The Economic Journal of Nepal, Vol. 18, No. 4, Issue 72, Oct.-Dec., 1995 © DOE-TU

Gini Coefficient and Kanel's Reduction

Nav R. Kanel*

INTRODUCTION

Lorenz curve and Gini coefficient (also known as Gini Concentration Ratio or Gini Index) are used to measure the inequality in the values of a variable. Inequality refers to the situation in which a particular variable under consideration does not show equality in its values. There are other measures of inequality that measure the dispersion in the values of a variable. Various measures of inequality (dispersion) such as range, relative mean deviation, variance, coefficient of variation, standard deviation of logarithms, quartile points etc. have been suggested in the literature and are in use. As these measures are relatively easier to calculate, many studies employ either of these measures to measure the inequality of the given data sets. As these measures have their own strengths and weaknesses regarding the measurement of inequality, none of them can fully solve the problems regarding the measurement of inequality.

Lorenz curve and Gini coefficient can solve some of the problems of measuring inequality which cannot be solved by any of the above-mentioned methods (measures). A detailed discussion on this issue is available elsewhere (Sen 1973, for example) and is, therefore, not discussed here. An interested reader might wish to consult the book for a detailed discussion on the topic of inequality measurement.

OBJECTIVES

The objective of this paper is to enquire into the characteristics of Gini coefficients by paying a revisit on the calculation and characteristics of Gini coefficient for any given data sets. Four theorems on the characteristics of Gini coefficient have been propagated; and their general enunciation and the theoretical proofs have also been provided. A hypothetical example is also used to provide empirical proof, wherever practical.

In my previous article (Kanel 1993), I have discussed about the relationship between the Lorenz curve and the Gini coefficient for a given data set. In that article I have also mentioned that Gini coefficient is the proportion of the total area (of the triangle) under the diagonal line that

---

* Dr. Kanel is Lecturer at Central Department of Economics, Tribhuvan University, Kirtipur, Kathmandu, Nepal.

lies in the area between the diagonal and the Lorenz curve. I have also shown the procedure to derive the formula to calculate the Gini coefficient from a Lorenz curve by measuring the area under the diagonal line and the Lorenz curve. To avoid repetition, I am skipping all the procedures that have already been discussed in that article. This article is, however, only an extension and continuation to that article. To resume continuity, an interested reader might wish to read that article first. Nevertheless, the main points and formulas underlying in Kanel (1993) are summarized in the following section. The derivations of the formulas, however, are not provided here to avoid repetition.

REVIEW OF KANEL (1993)

The Gini coefficient, G, is defined as the area between the Lorenz curve and the line of perfect equality(i.e., the diagonal line) divided by the area of the triangle under (above) the diagonal line. If the area between the diagonal line and the Lorenz curve is denoted by A and the area of the triangular figure below the Lorenz curve by B, the Gini coefficient can be specified in algebraic terms as follows:

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

When there is perfect equality in the data set (values of the variable under consideration), then the Lorenz curve will overlap with the diagonal line implying that the area A will be equal to zero. In this case the Gini coefficient will be equal to zero. On the other hand, if there is perfect inequality in the data set (only one observation carries the total of the variable, i.e., all the non-zero values of the variable are concentrated to only one observation), then the value of area B will be equal to zero. In this case the Gini coefficient will be equal to one. These are, therefore, the extreme values of Gini coefficient.

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

The higher the value of the Gini coefficient the higher will be the inequality. Similarly, lower value of the Gini coefficient indicates lower inequality. The formulas to calculate the Gini coefficient are as follows:

For Grouped Data:

$$G = \Sigma X_i Y_{i+1} - \Sigma X_{i+1} Y_i \qquad \ldots (3)$$

where, X_i denotes the *cumulative proportion* of the x- variable in the *ith* class interval, and

Y_i denotes the *cumulative proportion* of the y-variable in the *ith* class interval.

