Describing Income and Wealth Inequality Within a Population

36-313, Statistics of Inequality and Discrimination

2 September 2021 (Lecture 2)

Contents

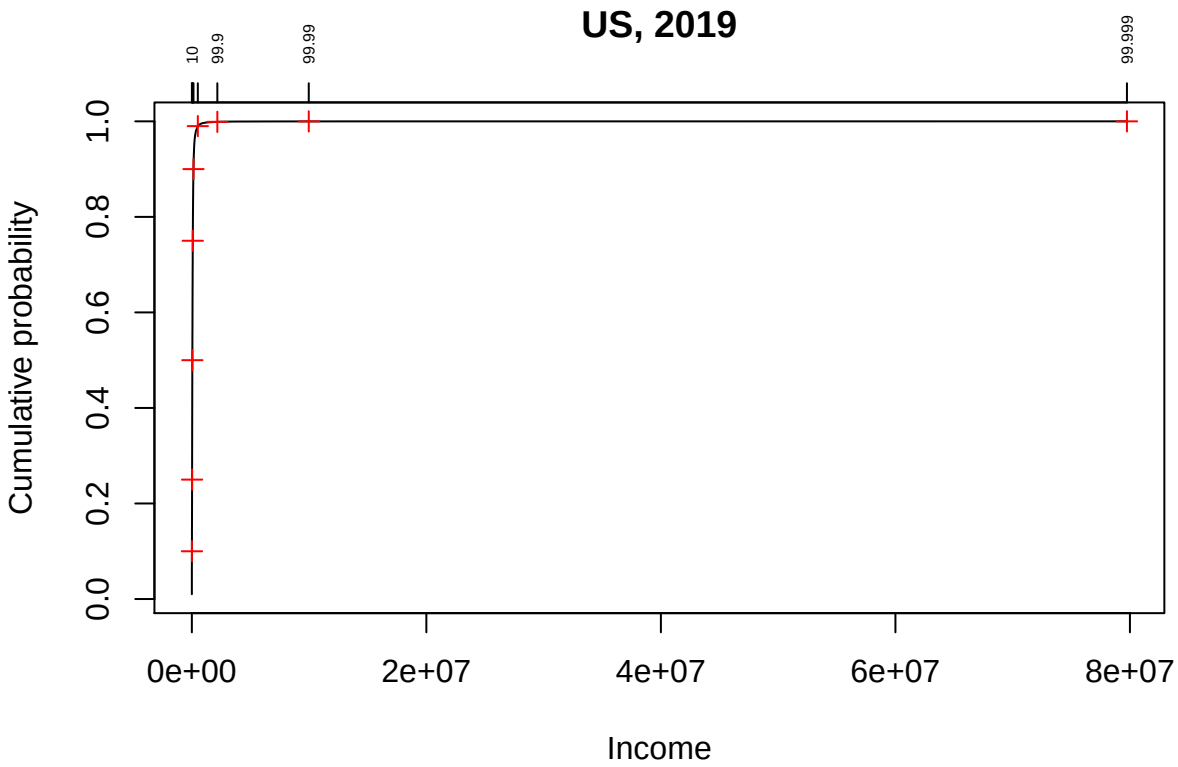
In this lecture, we're going to look at some of the data about the distribution of income and wealth within populations, and some of the tools used to describe those distributions. This will include summary measures of within-group inequality. Some of these statistical tools are going to be familiar from your earlier courses in data analysis. Others may be less familiar, because income and wealth distributions have some unusual features, and these call for special tools.

Income distributions

Cumulative distribution functions (CDF) for income

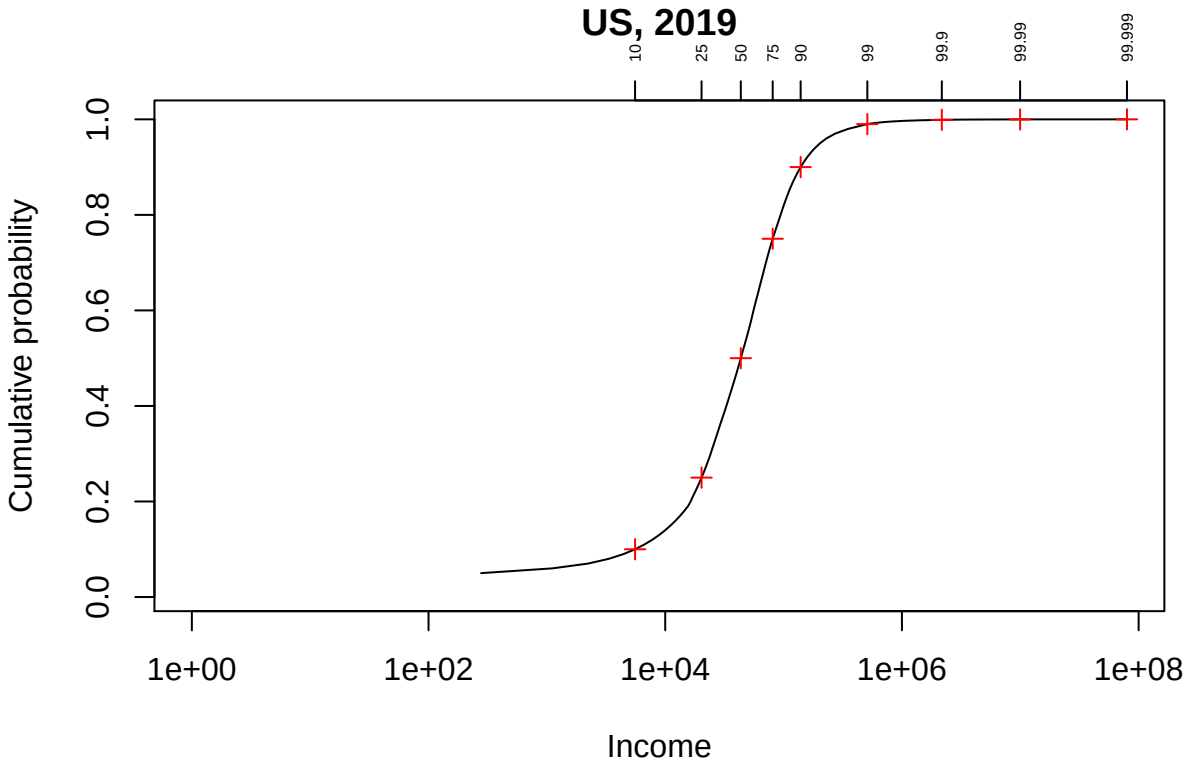
As in Lecture 1, I'm going to download some data from the World Inequality Database [https://wid.world/], but do so over multiple countries and multiple years.

Here again is the CDF for income in the US in 2019, with some round-number percentiles (10, 25, 50, 75, 90, 99, 99.9, 99.99, 99.999) highlighted in red.



You will notice that there is an *extremely* steep rise in this curve on the far left, implying that the probability density (pdf) is quite large there. But the curve keeps rising — higher and higher percentiles are further and further out to the right. Now this is true of any CDF, but if we were to plot the CDF of, say, a Gaussian distribution, it wouldn't look like this. The 90th percentile of a Gaussian is 1.3 standard deviations above the median; the 99th percentile is 2.3 standard deviations above the median; the 99.99th percentile, corresponding to the right-most point in the plot above, is only at 4.3 standard deviations above the median.

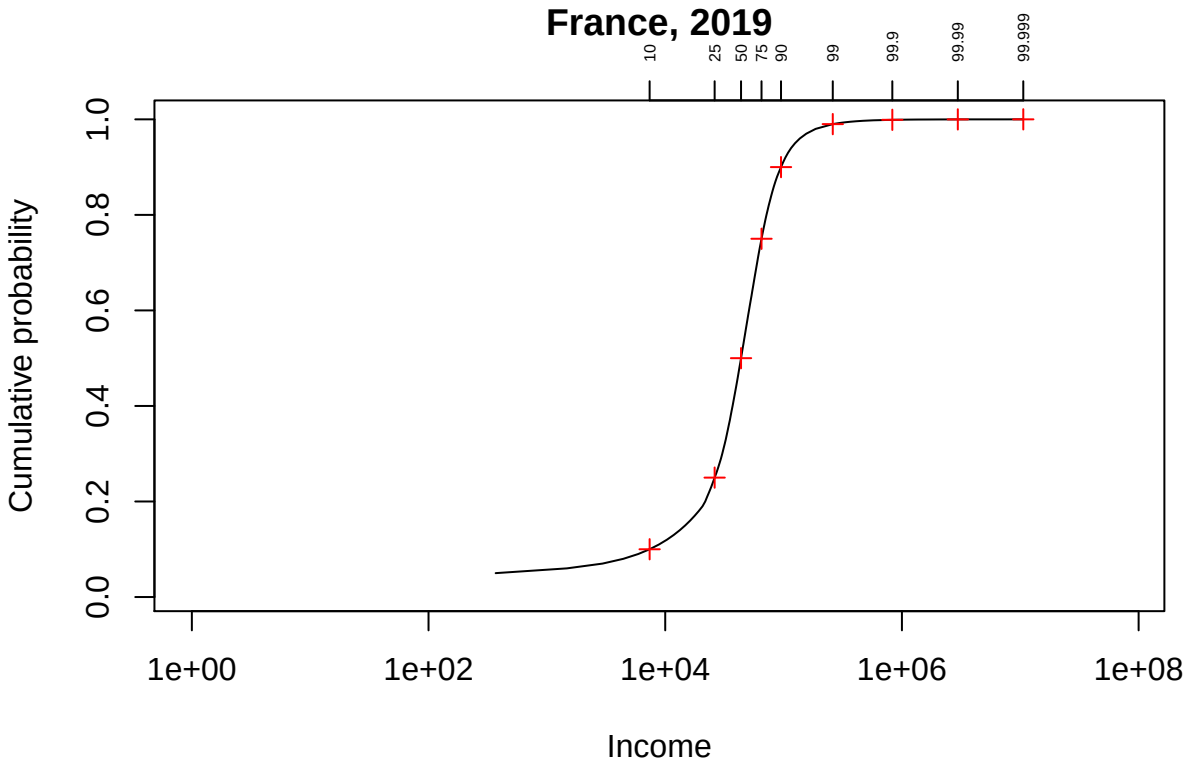
We can somewhat better see the details at the left — the comparatively-low-income region where most people actually live — if we use a logarithmic scale on the horizontal axis:



