12-2010

The Gini Index and Measures of Inequality

Frank A. Farris
Santa Clara University, ffarris@scu.edu

Recommended Citation

"FARRIS, Frank A. "The Gini Index and Measures of Equitability,"American Mathematical Monthly, December 2010: 851-864."

 DOI:10.4169/000298910X523344

The Gini Index and Measures of Inequality

Frank A. Farris

Abstract. The Gini index is a summary statistic that measures how fairly a resource is distributed in a population; income is a primary example. In addition to a self-contained presentation of the Gini index, we give two equivalent ways to interpret this summary statistic: first in terms of the percentile level of the person who earns the *average dollar*, and second in terms of how the lower of two randomly chosen incomes compares, on average, to mean income.

1. INTRODUCTION. You hear anecdotes all the time: The poorest 20% of the people on Earth earn only 1% of the income. A mere 20% of the people on Earth consume 86% of the consumer goods. Only 3% of the U.S. population owns 95% of the privately held land.

The Gini index offers a consistent way to talk about statistics like these. A single number that measures how equitably a resource is distributed in a population, the Gini index gives a simple, if blunt, tool for summarizing economic data. It allows us to illustrate how equity has changed in a given situation over time, such as how U.S. family income changed over the 20th century. (The poor got poorer over the second half.) We can also compare income or wealth across societies, and even analyze salary structures of organizations.

Being only a single summary statistic, the Gini index has been critiqued by social scientists [**2**]. It is true that no summary statistic can reveal all we need to know about the distribution of income, wealth, or land. Even so, the Gini index deserves to be better known in the mathematical community, as it continues to find application in new situations, from genetics [**7**] to astronomy [**1**].

In addition to a self-contained presentation of the Gini index, we give two equivalent ways to interpret this summary statistic: first in terms of the percentile level of the person who earns the *average dollar*, and second in terms of how the lower of two randomly chosen incomes compares, on average, to mean income. The first of these appears to be new; the second has appeared in the literature [**11**], but does not seem to be well known. Beyond the inherent mathematical interest, our story draws attention to the concept of inequity and offers readers tools to help them go beyond the factoids of the first paragraph.

2. DEFINING THE GINI INDEX. Though it is named for Italian statistician Corrado Gini (1884–1965), the Gini index can almost be glimpsed in the diagrams from a 1905 paper by M. O. Lorenz [**12**]. Gini's original work on the subject appeared in 1912 in Italian [**8**]; it is not easy to access. Fortunately, the paper by Lorenz is quite charming to read and gives an excellent historical snapshot of the seeds of this train of thought. The first sentence is memorable:

> There may be wide difference of opinion as to the significance of a very unequal distribution of wealth, but there can be no doubt as to the importance of knowing whether the present distribution is becoming more or less unequal.

doi:10.4169/000298910X523344

Let us define a *Lorenz curve*, the instrument Lorenz proposed for visualizing the distribution of a quantity in a population. Suppose that some quantity Q, which could stand for wealth, income, family income, land, food, and so on, is distributed in a population. If we imagine the population to be lined up by increasing order of their shares of Q (with ties being broken arbitrarily), then for any p between 0 and 1 the people in the first fraction p of the line represent the Q-poorest $100p\%$ of the population. We then call $L(p)$ the fraction of the totality of Q owned (or earned or controlled or eaten) by that fraction of the population. In summary:

> The Lorenz curve for a resource Q is the curve $y = L(p)$, where the Q-poorest fraction p of the population has a fraction $L(p)$ of the whole.

Using this vocabulary, the first sentence of the paper would be expressed as $L(.20) = .01$, where L is the Lorenz curve for world income. The variable p is called the *percentile variable*.

If everyone had exactly the same amount of Q, the order of our imaginary line-up would be completely arbitrary and we would say that $L(p) = p$, the curve of *perfect equitability*. In other situations where some fraction of the population all share equally in an amount of Q, our rule for an arbitrary order of that portion of the line results in a linear segment of the Lorenz curve. For instance, if everyone in the bottom half of the population owned an equal share of $1/4$ of the wealth, we would say that $L(p) = p/2$ for $0 \le p \le 1/2$, so that $L(1/2) = 1/4$. A purist might say that, in a population of N individuals, it only makes sense for p to take on values of the form k/N. In practice, we model Lorenz curves as being defined for all p, using linear interpolation whenever necessary. This requires us to say, for instance, that the poorest 10% of an individual earns 10% of that person's income, which is not too much of a stretch.

