

Calculus Challenge #9

Solutions due March 2, 2011

The Gini Index: Using Calculus to Measure Inequity

The data that economists use to quantify distribution of income is often presented in the form of Table 1. (see <http://www.census.gov/compendia/statab/2011/tables/1s0693.pdf>).

Fifth of Households	Percent of income	Fifths of Households	Percent of income
Lowest fifth	3.6	Lowest one-fifth	3.6
Second fifth	8.9	Lowest two-fifths	12.5
Third fifth	14.8	Lowest three-fifths	27.3
Fourth fifth	23.0	Lowest four-fifths	50.3
Highest fifth	49.8	Lowest five-fifths	100.0

Table 1: Percent distribution of aggregate income for 2000

Table 2: Cumulative percent distribution of aggregate income for 2000

The Gini Index

One measure employed by the economists is the ratio of the shaded areas *A* and *B* shown in Figure 4.

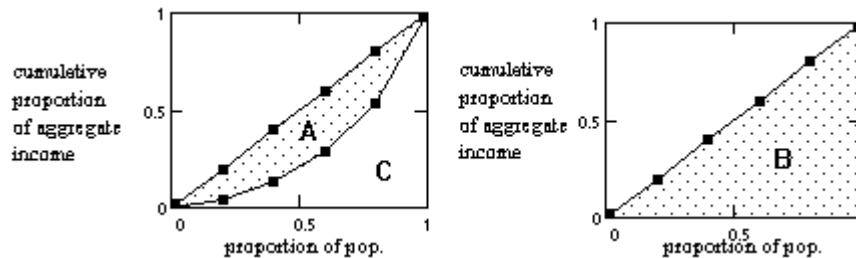


Figure 4: Areas to be computed

This ratio can have a value anywhere from 0, representing perfect equity, to 1, representing perfect inequity. The larger the ratio, the more inequitable the distribution of income. The area under the totally equitable distribution, *B*, is always one-half. To find the area of the shaded region *A*, we need to find the area between the line $y = x$ and the Lorenz curve.

Finding the Lorenz Curve using Least Squares

Since $(0, 0)$ and $(1, 1)$ are always points on the curves, a reasonable model for this data is a power function of the form $y = x^n$, with $n > 1$. We cannot use a power least squares procedure on our calculator to fit a power function to the data because a Lorenz curve must contain the point $(1,1)$ and a power least squares curve does not necessarily contain $(1,1)$. Also note that we only use the four coordinates $(0.2, 0.36)$, $(0.4, 0.125)$, $(0.6, 0.273)$, and $(0.8, 0.503)$ in our calculations, since, by using $y = x^n$ as our model, we guarantee that $(0,0)$ and $(1,1)$ fall on the curve.

a) Method 1 uses the fact that a log-log re-expression linearizes data that is modeled by a power function. Since $y = x^n$, we take the logarithm of both sides of the equation to obtain $\ln y = n \ln x$. We now can use our knowledge of calculus to find a least-squares estimate of n . Consider the linear equation $Y = nX$ (in our case $Y = \ln y$ and $X = \ln x$). Use the methods of calculus to minimize

$$S(n) = \sum_{i=1}^4 (Y_i - nX_i)^2 \text{ (remember } X_i \text{ and } Y_i \text{ are constants).}$$

Use this value of n for the Lorenz curve $y = x^n$ and find the ratio of area *A* to area *B* for the 2000 data. This represents the Gini index for 2000.

Finding the Lorenz curve:

Consider $Y = nX$. We want to minimize $S = \sum_{i=1}^4 (Y_i - nX_i)^2$, so $\frac{dS}{dn} = \sum_{i=1}^4 2(Y_i - nX_i) \cdot (-X_i)$ (remember,

we are differentiating with respect to n). If $\frac{dS}{dn} = 0$ then $\sum_{i=1}^4 X_i Y_i = n \sum_{i=1}^4 X_i^2$ and $n = \frac{\sum_{i=1}^4 X_i Y_i}{\sum_{i=1}^4 X_i^2}$.

