The Estimation of the Lorenz Curve and Gini Index

Author(s): Joseph L. Gastwirth

Source: *The Review of Economics and Statistics*, Aug., 1972, Vol. 54, No. 3 (Aug., 1972), pp. 306-316

Published by: The MIT Press

Stable URL: https://www.jstor.org/stable/1937992

THE ESTIMATION OF THE LORENZ CURVE AND GINI INDEX

Joseph L. Gastwirth *

I Introduction

MOST of the measures of income inequality are derived from the Lorenz curve; indeed Morgan (1962) states that the Gini index is the best single measure of inequality. The present article reviews some of the theoretical properties of the Lorenz curve, relates them to characteristics of the frequency function underlying the income distribution and develops methods for obtaining accurate bounds on the Gini index which do not depend on curve fitting. In the process we should also like to lay to rest some myths concerning the Gini index such as: (a) its relative insensitivity (Éltetö and Frigyes, 1968), (b) difficulty in computation (1968), and (c) problems related to the inclusion of negative incomes (Budd, 1970).

The basic idea of our approach is to obtain upper and lower bounds to the Gini index from data which are grouped in intervals and the mean income in each interval is known. The usual method (Morgan, 1962) of estimating the Gini index yields a lower bound by assuming that all incomes in any interval equal the average income. We derive an upper bound to the grouping correction (Goldsmith, et al., 1954, p. 10) and hence to the Gini index by distributing the income to maximize the spread within each group. On the 1967 Internal Revenue Service tax data, the difference between our bounds is less than 0.006. As most income distributions come from a frequency function (density) which decreases in the large income range, we develop improved bounds for the Gini index based on this assumption. Fortunately, this assumption can be checked from the data so that we can use the sharper bounds only for the appropriate intervals. Using this second method the difference between our bounds is $\leq .002$. Because Soltow (1965) detects a change in the Gini index of 0.8 of one per cent or about 0.003 or 0.004, our bound seems quite adequate. In section VI we extend our method to obtain upper and lower curves for the Lorenz curve.

After reviewing the basic properties of the Lorenz curve we proceed to derive bounds on the mean difference and Gini index. In section IV we analyze an actual sample and show that the method used by the Census Bureau (1967) often leads to estimates which are outside the mathematically possible bounds we derived. Finally, in an appendix we show that the Pareto law does not give a good fit to current United States tax data.

II Properties of the Lorenz Curve and Associated Measures of Inequality

Given a set of n ordered numbers, $x_1 \leq x_2 \leq \ldots \leq x_n$, the empirical Lorenz curve generated by them is defined at the points i/n, $i = 0, \ldots, n$, by $L(0) = 0$ and $L(i/n) = s_i/s_n$, where $s_i = x_1 + \ldots + x_i$. The empirical Lorenz curve, $L(p)$, is defined for all p in the interval (0,1) by linear interpolation and represents the fraction of the total variable measured (e.g., income) that the holders of the smallest p^{th} fraction possess.

For theoretical purposes it is useful to consider the numbers x_i as a sample drawn from a distribution function $F(x)$. Throughout this article we shall assume that $F(x)$ increases on its support (the values of x for which $0 < F(x) < 1$) and the mean μ of $F(x)$ exists. The first assumption implies that $F^{-1}(p)$ is well defined and is the population p^{th} quantile. Given any degree of freedom (d.f.) $F(x)$, the theoretical Lorenz curve corresponding to it is defined by

$$L(p) = \mu^{-1} \int_0^p F^{-1}(t)\, dt. \qquad (1)$$

In table 1 we present a short table of the Lorenz curves generated by several common distributions.

Received for publication November 11, 1970. Revision accepted for publication February 23, 1972.

* Research supported by the National Science Foundation Research Grant No. GP-20527 awarded to the Department of Statistics, The Johns Hopkins University.

TABLE 1. — THE LORENZ CURVES GENERATED BY SOME COMMON DISTRIBUTIONS

Distribution	C.D.F.	Lorenz Curve
Equal	$F(x) = \begin{cases} 0, x < \mu \\ 1, x \geqq \mu \end{cases}$	$L(p) = p$
Exponential	$F(x) = 1 - e^{-\lambda x}, x > 0$	$p + (1-p)\ln(1-p)$
Shifted Exponential	$F(x) = 1 - e^{-\lambda(x-a)}, x > a$	$p + (1+\lambda a)^{-1}(1-p)\ln(1-p)$
General Uniform	$F(x) = \dfrac{x-a}{\theta}, a < x < a + \theta$	$\dfrac{ap + \theta p^2/2}{a + \theta/2}$
Pareto	$F(x) = 1 - (a/x)^{\alpha}, x > a, \alpha > 1$	$1 - (1-p)^{(\alpha-1)/\alpha}$

Two simple facts concerning the Lorenz curve are derivable from formula (1) and the fact that $F^{-1}(t)$ is nondecreasing. We state *Lemma 1.* Let $L(p)$ be the Lorenz curve corresponding to a d.f. $F(x)$. Then $L(p)$ is convex and its derivative, $L'(p)$, equals one at $p = F(\mu)$.

The most common measure of inequality, the Gini index G, is the ratio of the area between the Lorenz curve $L(p)$ and the 45° line (see figure 1) to the area under the 45° line (which is 1/2). The area, A, between the Lorenz curve and the straight line is called the area of concentration.