If X_i and Y_i are considered as two columns, equation (3) can be written as the sum of all the (n-1) determinant values of X_i and Y_i taken two consecutive class intervals at a time. In other words, $\Sigma X_i Y_{i+1} - \Sigma X_{i+1} Y_i$ can be written as $\Sigma (X_i Y_{i+1} - X_{i+1} Y_i)$, which is nothing but the sum of all the determinant values of

$$\begin{vmatrix} X_i & Y_i \\ X_{i+1} & Y_{i+1} \end{vmatrix}, \quad \text{for } i = 1, 2, 3, \ldots, (n-1).$$

Obviously, there will be (n-1) such determinant values.

It should be noted from equation (3) that the variables x and y should be measured in proportions, not in percentages. However, if these variables are calculated as percentages, each of them have to be divided by 100 to change them into proportions from percentages. In that case the Gini coefficient will be calculated as :

$$G = 1/100 \, [\Sigma X_i Y_{i+1} - \Sigma X_{i+1} Y_i] \quad \text{percent}$$
$$\text{or, } G = 1/(100)^2 \, [\Sigma X_i Y_{i+1} - \Sigma X_{i+1} Y_i]. \qquad \ldots (4)$$

As Gini coefficient is the ratio of two areas, which are never negative, the value of Gini coefficient is always positive. We should take only the absolute value of the results even if we sometimes encounter with a negative value of G. Interchange of the variable names only changes the sign of G without affecting its absolute value. Therefore, assigning the variable names is also irrelevant so far as the calculation of Gini coefficient is concerned.

For Ungrouped Data:

$$G = (1+1/n) - 2/n^2 \mu \, [y_n + 2y_{n-1} + \ldots + ny_1] \qquad \ldots (5)$$
$$\text{for } y_1 \le y_2 \le \ldots \le y_n$$

where, n denotes the number of observations,

μ denotes the mean value of the variable (y), and

y_i denotes the variable value for the *ith* observation.

From equation (5) it is obvious that the data (values of the y variable) should be arranged in ascending order. But if the data are arranged in the descending order then the formula will change to

$$G = (1+1/n) - 2/n^2 \mu \, [y_1 + 2y_2 + \ . \ . \ . \ + ny_n]$$
$$\text{for } y_1 \geq y_2 \geq \ldots \geq y_n \qquad \ldots (6)$$

where n, μ and y_i have the same meaning as in equation (5).

A close scrutiny on these two equations, (5) and (6), will show that both of these equations are basically identical and hence yield the same result.

It should be noted that for ungrouped data, y_i denotes the value of the variable for the *i*th observation. Unlike in the case of grouped data, it is neither the proportion nor the percentage value of the *i*th observation.

An Illustrative Example

Consider the following pre-tax and post- tax income distributions:

Group	Income share	
	Pre-tax	Post-tax
Below 10%	2.8	3.3
20-30%	4.8	5.5
10-20%	3.1	3.6
40-50%	7.5	8.0
30-40%	5.9	6.5
60-70%	11.0	11.2
70-80%	13.1	13.2
80-90%	15.8	15.8
50-60%	9.2	9.5
Top 10%	26.8	23.4

Compute and interpret the Gini coefficients for the two data sets.

To solve this problem, first of all we have to rearrange the given data set in ascending order. In this case, the given data set can be written as:

Group	Income share	
	Pre-tax	Post-tax
Below 10%	2.8	3.3
10-20%	3.1	3.6
20-30%	4.8	5.5
30-40%	5.9	6.5
40-50%	7.5	8.0
50-60%	9.2	9.5
60-70%	11.0	11.2
70-80%	13.1	13.2
80-90%	15.8	15.8
Top 10%	26.8	23.4

Let X_i = *Cumulative percentage* of persons in the *i*th class interval, and
Y_i = *cumulative percentage* of income in the *i*th class interval.

This is a case of calculating Gini coefficients for grouped data sets. Therefore, we have to employ the corresponding formula to calculate the Gini coefficients for this given question. The corresponding worksheet to calculate the value of the Gini coefficients for the two data sets will be as follows.