Since $X_i = \ln(x_i)$ and $Y_i = \ln(y_i)$, we have $n = \frac{\sum_{i=1}^4 \ln(x_i) \cdot \ln(y_i)}{\sum_{i=1}^4 [\ln(x_i)]^2} \approx \frac{\sum_{i=1}^4 \ln(x_i) \cdot \ln(y_i)}{3.7406}$. The denominator

is a constant, since the values of x_i are the constants 0.2, 0.4, 0.6, and 0.8.

With the ordered pairs (0.2, 0.36), (0.4, 0.125), (0.6, 0.273), and

(0.8, 0.503), we have $n \approx \frac{\sum_{i=1}^4 \ln(x_i) \cdot \ln(y_i)}{3.7406} = 2.158$.

$$n \approx \frac{\sum_{i=1}^4 \ln(x_i) \cdot \ln(y_i)}{3.7406} = \frac{\sum_{i=0}^3 (\ln(x_i) \cdot \ln(y_i))}{3.7406} = 2.158$$

$x_i :=$	$y_i :=$
2	036
4	125
6	273
8	503

The Gini index is computed by finding the area between the Lorenz curve and $y = x$ and comparing this to 0.5, the area under the line of perfect equity. In general,

$$\text{Gini index} = \frac{\text{area bounded by Lorenz curve and } y = x}{\text{area of triangle for perfect equity}} = 2 \int_0^1 x - x^n dx = 1 - \frac{2}{n+1}.$$

Since we now have $n = 2.158$, we approximate the Gini index for 2000 as 0.367.

b) Another method is to fit a true least squares model. In this case, we have $S(n) = \sum (y_i - x_i^n)^2$.

What equation must be solved to find the value of n that minimizes S ? You will not be able to solve this equation analytically, so you will need to use a calculator or computer to find the value of n for 2000.

In this case, we want to differentiate $S = \sum (y_i - x_i^n)^2$

with respect to n , so $\frac{dS}{dn} = \sum 2(y_i - x_i^n)(-x_i^n) \ln(x_i)$

and we want to find the value of n for which

$0 = \sum (y_i - x_i^n)(-x_i^n) \ln(x_i)$. This equation requires

Newton's Method or similar numerical technique.

From the trace at right, we see the solution is

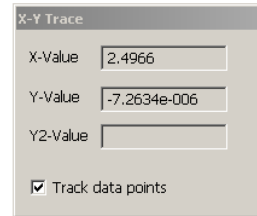
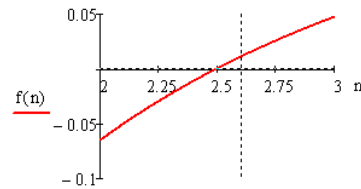
approximated by $n = 2.49665$. The Gini index is

again found by evaluating $1 - \frac{2}{n+1}$ at $n = 2.49665$, so

by this measure, the Gini index for 2000 is 0.428.

$x_i :=$	$y_i :=$	$0 = \sum (y_i - x_i^n)(-x_i^n) \ln(x_i)$
2	0.036	
4	0.125	
6	0.273	
8	0.503	$n := 2,2.0001..3$

$$f(n) := \sum_{i=0}^3 [(y_i - (x_i)^n) \cdot (-x_i) \cdot \ln(x_i)]$$



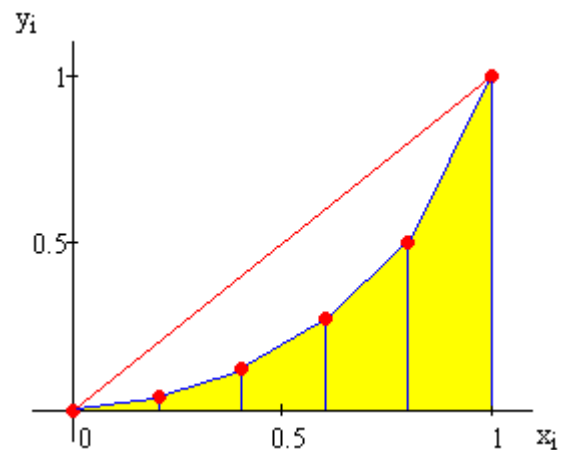
- c) Find the area using the trapezoidal rule with the points (0,0), (0.2,0.36), (0.4,0.125), (0.6,0.273), (0.8,0.503), and (1,1).