The Gini index is a quantity calculated from a particular Lorenz curve. It is defined as an integral that summarizes how much the Lorenz curve in question deviates from perfect equitability:

$$G := 2\int_0^1 [p - L(p)]\,dp. \tag{1}$$

The formula reveals why the Gini index sometimes appears in calculus books in the section on the area between two curves. The reason for the factor 2 is to scale the area in such a way that the Gini index varies between 0, perfect equitability where everyone has the same share of the good, and 1, where one person has everything.

I downloaded data for 2006 family income from the U.S. Census Bureau [**14**]. Modulo some details, described in another section, it is easy to use a spreadsheet program to create a Lorenz curve from the data and estimate the Gini index. This is shown in Figure 1.

The Gini index for this situation works out to be about .47, which agrees with the figure reported by the U.S. Census Bureau [**13**]. For perspective, the similar indices for Brazil and Denmark are about .58 and .24 respectively [**4**]. Does this fit with what you know about these societies? It is also instructive to compare Gini indices over time: The index for U.S. family income hit its 20th century low of around .36 in 1968. Does this match your understanding of social change in America over the last several decades?

The CIA World Factbook [**4**] reports that the Gini index of U.S. family income (for 2007) is .45, and other sources claim figures even below .40. These lower figures tend to be *adjusted* indices, taking income adjustments into account. For instance, in the

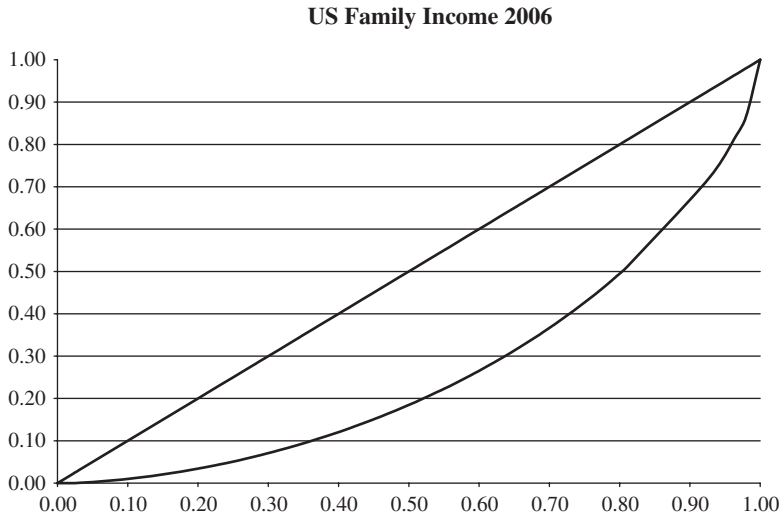

Figure 1. The Gini index is twice the area between the Lorenz curve and the curve of perfect equitability. For U.S. family income in 2006, the data leads to an estimate of $G \approx .47$.

U.S., the tax structure, which asks higher income citizens to pay a higher percentage of income so that various benefits (food stamps, Medicare, etc.) may be given to lower income residents, does in fact reduce inequity, but this feature of U.S. society would be missed if only raw income figures were used to compute a Gini index. When comparing Gini indices, we should take care to specify exactly what is being measured. We will side-step any deep analysis of this point in favor of mathematics, simply warning readers to be careful to compare Gini indices only when they are computed in the same way.

The Gini index need not be a grim topic. When I was learning how to calculate indices, Alex Rodriguez had just been hired by the N.Y. Yankees for an annual salary of $22 million. I downloaded the salary schedules for the Yankees and the Red Sox from the ESPN website and determined that the salary Gini for the Yankees during that season was .57, noticeably higher than for the more equitable Red Sox, at .52.

I found a contrasting example in the exam scores of my calculus students. Even with more than half the students below a mean of 84%, the Gini index for distribution of exam points was only .06. Does this mean that my grading is especially fair?

3. CONNECTING TO PROBABILITY. As we move toward calculating Gini indices, we must accept that economic data are almost always reported in aggregated form. Except in fanciful applications like salaries of baseball players where we have an income figure for each individual, we get tables of data where one column lists the (very large) number of people in a given range of incomes and another gives the mean income for this group. Income ranges are listed in convenient order from lowest to highest. A truncated version of the data appears in Table 1.

At the ends of the spectrum of 2006 U.S. family income, there are 2,533 thousand households with mean income $295 and 2,240 thousand households with mean income $448,687.

In general, let us name the entries in any such table as h_j units (households) that, on the average, have an amount x_j of our good Q (income). If our table has n rows, then j ranges from 1 to n and the order of the table means that $x_j < x_k$ when $j < k$.

Table 1. U.S. family income, 2006, aggregated.

