Modeling Income and Wealth Distributions I

36-313, Fall 2021

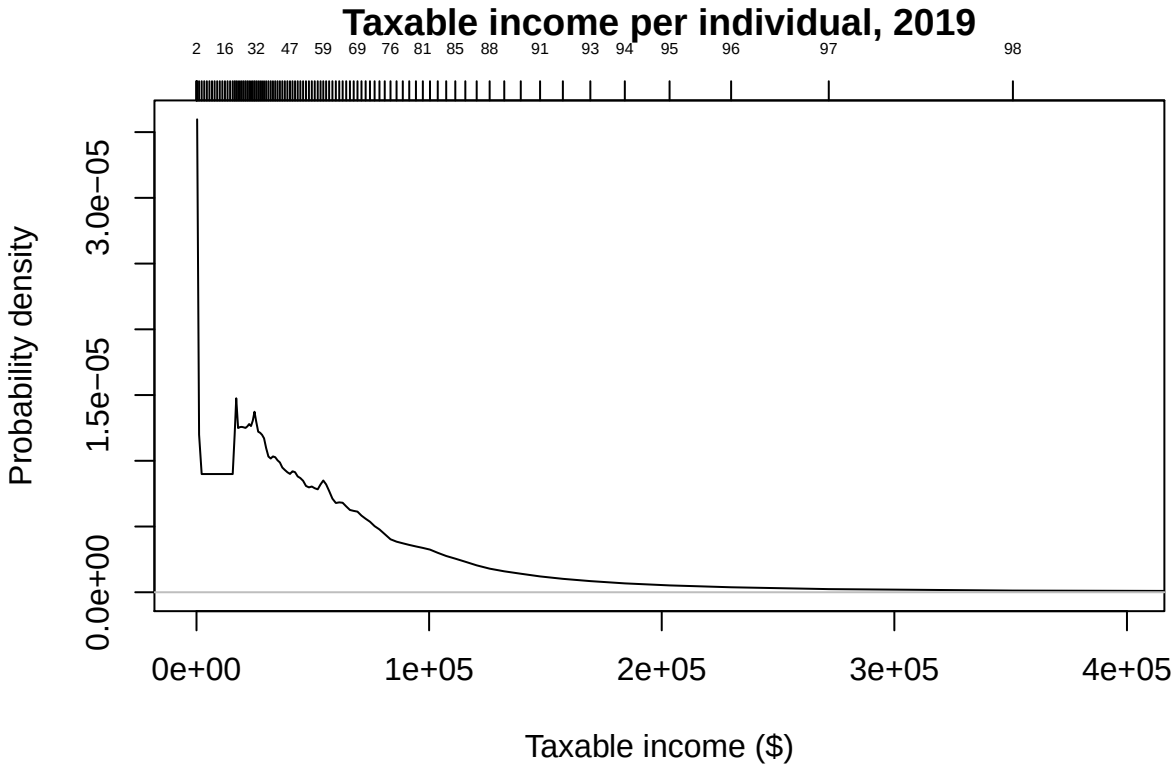
7 September 2021 (Lecture 3)