The area under the Lorenz curve can be approximated by area of the trapezoids. The Gini index can then be computed.

The area of the trapezoids is given by

$$0.2 \left(\frac{(0 + 2(0.036) + 2(0.125) + 2(0.273) + 2(0.503) + 1)}{2} \right)$$

This value is 0.2874, so the Gini index measured by this method is $\frac{0.5 - 0.2874}{0.5} = 0.4252$.

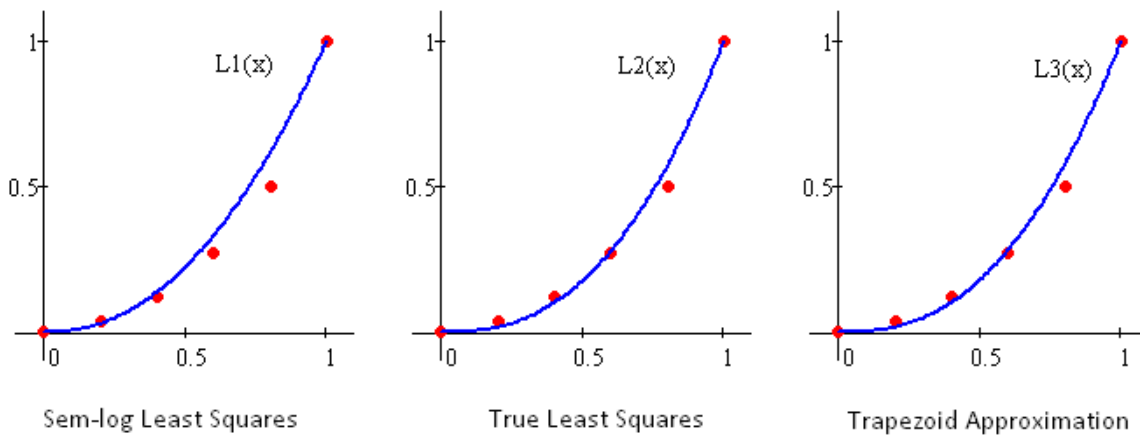


As was mentioned earlier, we cannot compare Gini indices unless they are created using the same techniques or metric. So, we must pick one choice and stick with it.

- d) Compare the points (0,0), (0.2,0.36), (0.4,0.125), (0.6,0.273), (0.8,0.503), (1,1) with the Lorenz curves found in a), b), and c). Which do you think fits better?

In a) we found $L_1(x) = x^{2.158}$, while in b) we found $L_2(x) = x^{2.49665}$. In c) we didn't find the Lorenz curve, going directly for the area using trapezoids, but we can backsolve our equation $G = 1 - \frac{2}{n+1}$ for n for

comparison. In this case $0.4252 = 1 - \frac{2}{n+1}$, so $n = 2.479$ and $L_3(x) = x^{2.479}$. The graphs below illustrate the three models.



The True Least Squares best matches the data, as it should, but is the most cumbersome to use. The semi-log model is easiest to use, but misses the data by the largest amount, so I choose the trapezoid approximation. You can make other choices for your solution.

Go to the US Census web-page <http://www.census.gov/compendia/statab/2011/tables/11s0693.pdf> (the table is replicated below).