Number of households (in thousands)	Average income
2,533	295
1,030	3,737
2,124	6,431
3,002	8,713
3,677	11,206
3,203	13,668
3,677	16,088
3,169	18,646
3,886	21,056
3,005	23,690
⋮	⋮
13,385	119,461
4,751	169,454
1,776	219,377
2,240	448,687

As a first step in calculating a Gini index from such a table, let us consider how to express the Lorenz curve, $L(p)$. First, we define numbers

$$N = \sum_{i=1}^{n} h_i \quad \text{and} \quad T = \sum_{i=1}^{n} x_i h_i.$$

In words, N is the size of the population and T is the total amount of the good Q. With this notation, the mean amount owned is T/N, which we call $\overline{X}$.

In our example of 2006 U.S. family income, N is about 116 million households, T is almost 8 trillion dollars, and the mean income is $\overline{X} \approx \$66{,}570$. The first entry in the table corresponds to a percentile value of $h_1/N \approx .02183$, the poorest 2% of the population. In general, the numbers

$$p_j = \frac{1}{N} \sum_{i=1}^{j} h_i$$

give us n (not necessarily equally spaced) points along the p-axis between 0 and 1. For convenience, we define $p_0 = 0$.

The Lorenz curve is easily calculated for these particular values of p: $L(0) = 0$ and

$$L(p_j) = \frac{1}{T} \sum_{i=1}^{j} x_i h_i, \quad 1 \le j \le n, \tag{2}$$

because this is the fraction of the total earned by the poorest fraction p_j.

The simplest way to fill in the rest of the Lorenz curve is by linear interpolation. When we do this, we are assuming that every one of the h_j units with mean amount x_j has an equal share in that amount. As we examine in more detail later, this always results in an *underestimate* of the Gini index, though a small one if our aggregation uses a fine partition.

It is a little cumbersome to calculate the Gini index (1) by direct integration from (2). Instead, we will reinterpret that equation, uncovering a surprising connection to probability density functions.

With a little algebra, we rewrite (2) in terms of the percentile variable and recognize the result as a Riemann sum:

$$\begin{aligned} L(p_j) &= \sum_{i=1}^{j} \frac{x_i}{T/N}(p_i - p_{i-1}) \\ &= \int_0^{p_j} s(p)\,dp, \end{aligned} \tag{3}$$

where

$$s(p) = \frac{x_j}{\overline{X}}, \quad \text{for } p_{j-1} < p \le p_j. \tag{4}$$

We call the function $s(p)$ the *share density*, because it tells us what share of the whole is owned by the portion of the population that falls in a given percentile range. In our example, the income of the poorest 2% is about 0.0044 of the mean income, while the group at the top, above the 99th percentile, have a share that is about 6.74 times the mean.

It seem reasonable to use (3) to define the Lorenz curve for every value of p, not just the numbers p_j, which accomplishes the linear interpolation we spoke of before. In fact, we prefer to think of the share density as the primary object here, from which details of the Lorenz curve can be derived. One result of this approach is that the nondecreasing nature of $s(p)$ establishes $L(p)$ as a convex function. For instance, the following inequalities demonstrate midpoint convexity:

$$\int_{p_1}^{\frac{p_1+p_2}{2}} s(\tilde{p})d\tilde{p} \le \int_{\frac{p_1+p_2}{2}}^{p_2} s(\tilde{p})d\tilde{p}, \quad \text{so } L\left(\frac{p_1+p_2}{2}\right) \le \frac{L(p_1)+L(p_2)}{2}.$$

We acknowledge that aggregation of data always leads to some error in using (3) to define the Lorenz curve, and hence in the Gini index from which it is computed. We address this briefly in a later section. It has also been treated widely in the economics literature. Indeed, the concept of share density is implicit in the work of Gastwirth from 1971 [**5**].

It may happen that a Lorenz curve for a given situation is known, or proposed theoretically as a function of some particular type [**6**]. In such a case, we could define the share density, perhaps only almost everywhere, as

$$s(p) = \frac{d}{dp}\,L(p). \tag{5}$$

Since $s(p) \ge 0$ and $\int_0^1 s(p)\,dp = L(1) = 1$, the function $s(p)$ fits the requirements of a *probability density function* (pdf). What experiment would lead to a random variable that has this pdf? We propose the following:

Pick a dollar earned by a U.S. household at random, assuming that every dollar is equally likely to be chosen. Record the value of the percentile variable, p, of the person who earned that dollar.

For this experiment, p is a random variable with density $s(p)$. To see this, look at (4). For instance, the probability that a dollar chosen at random was earned in the percentile range from p_{j-1} to p_j is exactly the fraction of those dollars in proportion to the whole, which is

$$\frac{x_j h_j}{T} = \frac{x_j}{T/N}(p_j - p_{j-1}) = \int_{p_{j-1}}^{p_j} s(p)\,dp.$$

Also, we note that a share density of $s(p) \equiv 1$ would indicate a perfectly equitable distribution, in which case each dollar has an equal chance of being earned in all percentiles.