	Pre-tax				Post-tax			
i	X_i	Y_i	$X_i Y_{i+1}$	$X_{i+1} Y_i$	X_i	Y_i	$X_i Y_{i+1}$	$X_{i+1} Y_i$
1	10	2.8	59	56	10	3.3	69	66
2	20	5.9	214	177	20	6.9	248	207
3	30	10.7	498	428	30	12.4	567	496
4	40	16.6	964	830	40	18.9	1076	945
5	50	24.1	1665	1446	50	26.9	1820	1614
6	60	33.3	2658	2331	60	36.4	2856	2548
7	70	44.3	4018	3544	70	47.6	4256	3808
8	80	57.4	5856	5166	80	60.8	6128	5472
9	90	73.2	9000	7320	90	76.6	9000	7660
10	100	100.0	–	–	100	100.0	–	–
Total	-	368.3	24932	21298	-	389.8	26020	22816

Since the variables have been measured in *percentages*, we have to use equation (4) to calculate the Gini coefficient.

The Gini coefficient for the pre-tax income distribution will be:

$$G_{pre\text{-}tax} = 1/(100)^2 (\Sigma X_i Y_{i+1} - \Sigma X_{i+1} Y_i)$$
$$= 1/(100)^2 (24932 - 21298)$$
$$= 1/(100)^2 (3634)$$
$$= 0.3634.$$

Similarly, the Gini coefficient for the post-tax income distribution will be:

$$G_{post\text{-}tax} = 1/(100)^2 (\Sigma X_i Y_{i+1} - \Sigma X_{i+1} Y_i)$$
$$= 1/(100)^2 (26020 - 22816)$$
$$= 1/(100)^2 (3204)$$
$$= 0.3204.$$

Since the Gini coefficient for the post-tax income distribution is lower than that for the pre-tax income distribution, it shows that the imposition of taxation in income helps to reduce income inequality to some extent.

ENUNCIATION AND PROOFS OF THEOREMS

Theorem 1 *(Kanel's Reduction)*

When the class interval of a grouped data set remains constant, the formula to calculate the Gini coefficient can be reduced to a much simpler form which is independent of the population variable.

For grouped data, we have $G = \Sigma X_i Y_{i+1} - \Sigma X_{i+1} Y_i$ as shown in equation (3). Here, X_i and Y_i denote the *cumulative proportions* of the x-variable and the y-variable in the *i*th class interval respectively.

Let x_i and y_i denote the *proportions* of the x-variable and the y-variable in the *i*th class interval respectively. Then X_i and Y_i are calculated as the *cumulative values* of these variables from the *first* class interval upto the *i*th class interval respectively. For example, $X_i = x_1+x_2+x_3 + \ldots + x_i$. If we have n groups in the given data set, then X_n will be the sum of all x_is upto the *last* class interval. Because the sum of all proportions (of a variable) has to be equal to one, the value of X_n is always equal to one. The same arguments hold good for Y_n as well.

When the groups are of equal intervals, the general case can be reduced to a special case. In such case, the worksheet to calculate the Gini coefficient will look as follows: (Here it is assumed that there are n groups in the given data set.)

i	x_i	X_i	y_i	Y_i	X_{i+1}	Y_{i+1}	$X_i Y_{i+1}$	$X_{i+1}Y_i$
1	1/n	1/n	y_1	Y_1	2/n	Y_2	1/n. Y_2	2/n.Y_1
2	1/n	2/n	y_2	Y_2	3/n	Y_3	2/n. Y_3	3/n.Y_2
3	1/n	3/n	y_3	Y_3	4/n	Y_4	3/n. Y_4	4/n.Y_3
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	n-1/n	Y_{n-1}	.	.
n-1	1/n	n-1/n	y_{n-1}	Y_{n-1}	n/n	Y_n	(n-1)/n.Y_n	n/n.Y_{n-1}
n	1/n	n/n	y_n	Y_n	——	——	——	——

Obviously, $X_n = 1$ and $Y_n = 1$, as explained earlier.