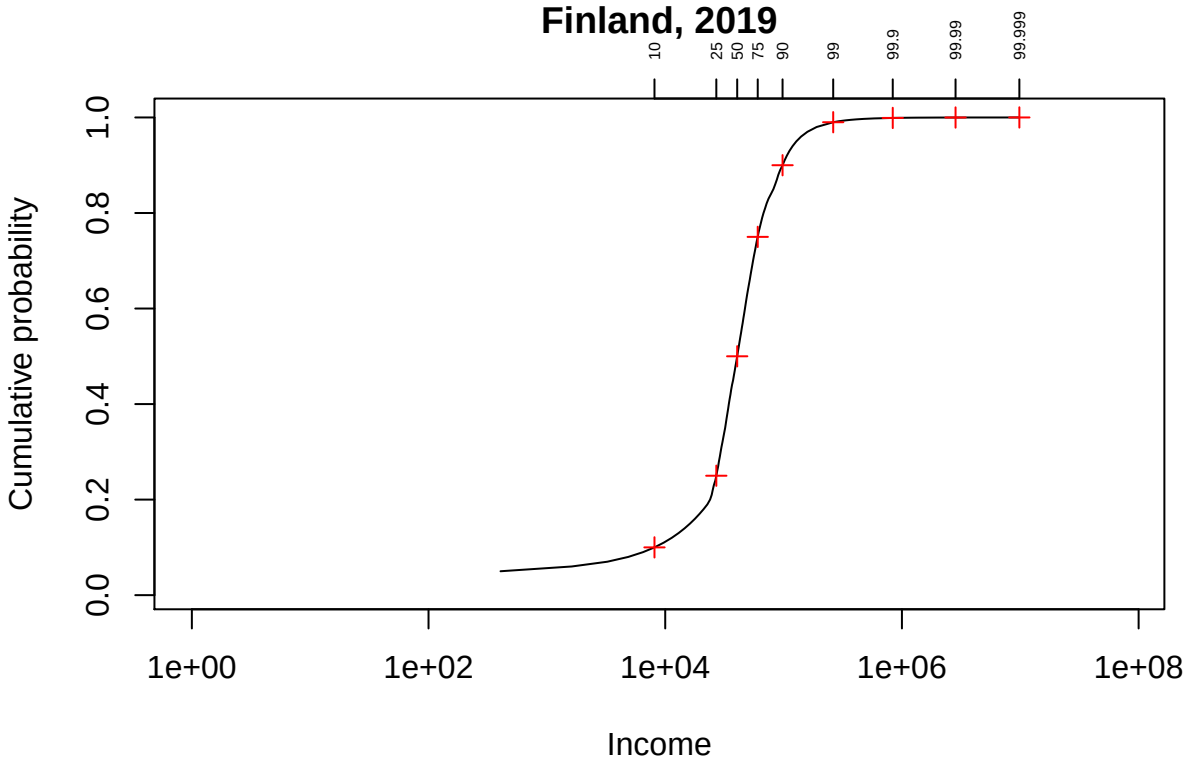
When looking at a plot like this, however, it's important to remember that equal lengths on the horizontal axis mark out equal *multiples* of income, not equal *amounts* of income.

In a general way, this sort of shape to income distributions is pretty typical under capitalism, meaning places where people buy and sell for a living, and most things, especially businesses, are private property. Here, for instance, is the comparable[1] curve for France in 2019:

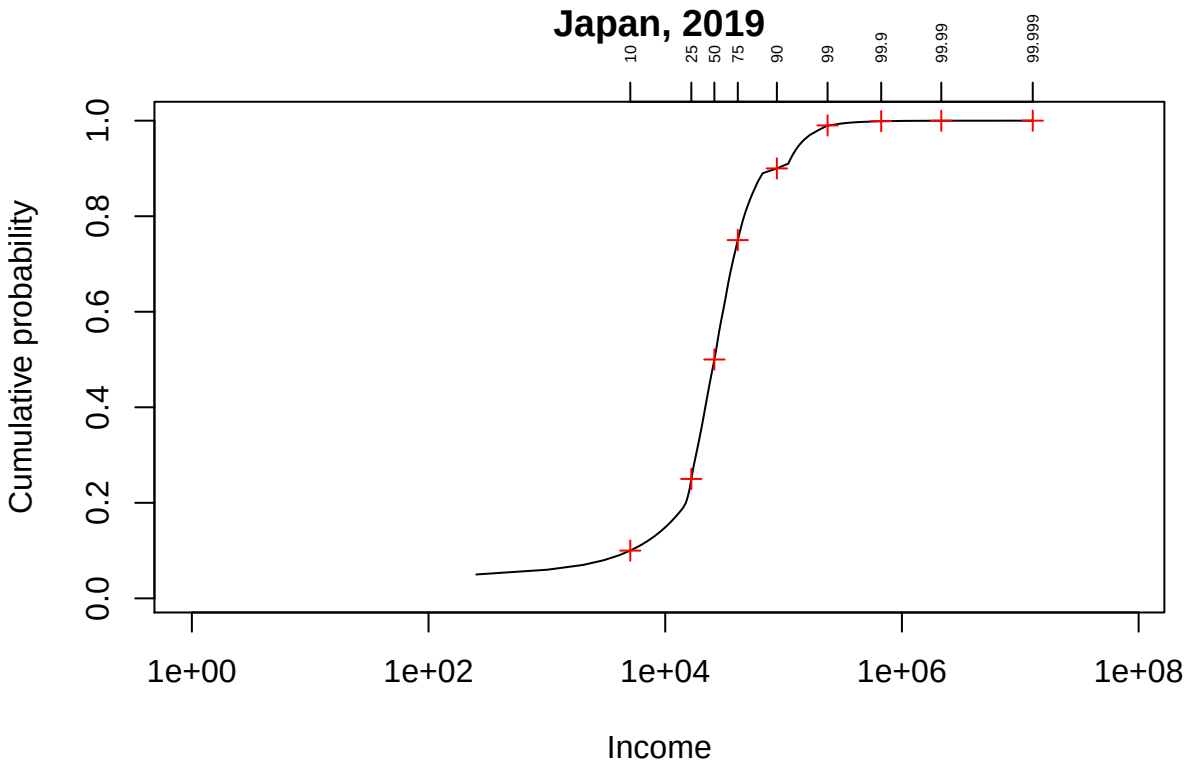
[1]The data for France are naturally in euros, not dollars. I could have plotted the CDF in euros, but it's interesting, in comparing these curves, to get a sense of how much money people in different countries make compared to each other. I could have made the comparisons just by using the **exchange rates**, the prices at which you can buy euros in dollars (or vice versa) at the bank or the airport. But different goods have different relative prices in different countries, even when you take the exchange rate into account. (In some places housing might be expensive relative to food, etc.) When economists talk about **purchasing power parity**, they're estimating how many euros (or whatever) it would take to purchase a standard "basket" of food, clothing, housing, etc., etc., in France, compared to the number of dollars it would take to buy the same "basket" in the US. This gives a fairer sense of the actual standard of living at different income levels than if we just used the exchange rates. I am taking the purchasing power parities I mention here from the Organization for Economic Cooperation and Development (OECD)'s compilation, [https://data.oecd.org/conversion/purchasing-power-parities-ppp.htm]. You will see these "PPP" conversion factors in my code.



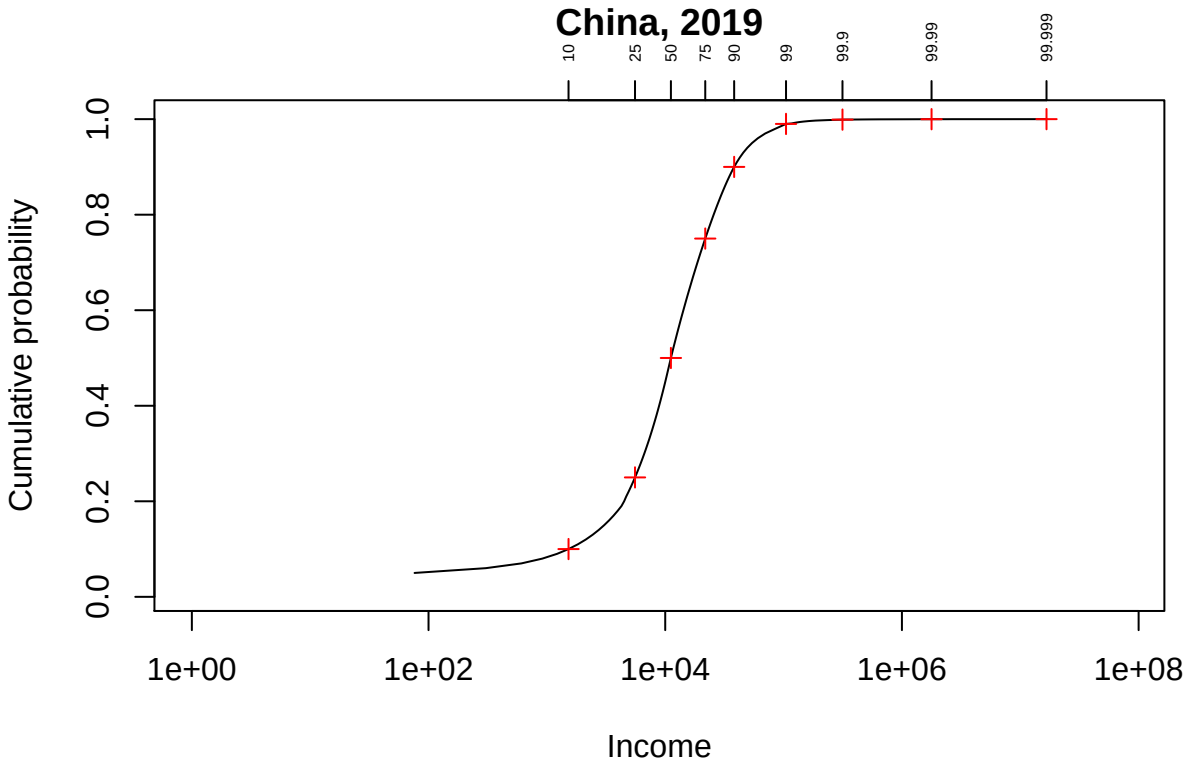
While this is the same general *kind* of shape as the US curve, you can see, if you look at the right-hand end of the curve, that the highest percentiles of the French income distribution are closer together than in the US, and that the largest incomes are not quite so large. The US does indeed have more inequality, in that sense, than France, or indeed than most other rich, democratic countries. Here for instance is Finland, which looks extremely similar to France, except perhaps even more compressed.



Nor is this just North America and Europe; here's Japan, another rich democracy:

