THE GINI INDEX

Robert F. Brown
Department of Mathematics
University of California
Los Angeles, CA 90095-1555
e-mail: rfb@math.ucla.edu

February 11, 2016

The "Occupy Wall Street" movement began in September, 2011 with a demonstration in a park near Wall Street in New York City, the symbol of the financial center of the United States. The demonstrators' motto: "We are the 99 percent" referred to the issue of income inequality. It is estimated that the top one percent of the American population in terms of income received approximately 20 percent of all pre-tax income. The Occupy movement claimed to represent the people who received the 80 percent that was left over.

Given that history, and the continuing discussion of income inequality among America's politicians, it's natural to ask: how do you describe income inequality and how do you measure it? If we can find good answers to these questions, then we should be able to answer such questions as: is income inequality in the United States really getting worse, as the demonstrators claimed, and how does the distribution of income in the United States compare to that of other countries?

A solution to the problem of how to describe the distribution of income among members of a population was proposed by the German economist Max Lorenz in 1905. In describing Lorenz's solution, it's convenient to convert percentages to proportions, that is, numbers between 0 and 1, by dividing the percent by 100 so that, for instance, 20 percent becomes .20. Lorenz used the usual (x, y)-graph of the plane. On the x-axis from 0 to 1 the points represent the proportion of the population ranked from the lowest to the highest with respect to income. Thus the point $x = .15$ would represent the 15 percent of the population with the least income, $x = .5$ the bottom half by income and $x = .99$ corresponds to the people in Occupy's motto.

On the y-axis, corresponding to each x value Lorenz placed the proportion of all the population's income that was received by the bottom x-proportion of the population. In 2006, the highest-earning 20 percent of the American population earned about 60 percent of all income so the bottom 80 percent, represented by $x = .80$, received 40 percent and thus the point $(.80, .40)$ appears on the curve. Figure 1 shows a typical "Lorenz curve". The curve connects $(0, 0)$ to $(1, 1)$ and, in this example, passes through the point $(.80, .40)$. For future reference, the figure also includes the straight line segment connecting $(0, 0)$ to $(1, 1)$. We will call this line the "equal distribution line".

Figure 1

Problem 1: Why do the points $(0, 0)$ and $(1, 1)$ lie on the Lorenz curve?

Problem 2: Explain why the line segment from $(0, 0)$ to $(1, 1)$ should be called the "equal distribution line".

Problem 3: Why does the Lorenz curve lie below the equal distribution line? For instance, why can't $(.25, .75)$ lie on the Lorenz curve?

The Lorenz curve describes the distribution of income among the members of the community, but how can you compare, for instance, the income distribution of the United States to that of Mexico? In 1912, the Italian statistician Corrado Gini proposed a way to describe the distribution of income by a single number. In Figure 2 the 45-degree line is labeled as (part of) the graph of the identity function $I(x) = x$ and the Lorenz curve is the graph of some function we'll call $L(x)$. The region of the plane between these two curves is labeled by Γ, the Greek capital letter "G", because it was the region of interest to Gini. Since the region beneath the 45-degree line is a triangle with height and base equal to one, and therefore its area is $\frac{1}{2}$, we can see that the area of Γ is no greater than $\frac{1}{2}$.

Figure 2

Problem 4: What distribution of income would make the area of Γ equal $\frac{1}{2}$?

Gini's idea was that the area of Γ would be a single number that could describe the extent of income inequality within the population. A large area would mean that the distribution is quite unequal because the Lorenz curve swoops up rapidly towards the right, indicating that the individuals with the highest incomes receive most of the total income, as in Figure 3.

Figure 3

On the other hand, a Lorenz curve near the equal distribution line, as in Figure 4, represents a distribution of income with little variability between the lowest and highest incomes. The greater the area of Γ, the greater the inequality of incomes.