We have,
$$G = (\Sigma X_i Y_{i+1}) - (\Sigma X_{i+1} Y_i)$$
$$= 1/n(Y_2 + 2Y_3 + 3Y_4 + \ldots + (n-1) Y_n)$$
$$- 1/n(2Y_1 + 3Y_2 + 4Y_3 + \ldots + nY_{n-1})$$
$$= - 1/n(2Y_1 + 2Y_2 + 2Y_3 + \ldots + 2Y_{n-1} - (n-1) Y_n)$$
$$= - 2/n(Y_1 + Y_2 + Y_3 + \ldots + Y_{n-1}) + (n-1)/n.1$$
[Because $Y_n = 1$.]
$$= (n-1)/n - 2/n (Y_1 + Y_2 + Y_3 + \ldots + Y_{n-1}).$$

Adding and subtracting $2Y_n/n$ in the above line, the expression will reduce to:

$$G = (1+1/n) - 2/n (Y_1 + Y_2 + Y_3 + \ldots + Y_n.) \qquad \ldots (7)$$

The above formula can be further simplified to:

$$G = 1/n (1 + n - 2. \sum_{i=1}^{n} Y_i) \qquad \ldots (7a)$$

This is a much simpler form of the original formula for calculating Gini coefficient. This expression has some notable characteristics. As it is independent of the x-variable we do not have to bother about the handling of the x-variable in this case. One is, therefore, completely relieved of the burden of finding the sums of the cross multiplications of the x variable and the y variable. Instead of the x variable, n, which is the number of class groups, enters into the picture rather in an uncomplicated manner. The value in the latter parenthesis in equation (7) is the sum of the cumulative proportions of the y-variable. To assert it more explicitly, it is the total of the cumulative proportions of the variable under consideration. This reduction of the Gini coefficient formula to a simpler form can be referred to as *Kanel's Reduction.*

It should be noted that both x and y were calculated as the proportions of their respective variables. But one might wonder whether he could use the above formula if he had already calculated the percentages of those variables instead of their proportions. There should not be any problem to apply *Kanel's Reduction* even with the percentages of the variables. As the x variable is out of the scene in the *Kanel's Reduction,* we need not to worry about that variable. The only remaining variable to treat for the required calculations is the y variable. Dividing the total of the cumulative percentages of the y variable (the sum of the Y variable) by 100 will change the percentages into proportions. After that, he can still use the *Kanel's Reduction* to get the required value.

Theorem 2

For grouped data, more groups yield larger value and, similarly, fewer groups yield smaller value of the Gini coefficient.

When we have more groups in a data set, the Lorenz curve will shift outwards (from the diagonal line). Hence the Gini coefficient will be larger. On the other hand, if we rearrange the given data set into fewer groups, the Lorenz curve will move towards the diagonal line yielding a smaller value of the Gini coefficient.

Theorem 3

Though the theoretical maximum value of the Gini coefficient is one, its empirical maximum value depends upon the number of observations; but it cannot exceed one.

The general formula for Gini coefficient (for grouped data) as shown in equation (3) is:

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

The formula for ungrouped data, as shown by equation (5) or (6), is derived as a special case of the above formula (Kanel 1993). Equation (2) states that the maximum and the minimum values of Gini coefficients are 1 and 0 respectively. These are, however, the theoretical extrema of G. Practically, the maximum value of Gini coefficient depends upon the number of observations. I will use the above general formula (for n classes) to prove the statement just made in the preceding sentence.

When the values of the variable are equally distributed, then α percent of the population will get α percent of the total value of the variable under consideration. In this case, x_i percent of the population will acquire x_i percent of the total y. But, we have assumed that x_i percent of the population will acquire y_i percent of the total y. Therefore, x_i will be equal to y_i, for all values of i. Hence, $X_i = Y_i$ and $X_{i+1} = Y_{i+1}$. Then both the terms of the right hand side of the above equation will be equal, as a result of which the value of G will be equal to zero.