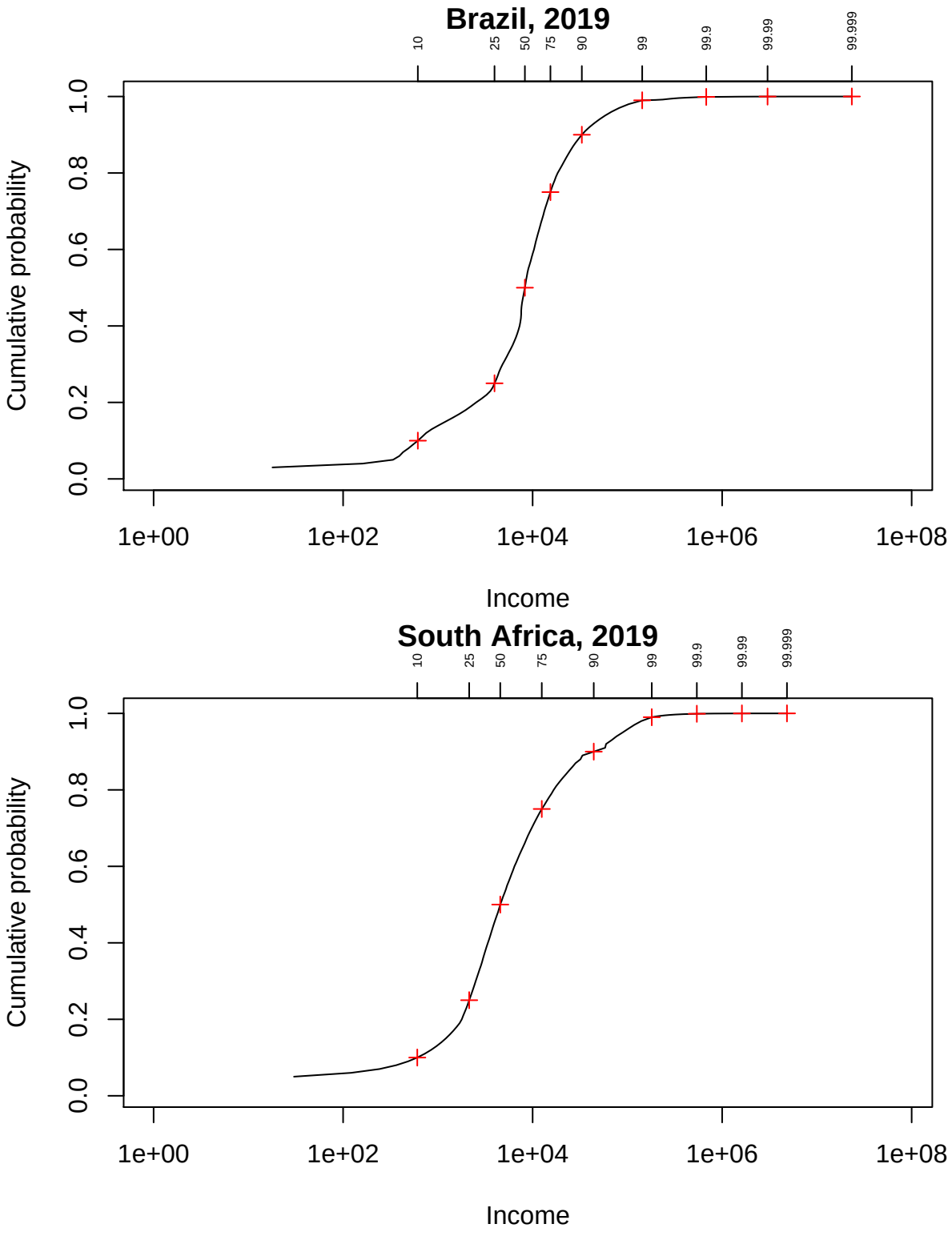


On the other hand, here are the figures for the People's Republic of China, which actually looks remarkably like the US curve, just shifted to the left:



I should also say that there are countries which have even more extreme inequality than the US; most of

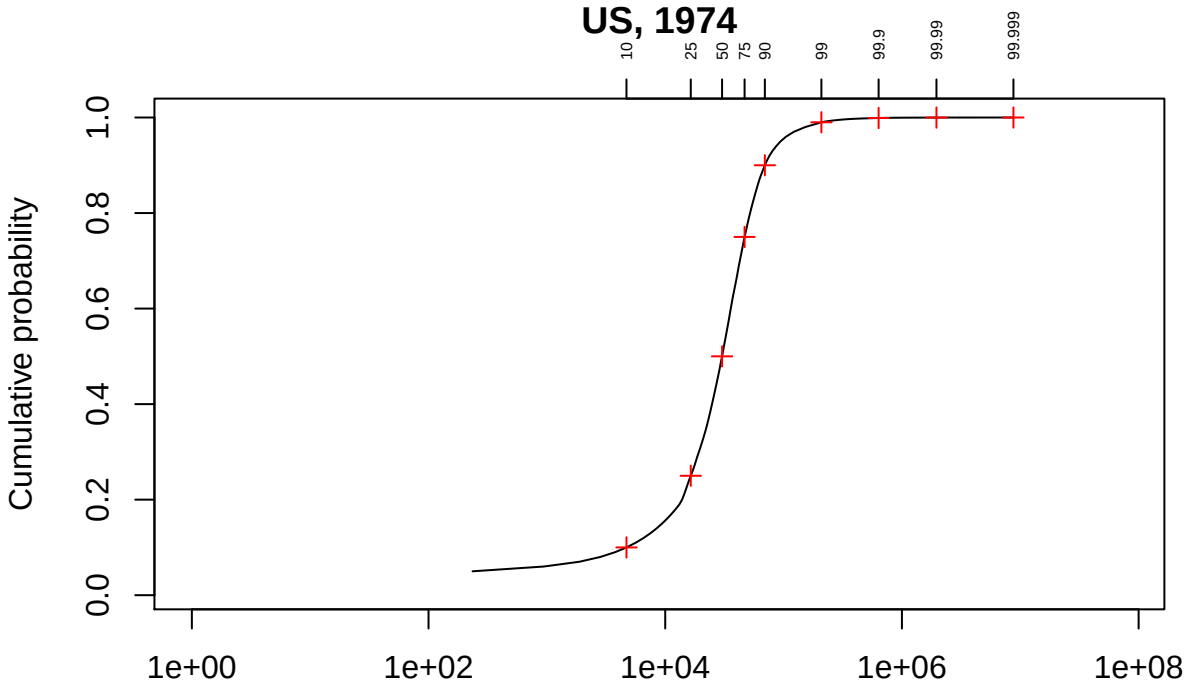
them tend to be poor countries which nonetheless have valuable natural resources. Brazil and South Africa have, in particular, long been famous for their levels of inequality, and from the figures you can see why.

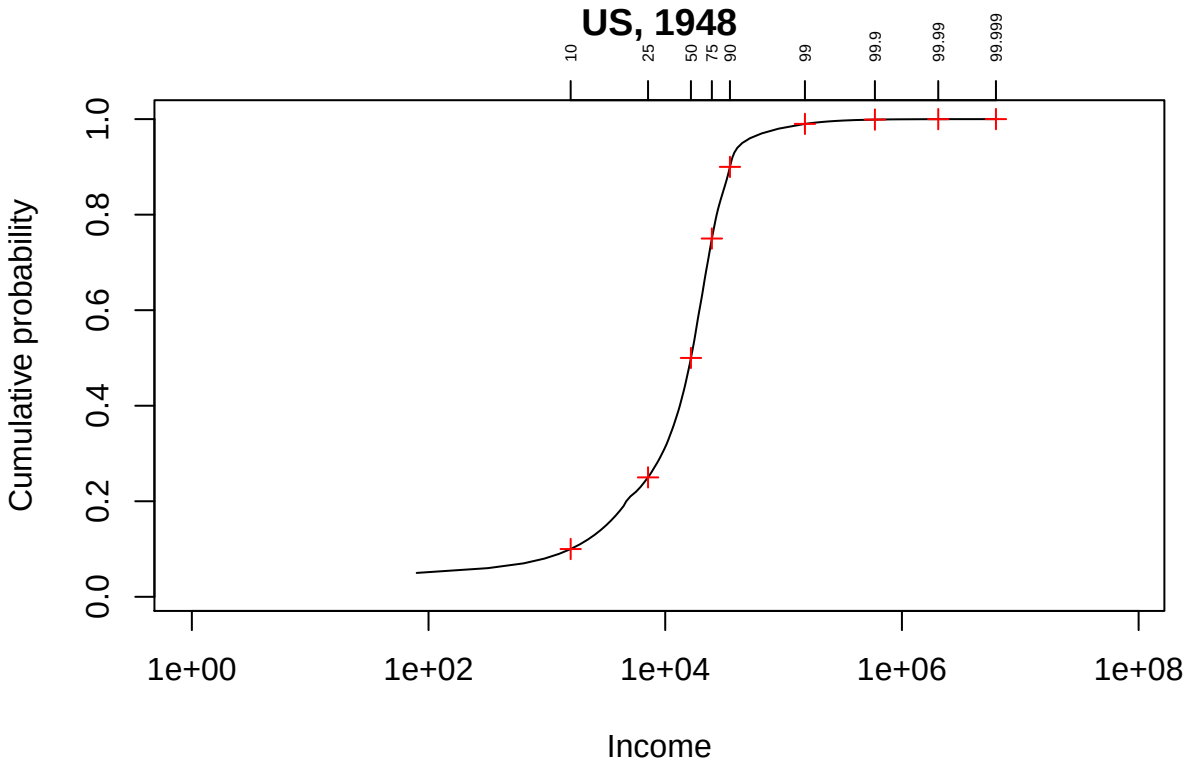


Finally, before concluding this section, I should say that the shape of these curves also changes over time. Good data for the US goes back[2] to the 1940s, so we can look at the curve in, say, 1974 (when I was born)

[2]These figures are "inflation-adjusted", rather than being in "nominal" dollars. The calculations which go into inflation

and indeed in 1948:





We can see that the right tail *used* to be much more compressed in the US. Of course America in 1948, and

adjustment over time are actually rather like those for purchasing power parity over space: trying to figure out how many currency units were needed to buy the same "basket" of goods and services in two different situations. You can imagine how this gets complicated as technology and living standards change.

even in 1974, had much lower over-all income than it does today. Whether a country can have an income distribution as compressed as America in 1974, or Finland in 2019, and income levels as high as American in 2019, is a deep question.

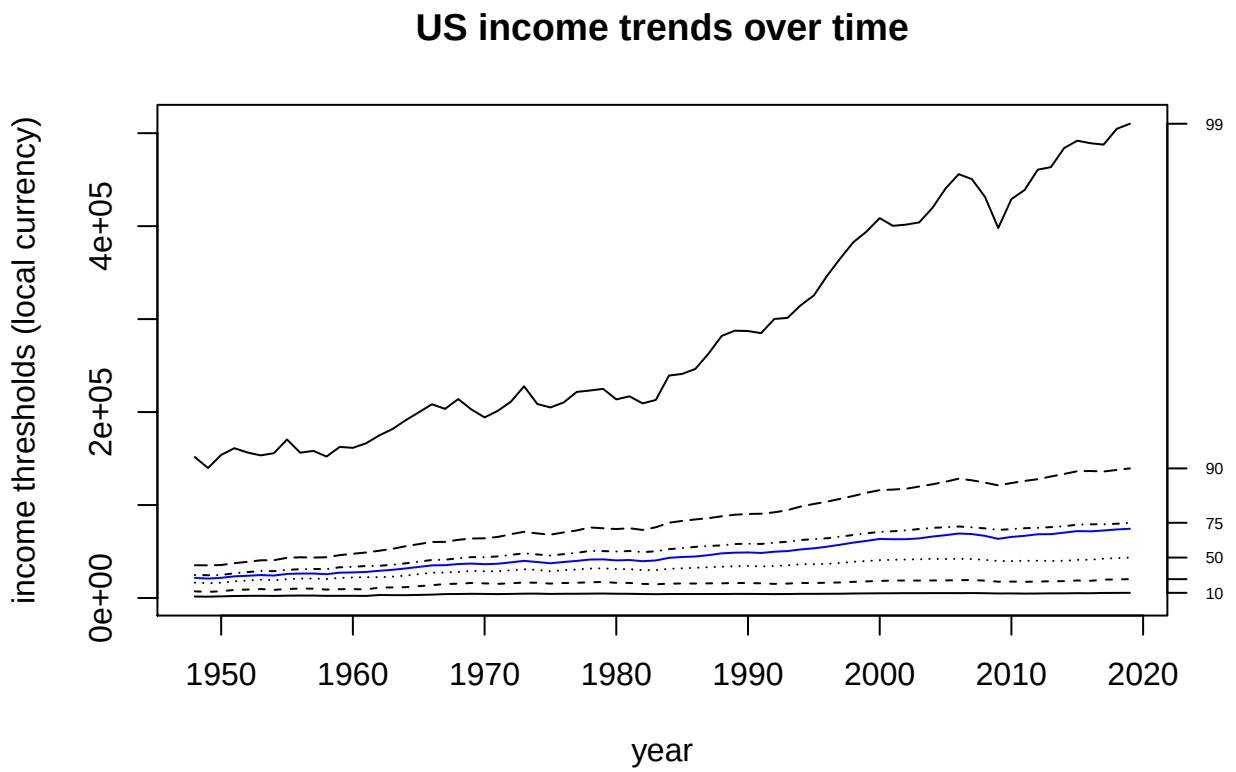
Central tendencies, and their trends over time

When we talk about **measures of central tendency**, we're talking about summary functions which give some idea of what a typical value might be for a member of the population[3]. The three most basic are:

- the **mode**: the most common value, or the one which maximizes the pdf[4];
- the **median**: the 50th percentile, where half the population lies to either side of that value, i.e. the x such that $F(x) = 0.5$; and
- the **mean** (or **arithmetic mean**[5]): the average, or **expectation value**, i.e. $\int xf(x)dx$.

