Handout 9: A primer on income distributions and inequality and poverty measures

ECON 81 Economic Development / Prof O'Connell / Spring 2018

1. **Generating a simulated sample of incomes for this handout.** The 4-panel Figure on page 4 of this handout shows an artificial data sample on incomes that I generated by taking random draws from a *lognormal* distribution, which is a type of probability distribution that fits many real-world income distributions quite well. The following explanation takes you through the panels in clockwise order.

   Panel 1 shows a simulation of income (y) for 2,500 individuals. I generated this sample by taking 2,500 random draws from a distribution in which the log of income $\ln(y)$ follows a normal distribution with a mean of 10.5 and a standard deviation of 1. I then converted the each draw on the log of income back into a level of income, rank-ordered the sample from the lowest income person to the highest-income person, and gave each person an ID code between 1 and 2,500 corresponding to their position in the rank order. **Panel 1** is simply a bar graph of all the incomes observed in the sample, with individuals ranked from 1 to 2,500 along the x-axis. Table 1 below provides summary statistics on this sample.

   To help you be confident that the data were indeed generated by an underlying normal distribution, **Panel 2** shows the frequency distribution of the log of income. Moving from left to right, the area of each bar in the panel corresponds to the proportion of the sample with $\ln(y)$ within the range covered by the width of the bar. This distribution has the characteristic bell-shape of a normal distribution. The sample mean and sample standard deviation of log income (Table 1) are quite close to their true values, which is a known property of random samples as the sample size gets large. Notice that the sample mean and sample median of log income are virtually identical: this makes sense given the large sample, because these two statistics are exactly the same in the underlying distribution (a normal distribution is symmetric around its mean).

   Panel 3 rearranges the information in Panel 1 into the frequency distribution of the level of income. Notice that the mean of almost $62,000 a year (a bit higher than the US mean) is much higher than the median. Most real-world income distributions (and any lognormal distribution) have this feature, reflecting the fact that the distribution of the level of income is skewed to the right.

   Table 1. *Summary statistics on the sample*

Variable	n	Mean	Median	Standard deviation	Min	Max
$\ln(y)$	2,500	10.491	10.481	1.028	7.051	14.201
$y\ (000's)$	2,500	61.747	35.616	88.276	1.154	1,470.013

2. **Constructing the *cumulative distribution of income*.** The final Panel in the Figure (Panel 4) takes the information in Panel 1 and makes two presentational changes to generate the ***cumulative distribution of income***. The cumulative distribution of income is a graph that shows, ***for any level of income (on the x-axis), the proportion of people in the sample with incomes at or below this level***.

   - First, switch the axes in order to measure income on the x-axis and individuals on the y-axis.[1]

[1] I apologize! I will continue to use "y-axis" to refer to the vertical axis in any Figure, even though income (y) may be measured on the x-axis.

- Then convert population ID numbers into sample proportions (in this case, by dividing by 2,500). So the y-axis now runs from the poorest percentile of the sample to the richest percentile, rather than from the poorest individual to the richest individual.

The cumulative distribution of income for a sample is therefore simply a graph that shows the proportion of the population (expressed as a percentage) with income less than or equal to the values shown on the x-axis.

3. **Determining the *poverty headcount ratio*.** Suppose that we identify poor households as those with income equal to or below some absolute threshold level $\boldsymbol{z}$. We can then read the poverty headcount ratio $\boldsymbol{H}$ directly from the cumulative distribution, as the proportion of the population with incomes at or below z. Similarly we can read median income straight from the cumulative distribution, as the income (on the x-axis) corresponding to a population percentile of 50 (on the y-axis).

4. **Changes in the headcount over time: growth effects and distributional effects.** Notice that we can alternatively calculate the poverty headcount ratio as the area underneath the normal distribution to the left of $\ln(z)$ in Figure 2. The latter property is very useful, because the shape of the distribution in Figure 2 depends on only 2 parameters: the mean of $\ln(y)$ and the standard deviation (or spread around the mean) of $\ln(y)$. What this means is that when we are comparing any two normal distributions of $\ln(y)$, we can conceptually break down the comparison into an effect due to a difference in the mean of income holding the standard deviation constant, and a difference in the standard deviation holding the mean constant – equivalently, into a combination of distribution-neutral economic growth and static redistribution.