The share density earns its keep from the following computation, in which we substitute the integral form of the Lorenz curve into the definition of the Gini index (1) and then switch the order of integration:

$$\begin{aligned} G &= 2\int_0^1 [p - L(p)]\,dp \\ &= 1 - 2\int_0^1 \int_0^p s(\tilde{p})\,d\tilde{p}\,dp \\ &= 1 - 2\int_0^1 (1-\tilde{p})s(\tilde{p})\,d\tilde{p} \\ &= 2\int_0^1 p\,s(p)\,dp - 1. \end{aligned} \tag{6}$$

This last integral is simply the expected value of our random variable with density $s(p)$. We use $\overline{p}$ for this expected value and call it the *percentile of the average dollar earned*. This proves a theorem that gives our first interpretation of the Gini index:

Theorem 1. *Suppose G is the Gini index associated with the Lorenz curve $L(p)$ and the share density is defined by $s(p) = L'(p)$ almost everywhere. Let $\overline{p}$ be the expected value of the random variable on* $[0, 1]$ *whose density function is $s(p)$. Then G and $\overline{p}$ are related by*

$$G = 2\overline{p} - 1 \quad \textit{and} \quad \overline{p} = \frac{G+1}{2}. \tag{7}$$

Let us apply this to our examples: The average dollar earned in the U.S. in 2006 was earned at a percentile level of $(.47 + 1)/2$, or above the 73rd percentile. For the Yankees, with salary Gini .57, the average dollar comes in above the 78th percentile. In my opinion, this gives a more visceral fact to share with the general public than just the value of an index that ranges between 0 and 1.

Calculating Ginis. Interpreting the Gini index in terms of the average dollar earned is also key to calculations. To compute $\overline{p}$, we break up the integral into pieces where

$s(p)$ is constant:

$$\begin{aligned}\overline{p} &= \int_0^1 p\, s(p)\, dp \\ &= \sum_{j=1}^{n} \frac{x_j}{\overline{X}} \int_{p_{j-1}}^{p_j} p\, dp \\ &= \sum_{j=1}^{n} \frac{x_j}{\overline{X}} \frac{p_j + p_{j-1}}{2} \Delta p_j, \end{aligned} \tag{8}$$

where, as usual, $\Delta p_j = p_j - p_{j-1}$. This form is ideal for entering into a spreadsheet and it was this version that produced the values reported for the examples.

We can rearrange (8) as

$$G = \left(\frac{1}{T} \sum_{j=1}^{n} x_j (p_j + p_{j-1}) h_j \right) - 1, \tag{9}$$

for the Gini index of the Lorenz curve obtained by linear interpolation of the data. With some work, this last expression for G can also be derived directly from the definition of the Gini index by applying the trapezoid rule to find the area under the piecewise linear Lorenz curve.

4. THE LOWER OF TWO INCOMES. Consider this experiment: Pick two households in the U.S. and record the lower of their two incomes; call the result Y, a random variable that takes values in $[0, \infty)$. An amusing computation shows that the expected value of Y, in ratio to the mean income, is the complement of the Gini index relative to 1. In symbols,

$$\frac{\overline{Y}}{\overline{X}} = 1 - G. \tag{10}$$

To prove this, we need to know the pdf for Y. This in turn requires the pdf for X, the random variable recording the income of a single household. This pdf is not so directly available from the data, as presented in government statistics. That data could be interpreted as requiring point probability masses placed at each aggregated mean income. In other words, we could place a point mass of $2{,}533/N$ at income $X = \$295$ and a point mass of $2{,}240/N$ at income $X = \$448{,}687$. This would be both inaccurate and mathematically cumbersome.

Instead, let us work theoretically, assuming a known piecewise smooth density function $f(x)$ $(0 \le x < \infty)$ for the random variable X. Knowing f allows us to compute the cumulative density function (cdf) for X:

$$F(x) = \mathcal{P}(X < x) = \int_0^x f(\tilde{x})\, d\tilde{x}.$$

A standard computation in order statistics gives a cdf for Y, $H(x)$, as follows:

$$\begin{aligned} H(x) &= \mathcal{P}(Y < x) = 1 - \mathcal{P}(Y > x) \\ &= 1 - \mathcal{P}(\text{first income} > x) \cdot \mathcal{P}(\text{second income} > x) \\ &= 1 - (1 - F(x)) \cdot (1 - F(x)). \end{aligned}$$

This gives the pdf for Y as $H'(x) = 2f(x)(1 - F(x))$, since $F'(x) = f(x)$ (almost everywhere). Therefore, the expected value of Y is

$$\overline{Y} = \int_0^\infty 2xf(x)(1 - F(x))\,dx. \tag{11}$$

Let us connect this to the Gini index. The percentile variable p is easily related to the cdf for X. For a specific value of x, the probability that an income chosen at random is less that x is exactly the size of the fraction of the population earning less than x. In symbols, this means that

$$F(x) = \int_0^x f(\tilde{x})d\tilde{x} = p,$$

which means that we can express x in terms of p and p in terms of x. The share density $s(p)$ is simply $x/\overline{X}$, recording the proportion of the mean income at percentile level p.