Of these three, the median and the mean are the most often used in studies of inequality, and the mode is comparatively neglected. This is because the median has a pretty good claim to being "typical" (half the population is at that level, half above), and the mean indicates what everyone would get if income was shared out exactly equally. The mode, by contrast, is harder to meaningfully interpret.

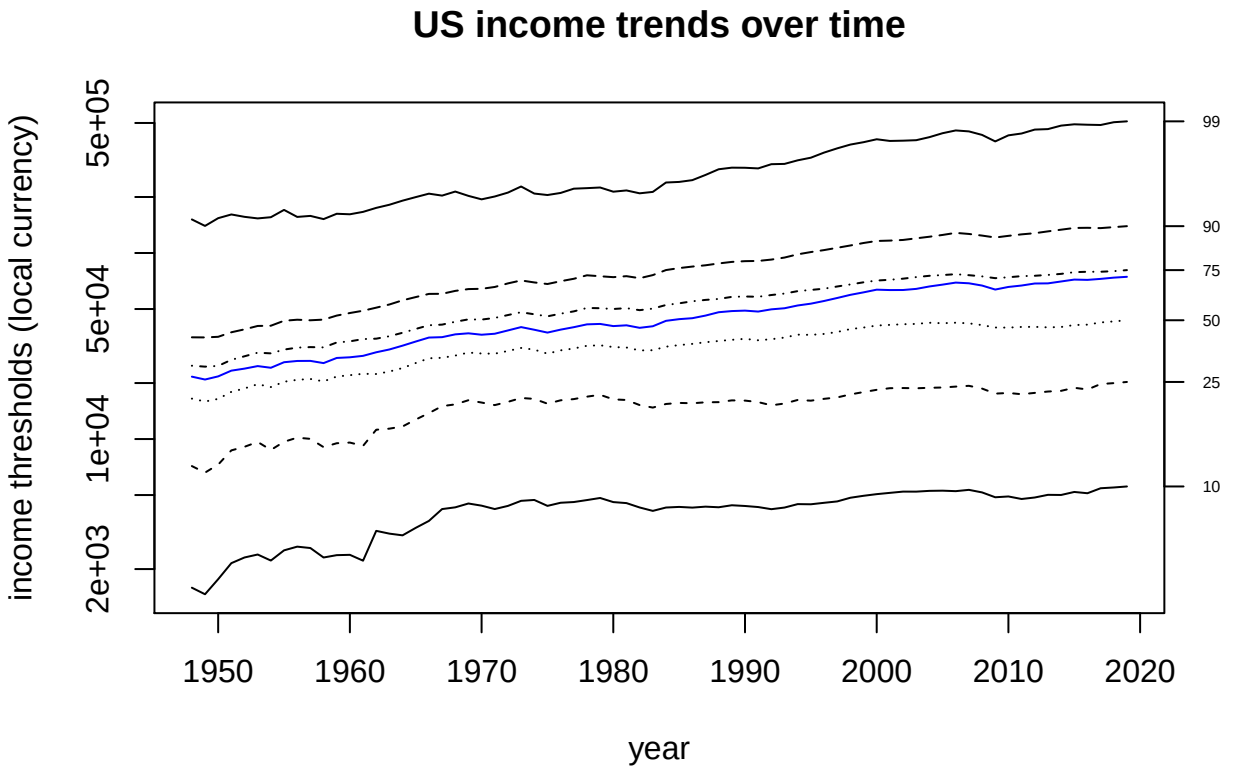
I am going to illustrate some trends over time for the median and the mode of income for the US, along with a few other percentiles to give a sense of scale. The figures are, again, adjusted for inflation. The first uses a linear scale on the vertical axis, the second a logarithmic scale.



[3]The name makes the most sense when thinking of something like a bell-curve distribution, where the mean, median and mode all coincide. With multi-modal distributions (among others), the idea of a "center" to the distribution makes less sense.

[4]If the pdf doesn't have a unique maximum, all the maxima are "modes". If the pdf has *local* maxima which are not also global maxima, some people call those values "local modes". In either case, the distribution is often called **multi-modal**.

[5]In addition to the arithmetic mean of a set of numbers $x_1, x_2, \ldots x_n$, namely $\frac{1}{n}\sum_{i=1}^{n} x_i$, the Ancestors also defined the **geometric mean**, $\left(\prod_{i=1}^{n} x_i\right)^{1/n}$, and the **harmonic mean**, $\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}\right)^{-1}$. Can you work out how to write the corresponding integrals with respect to the pdf?

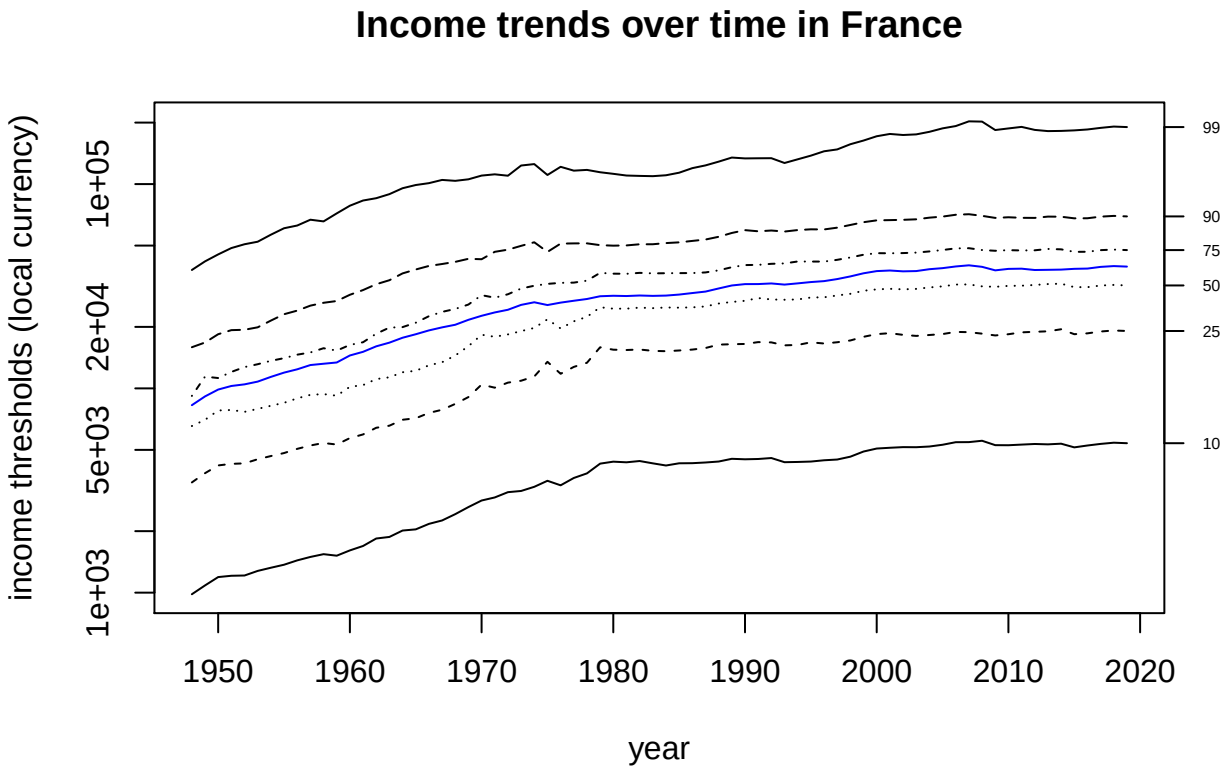


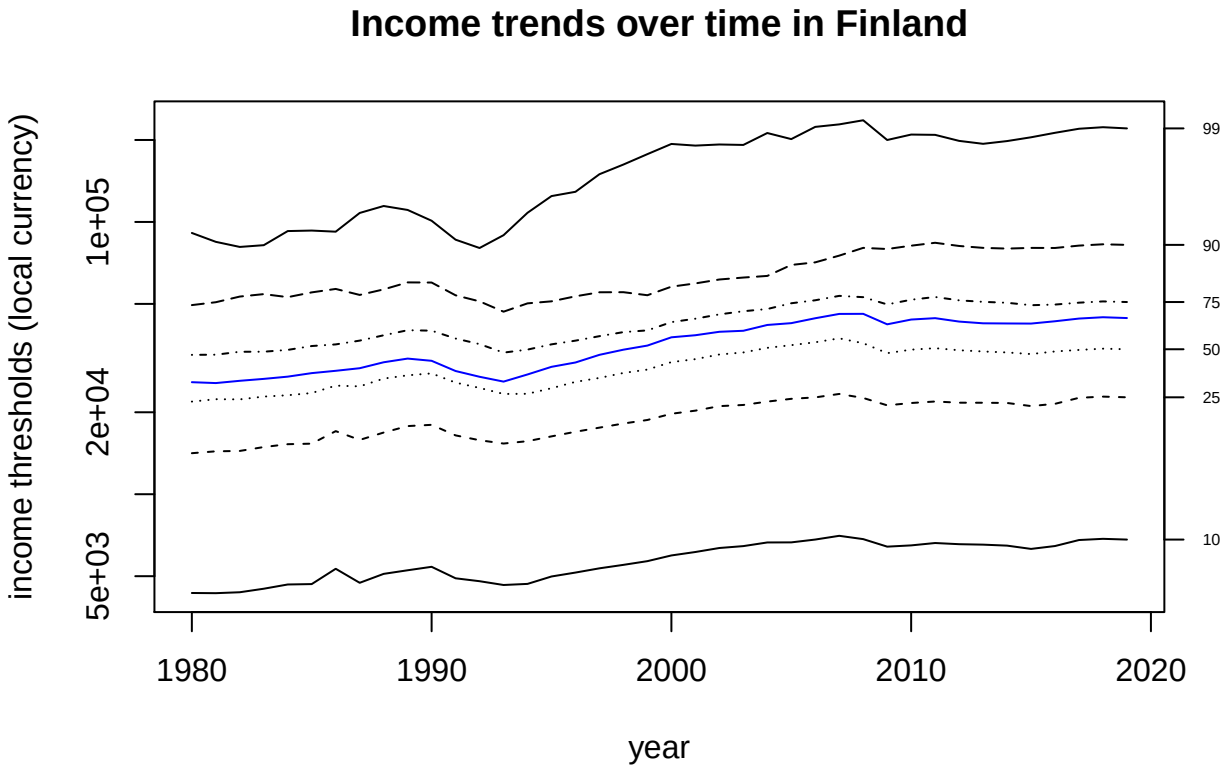
Here and in the following figures, the black lines are percentiles, while the blue line is the mean. You can see that all of these lines have generally climbed upwards between 1948, when the data begins for the US, and 2019, the last year when data is available, though with notable exceptions[6]. You can also see that up through (roughly) the 1970s, the lower percentiles were growing faster than the higher percentiles, while the mean and median were growing at roughly the same rate. This was a situation where those at the bottom of the income distribution were catching up with the rest, and income growth was very broadly shared. Since the 1970s, while income has continued to grow at every percentile, it has grown faster for the higher percentiles, comparatively little for the median, and even less for the lowest percentiles. The income distribution has become increasingly right-skewed and heavy-tailed, and the benefits of economic growth have disproportionately, but not exclusively, been felt at the top of the income distribution.