Contents

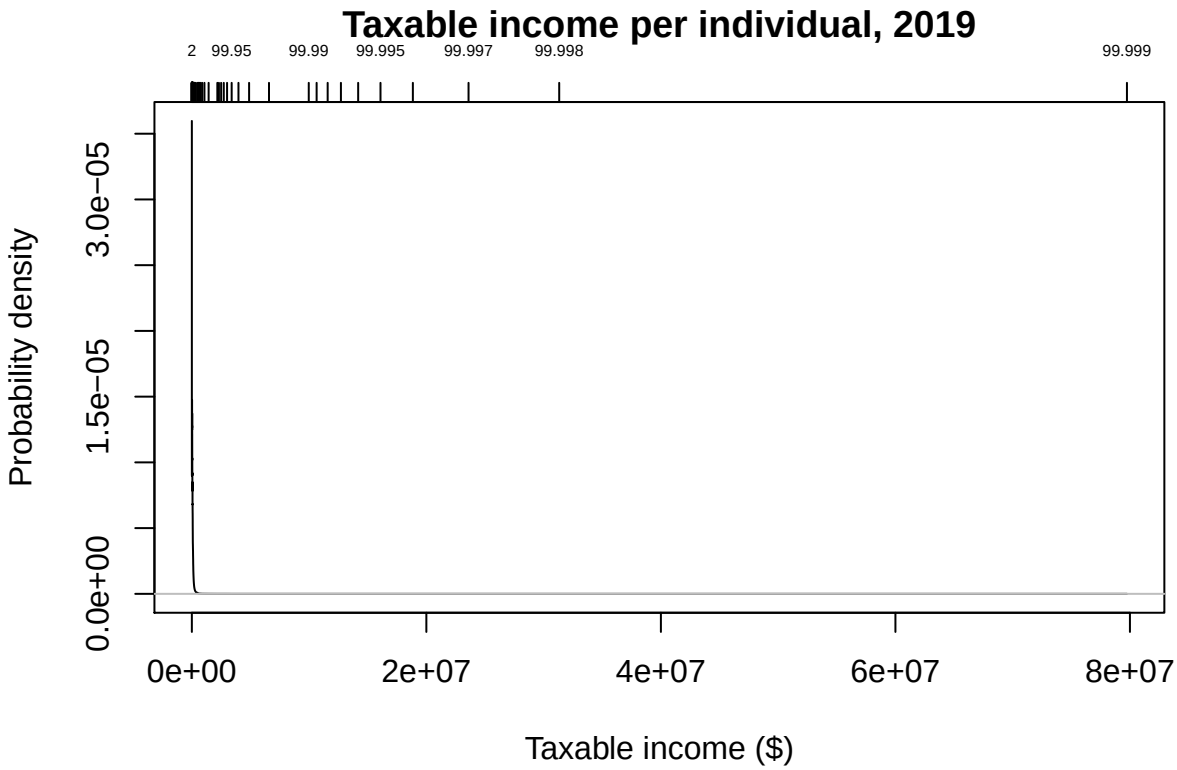
In this lecture, we're going to look at *models* for the distribution of income and wealth. These aren't facts about they world; they're ideas we've invented to help organize and make sense of those facts. Models are simplified, stylized pictures which deliberately suppress some details in order to highlight what are (hopefully) larger and more important patterns. To decide which stylized picture to draw, though, we need to pay some attention to the facts.

Reminders of what income distributions look like

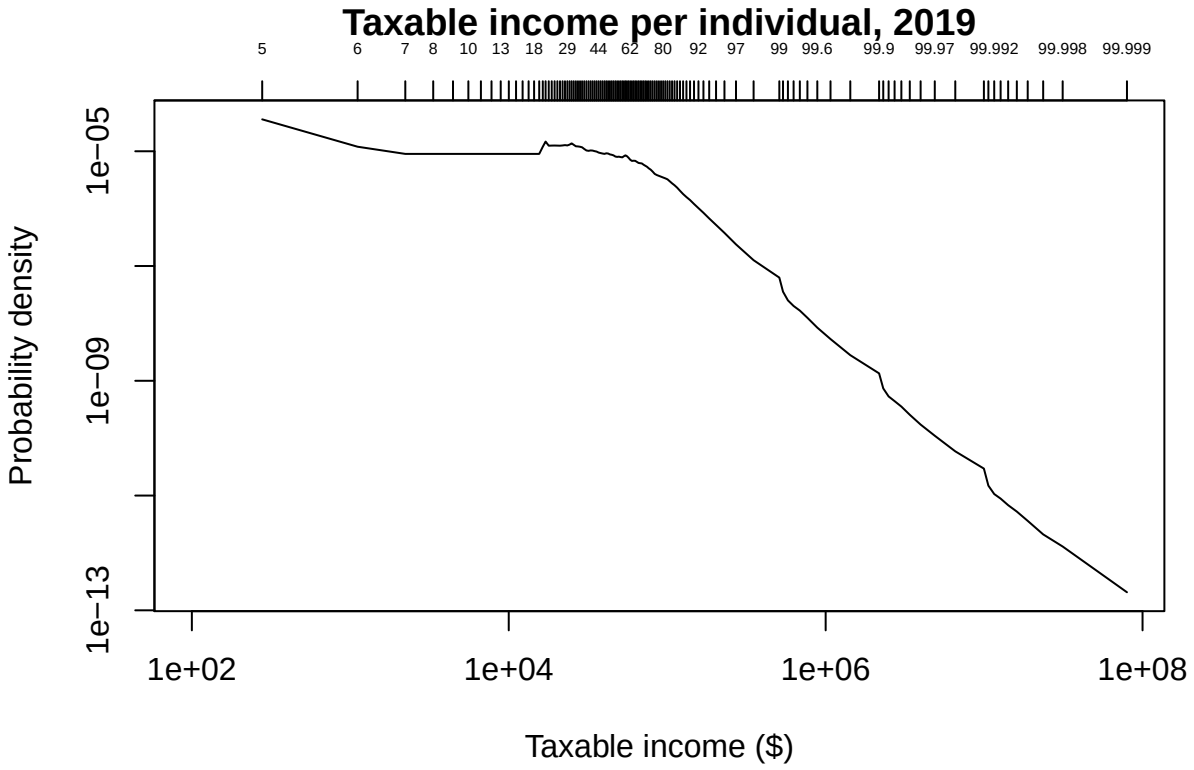
Here, as a reminder, is an approximation to the distribution of individual income for the US in 2019.



