Calculus Challenge #9 Solutions due March 2, 2011

The Gini Index: Using Calculus to Measure Inequity

The distribution of income in our society is a concept of ongoing interest to economists, politicians, public policy analysts, and other concerned individuals. In a capitalistic society such as the US, perfect equity in income distribution is neither possible nor desired. There would be no incentive to develop new products if you weren't able to capitalize on your invention. However, there is also a limit to how much of the total income should be controlled by a small group. Some suggest that this inequity in income distribution is playing an important role in the unrest apparent in Tunisia, Egypt, Yemen, and Bahrain. In the US, are the "rich getting richer, and the poor getting poorer" and is the "middle class disappearing" as some politicians suggest? And if so, how could you tell?

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/11s0693.pdf).

Fifth of Households	**Percent of income**
Lowest fifth	**3.6**
Second fifth	**8.9**
Third fifth	**14.8**
Fourth fifth	**23.0**
Highest fifth	**49.8**

Table 1: Percent distribution of aggregate income for 2000

Table 1 gives the percent of the total income of the United States earned by each fifth of the population, ordered by income. The procedure for determining the numbers in Table 1 can be thought of as follows.

> Each family and each unattached individual counts as one household with one income. Suppose all families and unattached individuals are lined up according to their earnings for the year, from least income to greatest income (that's Warren Buffett and Bill Gates on the far right). Starting at the least income level, we count off one-fifth of the total number of households. All of the households in this group constitute the lowest fifth income level. We add up all the income from all the households in this one-fifth and compare to the total income for everyone.

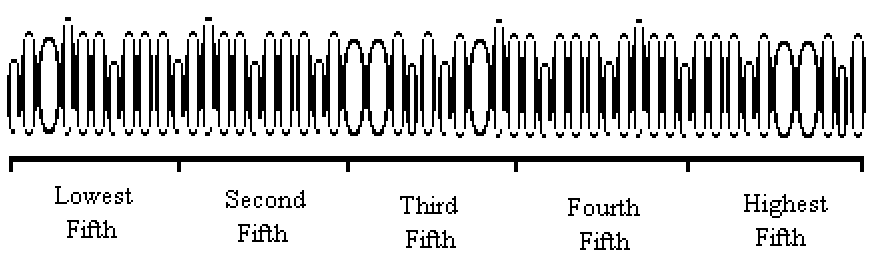

Figure 1: Lineup of population from least income to greatest income

> We continue counting households in order of increasing income and divide the number of households into the second, third, fourth, and highest fifth. The total income of each fifth, expressed as a percentage of the total income of all households, gives the percentage distribution for each fifth.

The data in Table 1indicates that in 2000, the lowest 20% of all households earned 3.6% of the total income earned. The second fifth earned 8.9% of all income, while the highest fifth of all households earned 49.8% of all income. In a perfectly equitable society, each 20% of households would earn 20% of the total income. Clearly, the distribution of income in 2000 was not equitable. How can you measure just how inequitable it is? Given data about income distribution in two different years, we would like a numerical way to compare the extent to which one income distribution is more or less inequitable than another. Fortunately, the standard method used by economists requires only a knowledge of differential and integral calculus!

Measuring Inequity

Economists have developed a very intuitive measure of this inequality of the distribution of incomes by considering the cumulative income. The cumulative data can be obtained by adding successive entries in the right-hand column of Table 1, yielding the data shown in Table 2. The lowest one-fifth of families earned 3.6% of the total income. The lowest two-fifths of families earned (3.6 + 8.9)% = 12.5% of the total income. The lowest three-fifths of families earned (3.6 + 8.9 + 14.8)% = 27.3% of the total income.

Fifths of Households	**Percent of income**
Lowest one-fifth	**3.6**
Lowest two-fifths	**12.5**
Lowest three-fifths	**27.3**
Lowest four-fifths	**50.3**
Lowest five-fifths	**100.0**

Table 2: Cumulative percent distribution of aggregate income for 2000

A graph of the data in Table 2 is obtained by plotting the cumulative proportional distribution of aggregate income versus the proportion of the population, as shown in the graph below in Figure 2. The percentages have been converted to decimal numbers, that is, the lowest two-fifths earning 12.5% of the aggregate income is represented by the point $(0.4, 0.125)$. The points (0,0) and (1,1) have been included because 0% of the households earn 0% of the income and 100% of the households earn 100% of the income. A curve that models data of the type (proportion of households, cumulative proportion of aggregate income) is called a ***Lorenz Curve.*** Such Lorenz curves always lie somewhere between the two curves $y = x$ and $y = 0$ since income distribution must fall somewhere between perfect equity and perfect inequity.

We can develop a measure of inequity by comparing the data for 2000 to the corresponding data for a perfectly equitable economy. If everything was equitable, then each fifth of the population would earn one-fifth of the income. The cumulative graph would be that of $y = x$, as shown in Figure 3.

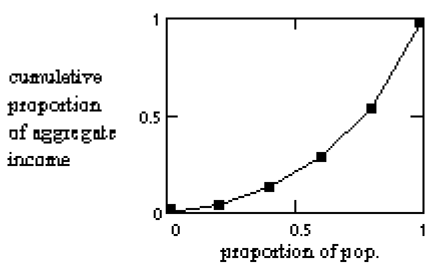

Figure 2: Graph of cumulative proportional distribution data

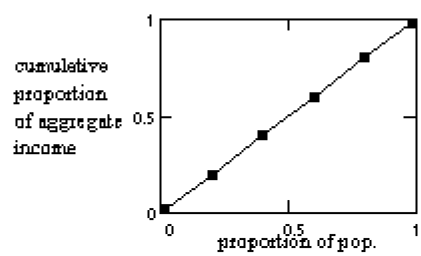

Figure 3: Perfectly equitable cumulative distribution data

How "far" from the graph of perfect equity shown in Figure 3 is the graph of the income distribution for the US in 2000 shown in Figure 2?

The Gini Index

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

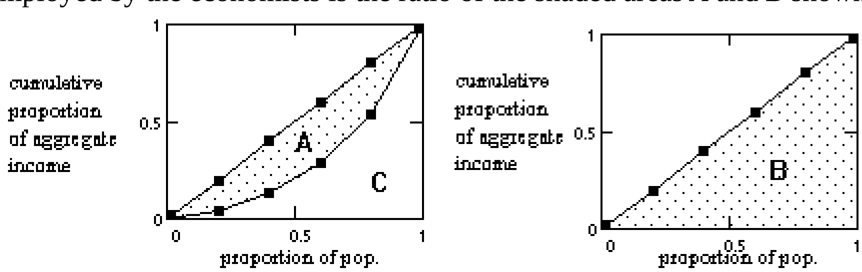

Figure 4: Areas to be computed

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

There are many ways of finding the equation for the Lorenz curve based on the data from Table 2. If you look on the web for Gini indices for various countries you will find that two estimates for the same year may differ. This is due to possible differences in their sources of information and to the method used to approximate the area *A*. Since we are interested in comparisons over time or among countries, as long as all indices are computed using the same techniques, everything works out fine.

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

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.

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 $(1,1)$.

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?

As was mentioned earlier, it really doesn't matter which method you use as long as you are consistent in your choice. It is the relative values of the Gini index that are important. If the Gini index is going up, the income distribution is getting less equitable. If it is going down, it is getting more equitable.

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) — Upper limit of each fifth				Income at selected positions (dollars)	Percent distribution of aggregate income					
		Lowest	Second	Third	Fourth	Top 5 percent	Lowest 5th	Second 5th	Third 5th	Fourth 5th	Highest 5th	Top 5 percent
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?