These trends are not unique to the US, though they're arguably more extreme here than in other rich democracies. Here, for instance, is the analogous plot for France and Finland[7].

[6] One of those exceptions is 2008–2010, the "Great Recession", when the world financial system ground to a halt and nearly brought the rest of the economy down with it. Most of you were probably too young to really follow what was going on at the time, but I recommend reading Eichengreen (2015) and/or Tooze (2018). But basically every point where one of those curves trends downwards is a historical episode with a name and a story that affected millions of people.

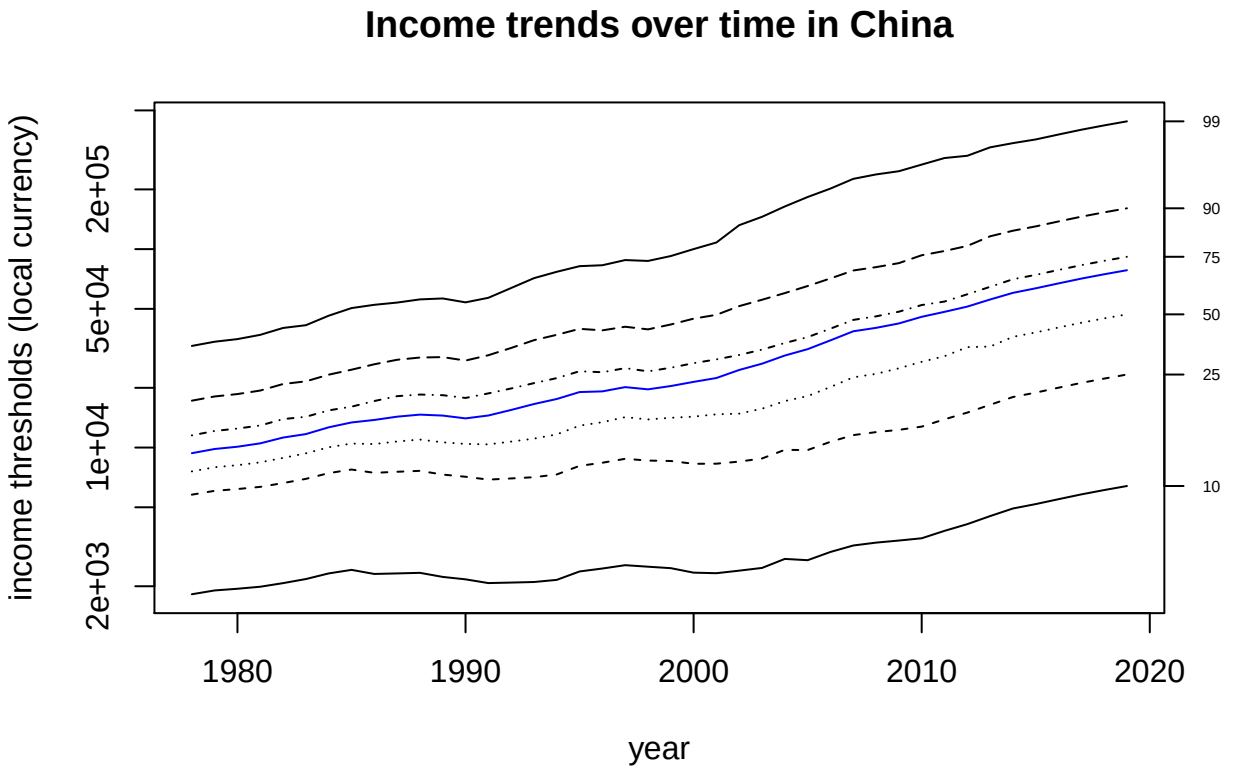
[7] Inflation-adjusted, but not adjusted for purchasing power, because I didn't feel like digging up that series, and anyway it hardly matters for the points about inequality within countries.





And here is China[8]:

[8]You will notice that the Chinese income data only goes back to the 1970s. Between the founding of the People's Republic in 1949 and the market reforms initiated in 1979 by Deng Xiaoping, it is very hard to get any reliable sense of income and its distribution in China, partly because monetary income just didn't mean as much in a command economy, partly because lots of basic facts were regarded as state secrets, and partly because even those who has access to state secrets had only a very vague notion of how things were going. In 1971, a high party official observed that different government departments "confidently"



That these have been the trends over your lifetime, and mine, and even that of my parents, is not in any serious dispute. (In the 1980s and 1990s there were people who tried to argue otherwise, but they were wrong and have almost all given up.) How to interpret these trends, and what, if anything, to do about it is another and much more contentious matter. *One* way to read this is to say that if we could have kept the pattern of growth that prevailed from the end of World War II to about 1980 (or about 1990 in the case of Finland), most people would in fact be substantially better off. Another interpretation is that there was, in fact, no way to maintain that growth pattern, and income at every percentile *has* continued to grow, so what's the cause for complaint, exactly? If you don't like either of these interpretations, there are others. These are matters which can't be settled by statistical evidence alone, though they can be *informed* by evidence.

Measures of inequality (I)

We have already looked at a number of measures of inequality, but let's be explicit about them. All of the following have claims to be ways of measuring inequality:

- The difference (or ratio) between mean and median income
- The ratio (or difference) between any two percentiles of income, say "P50/P10" (median to 10th percentile) or "P90/P10" (90th to 10th percentile).

In these contexts, ratios are more often used than differences. This is simply because high income values just are multiples of lower ones. In 2019 in the US, the difference between P90 and P10 was $\$1.3 \times 10^5$, but the ratio was 25, and people usually find it easier to grasp and compare such ratios than the absolute differences.

In addition, anything which we use as a measure of **dispersion** can also be used, in this context, as a measure of inequality. Thus the variance or standard deviation of incomes is, in itself, a measure of inequality. The **inter-quartile range**, the difference between the 75th and 25th percentile, is sometimes used as a measure

reported figures for the total population of China ranging from 750 million to 830 million people (Ellman 1978, 257n3). The idea that a government which didn't know the number of its own subjects to better than ±10% could meaningfully assess trends in the median income is absurd. Finally, such official statistics as were collected and reported were even more subject to political "corrections" than they are now.

of dispersion (it's "robust", in the sense of not being much affected by a few outliers), and so could be used as a measure of inequality. For the reason just explained, however, we'd be more likely to employ the ratio P75/P25.

Finally, you may have noticed that a bunch of the figures have used logarithmic scales to display dollar (or euro, etc.) amounts. This, too, is a reaction to the fact that high incomes are multiples, sometimes many multiples, of low incomes. Ratios of incomes correspond to differences in log incomes (because $\log \frac{a}{b} = \log a - \log b$). One also sometimes sees the standard deviation of log income used as a measure of inequality. This is especially convenient when we look at the log-normal distribution in the next lecture, but it still makes sense as a way of saying "how much scatter is there in income levels?" regardless of the distribution.

Concentration of income

When we talk about the **concentration** of income, we mean the way the rich receive a disproportionate share of the total income.

The mean income of is $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, where n is the number of individuals in the population, so the *total* income of the population is $n\overline{x}$. When we pick out any part of the population, say C, we can ask what share of the total number goes to members of the group C:

$$s(C) = \frac{\sum_{j \in C} x_j}{n\overline{x}} \tag{1}$$

Notice that if we write $\overline{x}_C$ for the mean income of members of the group c, and n_C for the number of individuals in that group, we get

$$s(C) = \frac{n_C}{n}\frac{\overline{x}_C}{\overline{x}} \tag{2}$$

So a group C will tend to get a big share if it is a big part of the total population, or if its average income is large compared to the population average.

Later in the course, we'll look at income shares by education, race, etc., but we can look at income shares defined by where people are in the income distribution

The Lorenz curve

The Lorenz[9] curve is most easily explained by classic thought experiment. Imagine we take all n individuals in the population and line them up in order of increasing income. (Break ties however you like.) Now for each p between 0 and 1 inclusive, we can take the first np individuals in line, and they'll be the bottom p^{th} quantile of the income distribution. We can also take the sum of all of their incomes, and divide it by the total income. The income share is thus

$$s(p) = \frac{\sum_{i=1}^{np} x_i}{n\overline{x}} \tag{3}$$

It's easy to convince yourself that $s(p) \leq p$, and that the only way we can have $s(p) = p$ everywhere is if everyone gets exactly the same income. (See Complementary Problem 1.)

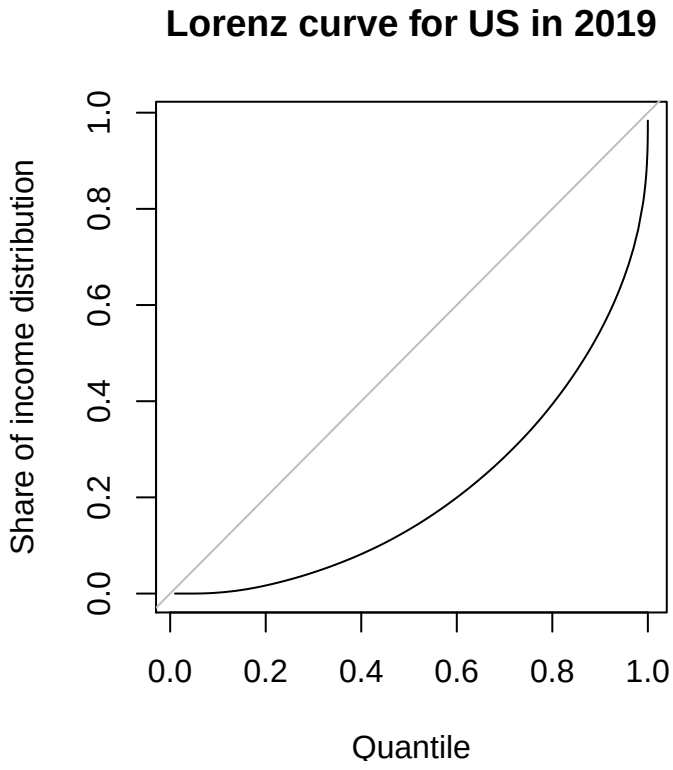
The **Lorenz curve** is a curve with p on the horizontal[10] axis (running from 0 to 1), and $s(p)$ on the vertical axis (again running from 0 to 1). It's a graphical summary of a *lot* of information about the income distribution. It's not everything (if all incomes magically doubled overnight, the Lorenz curve doesn't change), but it's almost everything.

[9] After the economist and statistician Max O. Lorenz, who introduced it in Lorenz (1905). (I'm not sure who named it the "Lorenz curve".)

[10] Interestingly, Lorenz's original paper puts income share on the horizontal axis, not the vertical axis. When the axes flipped around, I'm not sure.