We interpret (11) in terms of the variable p, using the substitution $p = F(x)$, $dp = f(x)\,dx$.

$$\begin{aligned}\overline{Y}/\overline{X} &= \int_0^\infty 2\frac{x}{\overline{X}} f(x)(1 - F(x))\,dx \\ &= \int_0^1 2s(p)(1 - p)\,dp \\ &= 2 - 2\overline{p} = 2 - 2\frac{G+1}{2} = 1 - G.\end{aligned} \tag{12}$$

This manipulation allows us to interpret any known Gini index in an approachable, conversational way: Assuming that the Gini index for U.S. family income is .47, we conclude that the lower of two U.S. family incomes, chosen at random, is about 53% of the mean; on the average, the poorer of two families earns only about half the national mean.

5. GINI ESTIMATES. Gini indices, in practice, must be computed from incomplete data. It helps to have some idea of the errors introduced when we infer a Gini index from partial data. This topic has received extensive treatment in the economics literature [**6**]; our self-contained treatment is meant to be accessible to mathematicians. One important way in which we depart from practical concerns is that we consider aggregate data, as for example in Table 1, as representing the exact averages for each group reported. In other words, we do not consider the reality that this data contains reporting errors. In this section, we make statements about the conclusions that can be drawn from the given data, assuming that it is accurate.

We begin with the case where our knowledge is limited to a single nontrivial point on a Lorenz curve. This is the situation of the first sentences of the paper. For instance, knowing that 20% of the people on Earth consume 86% of the consumer goods, what can we say about the Gini index?

Proposition 1. *If G is the Gini index associated with a Lorenz curve $L(p)$ and we know that $L(a) = b$, where $0 < b < a < 1$, then*

$$a - b \le G < 1 - 2b(1 - a). \tag{13}$$

Smaller upper bounds are possible, but less easy to state. For instance,

$$\text{if} \quad b < \frac{1}{2} < a, \quad \text{then} \quad G \leq 1 - 4b(1-a).$$

Proof. The least possible area between $y = p$ and $y = L(p)$ occurs when $L(p)$ is the piecewise linear function shown in Figure 2, which simply connects $(0, 0)$ to (a, b) and then to $(1, 1)$. In that case, the Gini index is easily calculated to be $1 - (ab + (1-a)(1+b)) = a - b$.

The rather blunt upper bound arises from geometric considerations. In Figure 2, the area of the b by $1 - a$ rectangle in the lower right corner is clearly off limits. The first estimate follows by removing twice this area from the highest possible Gini of 1.

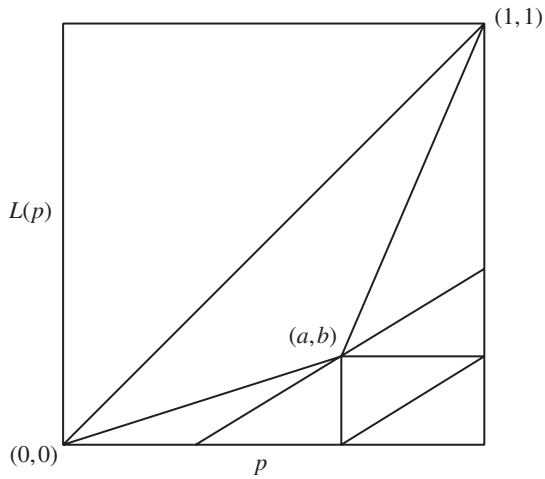

Figure 2. Estimates for the Gini index obtained from a single data point are possible, but not especially accurate.

The better estimate comes from recognizing that Lorenz curves must be convex. Extreme behavior arises from linear functions, at the edge of convexity; some readers may recognize that we are talking about *support lines* for the Lorenz curve.