An alternative formula for the Gini index, G, is based on the mean difference, Δ, of the underlying d.f. $F(x)$ and is given in
Lemma 2: (Kendall and Stuart, 1963). The Gini index of the Lorenz curve $L(p)$ generated by a d.f. $F(x)$ is $\Delta/(2\mu)$, where

$$\begin{aligned} \Delta &= \int_{-\infty}^{\infty}\int_{\infty}^{\infty} |x-y|dF(x)dF(y) \\ &= 2\int F(x)[1-F(x)]dx \\ &= 4\int x[F(x)-1/2]dF(x). \qquad (2) \end{aligned}$$

The formula $G = \Delta/(2\mu)$ shows that the Gini index measures *relative inequality* as it is the ratio of a measure of dispersion, the mean difference, to the average value (μ). Other measures are the coefficient of variation σ/μ, and half the relative mean deviation $\delta/2\mu$, where $\delta = \int |x-\mu|dF(x) = 2\int_{\mu}^{\infty} (x-\mu)dF(x)$ is the mean deviation. The relative mean deviation (δ/μ) is related to several other measures of relative inequality which we now review.

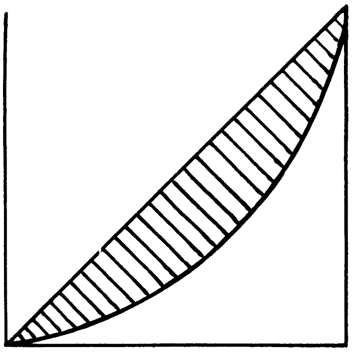

FIGURE 1. — A LORENZ CURVE (THE SHADED AREA IS THE AREA OF CONCENTRATION)

The area A is the area under the curve $p - L(p)$. As $L(p)$ is convex, $p - L(p)$ is concave and vanishes at $p = 0$ or 1. Thus, there is a value p' (called the point of *maximum discrepancy* between the Lorenz curve and line of equality) satisfying

$$p' - L(p') \geqq p - L(p) \quad \text{for all } p. \qquad (3)$$

The relevance of the point of maximum discrepancy is stated in
Lemma 3: The point p' equals $F(\mu)$ and the value of the maximum discrepancy, $p' - L(p')$, equals $\delta/(2\mu)$.

The lemma shows that p' is the fraction of the population receiving less than the "average" while the value of the maximum discrepancy, $p' - L(p')$, is half the relative mean deviation. In 1951, Shutz proposed to measure inequality by comparing the derivative, $L'(p)$, of the Lorenz curve to the derivative (one) of the line of total equality. His measure, S, is the area between $L'(p)$ and 1 in the region $(0,p')$, which reduces to $\delta/(2\mu)$.

The measure $\delta/(2\mu)$ was proposed by Yntema (1933) and Pietra in the 1930's. Pietra measured inequality by the area of the largest triangle that could be inscribed in the area of concentration. The ratio of the area of the Pietra triangle to the area $(1/2)$ under the 45° line of "perfect equality" equals half the relative mean deviation. Yntema's measure essentially is $p' - L(p')$ or $\delta/(2\mu)$. More recently Éltetö and Frigyes (1968) proposed related measures based on the Lorenz curve. We now study the estimation of the Gini index in detail as the same ideas are extended in section VI to derive bounds for the entire Lorenz curve from which one can obtain bounds on the other measures.

III Useful Inequalities for the Mean Difference and Gini Index

In this section we derive several bounds on Δ and G. By assuming weak properties of the d.f. $F(x)$ or its derivative $f(x)$, the density function, we can often obtain bounds which we use in the next section to estimate the Gini index.

Our first result gives a general bound on the mean difference for any d.f. $F(x)$ supported on a finite interval (a,b). From formula (3), inequality 105 in Hardy et al. (1952) and the fact that $F(x)$ increases, one can prove

Lemma 4: For any d.f. $F(x)$ supported on (a,b) with mean μ,

$$0 \leq \Delta \leq 2(\mu-a)(b-\mu)/(b-a). \qquad (4)$$

For intervals which are "open ended," using the second formula for Δ in (2) and the fact that $F(x) \leq 1$, one can derive

Lemma 5: If $F(x)$ is a d.f. supported on $[a, \infty]$ with finite mean μ, then $\Delta \leq 2\ (\mu - a)$.

Remark: For the Pareto law with parameter α, $\Delta/2\mu = (2\alpha-1)^{-1}$ and approaches one as α approaches one from above, so the bound is sharp.

Both the upper and lower bounds of *Lemma 4* can be strengthened if one is willing to assume that the density function decreases, i.e., the d.f. $F(x)$ is concave, in the interval (a,b). A bound which is derived from a result of Gauss (Kendall and Stuart, 1963, p. 92) is stated in

Lemma 6: Let $F(x)$ be a concave d.f. supported on (a,b); then

$$\Delta \geq (4/3)(\mu-a)^2/(b-a). \qquad (5)$$

A further improvement can be obtained by showing that the uniform distribution on any interval of the form $(a, 2\mu-a)$ has the smallest mean difference in the class of all concave d.f.'s supported in (a, ∞) with mean μ. In order to establish this result we require a modification of a result of Chow and Studden (1969).

Lemma 7: Let h and g be two nonincreasing functions on the real line such that

$$\int_0^\infty h(x)dx = \int_0^\infty g(x)dx \qquad (6)$$

and suppose that $h-g$ changes sign once from plus to minus, i.e., t_0 such that

$$[h(t)-g(t)](t-t_0) \leq 0 \text{ for all } t. \qquad (7)$$

Then, for any concave function ϕ,

$$\int \phi[g(t)]dt \geq \int \phi[h(t)]dt. \qquad (8)$$

Our main application is

Theorem 1: Let $F(x)$ be a concave d.f. supported on an interval (a, ∞) and let $F(x)$ have mean μ. Let $F_0(x)$ be the uniform d.f. on $(a,\ 2\mu - a)$ with the same mean μ as F has. Then the mean difference of $F(x)$ is greater than or equal to the mean difference, $2(\mu - a)/3$, of $F_0(x)$.