   In terms of impacts on the headcount ratio, distribution-neutral growth reduces poverty because everyone's income rises by the same proportion, meaning everyone has exactly the same increment to $\ln(y)$. The distribution of $\ln(y)$ therefore slides uniformly to the right in Panel 2 with no change in its shape, implying that the area beneath the curve to the left of $\ln(z)$ definitely shrinks (equivalently, the new cumulative distribution of income in Panel 4 would lie everywhere below the old one, generating the same reduction in the headcount ratio).

   As long as the poverty rate is below 50%, changes in inequality also have an unambiguous impact on the headcount ratio: a reduction in inequality while holding the mean constant narrows the distribution of $\ln(y)$ and therefore reduces the headcount ratio, while an increase in inequality widens it and increases inequality.

   We will look at this decomposition exercise in class. In reality, of course, both forces – growth and redistribution – are always present over time as the economy responds to innovations, shocks, policies, and other influences on income. In the USA, average incomes have risen since the early 1980s, tending to reduce the poverty rate, but inequality has been rising, tending to increase the poverty rate. On balance in the USA, the official poverty rate has risen slightly. China has had the same combination of forces in operation since around 1990 – positive economic growth along with increases in inequality – but the growth in average incomes in China has been so rapid that the poverty headcount ratio has fallen dramatically.

5. **Constructing the *Lorenz curve***. If we are primarily interested in inequality rather than poverty, it is useful to represent the data in a slightly different format. Going back to **Panel 1**, suppose that we continue to measure population along the x-axis, but instead of measuring individual income on the

y-axis, we measure the *cumulative* income earned by all individuals up to and including the individual in question. We then convert both axes to percentiles, by dividing the x-axis units by 2,500 (and multiplying by 100 to convert to percentages) and dividing the cumulative incomes along the y-axis by total income in the sample (and multiplying by 100). This generates **the Lorenz curve**, which is simply ***a rank-ordered plot of cumulative income shares against cumulative population shares***.

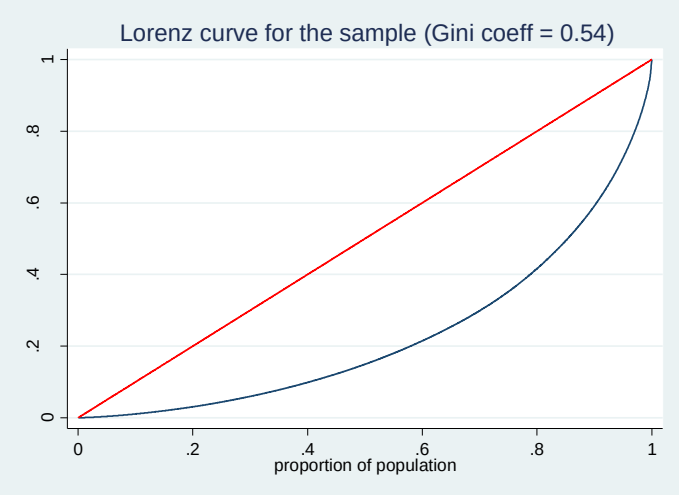
Notice that the Lorenz curve is scale-free, in the sense that its shape is completely independent of the mean of income and also of the size of the population. It also has the very desirable feature that it bows outwards – implying an unambiguous increase in inequality – whenever a new distribution can be generated from an old one through a series of regressive transfers, i.e., a series of transfers of income from poorer person(s) to richer person(s).

6. **Calculating and interpreting the Gini coefficient**. The **Gini coefficient** is a commonly-used measure of inequality, defined as ***the average difference between any two incomes in the population, expressed as a ratio to average income***. Just "FYI" the Gini can be written as follows:

$$G = \frac{\frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|y_i - y_j|}{\bar{y}},$$