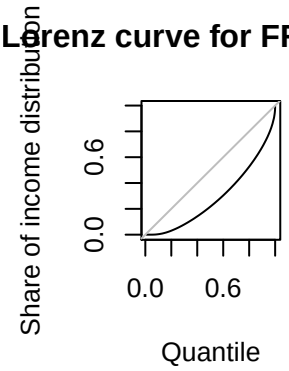
Examples of Lorenz Curves

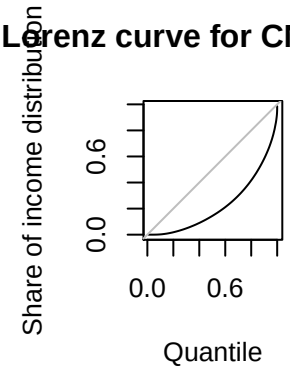
The same data source I've been using to give percentiles of the income distribution also has income shares.

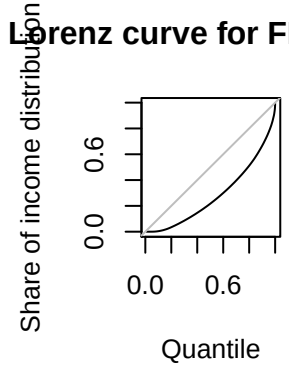


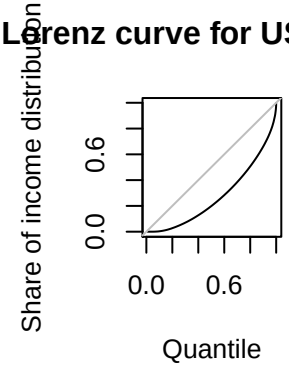
I have included the diagonal, 45-degree line in the plot as a kind of reference. Remember that if everyone got exactly the same income, equal to the mean income, the Lorenz curve *would* be the 45 degree line. As the Lorenz curve bows away from the 45 degree line, the income distribution becomes more unequal. Specifically it becomes more *concentrated* among those making the highest incomes — they receive an increasing proportion of all the income. The theoretical limit would be if exactly one person were to receive all the income, and everyone else got nothing; this would give a Lorenz curve which followed the bottom and the right edges of the plot.

Here is a little mosaic of other Lorenz curves, for other countries and/or times:









A note on calculating the Lorenz curve

The easiest way to find the Lorenz curve is to have someone else do it. Less flippantly, you may be able to find data sources which have already calculated income shares, and use them to plot the curve — that's what I've done here.

The next-easiest situation is to have access to *individual*-level data on incomes. Then you sort them in increasing order, and for each $p \in [0,1]$, you plot

$$\frac{\sum_{i=1}^{np} x_i}{n\overline{x}}$$

Unfortunately, to have individual-level data on the income of everyone in a substantial population, you need to be a trusted employee of an official statistical agency of a competent government. Those of us on the outside of these agencies will however sometimes have access to individual-level income data from *samples*, and we can calculate the Lorenz curves of our samples. (In a later homework, you'll work with such sample data on individual incomes.)

Once we no longer have individual-level data, but just aggregates, we have to resort to some approximation, depending on the kind of aggregates we have.

- One common situation is to have the *average* income level of, say, every percentile in the distribution. So we'd know the average income of everyone between the 0th and the 1st percentile, the (higher) average income of everyone between the 1st and the 2nd percentile, etc. Call these averages $a_1, a_2, \ldots a_{100}$. The over-all mean is then the average of these a_i, $\overline{x} = \frac{1}{100}\sum_{i=1}^{n} a_i$. (Why?) This will let us make a Lorenz curve, or at least 100 points on the curve. The fraction of the total income received by the bottom 1% is going to be $\frac{a_1/100}{\overline{x}}$. (Why?) The fraction of the total income received by the the bottom 2% is going to be $\frac{a_1/100+a_2/100}{\overline{x}}$. (Why?) In general, the fraction received by those at the pth percentile will be $\frac{\sum_{i=1}^{p} a_i/100}{\overline{x}}$. (Why?) We don't know exactly what value to put for the Lorenz curve in between

these percentiles, but we do know where the curve should be *at* those percentiles. In between we can approximate it with flat lines or straight lines and it usually makes little difference, at least visually. (There's nothing magic on having average incomes by percentiles, but I'll let you work out how to modify this if you have average incomes for wider or narrow slices of the population.)

- Another common situation is to know the *threshold* income for each percentile. This gives us an upper and a lower limit on the average income of people in that percentile. If they're close enough, the mid-point between them is often a reasonable guess at the average. (Also, the Lorenz plot can absorb a fair amount of "slop" in the numbers without visually changing too much.) But once we have averages we're back to the situation of the previous paragraph.
  The one place where this approach needs some extra work is that the threshold for the 100th percentile is often not available, because official agencies are reluctant to report the absolute maximum amount that anyone made. (You can imagine their reasons.) One often just makes something (reasonable) up for the average in the last percentile. (If you have the over-all average, and you've worked out averages for every other percentile, you can work out what the average for the last percentile must be. [Why?])

The Gini coefficient or index

The Lorenz curve is nice, but we often want a one-number summary of income inequality. A quantity derived from the Lorenz curve is often used here, which is to say the **Gini coefficient** or **Gini index**. This is motivated by what I just said about the Lorenz curve under perfect equality (it follows the diagonal) and under maximum possible inequality (it follows the edge of the square). So

$$G \equiv \frac{\text{area between diagonal and Lorenz curve}}{\text{area under diagonal}} \tag{4}$$

This will be 0 for perfect equality, and 1 for maximum possible inequality. In between, the bigger it gets, the further the Lorenz curve is departing from equality.

Some notes on calculating Gini indices from data

The area under the triangle is plain 1/2 (because it's half the unit square). So another way to say this is just

$$G = 2(\text{area between diagonal and Lorenz curve}) \tag{5}$$

Similarly,

$$\text{area between diagonal and Lorenz curve} = \frac{1}{2} - \text{area under Lorenz curve} \tag{6}$$

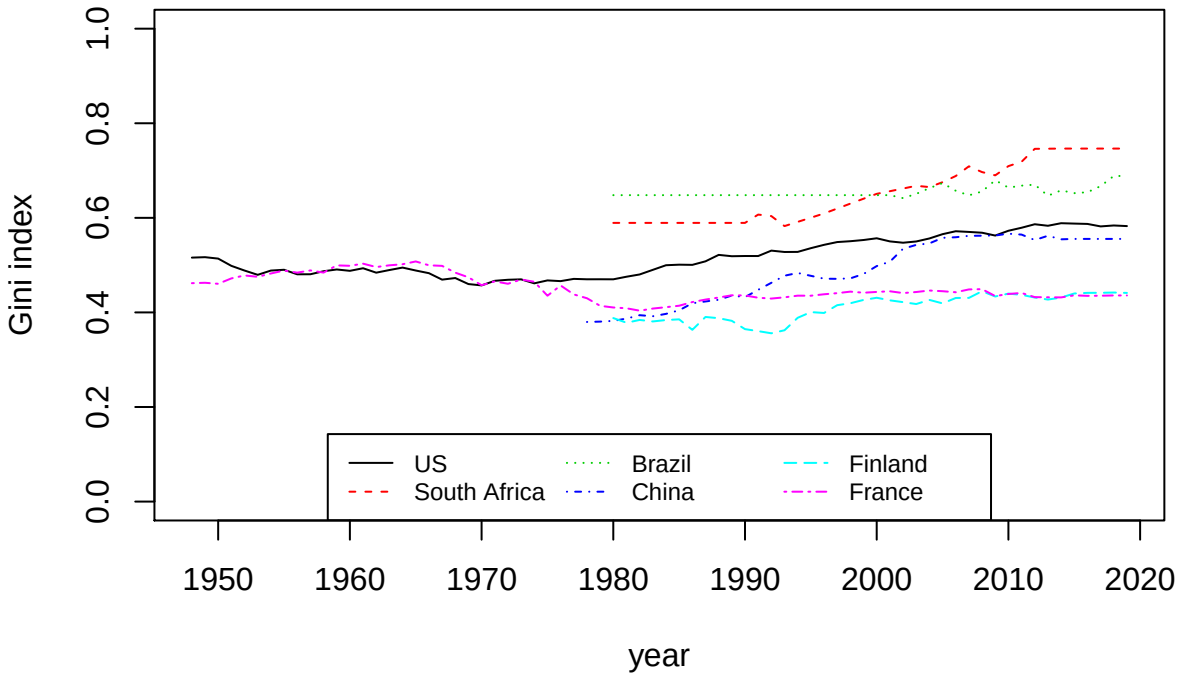
so

$$G = 1 - 2(\text{area under Lorenz curve}) \tag{7}$$

which is often the easiest way to calculate G.

- If the Lorenz curve is really made up of a lot of little steps, a **step function**, the area under the curve is a sum of rectangles, and the area of rectangles is easy to calculate, it's height times width. In my examples, the data source gives me income shares for each percentile of income from 0 to 99, then tenths of a percent above 99, then hundredths of a percent above 99.9, etc. These percentages tell me the widths of each rectangle, and the heights are the shares, so it's easy for me, or rather for my computer, to sum up the areas.
- If the Lorenz curve is on the contrary a bunch of straight line segments, the area under the curve is a sum of a bunch of trapezoids, each one a rectangle plus a right triangle. But a little geometry says that the area of this trapezoid is the *average* height times the width.

The WID contains calculations of the Gini indices for lots of countries and years, using as close to the full, individual-level data as allowed, so we can look at how those have changed. Also, Gini indices are directly comparable across countries and across time, without worrying about currencies or prices.



This plot suggests a number of correct things:

0. Gini indices change over time, but slowly.
1. Typical values for Gini indices are somewhere in the range 0.3–0.5 or even narrower. (At least, this is true of developed countries in modern times.) The US's 2019 Gini index, 0.58$, is quite high by the standards of rich democracies, at least since WWII. (Compare the US levels to those of France or Finland.) Historically, as today, the US has usually had among the highest Gini indices among rich democracies.
2. Within the US, as in most other rich democracies, the trend was decreasing from about WWII to, roughly, 1980, and then reversed.
3. Many poorer countries, such as Brazil or South Africa, have much higher levels of inequality, as measured by the Gini index, than the US or any other rich democracy.

Measures of inequality (II)

Just to be very clear, the Gini index is a measure of inequality; it's just one that has to be calculated via the Lorenz curve first.

There are other measures of inequality based on the Lorenz curve and/or the whole of the income distribution. Some of these have a richer economic interpretation, for instance, based on notions of decreasing marginal utility or welfare of money[11]. But these are often more controversial than the Gini index, which is *just* about

[11] Going from an income of $0 a year to $1000 a year improves your life tremendously. Going from $10,000 a year to $11,000 a year is still a pretty big deal. Going from an income of $1,000,000 a year to $1,001,000 is not a big deal. (Or so I'm told.) This suggests that each additional ("marginal") dollar of income does less and less to increase your "utility" or "welfare" or "general lot in life". (Moral philosophers draw subtle distinctions between these concepts, but the philosophy department is down the stairs and to the left.) If we expressed this by saying something like $U = \log X$, we could look, not just at the distribution of income, but the distribution of welfare, and how unequal welfare is across the population. This could even lead to the conclusion that straight-up *taking* money from high income people and giving it to poor people could improve average welfare, even if some money evaporated in the process. In the next lecture, we will meet the great economist Vilfredo Pareto as one of the first people to seriously study the dstribution of high levels of income and wealth. It is no accident that Pareto *also* developed a theory of how to find economic optima *without* being able to compare levels of utility across people. The theory of "Pareto improvements" and "Pareto optima" is still very useful, even if you don't share Pareto's aversion to taxing the rich. (Conversely, even if you *do* want to tax the rich, you could agree with Pareto that there's something very fishy about pretending we can numerically compare utility across people.)

who has what share of the total, and so they're less universally used.