Proof: The result follows once it is shown that

$$\int_0^\infty F(x)[1-F(x)]dx \geq \int_0^\infty F_0(x)[1-F_0(x)]dx, \qquad (9)$$

where $F_0(x) = 0$ when $x < a$, $(x-a)/(2\mu-a)$ when $a \leq x < 2\mu-a$ and one when $x \geq 2\mu-a$. In *Lemma 7*, set $g(x) = 1-F(x)$ and $h(x) = 1-F_0(x)$. Since $F_0(x)$ and $F(x)$ have the same mean μ, condition (19) holds. As $F(x)$ is concave, g is convex and $h(x)$ is a straight line so that h crosses g exactly once from above. The function $\phi(t) = t(1-t)$ is concave so that (8) implies that

$$(1/3)(\mu-a) = \int F_0(t)[1-F_0(t)] \leq \int F(t)[1-F(t)]dt. \qquad (10)$$

By analogous methods we can obtain

Theorem 2: If $F(x)$ is concave on (a,b), then

$$2/3\ (\mu-a) \leq \Delta \leq 2(\mu-a)(b-a)^{-1} [(b-\mu) - 1/3\ (\mu-a)]. \tag{11}$$

It should be noted that in contrast to Theorem 1 the upper bound (11) depends on the finiteness of the interval (a,b). When $F(x)$ is convex on (a,b), the mean difference is bounded by

$$(2/3)(b-\mu)(b-a)^{-1} \leq \Delta \leq 2/3\ (b-\mu)(b-a)^{-1} [4(\mu-a) - (b-a)]. \tag{12}$$

For the open ended interval (a, ∞) the upper bound for Δ given in *Lemma 5* cannot be improved. Consideration of densities with a decreasing hazard rate permits strengthening the lower bound of Theorem 1. As the Pareto law and the tails of the lognormal (Barlow and Proschan, 1965) and Fisk-Champernowne Law (Fisk, 1961 and Champernowne, 1952) have this property, it is a reasonable assumption. We recall the *Definition*: A d.f. $F(x)$ supported on (a,∞) has the decreasing hazard rate (D.H.R.) property if $-\log[1 - F(x)]$ is concave for $x \geq a$. When the density function $f(x) = F'(x)$ exists, then the function

$$q(x) = f(x)[1-F(x)]^{-1} \tag{13}$$

is nonincreasing.

Using *Lemma 5* of Hardy et al., (1952) one can prove

Theorem 3: If $F(x)$ is a d.f. on (a, ∞) with the D.H.R. property, density $f(x)$ and finite mean μ, then

$$(\mu-a) \leq \Delta \leq 2(\mu-a). \tag{14}$$

IV Estimation of the Gini Index

Given a Lorenz curve $L(p)$ the standard method (Census, 1967 and Morgan, 1962) of estimating the Gini index is to approximate the area of concentration by choosing k fractiles $0 = p_0 < p_1 < p_2 < \ldots < p_k < p_{k+1} = 1$ and computing the area of the polygon with vertices $(0,0)$, $(p_1,L(p_1)), \ldots, (p_k,L(p_k))$ and $(1,1)$. This procedure leads to an under-estimate of both A and G since the straight line connecting $(p_i,L(p_i))$ to $(p_{i+1},L(p_{i+1}))$ lies above the convex curve $L(p)$. Thus, the standard procedure yields the following lower bound for G:

$$G \geq 1 - \sum_{i=0}^{k} (p_{i+1} - p_i)[L(p_i) + L(p_{i+1})]. \tag{15}$$

In order to assess the accuracy of (15) we need an upper bound for G. As the Lorenz curve is convex, its derivative increases and by constructing the tangents to the curve at the points $(p_i,L(p_i))$ we can bound the curve from below. This approach is developed in section VI. For our purposes an approach based on the formula $G = \Delta/(2\mu)$ is more convenient.

Given n ordered numbers which are grouped (preserving the ordering) into $(k+1)$ subgroups

$$x_1, \ldots, x_{m_1};\ x_{m_1+1}, \ldots, x_{m_2};\ \ldots;\ x_{m_k+1}, \ldots, x_n,$$

where $m_1 = np_1$, $m_2 = np_2, \ldots, m_k = p_k n$ and $0 < p_1 < p_2 < \ldots < p_k < p_{k+1} = 1$, then the empirical mean difference Δ^* of the original numbers equals (Yntema, 1933)

$$\Delta^* = \binom{n}{2}^{-1} \sum_{i<j}\sum |x_i - x_j| = \sum_{i\neq j}\sum \gamma_i\gamma_j|\mu_i - \mu_j| + \sum_{i=1}^{k+1} \gamma_i^2\, \Delta^*_i, \tag{16}$$

where μ_i is the mean of the i^{th} group and Δ^*_i is the mean difference of the i^{th} group, and γ_i is the proportion of observations in the i^{th} groups (i.e., $\gamma_1 = p_1$, $\gamma_2 = p_2 - p_1, \ldots, \gamma_{k+1} = 1 - p_k$). Of course, $G = \Delta^*/(2\mu)$. Formula (16) shows that the mean difference of the original numbers can be regarded as the sum of the "mean difference between groups" and a correction term which weights the mean difference (Δ^*_i) *within* each group by the factor γ_i^2.

When all the observations in each of the $(k+1)$ groups are equal, the Gini index reduces (after some algebra) to the lower bound (15). Thus, the standard method of estimating the Gini index neglects the differences in income within the groups and underestimates G by

$$D = (2\mu)^{-1} \sum \gamma_i^2\, \Delta^*_i. \tag{17}$$

The factor D is known as the "grouping correction" (Goldsmith et al., 1954) and almost all the interpolation formulas attempt to estimate it by fitting a specific density in each interval. The *main idea* of our approach is to obtain bounds on D under *minimal* assumptions on the density function in each interval. Let all the observations in the i^{th} group lie between a_{i-1} and a_i and let μ_i be the group mean. *Lemma 4* yields bounds on Δ^*_i, the mean difference of the observations in the i^{th} group, *regardless of the form of the underlying density.* Thus, the Gini index always lies between

$$GL = (2\mu)^{-1} \sum_{i \neq j} \sum \gamma_i \gamma_j |\mu_i - \mu_j| \leq G \leq GL + \bar{D} = GU \qquad (18)$$

where

$$\bar{D} = \mu^{-1} \sum_{i=1}^{k+1} \gamma_i^2 (\mu_i - a_{i-1}) \times (a_i - \mu_i)(a_i - a_{i-1})^{-1}. \qquad (19)$$