When there is perfect inequality in the distribution of the values of the variable, the values of the y_i column will be different. The value of y_i will be equal to one for i = n, because the last group will acquire all the non-zero values of y; and y_i will be equal to zero for other values of i. The values of the Y_i column, which are the cumulative values of y_1 to y_i, will be the same as for the y_i column. The values of the x_i and the X_i columns will not be affected. From these values, we can find the X_{i+1} and the Y_{i+1} columns. From the calculation of all these columns there should not be any problem to find out the values of $\Sigma X_i Y_{i+1}$ and $\Sigma X_{i+1} Y_i$. A worksheet well help to understand the steps mentioned in this paragraph. From the worksheet it can be shown that $\Sigma X_i Y_{i+1}$ will be equal to $X_n - 1$ and $\Sigma X_{i+1} Y_i$ will be equal to zero.

Therefore,
$$\begin{aligned} G &= \Sigma X_i Y_{i+1} - \Sigma X_{i+1} Y_i \\ &= X_{n-1} - 0 \\ &= X_{n-1}. \end{aligned}$$

But what is the value of X_{n-1} ? It is the cumulative value of the x-variable from x_1 to x_{n-1}. The value of X_n, which is the cumulative value of x_1 to x_n, is obviously equal to 1.

Therefore, $X_{n-1} + x_n = Xn = 1$

or, $X_{n-1} = 1 - x_n$.

Hence, $G = 1 - x_n$.

This upper value of G is obviously not equal to one unless x_n is equal to zero. Here x_n is the proportion of the population that acquires all the non-zero values of the variable for which one is trying to calculate the Gini coefficient. Therefore, $(1-x_n)$ is the proportion of the population that gets nothing of the y variable. This is an interesting alternate interpretation of Gini coefficient that follows from the above finding.

Obviously, x_n cannot be equal to zero. If it is equal to zero, then who gets all those values of the y variable ? Therefore, G cannot be equal to one. However, G will approach to one as x_n approaches to zero.

Similarly, it can be shown that the maximum value of G for ungrouped data and grouped data with equal class intervals is (1-1/n), where n is the number of observations or the number of class intervals respectively. The value of x_n for ungrouped data is 1/n, because the last person will get all the values of the y variable. Similarly, the value of x_n for grouped data with equal class intervals will also be equal to 1/n, because the last group carries 1/n proportion of the total population. As G = 1 - 1/n, the value of G increases as n increases. For its maximum value, G will approach to one as n approaches to infinity.

It follows from the above discussion that the value of $\sum_{i=1}^{n} Y_i$, i.e. , $(Y_1+Y_2+Y_3+ \ldots +Y_n)$ will be maximum when the y variable is equally distributed. In this case, its value will be (n+1)/2. On the other hand, the value of $\sum_{i=1}^{n} Y_i$ will be minimum when the y variable is distributed absolutely unequally. In this case, its value will be 1. Its proof has been left to the readers.

Theorem 4

When the class interval of a grouped data set remains constant, the formulas to calculate the Gini coefficient for grouped data and ungrouped data are virtually the same.

This theorem refers to the congruity between *Kanel's Reduction* for grouped data and the formula for ungrouped data.

From *Kanel's Reduction*,

$$G = (1+1/n) - 2/n\,(Y_1 + Y_2 + Y_3 + \ldots + Y_n). \qquad \ldots (7)$$

For ungrouped data,

$$G =(1+1/n) - 2/n^2 \mu\,[y_n + 2y_{n-1} + \ldots + ny_1] \qquad \ldots (5)$$

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

It should be noted that y_is in equation (5) are the absolute values, not the proportions as we have been treating so far, of the y variable.

Therefore, first of all we should change the absolute values of the variable to corresponding proportions. To do so, we have to divide the absolute values by the sum of the y variable, which is given by $n\mu$. Doing so, equation (5) will reduce to:

$$G = (1+1/n) - 2/n\,[y_n + 2y_{n-1} + \ldots + ny_1], \qquad \ldots (5a)$$

where y_is are now measured as the proportions.

Now the terms in the bracket of the above equation can be written as:

$$y_n + 2y_{n-1} + \ldots + ny_1$$

$$= y_1$$
$$+ y_1 + y_2$$
$$+ y_1 + y_2 + y_3$$
$$+ y_1 + y_2 + y_3 + y_4$$

.

.