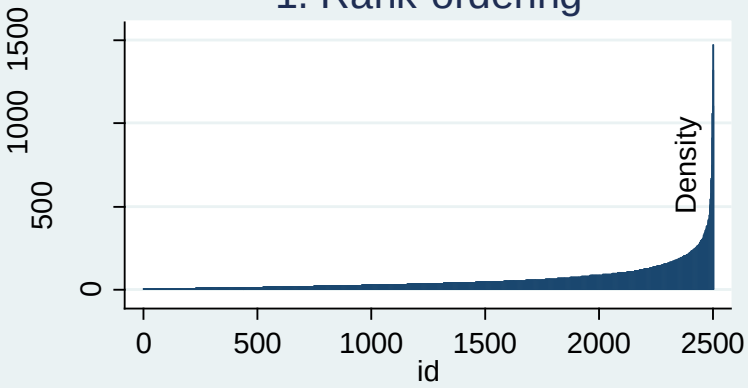
The Gini coefficient is intimately connected to the Lorenz curve: it can be calculated as the ratio of the area of the "lens" formed between the 45-degree line and the Lorenz curve to the area of the full triangle below the 45-degree line. This means that the Gini coefficient is *Lorenz-consistent* – i.e., it will always be higher for a distribution whose Lorenz curve is more bowed out over some portion(s) of the sample (and is not more bowed in over any portion!) than the Lorenz curve for some other distribution. The figure directly below shows the Lorenz curve for our sample, which has a Gini coefficient of 0.5403.

Any single measure (like the standard deviation of the log of income, or the Gini coefficient) obviously compresses the information in the income distribution ferociously. In practice it is not always obvious from looking at the full distribution whether inequality has risen or fallen: in class we will look at "Lorenz crossings", where inequality worsens in some portion(s) of the distribution and improves in others. In such cases the Gini will rise or fall but the underlying story may be complicated, and it will always be useful to look at changes in the whole Lorenz curve if possible.

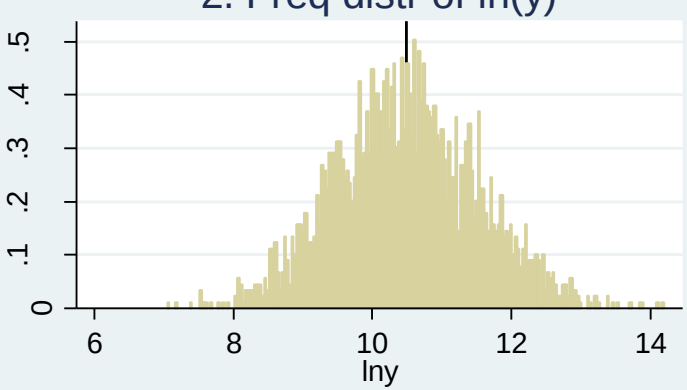


Distribution of 25,000 random draws on ln(y)

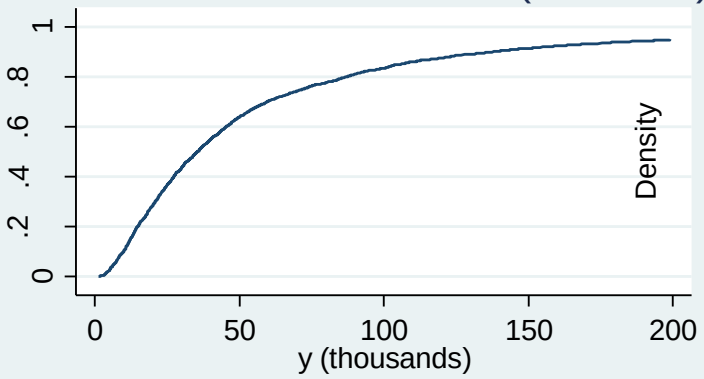
1. Rank-ordering



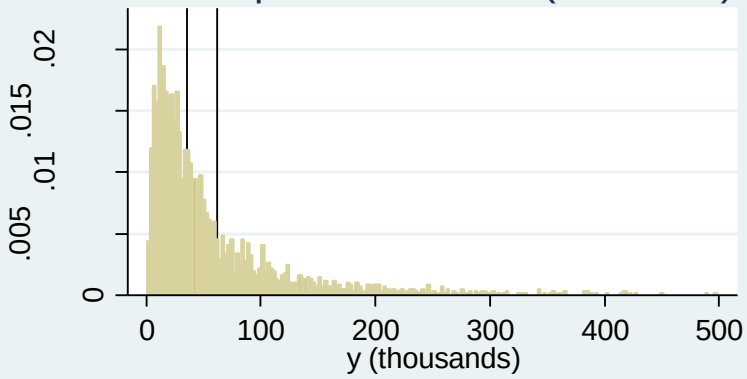
2. Freq distr of ln(y)



4. Cumul distr of income (to $200K)



3. Freq dist of income (to $500K)



Vertical lines show the mean and median of the frequency distributions.