One interesting application of formula (17) is its use in designing the grouping intervals needed to obtain a desired degree of accuracy. Since the theoretical p^{th} quantile of the population is $F^{-1}(p)$, setting $a_i = F^{-1}(p_i)$ one can use (17) to determine the choice of fractiles $\{p_i\}$ or population quantiles $\{a_i\}$ which minimize the "grouping correction." An example of this type of result is

Proposition 1: If the underlying d.f. $F(x)$ is the uniform distribution on an interval (a,b), then the optimum bounds using k-fractiles $(0 = p_0 < p_1 < \ldots < p_k < 1)$ or $(k+1)$ groups are achieved when $p_i = i/(k+1)$ and

$$D = (1/2)\frac{(b-a)}{(b+a)}\frac{1}{(k+1)^2}. \qquad (20)$$

The above result implies that the standard practice of using quintiles or deciles, i.e., *equally spaced* fractiles $\{p_i\}$, is *not* an *optimal* choice for income distributions which typically are skewed to the right.

In practice the number of groups required to achieve close bounds on G is rather large (at least 20) as the group boundaries (a_i) are not chosen with the purpose of minimizing D or $\bar{D}$. So far, no prior knowledge concerning the shape of the income distribution has been used. As most frequency functions that have been fit to income data decrease in the high income range, we can sharpen the bounds on the "within group mean differences," Δ^*_i, by using the bound (derived from (11))

$$(2/3)(\mu_j - a_{i-1}) \leq \Delta^*_i \leq 2(\mu_i - a_{i-1})(a_i - a_{i-1})^{-1} [(a_i - \mu_i) - 1/3\,(\mu_i - a_{i-1})] \qquad (21)$$

on the intervals (a_{i-1}, a_i) on which the density decreases. Finally, one can test the assumption that the frequency function decreases in (a_{i-1}, a_i) by requiring that $\mu_i < (1/2)(a_{i-1} + a_i)$ and that

$$n_{i-1}(a_{i-1} - a_{i-2})^{-1} > n_i(a_i - a_{i-1})^{-1} > n_{i+1}(a_{i+1} - a_i)^{-1}, \qquad (22)$$

where n_i is the number of observations in the interval (a_{i-1}, a_i). In the last interval (a_k, ∞) one may be willing to use bounds obtained from Theorem 3.

TABLE 2. — C.P.S. INCOME DISTRIBUTION IN TEN GROUPS

Income Interval (thousands of dollars)	Per Cent of Families	Per Cent of Income	Mean Income (dollars)
0–1⁻	4.824	.323	541.41
1–2⁻	8.253	1.492	1,463.63
2–3⁻	7.215	2.179	2,445.72
3–4⁻	6.902	2.931	3,438.90
4–5⁻	6.615	3.625	4,437.32
5–6⁻	7.598	5.068	5,401.18
6–7⁻	7.847	6.195	6,392.92
7–10⁻	21.404	21.950	8,304.54
10–15⁻	19.111	28.094	11,904.33
over 15	10.241	28.154	22,261.50

In order to judge our method we tested it on data, presented in table 2, given to us by Dr. Benjamin Tepping of the Census Bureau. He computed the exact Gini index for the CPS sample (1968) and formed two groupings (into 10 and 28 subgroups) of the data.

The Gini index computed by Dr. Tepping from the entire sample of approximately 60,000 incomes was 0.4014. We computed the Gini index using three different procedures. Method 1 used the crude bounds (Mendershausen, 1946). No matter what the underlying density is, the Gini index of the given numbers *must* lie between these bounds. Method 2 tests for decreasing density and replaces the crude bound on the within group mean differences used in Method 1 by Soltow (1960) where appropriate. The third method was the same as Method 2 except that we assumed that the

TABLE 3. — ESTIMATES AND BOUNDS ON THE GINI INDEX

	Data in 10 Groups		Data in 28 Groups	
Procedure Method No.	GL	GU	GL	GU
1	.3883	.4083	.4001	.4020
2	.3928	.40605	.4005	.40175
3	.3975	.40605	.40055	.40175
4	.4009	.4009	.40525	.40525

TABLE 4. — "ESTIMATES" OF THE GINI INDEX FROM IRS DATA (POSITIVE INCOME)

Year	Number of Groups	Method 1		Method 2		Method 4
		GL	GU	GL	GU	
1955	26	.4256	.4283	.4266	.4278	.4329
1956	25	.4254	.4281	.4265	.4276	.4349
1957	25	.4253	.4280	.4263	.4275	.4346
1958	25	.4307	.4335	.4318	.4330	.4407
1959	25	.4351	.4379	.4363	.4374	.4454
1960	25	.4336	.4366	.4351	.4360	.4442
1961	29	.4412	.4433	.4421	.4429	.4465
1962	29	.4401	.4422	.4411	.4417	.4481
1963	29	.4423	.4443	.4433	.4438	.4501
1964	18	.4440	.4492	.4465	.4482	.4670
1965	18	.4504	.4559	.4530	.4549	.4747
1966	19	.4536	.4596	.4565	.4584	.4734
1967	19	.4574	.4638	.4608	.4624	.4785
1968	21	.4622	.4686	.4659	.4670	.4721
1969	21	.4597	.4669	.4638	.4651	.4669

density had a decreasing hazard rate in the last interval. The fourth method studied (The Census Bureau's) does not use the means of each group but assumes that the mean income of each group is at the midpoint of the interval and fits a Pareto-tail to the last (open-ended) interval. In table 3, we present the results of computing our bounds on Dr. Tepping's data using both the 10 group decomposition and another grouping into 28 intervals.