To get the largest possible Gini index, we should cut the smallest possible area from below $y = x$. It turns out that among all lines through (a, b), the one that forms the smallest triangle (together with the x-axis and the line $x = 1$) is parallel to the diagonal of the b-by-$(1 - a)$ rectangle mentioned earlier. The only issue is whether this line crosses $y = x$. As long as $b < \frac{1}{2} < a$, this line cuts a triangle of area $2b(1 - a)$ from the large triangle, resulting in a Gini index of $1 - 4b(1 - a)$. (Alert readers may recognize that we are talking about a discontinuous Lorenz curve with $\lim_{p \to 1^-} L(p) < 1$, which requires the share density to have a point mass at $p = 1$. This models a situation where a very small portion of the population has a very large share of Q.) ■

An economic interpretation of this proposition holds some interest. The lowest possible Gini estimate comes from assuming that a fraction b of the good Q is distributed absolutely equally through the poorest fraction a of the population, with the remaining portion shared equally among the remainder.

For instance, when we conclude from the estimate about consumer goods (where $L(.80) = .14$) that $.66 \leq G$, the value .66 would arise from the poorest 80% sharing equally in 14% of the goods—an unlikely situation.

The highest possible Gini index consistent with our single data point is $1 - 4(.14)(.2) = .888$. This corresponds to a distribution under which 60% have no goods at all, 40% have an equal share in 28%, and one person has all the remaining 72%—again, unlikely. (As mentioned earlier, this requires the share density to have a point mass at $p = 1$.) The range of the estimate is wide, but we can still say that consumer goods are less equitably distributed than U.S. family income.

Golden [**9**] offers a related estimate, based on knowledge of the Lorenz curve at a point that is known to be farthest from the line of perfect equity, in which case the relevant support line has slope 1. This is done in the special case of quintile data. Readers may wish to derive this estimate as an exercise.

Our simple estimate becomes more useful when we apply the same geometric ideas to individual summands in our trapezoid rule approximation for the Gini index, (9). Let us use G_T to denote the Gini index for the Lorenz curve obtained by linear interpolation of aggregated data and find estimates for the Gini index for the situation where we have one data point for every individual in the population. In each term of the sum, the simplest estimate we can make is that the Gini index could increase by twice the area of a triangle with base $p_j - p_{j-1}$ and height $L(p_j) - L(p_{j-1})$. Working out formulas for these gives a potential positive contribution to the error of $x_j h_j^2/(TN)$. We have proved a proposition:

Proposition 2. *If G_T is the approximation for a Gini index from* (9)*, the actual index, G, satisfies*

$$G_T \leq G < G_T + \frac{1}{TN}\sum_{j=0}^{n} x_j h_j^2. \tag{14}$$

In our computations for U.S. family income, the error is bounded by 0.042 and we conclude that $.467 \leq G < .509$.

For a more precise upper bound on the Gini index, we now focus on one particular interval in which mean data has been given. In terms of the notation established earlier, we are talking about the population between p_{j-1} and p_j, where at first we assume that $1 < j < n$, leaving discussion of the first and last intervals for later.

Over this interval, the Lorenz curve has slope s_j, which is intermediate between the slopes on the left, s_{j-1}, and right, s_{j+1}. This is shown in Figure 3 where segments are labeled with their slopes. The most extreme Lorenz curve that still connects points $(p_{j-1}, L(p_{j-1}))$ and $(p_j, L(p_j))$ consists of the two line segments shown, continuing the line of slope s_{j-1} as far as the point where it meets the line of slope s_{j+1}. Call the p-coordinate of this point of intersection p^*.

To understand this in economic terms, recall that in this portion of the population, h_j households have an average income of x_j, which gives them a share density of $s_j = x_j/\bar{X} = x_j/(T/N)$. (Although our notation favors the interpretation of family income, the discussion applies to any situation.) To achieve the extreme Lorentz curve mentioned, we could redistribute the $h_j x_j$ dollars earned, taking away income from one group to push a fraction $p^* - p_{j-1}$ down one bracket to have share density s_{j-1}, pushing the remainder up to a share density of s_{j+1}. Note that this results in one fewer income bracket than the data original suggested.

The effect on the Gini index would be to increase it by twice the area of the triangle in Figure 3.

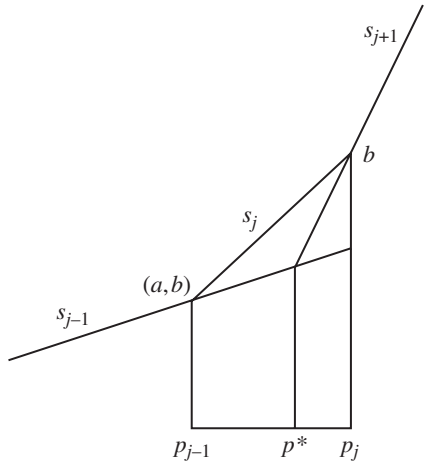

Figure 3. Better estimates for the Gini index involve an interval-by-interval analysis.

