Press (and New York: Norton).

UNDP, (1993), *Human Development Report 1993*, UNDP.