From table 3, it is seen that the first three of our methods give bounds which bracket the true value (0.4014). Moreover, with a large number of groups the interval obtained by Method 1, $0.4001 < G < 0.4020$, is of sufficient accuracy to detect small changes in the Gini index. When the data was grouped in 10 intervals the bounds derived from Method 1 differed by 0.02; however, the bounds using Methods 2 and 3 reduced this difference to 0.01325 and 0.00855, respectively. The value of using some extra assumptions is apparent if the data is coarsely grouped.

One interesting result of our study is the rather poor performance of Method 4 which does not use the individual group means. That estimate was more accurate in the case of 10 groups than in the case of 28 groups. Indeed, the estimated Gini index using 28 groups lies outside the bounds given by Method 1 and is, therefore, an impossible value.

As a next step we included negative incomes in our study and used 29 intervals. As long as the average income of the population is positive, this causes no difficulty in the computation of our bounds. Using Method 1, the Gini index of the Census Bureau's data was bounded by $0.4024 < G < 0.4039$ while Method 2 yielded the bounds $0.4027 < G < 0.4037$. The inclusion of negative income increases both sets of bounds by about 0.002 which is slightly larger than the difference between the bounds.

V Analysis of Income Tax Data

The IRS summarizes income tax data by grouping the data into intervals and estimating the average income of each group. By estimating the Gini index directly and deriving bounds on the "grouping correction" we avoid the problem noted by Budd (1970): "Published size distributions give frequencies for dollar income size brackets that remain relatively constant from year to year; as a result, sizes and positions of the quantile readings for relative distributions derived from them vary considerably."

In table 4 we present an analysis of all income tax returns for the years 1955 thru 1969 which reported a positive adjusted gross income. The second column gives the number of grouping intervals wherein the data was summarized. Columns 3 and 4 give the upper and lower bounds for the Gini index using Method 1 while the bounds derived from Method 2 are given in columns 5 and 6. Column 4 gives the single estimate derived by Method 4.

The figures in table 4 show how the number of groups affects the accuracy of the estimated Gini index. Since the standard procedure is the lower bound (*GL*) of Method 1 we note that when the number of groups is large (25 or more), it underestimates the true Gini index by only about 0.002 or 0.003. When the number of groups used was 18 or 19, this estimate was short by about 0.005 or 0.006. Since Soltow (1960) was trying to detect a change in the Gini index of about one-half of one per cent a year, it appears that Morgan's suggestion that 8 groups would suffice (1962, p. 28) is not correct. The difference between the bounds derived by Method 2, however, always differed by less than 0.002 and usually one can distinguish between years. The years 1955 thru 1957 seem to have a "nearly constant" Gini index; however, since 1962 the Gini index appears to have increased year by year.

Another result of this analysis is the poor performance of Method 4 which always was larger than the upper bound given by Method 1. This occurs because the typical frequency function of income (Lydall, 1968) is unimodal, rising up to a modal value and decreasing thereafter. In the low-income range, the frequency function rises so the true group mean is greater than the midpoint of those intervals. The reverse is true for intervals past the modal income. Thus, the procedure underestimates the income of the low income groups and overestimates the income in the higher brackets, thereby overestimating the "relative inequality."

Finally we mention that Method 4 shows that the Gini index fell in 1968 and 1969 while our method shows the index didn't fall until 1969. Since the recent boom picked up steam in 1967–1968 it is reasonable that inequality increased (on IRS data) in 1968 as more low income recipients earned enough money to file returns and report their incomes.

In order to compare our bounds with the recent estimates of Budd (1970) we had to include negative incomes. In the computation of our bounds we used only the crude bound (14) on the within group mean difference of the negative income group. Before presenting the various bounds on IRS data we recall that Budd's estimates are based on interpolating a special kind of Lorenz curve and then computing the area of concentration. He fits a function which is a polynomial for $p \leq p_0$ (where p_0 is determined from the data and is usually in the range 0.7 to 0.8) and a function suggested by Gini (1936), which is equivalent to fitting a Pareto tail to the upper incomes.

TABLE 5. — "ESTIMATES" OF THE GINI INDEX FROM IRS DATA (NEGATIVE INCOME INCLUDED)

		Method 1		Method 2	
Year	Number of Groups	GL	GU	GL	GU
1955	27	.4350	.4377	.4360	.4372
1956	26	.4339	.4366	.4349	.4361
1957	26	.4343	.4370	.4354	.4366
1958	26	.4396	.4423	.4406	.4418
1959	26	.4463	.4491	.4475	.4486
1960	26	.4426	.4456	.4440	.4450
1961	30	.4498	.4519	.4508	.4515
1962	30	.4487	.4507	.4497	.4503
1963	30	.4519	.4539	.4528	.4534
1964	19	.4533	.4585	.4558	.4574
1965	19	.4586	.4641	.4612	.4630
1966	20	.4622	.4682	.4651	.4670
1967	20	.4655	.4719	.4688	.4705
1968	22	.4699	.4764	.4736	.4748
1969	22	.4678	.4750	.4718	.4732

It is interesting to contrast our results to those of Budd. His estimates of the Gini index for the years 1955, 1960, 1964, 1966 and 1967 were 0.435, 0.443, 0.461, 0.464 and 0.468. The most significant result is that the value for 1964 (0.461) lies outside the interval (given by Method 1) in which the Gini index must lie. The estimates for the other years lie inside the intervals obtained by Method 1 but not by Method 2.

Since our approach enables one to detect rather small changes in the Gini index without fitting curves and as the computer time to run our program for both Methods 1 and 2 for thirteen years of data was under eighteen seconds on the IBM 360, it should be of practical use.