A simple computation, writing the area as the difference of two triangles whose base is labeled b in the diagram, shows the maximal increase in the Gini index to be

$$\frac{(s_j - s_{j-1})(s_{j+1} - s_j)}{(s_{j+1} - s_{j-1})}(p_j - p_{j-1})^2 = \frac{1}{NT}\frac{(x_j - x_{j-1})(x_{j+1} - x_j)}{(x_{j+1} - x_{j-1})}h_j^2. \tag{15}$$

At the lowest endpoint, there is no lower income bracket to push population into; in the highest bin, there is no possibility to redistribute upward, so the estimates work out a bit differently. Still, if we set $x_0 = 0$ and $x_{n+1} = x_n$, the formulas are correct.

Of course, pushing population from one bracket into neighboring brackets affects the analysis of those regions; we cannot capture this extra area in every interval while still maintaining a convex curve. Even so, we know that the Gini index cannot be increased by more than twice the sum of areas of all these triangles. We have sketched the proof of a theorem:

Theorem 2. *If a Lorenz curve is generated from aggregate data, where the jth bin consists of h_j individual units in possession of x_j units of the resource Q, then the actual Gini index G, the one that would result from analyzing every individual's portion, must satisfy*

$$G_T \leq G < G_T + \frac{1}{NT}\sum_{j=1}^{n}\frac{(x_j - x_{j-1})(x_{j+1} - x_j)}{(x_{j+1} - x_{j-1})}h_j^2, \tag{16}$$

where we set $x_0 = 0$ and $x_{n+1} = x_n$.

Applying this analysis to U.S. family income from 2006 gives $.467 \leq G < .472$, showing that .47 is the correct Gini index to two-digit accuracy. (Remember that this is a raw Gini index, ignoring the effect of taxes, Medicare, and Social Security.)

Efforts to find the best upper bound have shown this to be a complicated question. Experiments attempting to minimize total area using a variable support line at each data point suggest that the minimum is realized only at endpoints of the intervals of possible slopes, except in simple cases like the example with a single point. I conjecture that the largest possible Gini index consistent with data from $2n$ bins arises by

applying the redistribution method outlined above in alternate bins, pushing population into neighboring bins to create n new bins and a maximally less equitable distribution.

6. HIGHER ORDER GINIS. Atkinson [**2**] rightly pointed out in 1970 that the Gini index is no universal measure of society. Sometimes it helps little in judging whether one distribution of income is preferable to another. It is easy to give a mathematical reason for this: Many Lorenz curves give rise to the same Gini index.

Atkinson uses utility functions to weight income disparities, asking those who would judge inequity: Is extreme poverty more socially harmful than extreme wealth? Of course, no single utility function will serve all purposes.

My response on reading Atkinson was to place the Gini index as first in a family of indices, each weighting the percentile range differently. Introducing a weighting factor of $(1 - p)^{k-1}$ for $k \geq 1$ yields an index where extreme poverty is weighted more for higher values of k. In fact, this idea originally appeared in a paper by Kakwani in 1982 [**10**], and many other economists and social scientists have taken the ball and run with it.

We define the kth Gini index (or perhaps kth Gini poverty index) by the formula

$$G_k := k(k+1) \int_0^1 (p - L(p)) \cdot (1-p)^{k-1}\, dp.$$

The factor $k(k+1)$, analogous to the 2 in Gini's original definition, forces each index in the sequence to lie between 0 and 1. Note that G_1 is simply G from (1).

It is a simple matter to mimic earlier computations to produce a spreadsheet-friendly formula to approximate G_k from aggregated data. Probabilists may recognize that we are really talking about moments of the share density. An analog of (9) is

$$G_k = 1 - \frac{1}{\overline{X}} \sum_{j=1}^{n} x_j \left((1 - p_{j-1})^{k+1} - (1 - p_j)^{k+1} \right), \tag{17}$$

though the resemblance may not be immediately evident.

Using (17) with data from U.S. family income suggests that G_2 is about .61. What does this mean? To answer this, we return to order statistics.

As before, the random variable X denotes the income of a single household chosen at random; the pdf for X is $f(x)$ and the cdf is $F(x)$. Now we independently choose $k+1$ households and record the *lowest* of their incomes as $Y_k^{\min}$. A computation just like the one that led to (11) shows that the expected value of $Y_k^{\min}$ is

$$\overline{Y}_k^{\min} = (k+1) \int_0^\infty x(1 - F(x))^k f(x)\, dx.$$

The reasoning that led us to (12) in this case gives

$$\frac{\overline{Y}_k^{\min}}{\overline{X}} = 1 - G_k. \tag{18}$$

If the second-order Gini index of U.S. family income is $G_2 = .61$, this means that, on average, the lowest of *three* incomes randomly selected is only 39% of the mean income.

We should mention that (18), though new to the author, appears in a paper by Kleiber and Kotz [**11**]. We hope that MONTHLY readers find this presentation easier to follow than anything in the extensive economics literature on the subject.