Table 693. Share of Aggregate Income Received by Each Fifth and Top 5 Percent of Households: 1970 to 2008

[Households as of March of the following year, (64,778 represents 64,778,000). Income in constant 2008 CPI-U-RS-adjusted dollars. The shares method ranks households from highest to lowest on the basis of income and then divides them into groups of equal population size, typically quintiles. The aggregate income of each group is then divided by the overall aggregate income to derive shares. Based on the Current Population Survey, Annual Social and Economic Supplement (ASEC); see text, this section and Section 1, and Appendix III. For data collection changes over time, see <<http://www.census.gov/hhes/www/income/data/historical/history.html>>]

Year	Number of households (1,000)	Income at selected positions (dollars)					Percent distribution of aggregate income					
		Upper limit of each fifth				Top 5 percent	Lowest 5th	Second 5th	Third 5th	Fourth 5th	Highest 5th	Top 5 percent
		Lowest	Second	Third	Fourth							
1970.....	64,778	18,250	34,960	50,849	72,548	114,678	4.1	10.8	17.4	24.5	43.3	16.6
1980.....	82,368	18,604	34,889	53,488	78,316	126,035	4.2	10.2	16.8	24.7	44.1	16.5
1990.....	94,312	19,962	37,787	57,810	88,161	151,310	3.8	9.6	15.9	24.0	46.6	18.5
1995 ¹	99,627	20,201	37,756	58,922	91,359	158,521	3.7	9.1	15.2	23.3	48.7	21.0
2000 ^{2,3} ...	108,209	22,405	41,260	65,233	102,232	181,568	3.6	8.9	14.8	23.0	49.8	22.1
2001.....	109,297	21,854	40,515	64,456	101,549	183,030	3.5	8.7	14.6	23.0	50.1	22.4
2002.....	111,278	21,442	39,946	63,625	100,552	179,525	3.5	8.8	14.8	23.3	49.7	21.7
2003.....	112,000	21,053	39,803	63,747	101,693	180,425	3.4	8.7	14.8	23.4	49.8	21.4
2004 ⁴	113,343	21,072	39,525	62,955	100,311	179,133	3.4	8.7	14.7	23.2	50.1	21.8
2005.....	114,384	21,151	39,704	63,593	101,141	183,081	3.4	8.6	14.6	23.0	50.4	22.2
2006.....	116,011	21,395	40,338	64,073	103,619	185,824	3.4	8.6	14.5	22.9	50.5	22.3
2007.....	116,783	21,071	40,602	64,382	103,842	183,801	3.4	8.7	14.8	23.4	49.7	21.2
2008.....	117,181	20,712	39,000	62,725	100,240	180,000	3.4	8.6	14.7	23.3	50.0	21.5

¹ Data reflect full implementation of the 1990 census-based sample design and metropolitan definitions, 7,000 household sample reduction, and revised race edits. ² Implementation of Census 2000-based population controls. ³ Implementation of a 28,000 household sample expansion. ⁴ Data have been revised to reflect a correction to the weights in the 2005 ASEC.

Source: U.S. Census Bureau, *Income, Poverty and Health Insurance Coverage in the United States: 2008*, Current Population Reports, P60-236RV, and Historical Tables—Tables H1 and H2, September 2009. See also <<http://www.census.gov/hhes/www/income/income.html>> and <<http://www.census.gov/hhes/www/income/data/historical/household/index.html>>.

e) Using whichever method you like best, find the Gini index for 1970, 1980, 1990, 2000, and 2008. Which decade saw the greatest change in the Gini Index?

The formula for the trapezoid rule is simplified nicely:

$$\frac{0.5 - 0.2 \left(\frac{(0 + 2(P_{0.2}) + 2(P_{0.4}) + 2(P_{0.6}) + 2(P_{0.8}) + 1)}{2} \right)}{0.5} = 0.8 - 0.4(P_{0.2} + P_{0.4} + P_{0.6} + P_{0.8}).$$

So, we need to first compute the cumulative proportions $P_{0.2}$, $P_{0.4}$, $P_{0.6}$, and $P_{0.8}$, then compute $G = 0.8 - 0.4(P_{0.2} + P_{0.4} + P_{0.6} + P_{0.8})$ for each year. The table below gives the results.