VI The Comparison of Lorenz Curves

Given two Lorenz curves $L_1(p)$ and $L_2(p)$, the distribution generating $L_2(p)$ is more concentrated than the distribution generating $L_1(p)$ whenever $L_1(p) \leq L_2(p)$ for all p. This implies that the corresponding Gini indexes G_1 and G_2 satisfy $G_2 \leq G_1$. In this section we study conditions on the underlying d.f.'s F_1 and F_2 which allow us to conclude that one Lorenz curve lies above another. It turns out that the d.f.'s generating the extreme Gini indices also generate the extreme Lorenz curves. This allows us to generalize the methods of section 4 to obtain bounding curves to the true Lorenz curve for grouped data.

In order to compare the "relative spread" of two positive r.v.'s X_1 and X_2 with corresponding d.f.'s F_1 and F_2 and Lorenz curves L_1 and L_2 it is convenient to standardize them by requiring that their means be equal. Since all the proofs of the results of this section are based on the methods used in (Barlow and Proschan, 1965) we omit them and only state the generalizations of the results of section III.

Our first result generalizes *Lemma 4* and yields bounds on the Lorenz curve for arbitrary distributions. Specifically, we have

Theorem 4: Let $F(x)$ be a d.f. with mean μ and support (a,b). Then its Lorenz curve, $L(p)$, satisfies

$$B(p) \leq L(p) \leq p, \tag{23}$$

where

$$B(p) = \begin{cases} \mu^{-1}(ap) & , p < r \\ \mu^{-1}(ar) + \mu^{-1}b(p-r), & p > r \end{cases} \tag{24}$$

and r is determined by the relation $ra+(1-r)b = \mu$. The r.v. X generating the Lorenz curve $B(p)$ takes on the value a with probability r and the value b with probability $(1-r)$.

For densities which decrease on finite intervals the analogs of Theorems 1 and 2 are given by *Theorem 5*: Let $F(x)$ be a concave d.f. supported on (a,b) with mean μ and let $U(x)$ be the uniform density on $(a, 2\mu-a)$ and

$$B(x) = \begin{cases} 0, & , a \geqq x \\ 1 - \dfrac{2(\mu-a)}{(b-a)} + \dfrac{2(\mu-a)}{(b-a)}\dfrac{(x-a)}{(b-a)} & , a \leqq x < b. \\ 1, & , b \leqq x \end{cases} \tag{25}$$

Then the Lorenz curve generated by the d.f. $F(x)$ obeys

$$B(p) \leq L(p) \leq [ap + (\mu-a)p^2]\mu^{-1}, \tag{26}$$

where

$$B(p) = \begin{cases} \mu^{-1}ap & , \quad p \leq f \\ f\mu^{-1}a + (1-f)\mu^{-1}\mu^*\left[a\mu^{*-1}\left(\dfrac{p-f}{1-f}\right) + \mu^{*-1}(\mu^*-a)\left(\dfrac{p-f}{1-f}\right)^2 \right], & p > f \end{cases} \tag{27}$$

and

$$f = 1 - 2(\mu-a)/(b-a) \text{ and } \mu^* = (a+b)/2.$$

In order to generalize Theorem 3 we recall the concept of a "matching exponential distribution" for any d.f. F with D.H.R. supported on (θ,∞) with mean μ. This is the exponential density with the same mean and support as F. Its d.f. is

$$E(x) = 1 - \exp\{-(x-\theta)/(\mu-\theta)\},\ x > \theta, \tag{28}$$

and its Lorenz curve is

$$L(p) = p + (1-\theta\mu^{-1})(1-p)\ln(1-p). \tag{29}$$

The appropriate generalization of Theorem 3 is

Theorem 7: Let $H(x)$ be a D.H.R. law on (θ,∞) with density $h(x)$ and let $E(x)$ be its matching exponential. Then the Lorenz curve $L(p)$ generated by $H(x)$ satisfies $L(p) \leq L^*(p)$, where $L^*(p)$ is the Lorenz curve generated by $E(x)$.

Remark: When comparing a D.H.R. law with its matching exponential law it is essential that both the *origins* and the *means* of both d.f.'s be identical. Theorem 7 only gives an "upper bound" to the Lorenz curve. Unfortunately no good "lower bound" exists because the family of r.v.'s supported on $(0,\infty)$ with mean μ given by the d.f.'s

$$G_\epsilon(x) = 1 - \epsilon e^{-\alpha x},\ x > 0 \tag{30}$$

where $\alpha = \epsilon\mu^{-1}$, have the D.H.R. property. Each d.f. of this family corresponds to a r.v. which equals 0 with probability $1 - \epsilon$ and is an exponential r.v. with mean $\mu\epsilon^{-1}$ with probability ϵ. As ϵ approaches 0, most of the population receives no income while the small fraction, ϵ, receiving income get large incomes. Thus the Lorenz curves generated by $G_\epsilon(x)$ approach the most extreme Lorenz curve possible, namely, $L(p) = 0$ for $p < 1$ and $L(1) = 1$.

We now discuss how Theorems 4 and 5 can be used to obtain bounding curves to the true Lorenz curve in all but the open-ended group. As usual, assume that the data have been grouped into intervals where boundaries are given by $0 = a_0 < a_1 \ldots < a_k < a_{k+1}$, the mean income in the interval (a_{i-1},a_i) is μ_i, the number of incomes in the interval (a_{i-1},a_i) is n_i and the fractiles $p_i = F(a_i) = (n_1 + \ldots + n_i)/(n_1 + \ldots + n_{k+1})$ are used to estimate the Lorenz curve. Thus, we have $(k+1)$ values, $L(p_i)$, of the Lorenz curve. In the region (p_{i-1},p_i) corresponding to incomes in the interval (a_{i-1},a_i) Theorem 4 yields the upper boundary line

$$L(p_{i-1}) + [L(p_i) - L(p_{i-1})]\frac{p-p_{i-1}}{p_i-p_{i-1}} \qquad (31)$$

and, if the density can be assumed to decrease in (a_{i-1},a_i) Theorem 5 yields the upper boundary curve

$$L(p_{i-1}) + \frac{[L(p_i) - L(p_{i-1})]}{\mu_i} \times$$
$$\left\{ a_{i-1}\frac{p-p_{i-1}}{p_i-p_{i-1}} + (\mu_i - a_i)\frac{(p-p_{i-1})^2}{(p_i-p_{i-1})^2} \right\}. \qquad (32)$$