Some asides / minor points

"The 1%"

Some of you may remember the "Occupy Wall Street" movement of 2011, which introduced the slogan "We are the 99%", and the catch-phrase of talking about "the 1%" as a short-hand for the wealthy elite at the top of society. (Some of you won't remember it, but Occupy was a big deal at the time, and set the stage for a lot of what's come since.) This is, when you think about it, a very curious kind of slogan. Earlier movements against economic inequality had talked about their opponents as "capitalists", "the bourgeois", "the establishment", "exploiters", etc. None of them, so far as I'm aware, used this sort of numerical figure[12]. Where did it come from?

The answer, so far as I can work out, was from economists doing statistical studies on income shares. This is an old topic — remember that Lorenz introduced the Lorenz curve in 1905 — but it was revitalized in the early 2000s by a new generation of economists, in particular Thomas Piketty and Emmanuel Saez. One of their joint papers, Piketty and Saez (2003), in particular made a very large impression on other economists, on journalists covering economics, and through them on activists concerned with inequality. The main contribution of that paper was new estimates of top income shares extending over almost the whole of the 20th century for the US, based directly on tax data. The centerpiece of the paper, widely reproduced, was Figure II:

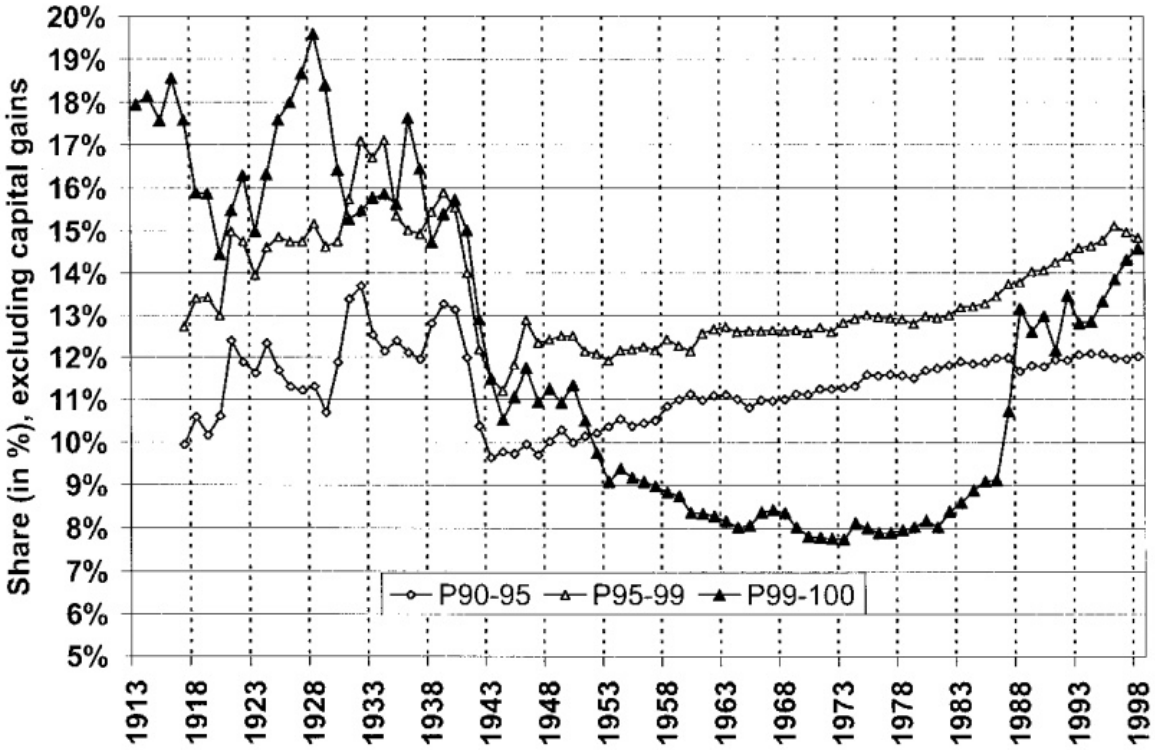


FIGURE II
The Income Shares of P90–95, P95–99, and P99–100, 1913–1998

[12] Anyone who can find a counter-example from before 2003 will get 10 points extra credit on HW 1.

The explosive finding here wasn't that the top 1% claimed a very high income share before World War II (and especially before the Great Depression). It was that starting around 1980, the income share of the 1% had started climbing and by 1998 was back to pre-war levels. Even more, the next 4% and the 5% below that *weren't* seeing similar increases in their income shares — it was really only the top 1% which was claiming more and more of the total income. Many people found the *level* of inequality reported by these figures shocking. It also weakened the case for many explanations of rising inequality as functional or adaptive[13].

Following this paper, Piketty went on to launch the World Top Incomes Database (WTID), which ultimately evolved into the World Incomes Database we're using as our data source in this lecture. That whole project is a very direct extension of the one in Piketty and Saez (2003).

The immediate source from which Occupy took the idea of the 1% seems to have been a May 2011 essay by the economist Joseph Stiglitz[14]. Stiglitz opens that essay by giving income and wealth shares for the top 1% — he doesn't cite any particular sources, but is clearly relying on either WTID or a *very* similar resource. Throughout the essay, Stiglitz refers to "the top 1 percent" (or, occasionally, "the upper 1 percent"), not "the 1 percent", and, just once, to "the other 99 percent". The transition to "the 1%" was made by the anonymous headline writer at *Vanity Fair*; the transition to "the 99%" seems to have been the contribution of the anthropologist David Graeber, who was also an anarchist activist involved in some of the initial organizing for Occupy.

Expected values and CDFs

The question came up in lecture about whether there was any meaning to the area under the CDF curve, and I waffled that there was some kind of relationship between integrating the CDF and the expected value, but that I'd get it wrong from memory. Well, here it is, not from memory.

Suppose the random variable X is non-negative, so $\mathbb{P}(X \geq 0) = 1$. Then

$$\mathbb{E}(X) \equiv \int_0^\infty x f(x) dx \tag{8}$$

At this point, we invoke the calculus trick called "integration by parts":

$$\int_a^b u(x)v'(x)dx = u(b)v(b) - u(a)v(a) - \int_a^b u'(x)v(x)dx \tag{9}$$

where primes indicate derivatives[15]. Make the following identifications:

$$a = 0 \tag{10}$$
$$b = \infty \tag{11}$$
$$u(x) = x \tag{12}$$
$$v'(x) = f(x) \tag{13}$$

[13]For instance, many economists give a lot of weight, when explaining the increase in inequality in recent decades, to the idea of "skill-biased technical change" — that shifts in technology like the introduction of computers and global supply chains have made it relatively more valuable for people in rich countries to have certain kinds of skills (like an advanced knowledge of computer programming or industrial design) and less valuable to have other abilities (like wrestling bags off ships and on to trucks). (Cf. Levinson (2006)). Since those rewarded skills are relatively rare, their possessors have benefitted disproprotionately, increasing inequality. (The healthy response of a market economy to a situation like this would be for more people to acquire the relatively-rare-but-valuable skills, i.e., "learn to code!"; the goal of such a process is to make being a software engineer and working retail more nearly equally rewarding.) There is probably something to this story, but the difficulty, in light of Piketty and Saez, is that it's hard to specify any way in which the skills of the 1% are *that* much better than those of the rest of the top 5%, or even the rest of the top 10%.

[14]Joseph Stiglitz, "Of the 1%, by the 1%, for the 1%", *Vanity Fair* May 2011, [https://www.vanityfair.com/news/2011/05/top-one-percent-201105]. Full disclosure, I should perhaps say that Stiglitz used to be my father's boss a long time ago.

[15]Remember the "product rule", that $\frac{d}{dx}(u(x)v(x)) = u(x)v'(x) + u'(x)v(x)$, so $u(x)v'(x) = \frac{d}{dx}(u(x)v(x)) - u'(x)v(x)$. Now integrate both sides of that last equation, from $x = a$ to $x = b$.

Then

$$u'(x) = 1 \tag{14}$$
$$v(x) = F(x) - 1 \tag{15}$$

(The reason why I chose $F(x) - 1$ rather than $F(x)$, or $F(x) - c$ for any other c, will be clear in a moment.) Plugging in[16],

$$\mathbb{E}(X) = \infty(F(\infty) - 1) - 0(F(0) - 1) - \int_0^\infty (F(x) - 1)dx \tag{16}$$
$$= 0 - 0 + \int_0^\infty (1 - F(x))dx \tag{17}$$
$$= \int_0^\infty \mathbb{P}(X > x)\, dx \tag{18}$$

This isn't the area under the CDF, but the area under what's called the **complementary CDF**, or **upper CDF**, or **survival function**.

Pre-tax income, taxes and transfers

All the data in today's notes are for what's often called "pre-tax" income. More strictly, they're income before "taxes and transfers", which calls for some explanation.

Governments tax away some portion of people's incomes, but governments also give things to people. Sometimes they give money[17], but they also provide services[18]. The data we've been plotting don't try to account either for the taxes or the transfers, though these are a substantial part of the economy in every rich democracy[19]. If we wanted to make a serious comparison of the distribution of *living standards* between, say, the US and Finland, or between the US in 2019 and the US in 1974, it'd be important to take taxes and transfers into account. But doing so carefully is hard, and would get us into economics rather than just into statistics. Suffice it to say that when people have tried to do this sort of comparison, the conclusion is that we end up with the same sort of distribution, but the European "social democracies" (like Finland) have even more compressed income distributions once taxes and transfers are taken into account.

What about non-capitalist economies?

We've been comparing income distributions in market economies. (Even modern China is a market economy.) There have been other types of economy in the world, and perhaps they will be more common again in the future. It's natural to wonder whether they, too, will have income distributions like this. This is actually a hard question to answer, even about the historical non-market economies, let alone speculative future economies.

[16] If you're worried about multiplying zero and infinity to get 0, substitute a large but finite upper limit b for infinity in the integral, and take the limit $b \to \infty$.

[17] Examples include: emergency assistance like unemployment insurance, or 2020's stimulus, old-age pensions (US "Social Security"), disability payments, EBT ("food stamps") and tax credits. "Refundable" tax credits, like the US "earned income tax credit" (EITC) mean that some people end up paying *negative* taxes. Even when tax credits aren't fully refundable, selectively reducing some people's taxes (e.g., because they're paying a mortgage) but not others amounts to an implicit transfer. The US government is especially fond of using the tax code to transfer money from some people to others without explicitly writing checks (Mettler 2011). It also writes a lot of checks!

[18] Medical services are a big part of this in many countries, including the US (Medicare for old people, Medicaid for the poor, the Veteran's Administration medical system...). One can debate whether services like public education, police, firefighters, etc., should be counted as part of the tax-and-transfer system.