.

$$+ y_1 + y_2 + y_3 + \ldots + y_{n-1}$$
$$+ y_1 + y_2 + y_3 + \ldots + y_{n-1} + y_n$$
$$= Y_1 + Y_2 + Y_3 + \ldots + Y_n,$$

because $Y_i = y_1 + y_2 + \ldots + y_i$ as defined earlier.

Hence, equation (5a) can be written as :

$$G = \left(1 + \frac{1}{n}\right) - 2/n\,(Y_1 + Y_2 + Y_3 + \ldots + Y_n),$$

which is the duplication of equation (7). Hence the proof.

As stated in equation (7a), the above formula for both the cases can be written in another form as:

$$G = 1/n\,(1 + n - 2.\sum_{i=1}^{n} Y_i). \qquad \ldots (7a)$$

EMPIRICAL PROOFS OF THE THEOREMS

Only Theorems 1 and 2 are relevant for the empirical proofs. Consider the afore-mentioned illustrative example.

Theorem 1 ***(Kanel's Reduction)***

Since the illustrative example contains equal class intervals, one can use *Kanel's Reduction* (Equation 7) to compute the Gini coefficients for the given data sets.

The Gini coefficient for the pre-tax income distribution using *Kanel's Reduction* will be:

$$G_{pre-tax} = 1/n\,(1+n-2.\sum_{i=1}^{n} Y_i)$$

Here, n (number of observations) = 10, and

$\Sigma Y_i = (Y_1 + Y_2 + Y_3 + \ldots + Y_n) = 368.3$ (in percentages),

or 3.683 (in proportions).

Therefore, $G_{pre-tax} = 1/10\,(1 + 10 - 2^* 3.683)$

$= 1/10\,(11 - 7.366)$

$= 0.3634.$

Similarly, the Gini coefficient for the post-tax income distribution using *Kanel's Reduction* will be:

$$G_{post-tax} = 1/n\,(1+n-2.\sum_{i=1}^{n} Y_i)$$

Here, n (number of observations) = 10, and

$\Sigma Y_i = (Y_1 + Y_2 + Y_3 + \ldots + Y_n) = 389.8$ (in percentages),

or 3.898 (in proportions).

Therefore, $G_{post-tax} = 1/10\,(1 + 10 - 2^* 3.898)$

$= 1/10\,(11 - 7.796)$

$= 0.3204.$

These results coincide with the results obtained by using the general formula as shown in the illustrative example, but in a much easier and faster way.

Theorem 2

To prove this theorem, I will use the data that are used in the illustrative example. Suppose we reduce the class groups to 5 from 10.

To do so, I will regroup the data set by taking 20 percent populations in each group. By doing so, the new data set will also have equal class intervals. In this case I can use *Kanel's Reduction* rather than the general formula. Then the corresponding income distributions for the pre-tax and post- tax situations will be as follows:

Group	Income share	
	Pre-tax	Post - tax
Below 20%	5.9	6.9
20-40%	10.7	12.0
40-60%	16.7	17.5
60- 80%	24.1	24.4
Top 20%	42.6	39.2

Then the cumulative values of the percentages can be written as:

i	Y_i pre-tax	Y_i post-tax
1	5.9	6.9
2	16.6	18.9
3	33.3	36.4
4	57.4	60.8
5	100.0	100.0
Total	213.2	223.0 (in percentages)
Total	2.132	2.230 (in proportions)

Using *Kanel's Reduction,*

$$G_{pre-tax} = 1/5\ [\ 1 + 5 - 2 * (\ 2.132)] = 1/5\ (6 - 4.264) = 0.3472.$$

Similarly,

$$G_{post-tax} = 1/5\ [1 + 5 - 2 * (2.230)] = 1/5\ (6 - 4.460) = 0.3080.$$

Both these results are smaller than the results obtained from the illustrative example. We would have got the same results had we used the general formula (Equation 3) to solve the above problem.

SELECTED REFERENCES

Kanel, Nav R. (1993) "Lorenz Curve and Gini Coefficient: Conceptual Considerations," *The Economic Journal of Nepal* Vol.16 No.4, Central Department of Economics, Tribhuvan University, Kirtipur, Kathmandu, Nepal.

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