To obtain the corresponding lower boundaries to the Lorenz curve, we note that the lower bound given by Theorem 4 corresponds to giving a fraction f of the population income a_{i-1} and a fraction $(1-f)$ of income a_i where

$$f = 1 - (\mu_i - a_{i-1})(a_i - a_{i-1})^{-1}. \qquad (33)$$

Letting $p^*_i = p_{i-1} + f(p_i - p_{i-1})$, the Lorenz curve is bounded from below in $(p_i - p_{i-1})$ by the lines

$$L(p_{i-1}) + [L(p^*_i) - L(p_{i-1})]$$
$$(p-p_{i-1})(p^*_i - p_{i-1})^{-1},$$
$$p_{i-1} < p < p^*_i \qquad (34a)$$

and

$$L(p^*_i) + [L(p_i) - L(p^*_i)](p-p^*_i)(p_i-p^*_i)^{-1},$$
$$p^*_i \leqq p \leqq p_i \qquad (34b)$$

where $L(p^*_i) = L(p_{i-1}) + \mu^{-1}fa_{i-1}$ and μ is the mean income of the entire population, i.e., $\mu = \Sigma\, n_i\mu_i/\Sigma n_i$. If the density can be assumed to decrease in (a_{i-1},a_i) Theorem 5 implies that the most "unequal" distribution of income is generated by giving a fraction f of the population the lowest income possible, a_{i-1}, and assuming that the remaining income is distributed according to a uniform law on (a_{i-1},a_i). Then f satisfies the equation

$$f = 1 - 2(\mu_i - a_{i-1})/(a_i - a_{i-1}). \qquad (35)$$

Defining the interpolation point p^*_i now by $p^*_i = p_{i-1} + f(p_i - p_{i-1})$ the Lorenz curve in (p_{i-1},p^*_i) is bounded below by

$$L(p_{i-1}) + [L(p^*_i) - L(p_{i-1})]$$
$$(p-p_{i-1})(p^*_1 - p_{i-1})^{-1} \qquad (36a)$$

where $L(p^*) = L(p_{i-1}) + \mu^{-1}fa_{i-1}$. In the region (p^*_i,p_i), the underlying density is the uniform law with mean $\mu^*_i = (a_{i-1}+a_i)/2$ so Theorem 5 yields the lower boundary

$$L(p^*_i) + \frac{[L(p_i) - L(p^*_i)]}{\mu^*_i} \times$$
$$\left\{ a_{i-1}\frac{(p-p^*_i)}{(p_i-p^*_i)} + (\mu^*_i - a_{i-1})\frac{(p-p^*_i)^2}{(p_i-p^*_i)} \right\}. \qquad (36b)$$

Remarks: (a) The reader should note that the fraction f and the point p^*_i of interpolation are different in the two formulas (34) and (36). Indeed in the case of decreasing densities the f of (33) is larger than the f of (35). (b) The bounds on the Lorenz curve generated by (31) and (34) generate very general boundary curves. They once were drawn by Hanna et al. (1948) but because the number of groupings used was very small the entire idea seems to have been dropped. (c) The bounds given by (31) and (34) also allow us to derive bounds on the derivative or slope of the Lorenz curve which can be applied to obtain bounds on other measures of inequality. (d) The bounds obtained for those intervals where the density decreases, (32) and (36), are an improvement over the corresponding lines (31) and (34) as the curve (32) lies below (31) while (36) is above (34).

VII Conclusions and Future Problems

This paper shows that the Gini index can be accurately estimated without fitting curves to data whenever the data is grouped properly. Nevertheless, some future problems remain.

From a statistical viewpoint we need to assess the effect that estimating the group means has on our bounds. The IRS sample is so large we ignored this in our analysis.

For analytical purposes the variation in the concept of income measured may be severe (see Budd, 1970). We saw that the effect of including negative incomes was much greater on IRS data than on the Census data. This suggests that we might obtain a better picture of what is happening if we consider just wage incomes in industrial areas. Aggregate incomes from tax returns cover so many workers in so many incomparable jobs that they may not be as accurate an indicator as one would like. Indeed the Census Bureau (1967) found that the Gini index varies greatly among different occupations.

Finally it is a pleasure to thank Dr. Benjamin Tepping of the Bureau of the Census not only for providing data for our study but for his constant encouragement during the course of this investigation. Also I wish to thank Mr. J. T. Smith and Mr. David Kasik of The Johns Hopkins University who wrote the computer programs for all our procedures and tests. Their cooperation exemplified not only the idea that teaching and research are mutually related but that they are fun.

APPENDIX

Does the Pareto Law Fit United States Income Data?

Since the Census Bureau fits a Pareto law to the open-ended group, we made a comparison between the average income assigned to the over-$20,000 group by this procedure and the value estimated by the IRS from tax data. It appears that the Pareto fit was not bad as late as 1955 but is no longer appropriate. Our purpose is not to determine whether the Pareto law can be fitted to the "tail" of the income distribution but to justify the desirability of our approach which avoids fitting curves and estimating parameters.

Recall that if income is distributed according to a Pareto law on (A, ∞), the proportion, Q, of the total population with incomes $\geqq x$ is $(A/x)^{a}$ so that

$$a = \ln Q \,/\, (\ln A - \ln x) \qquad (37)$$

and the mean income of the group receiving at least x is $x\ a/(a-1)$.