Figure 4

In order to present the area of Γ as a proportion of the possible area, that is $\frac{1}{2}$, Gini divided the area by $\frac{1}{2}$, which multiplies it by 2, so it is on a scale that runs between 0 and 1. The result came to be called the "Gini index" so, formally,

$$\text{Gini index} = 2 \cdot \text{area}(\Gamma).$$

The Gini index permits us to compare income distributions among the countries of the world. The data is presented in the "World Fact Book" Web site that is maintained by the Central Intelligence Agency (CIA) of the United States:

https://www.cia.gov/library/publications/the-world-factbook/ rankorder/2172rank.html

Here are some of the Gini indices we find there:

Brazil .52; China .47; Dominican Republic .46; Germany .27; Mexico .48; Russia .42; United States .45

The numbers run from the mid .20s, for instance Sweden is .25, to the low .60s, for instance South Africa is .63. It is worth noticing that the communist economy of China and the capitalist economy of the United States appear to lead to similar distributions of income and that the much poorer country than either of them, the Dominican Republic, distributes incomes in much the same way. The relative wealth of a country seems to have no correlation with income distribution: Haiti shares an island, and also low incomes, with the Dominican Republic, but its distribution is considerably more unequal as described by a Gini index of .59. Incomes are also low in Albania, yet its index is of .27 suggests a much more equal distribution.

The Gini index can also be used to plot trends in income distribution within a single country. Here are some figures for the Gini index of the United States since the middle of the last century:

1950: .38; 1970: .39; 1980: .40; 1990: .43; 2000: .46; 2006: .47; 2011: .48

and they appear to support the claims of the Occupy demonstrators.

The Gini index calculations in the CIA World Factbook are based on extensive statistical data, either calculated or estimated depending on what is available from each country. However, even with very limited information, we can approximate the value of the Gini index. It's a little easier to do it in a somewhat indirect way, as illustrated by Figure 5.

Figure 5

The region below the Lorenz curve, which we have labeled with Λ, the Greek "L", that is, what is "left over" after taking away the Gini region Γ. We can calculate the Gini index

$$G = 2 \cdot area(\Gamma)$$

if we know the area of Λ because

$$area(\Gamma) + area(\Lambda) = \frac{1}{2}$$

and therefore

$$G = 1 - 2 \cdot area(\Lambda).$$

We can make a rough estimate of the Gini index even from a single observation. Suppose it is estimated that, in some country, the top 20 percent of income earners receive 60 percent of the total income for that country. Thus the other 80 percent of the population shares the remaining 40 percent of the income and we know that the point (.80, .40) lies on the Lorenz curve. Since (0, 0) and (1, 1) also lie on that curve, we'll connect these three points by line segments, as in Figure 6.

Figure 6

Problem 5: Given the Lorenz curve consisting of two line segments in Figure 5, calculate the area of Λ and use that information to compute the Gini index.

In general, if we know that the proportion a of the lowest earners receives a proportion b of the total income, that means that the points (a, b) lies on the Lorenz curve and we can approximate that curve by line segments as in Figure 7.

Figure 7

Problem 6: Calculate the area of Λ in Figure 6 and then write a general formula for this "one-point-estimate" of the Gini index in terms of a and b.

Income distributions that are qualitatively very different can have the same Gini index.

Problem 7: Describe two income distributions that have very different Lorenz curves and yet the same Gini index. The one-point-estimates can give you a good approximation to their shapes.

We can get a more accurate estimate of the Gini index if we know two points on the Lorenz curve, for instance $(.8, .4)$ and $(.99, .8)$. In Figure 8 we connected the four points of the Lorenz curve by line segments assuming we knew points (a, b) and (c, d).

Figure 8

Problem 8: Calculate the area of Λ in Figure 7 to show that the general formula for this "two-point-estimate" of the Gini index is $ad - bc + c - d$.

Problem 9: Use the answer to Problem 8 to make a two-point-estimate of the Gini index if $(a, b) = (.8, .4)$ and $(c, d) = (.99, .8)$.

Problem 10: Explain how it could happen that the income of everyone in a country increases and yet the Gini index also increases.

The Gini index can also be used to measure the distribution of wealth among the population of a country. While a person's (annual) "income" is the amount of money they receive during the year, their "wealth" is the value of things that they own. Thus people's paychecks and stock dividends count as income whereas the value of their cars, if they own them, and their equity in their houses are part of their wealth. Since people who are rich tend to have many valuable possessions whereas poor people possess little, it should not surprise us that inequality of wealth is much greater than inequality of income and therefore the Gini index for wealth is usually a highe number than the Gini index for income of the same population.

Problem 11: According to Wikipedia, the richest 2 percent of Americans, in terms of wealth, own 87.7 percent of the total wealth of the country. Use this information to make a one-point-estimate of the Gini index for wealth in the United States.