7. CLOSING THOUGHTS. Higher-order Gini indices can be useful in calibrating models. Social scientists (and authors of calculus texts) often model Lorenz curves with a variety of *Pareto* functions, which are convex combinations of power functions, such as

$$L(p) = ap + (1 - a)p^b.$$

Since this model has two free parameters, a and b, it is natural to calibrate it to match values of G_1 and G_2 derived from data. The degree to which this model fits the data can be judged by the difference between the value of G_3 calculated from the model and the one calculated from the data using (17).

Though it arose in the study of poverty, the Gini index is a flexible idea that deserves to be better known. As mentioned in the introduction, scientists in many fields [**1, 7**] have found occasion to apply the Gini index. These are certainly not the only examples, and the reader may enjoy finding others.

A search of the literature turns up many efforts to explain away the Gini index as inaccurate or incomplete. It seems to me that none of these objections should prevent us from using the Gini index to analyze data for ourselves and share the results with those we know. It matters to me to know a few summary statistics, though they be mere summary statistics, and to know how to relate them to average outcomes of thought experiments about who earns the average dollar and about the poorer of two households.

Our country is rich with diverse opinions about what one ought to do about economic facts, but perhaps we are not sufficiently armed with facts that we have checked for ourselves. I hope that this article will inspire you to dig into the mountains of data that are available and refine some summary statistics for yourself.

ACKNOWLEDGMENTS. The author is grateful for the support that the MAA provides to editors, allowing them to supervise a generous and thorough review process, which greatly improved this paper.

REFERENCES

1. R. Abraham, S. van den Bergh, and P. Nair, A new approach to galaxy morphology, I: Analysis of the Sloan digital sky survey early data release, *Astrophysical Journal* **588** (2003) 218–229. doi:10.1086/373919
2. A. B. Atkinson, On the measurement of inequality, *J. Econom. Theory* **2** (1970) 244–263. doi:10.1016/0022-0531(70)90039-6
3. ———, On the measurement of poverty, *Econometrica* **55** (1987) 749. doi:10.2307/1911028
4. *Central Intelligence Agency World Factbook*, available at https://www.cia.gov/library/publications/the-world-factbook/geos/us.html, accessed April 20, 2008.
5. J. L. Gastwirth, A general definition of the Lorenz curve, *Econometrica* **39** (1971) 1037–1039. doi:10.2307/1909675
6. ———, The estimation of the Lorenz curve and Gini index, *Rev. Econom. Statist.* **54** (1972) 306–316. doi:10.2307/1937992
7. D. Gianola, M. Perez-Enciso, and M. A. Toro, On marker-assisted prediction of genetic value: Beyond the ridge, *Genetics* **163** (2003) 347–365.
8. C. Gini, Variabilità e mutabilità; reprinted in *Memorie di Metodologica Statistica*, E. Pizetti and T. Salvemini, eds., Libreria Eredi Virgilio Veschi, Rome, 1955.
9. J. Golden, A simple geometric approach to approximating the Gini coefficient, *Journal of Economic Education* **39** (2008) 68–77. doi:10.3200/JECE.39.1.68-77
10. N. Kakwani, On a class of poverty measures, *Econometrica* **48** (1980) 437–446. doi:10.2307/1911106
11. C. Leiber and S. Kotz, A characterization of income distributions in terms of generalized Gini coeefficients, *Social Choice and Welfare* **19** (2001) 789–794. doi:10.1007/s003550200154
12. M. O. Lorenz, Methods of measuring the concentration of wealth, *J. Amer. Statist. Assoc.* **9** (1905) 209–219.

13. C. DeNavas-Walt, B. D. Proctor, and J. Smith, *Income, Poverty, and Health Insurance Coverage in the United States: 2006*, Current Population Report, U.S. Census Bureau, August, 2007.
14. U.S. Census Bureau, *Current Population Survey (CPS)*, available at `http://pubdb3.census.gov/macro/032007/hhinc/new06_000.htm`, accessed April 20, 2008.
15. B. H. Webster, Jr. and A. Bishaw, *Income, Earnings, and Poverty Data From the 2006 American Community Survey*, American Community Survey Report, U. S. Census Bureau, August, 2007.

FRANK A. FARRIS received his B.A. from Pomona College in 1977 and his Ph.D. from M.I.T. in 1981. He has taught at Santa Clara University since 1984 and recently finished a second stint as editor of *Mathematics Magazine*. His article "The Edge of the Universe" in *Math Horizons* was honored with the Trevor Evans Award.
Department of Mathematics and Computer Science, Santa Clara University, Santa Clara, CA 95053
ffarris@scu.edu