In table 6, we present the estimated value of a and the mean income received by those earning at least $20,000 (i.e., $x = 20{,}000$) when the value A, specifying the starting point of the Pareto fit, is taken as $10,000 and $15,000. The last column is the actual mean of the group obtained from the IRS data. It is interesting that the estimates using $15,000 as the origin are slightly closer to the true value in recent years than the estimates using $10,000; however, the fit in recent years leaves a lot to be desired.

As another example of the arbitrariness of the Pareto fit to the tail of income data, we estimated a from the CPS-P-60 series No. 59 (April 18, 1969) data. We set $x = \$25{,}000$ and the origin A of the Pareto law at $12,000 and $15,000. When $A = \$12{,}000$, the estimate of a was 2.04316 which yielded an estimate of the average income of $48,965 in the group receiving at least $25,000. When $A = \$15{,}000$, the estimate of a was 2.38315 yielding an estimated group mean of $43,126 for those incomes greater than $25,000. This illustrates how sensitive a Pareto tail is to the *choice* of the *origin* of the *fit*. This problem is especially severe with economic data since the grouping intervals are not determined with the objective of fitting a curve to the last group. Thus, the "ideal origin" may be in the middle of a group.

TABLE 6. — THE PARETO FIT TO IRS DATA

$A=15,\ x=20$			$A=10,\ x=20$		Real Data
Year	Alpha	$\bar{X}$	Alpha	$\bar{X}$	
1967	2.7702	31298.00	2.7549	31396.74	37524.98
1966	2.7172	31646.61	2.8226	30973.69	37323.54
1964	2.4094	34190.23	2.7205	31624.68	37232.96
1963	2.5047	33292.03	2.7959	31136.73	36678.53
1961	2.2991	35395.74	2.6294	32274.78	38195.41
1960	2.2830	35588.41	2.6356	32228.01	37630.12
1959	2.1518	37363.87	2.5146	33205.04	38918.30
1957	1.9453	41158.25	2.2730	35710.69	38560.75
1956	1.7757	45784.49	2.1430	37498.46	39252.67
1955	1.7119	48094.72	2.0000	40000.00	39646.30

REFERENCES

Aitcheson, J., and J. A. C. Brown, *The Lognormal Distribution* (Cambridge: Cambridge University Press, 1957).

Barlow, R. E., and F. Proschan, *Mathematical Theory of Reliability* (New York: John Wiley and Sons, Inc., 1965).

Bowman, M. J., "A Graphical Analysis of Personal Income Distribution in the United States," *American Economic Review*, 35 (Sept. 1945), 606–628.

Budd, E. C., "Postwar Changes in the Size Distribution of Income in the United States," *American Economic Review*, 60 (May 1970), 247–260.

Bureau of the Census, *Trends in the Income of Families and Persons in the United States*, 1947–1964, Technical Paper No. 17 (U.S. Government Printing Office, 1967).

Champernowne, D. G., "The Graduation of Income Distributions," *Econometrica*, 20 (1952), 591–615.

Chow, Y. S. and W. J. Studden, "Monotonicity of the Variance Under Truncation and Variations of Jensen's Inequality." *Annals of Mathematical Statistics*, 40 (June 1969), 1106–1108.

Éltetö, Ö., and E. Frigyes, "New Income Inequality Measures as Efficient Tools for Causal Analysis and Planning," *Econometrica* 36 (Apr. 1968), 383–396.

Fisk, P. R., "The Graduation of Income Distributions," *Econometrica* 29 (Apr. 1961), 171–185.

Gini, C., "On the Measure of Concentration with Special Reference to Income and Wealth," Abstracts of papers presented at the Cowles Commission Research Conference on Economics and Statistics (Colorado Springs: Colorado College Press, 1936).

Goldsmith, S., G. Jaszi, H. Kaitz, and M. Liebenberg, "Size Distribution of Income Since the Mid-Thirties," this REVIEW, 36 (Feb. 1954), 1–32.

Hanna, F. A., J. A. Pechman, and S. M. Lerner, *Analysis of Wisconsin Income*. Conference on Research in Income and Wealth, 9 (New York: National Bureau of Economic Research, 1948).

Hardy, G. H., J. E. Littlewood, and G. Polya, *Inequalities* (Cambridge: Cambridge University Press, 1952).

Internal Revenue Service, *Statistics of Income: Individual Income Tax Returns* for 1955 thru 1967. (U.S. Government Printing Office).

Kendall, M. G., and A. Stuart, *The Advanced Theory of Statistics* 1. 2nd ed. (London: Charles Griffen and Company, 1963).

Liebenberg, M. and H. Kaitz, "An Income Size Distribution from Income Tax and Survey Data, 1944," in *Studies in Income and Wealth*, 13 (Cambridge: Riverside Press for NBER 1951), 443–444.

Lydall, H. F., *The Structure of Earnings* (Oxford: Clarendon Press, 1968).

Mendershausen, H., *Changes in Income Distribution During the Great Depression*. Studies in Income and Wealth, 7 (New York: H. Wolff for NBER 1946).

Morgan, J., "The Anatomy of Income Distribution," this REVIEW, 44 (Aug. 1962), 270–282.

Schutz, R. R., "On the Measurement of Income Inequality," *American Economic Review* (Mar. 1951), 107–122.

Soltow, L., "The Distribution of Income Related to Changes in the Distribution of Education, Age and Occupation," this REVIEW, 42 (Nov. 1960), 450–454.

———, "The Share of Lower Income Groups in Income," this REVIEW, 47 (Nov. 1965), 429–433.

Taguchi, T., "Concentration-Curve Methods and Structures of Skew Populations," *Annals of the Institute of Statistical Mathematics* 20 (1968).

Yntema, D., "Measures of the Inequality in the Personal Distribution of Wealth or Income," *Journal of the American Statistical Association* 28 (1933), 423–433.