Problem 12: Wikipedia also states that the richest 1 percent of Americans own 37.1 percent of the total wealth. Combine this information with that of Problems 8 and 11 to make a two-point-estimate of the Gini index for wealth in the United States.

In theory, the Lorenz curve is constructed by ordering all the people in a country from the individual with the lowest income to the person with the highest income. Let x_n be the n-th person in this ordering and $f(x_n)$ that person's income. The point on the Lorenz curve corresponding to x_n would then be $(x_n, L(x_n))$ where

$$L(x_n) = f(x_1) + f(x_2) + \cdots + f(x_n).$$

We can assume no person has zero income since even if they live by charity there is some value to what they receive. Thus we have $f(x_{n+1}) > f(x_n) > 0$ for all n. If we extend the points $(x_n, f(x_n))$ to the graph of a smooth function $f(x)$ for $0 \leq x \leq 1$, as Max Lorenz did back in 1912, then $f(x)$ is strictly increasing and thus the derivative of $f(x)$ is positive for all x. Lorenz then extended the definition of the function L in the corresponding way by defining $L(x)$ for $0 \leq x \leq 1$ to be the area under the graph of $f(x)$ from 0 to x. The Fundamental Theorem of Calculus then tells us that the derivative of $L(x)$ is $f(x)$ and therefore the second derivative of $L(x)$ is the derivative of $f(x)$ which, as we pointed out, is always positive. Therefore, the function $L(x)$ is concave up, as in Figure 9.

Figure 9

Problem 13: If we base our calculation of the Gini index on a one-point-estimate, will that number always be larger or always smaller than the correct value of the Gini index, or will it sometimes be larger and sometimes smaller?

A "mathematical model" of the Lorenz curve is a function $L(x)$ whose graph traces out the points of the curve. A simple candidate for such a model is the function $L(x) = x^p$ for $p > 2$, for instance $L(x) = x^4$ as in Figure 10. The function $L(x) = x^p$ for $p > 2$ has the properties $L(0) = 0$ and $L(1) = 1$ and $L'(x) = p(x^{p-1})$, which could be a model of the income function $f(x)$. Also, $L''(x) = p(p-1)x^{p-2} > 0$ so the graph is concave up, as required.

Figure 10

Problem 14: For what value of p would the point $(.8, .4)$ lie on the graph of $L(x) = x^p$?

Problem 15: For what value of p would the point (a, b) lie on the graph of $L(x) = x^p$?

The region we called Λ is the area under the curve $L(x)$ so, by the other part of the Fundamental Theorem of Calculus, can be calculated by

$$area(\Lambda) = \mathcal{L}(1) - \mathcal{L}(0)$$

where $\mathcal{L}(x)$ denotes the anti-derivative of $L(x)$.

Problem 16: Use the anti-derivative of $L(x) = x^p$ to find a formula, in terms of p, for the Gini index.

Problem 17: Show that the value of the Gini index based on the estimate of the value of p from Problem 14 is much higher than the Gini index for for income of .48 in the United States.

Even granting that a one-point-estimate is very inaccurate, the answer to Problem 17 is so far off, it suggests that the function $L(x) = x^p$ is not a good model for the Lorenz curve and in fact it is not used for that purpose. A somewhat better model is

$$L(x) = qx + (1 - q)x^p$$

where $p > 2$ and $0 < q < 1$. In Figure 11 we see the case $p = 4$ and $q = \frac{1}{2}$.

Figure 11

Problem 18: Show that this formula for $L(x)$ has the properties $L(0) = 0, L(1) = 1, L'(x) > 0$ and $L''(x) > 0$.

We can have a computer prepare a table for various values of the power p and a single reference point (a, b) in order to select the parameters p and q that determine this form of the function $L(x)$.

Problem 19: Show that if p and (a, b) are given, then the value of q is determined.

Problem 20: Write the formula for $L(x)$ in terms of p, a and b.

Problem 21: Write the formula for the Gini index in terms of p, a and b.

I thank Dillon Zhi, who prepared the figures, and Donald Laackman for testing some of this material in the Los Angeles Math Circle. Also, I thank Olga Radko for giving me the opportunity to contribute to this volume.