[19] The total US national income in 2019 was just under $21.5 trillion. (This is "gross domestic product", GDP.) The budget of the US federal government was $4.4 trillion. State and local governments add approximately another $4 trillion [https://www.census.gov/data/datasets/2019/econ/local/public-use-datasets.html]. Subtracting the $760 billion that the Federal government gave to state and loca governments, we're still looking at a total expenditure of above $7.6 trillion, which is 35% of the national income. And the US has a comparatively *small* government by the standards of rich democracies.

The best-known non-market economies are those of the Communist countries of the 20th century, starting with the USSR in 1917 and ending with the USSR in 1991. These economies still had money, so one could look at the distribution of money incomes, and those were indeed much more compressed than in capitalist economies. People who have done sound find, for instance, Gini indices of 0.275–0.290 in the USSR in the 1980s (Ellman 2014, Table 7.5, p. 274), which is indeed very low.

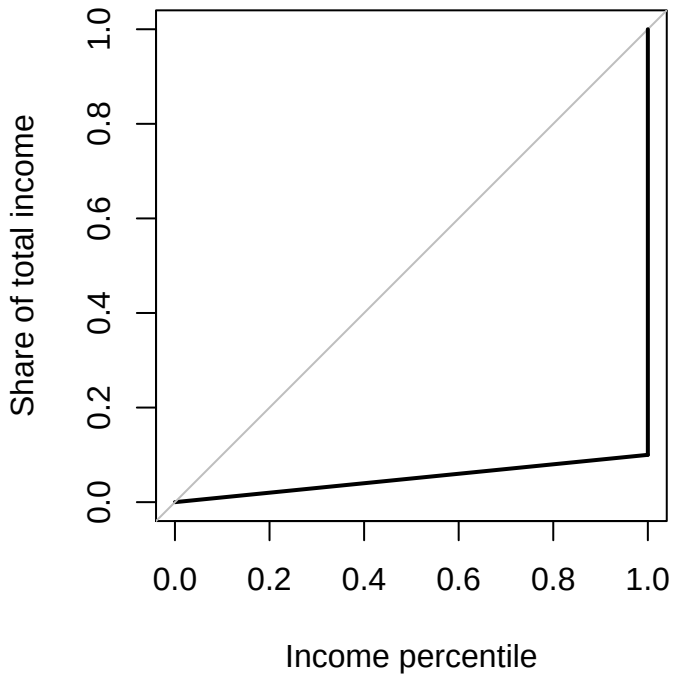
But money just didn't mean all that much in those economies — you couldn't (say) go out and buy a car (or bread) just because you had the cash for the price of a car (or bread) in your wallet. For that matter, there were ways of getting cars (and bread) even if you didn't have the money but you did have connections to people in the Communist Party (or the underworld). So trying to figure out the distribution of real income under Communism is hard. One of the best reviews of the subject, Ellman (2014), ch. 7, table 7.7, suggests that in the 1970s the USSR and its satellites in Eastern Europe were as unequal, or even a bit more unequal, than capitalist countries like Sweden or Canada, though less unequal than the US.

If we try to look at even older economies, like Europe in the Middle Ages, we again have the difficulty that lots of their economic life just didn't involve money. (Most people were peasant farmers, and most of what they ate, wore and used was stuff they'd grown and woven themselves. Even most of the rent they paid to their lords was "in kind", not in money.) We can say that basically ever since people invented cities, *some* people have been much richer than others, and that ever since they invented writing they've been *complaining* about that (and/or trying to calculate just how much richer someone else is), but meaningful statements about the exact distribution are hard.

What is the *real* maximum Gini index?

I said above that the theoretical maximum for the Gini index is 1, when one person gets all the income and infinitely many other people get an income of zero. This results in a Gini index of 1.

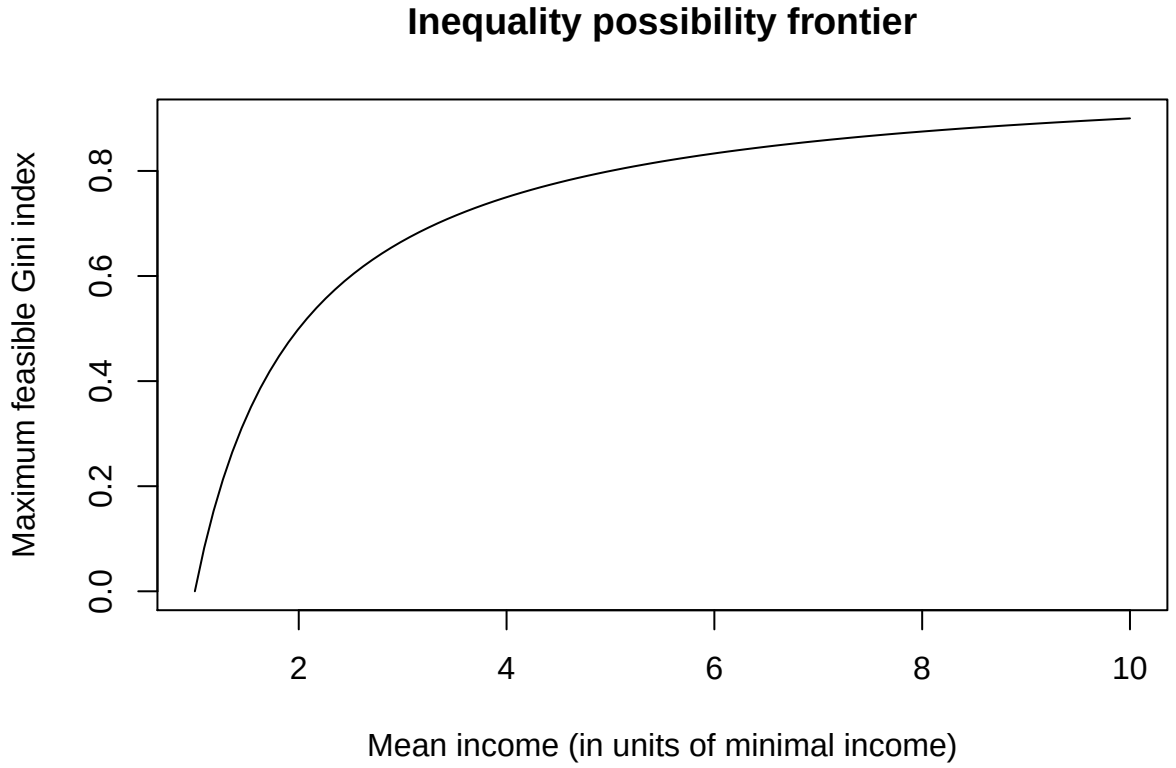
Now, *actually* getting zero income is not very sustainable in a market economy, certainly not at the household level. If you imagine that there is a certain minimum level of income needed to keep a household alive, the maximum feasible level of inequality is when every household gets exactly that minimum, and *one* household gets everything above the minimum. The resulting Lorenz curve would look like this, in the limit of a very large population:



Let's be a little more concrete. Say that the minimum per unit is m, there are $n+1$ units in the population, and average income $\overline{x}$. There are n units which all receive income m, and one unit which receives income $n(\overline{x}-m)$. The Lorenz curve then has three corners, $(0,0)$, $(\frac{n}{n+1}, \frac{m}{\overline{x}})$ and $(1,1)$. A little algebra (see Complementary Problem 2) says that in the limit of large population size ($n \to \infty$), the Gini index for this Lorenz curve will approach

$$1 - \frac{m}{\overline{x}}$$

which looks like so:



In words, a very poor society just *can't* be too unequal, because everyone just barely has enough to survive. It's only when a society gets richer that inequality becomes possible, and the richer it is, the more inequality it can afford.

The reasoning above is entirely from an extremely ingenious paper, Milanovic, Lindert, and Williamson (2011). This paper also uses the very scattered data sources we have about income in ancient societies to try to guess at their mean incomes and their Gini coefficients. To the extent we can trust their estimates, they suggest that many pre-industrial societies had lower Gini coefficients than rich democracies today, but were much closer to their maximum possible Ginis — they were about as unequal as their poverty allowed. This is a very plausible conclusion, but also one which, as I've indicated, involves a *lot* of guesswork.

What about wealth?

We've been looking at **income**, how much money people make over some course of time, in this case a year. It's also interesting to look at the distribution of **wealth**, how much people's property is worth. This is usually defined as **net worth**, meaning the price of what people own *minus* the amounts they owe. In most contemporary economies, lots of people have zero or even negative net worth. (For instance, it's not uncommon for people to graduate from prestigious American schools with high incomes but negative net worth, because of large student loans.) That being said, the shape of the wealth distribution is typically like what we've been seeing for income distributions, with most people having fairly modest amounts of wealth, and a substantial minority owning orders of magnitude more. In fact, wealth distributions are typically *more*

unequal and concentrated, by any of the measures of inequality and concentration which we'll cover, than are income distributions. (You'll get a chance to work with wealth distributions in a later homework.)

Complementary Problems

These are just to think through and/or practice on, not hand in.

1. Suppose we have a population of n individuals, sorted in increasing order, so $0 \leq x_1 \leq x_2 \leq \ldots \leq x_n$. Write $\overline{x}_p$ for the average of the first n_p individuals.
   a. Explain why $\overline{x}_p \leq \overline{x}$.
   b. Define $s(p)$ as in the section on the Lorenz curve. Using (a), explain why $s(p) \leq p$.
   c. Suppose that all incomes are exactly equal; show that $s(p) = p$.
   d. Suppose that not all incomes are equal. Show that there must be *some* value of p where $s(p) < p$.
   e. Do we need the assumption that all the values of x_i are strictly > 0?
2. Read the section on "What is the *real* maximum Gini index?" if you haven't already.
   a. Show that the area under the Lorenz curve is
   $$\frac{1}{2}\frac{n}{n+1}\frac{m}{\overline{x}} + \frac{1}{n+1}\left(\frac{1+m/\overline{x}}{2}\right)$$
   b. Explain why for large populations, this will approach $\frac{1}{2}\frac{m}{\overline{x}}$.
   c. Find the Gini index in the large population limit.

References

Eichengreen, Barry. 2015. *Hall of Mirrors: The Great Depression, the Great Recession, and the Uses — and Misuses — of History*. Oxford: Oxford University Press.

Ellman, Michael. 1978. "The Fundamental Problem of Socialist Planning." *Oxford Economic Papers* 30:249–62. https://www.jstor.org/stable/2662890.

———. 2014. *Socialist Planning*. Third. Cambridge, England: Cambridge University Press. https://doi.org/10.1017/CBO9781139871341.

Levinson, Marc. 2006. *The Box: How the Shipping Container Made the World Smaller and the World Economy Bigger*. Princeton, New Jersey: Princeton University Press.

Lorenz, M. O. 1905. "Methods of Measuring the Concentration of Wealth." *Publications of the American Statistical Association* 9 (70):209–19. https://doi.org/10.2307/2276207.

Mettler, Suzanne. 2011. *The Submerged State: How Invisible Government Policies Undermine American Democracy*. Chicago: University of Chicago Press.

Milanovic, Branko, Peter H. Lindert, and Jeffrey G. Williamson. 2011. "Pre-Industrial Inequality." *The Economic Journal* 121:255–72. https://doi.org/10.1111/j.1468-0297.2010.02403.x.

Piketty, Thomas, and Emmanuel Saez. 2003. "Income Inequality in the United States, 1913–1998." *Quarterly Journal of Economics* 118:1–39. https://doi.org/10.1162/00335530360535135.

Tooze, Adam. 2018. *Crashed: How a Decade of Financial Crises Changed the World*. New York: Viking.