$Y70_i :=$ <table border="1"> <tr><td>0.041</td></tr> <tr><td>0.108</td></tr> <tr><td>0.174</td></tr> <tr><td>0.245</td></tr> </table>	0.041	0.108	0.174	0.245	$Y80_i :=$ <table border="1"> <tr><td>0.042</td></tr> <tr><td>0.102</td></tr> <tr><td>0.168</td></tr> <tr><td>0.247</td></tr> </table>	0.042	0.102	0.168	0.247	$Y90_i :=$ <table border="1"> <tr><td>0.038</td></tr> <tr><td>0.096</td></tr> <tr><td>0.159</td></tr> <tr><td>0.240</td></tr> </table>	0.038	0.096	0.159	0.240	$Y00_i :=$ <table border="1"> <tr><td>0.036</td></tr> <tr><td>0.089</td></tr> <tr><td>0.148</td></tr> <tr><td>0.23</td></tr> </table>	0.036	0.089	0.148	0.23	$Y08_i :=$ <table border="1"> <tr><td>0.034</td></tr> <tr><td>0.086</td></tr> <tr><td>0.147</td></tr> <tr><td>0.233</td></tr> </table>	0.034	0.086	0.147	0.233
0.041																								
0.108																								
0.174																								
0.245																								
0.042																								
0.102																								
0.168																								
0.247																								
0.038																								
0.096																								
0.159																								
0.240																								
0.036																								
0.089																								
0.148																								
0.23																								
0.034																								
0.086																								
0.147																								
0.233																								
$P70_i := \sum_{k=0}^i Y70_k$	$P80_i := \sum_{k=0}^i Y80_k$	$P90_i := \sum_{k=0}^i Y90_k$	$P00_i := \sum_{k=0}^i Y00_k$	$P08_i := \sum_{k=0}^i Y08_k$																				
$P70_i =$ <table border="1"> <tr><td>0.041</td></tr> <tr><td>0.149</td></tr> <tr><td>0.323</td></tr> <tr><td>0.568</td></tr> </table>	0.041	0.149	0.323	0.568	$P80_i =$ <table border="1"> <tr><td>0.042</td></tr> <tr><td>0.144</td></tr> <tr><td>0.312</td></tr> <tr><td>0.559</td></tr> </table>	0.042	0.144	0.312	0.559	$P90_i =$ <table border="1"> <tr><td>0.038</td></tr> <tr><td>0.134</td></tr> <tr><td>0.293</td></tr> <tr><td>0.533</td></tr> </table>	0.038	0.134	0.293	0.533	$P00_i =$ <table border="1"> <tr><td>0.036</td></tr> <tr><td>0.125</td></tr> <tr><td>0.273</td></tr> <tr><td>0.503</td></tr> </table>	0.036	0.125	0.273	0.503	$P08_i =$ <table border="1"> <tr><td>0.034</td></tr> <tr><td>0.12</td></tr> <tr><td>0.267</td></tr> <tr><td>0.5</td></tr> </table>	0.034	0.12	0.267	0.5
0.041																								
0.149																								
0.323																								
0.568																								
0.042																								
0.144																								
0.312																								
0.559																								
0.038																								
0.134																								
0.293																								
0.533																								
0.036																								
0.125																								
0.273																								
0.503																								
0.034																								
0.12																								
0.267																								
0.5																								
$G70 := 0.8 - 0.4 \left(\sum_i P70_i \right)$	$G80 := 0.8 - 0.4 \left(\sum_i P80_i \right)$	$G90 := 0.8 - 0.4 \left(\sum_i P90_i \right)$	$G00 := 0.8 - 0.4 \left(\sum_i P00_i \right)$	$G08 := 0.8 - 0.4 \left(\sum_i P08_i \right)$																				
$G70 = 0.368$	$G80 = 0.377$	$G90 = 0.401$	$G00 = 0.425$	$G08 = 0.432$																				

The largest Gini index, indicating this most inequitable distribution, is in 2008. The greatest increase came in the 1980's and 1990's, each with an increase of 0.024.

If you are taking a government, economics, history, or statistics course, and need a topic for a final project, comparing Gini indices through history for Republican and Democratic presidencies is a very nice project. You will need to search the web for the data necessary, but the results may be surprising.