If we zoom out to the full range, this looks pretty much like a spike, because the range is so great:



On the other hand, the little numbers on the upper edge of the plot show percentiles of income, so it's not that there's *no* probability density out around an income of (say) 20 million dollars a year. We can get a bit better sense of what's going on by using a logarithmic scale on both axes:



What should we take away from these plots?

1. There's fairly substantial probability density for a wide range around the median income, maybe from an order of magnitude below it to an order of magnitude above it. (Look at the close bunching of the quantile marks in the log-scale plot.)
2. The distribution is highly right-skewed: the mean is much larger than the median.
3. The distribution is heavy tailed: high quantiles are many orders of magnitude larger than the median, and very high quantiles are orders of magnitude larger than merely high quantiles.
4. At very high incomes, the log-log plot of density looks like a straight line. This is what we'd see if the pdf were shrinking like some power of income, $f(x) \propto x^{-\alpha}$, since then $\log f(x) = -\alpha \log x + (\text{intercept})$.

All of these are pretty common features of income and wealth distributions. (See Complementary Problem 1.) When we introduce models, it'd be nice if they could account for, or at least reproduce, some of these features. While there isn't, unfortunately, one simple model which does all of this, there are two simple models each of which captures *part* of what's going on.

The Log-Normal Distribution

You remember the Gaussian or "normal" distribution; it's conveniently defined by its pdf,

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

There are two adjustable knobs in this, the **parameters** μ and σ^2. We write this $X \sim \mathcal{N}(\mu, \sigma^2)$, where the symbol $\sim$ is read "is distributed as".

Some calculus, which I won't reproduce, shows that with this pdf, $\mathbb{E}(X) = \mu$ while $\mathrm{Var}(X) = \sigma^2$. The highest probability density, the mode, occurs when $x = \mu$. As x moves away from μ in either direction, the pdf falls off, because $(x-\mu)^2 \geq 0$. Initially, the fall-off is slow, because if $|x-\mu|/\sigma < 1$, then $(x-\mu)^2/\sigma^2 < |x-\mu|/\sigma$. (Incidentally, this algebra explains why "close to μ" and "far from μ" are measured with respect to the

standard deviation σ.) But conversely, once $|x-\mu|/\sigma$ becomes large, the pdf goes to zero very rapidly. The familiar conclusion is that with high probability, a random X drawn from this distribution will be found near μ, meaning within $\pm$ a few multiples of σ of μ.

The log-normal distribution is just what it sounds like: X has the log-normal[1] distribution when $\log X$ has a normal or Gaussian distribution[2]. We write this as $X \sim \mathcal{LN}(\mu, \sigma^2)$, with the understanding that now $\mathbb{E}(\log X) = \mu$, $\mathrm{Var}(\log X) = \sigma^2$. What does this mean in practice?

Some properties of log-normals

First, log is a strictly increasing function, so

$$\mathbb{P}(X \leq b) = \mathbb{P}(\log X \leq \log b)$$

and thus

$$\mathbb{P}(a < X \leq b) = \mathbb{P}(\log a < \log X \log b)$$

One immediate implication is that quantiles of X are just the exponentials of the quantiles of $\log X$. In particular, the median of X is e^{μ}. (Why does that follow?)

Another implication of this basic relationship is the following. We know that it's very improbable for $\log X$ to be more than a few multiples of σ away from μ. Let's write this formally:

$$\mathbb{P}(\mu - k\sigma < \log X \leq \mu + k\sigma) = \text{nearly } 1$$

where k is just a convenient algebraic symbol for "a few". But this implies

$$\mathbb{P}\left(e^{\mu}(e^{-\sigma})^k < X \leq e^{\mu}(e^{\sigma})^k\right) = \text{nearly } 1$$

So it's very likely that X is within a *factor* of a few powers of e^{σ} of e^{μ}. The range probable range of X, in other words, is multiplicative, not additive. (This is already promising in view of the empirical pdf of income.)

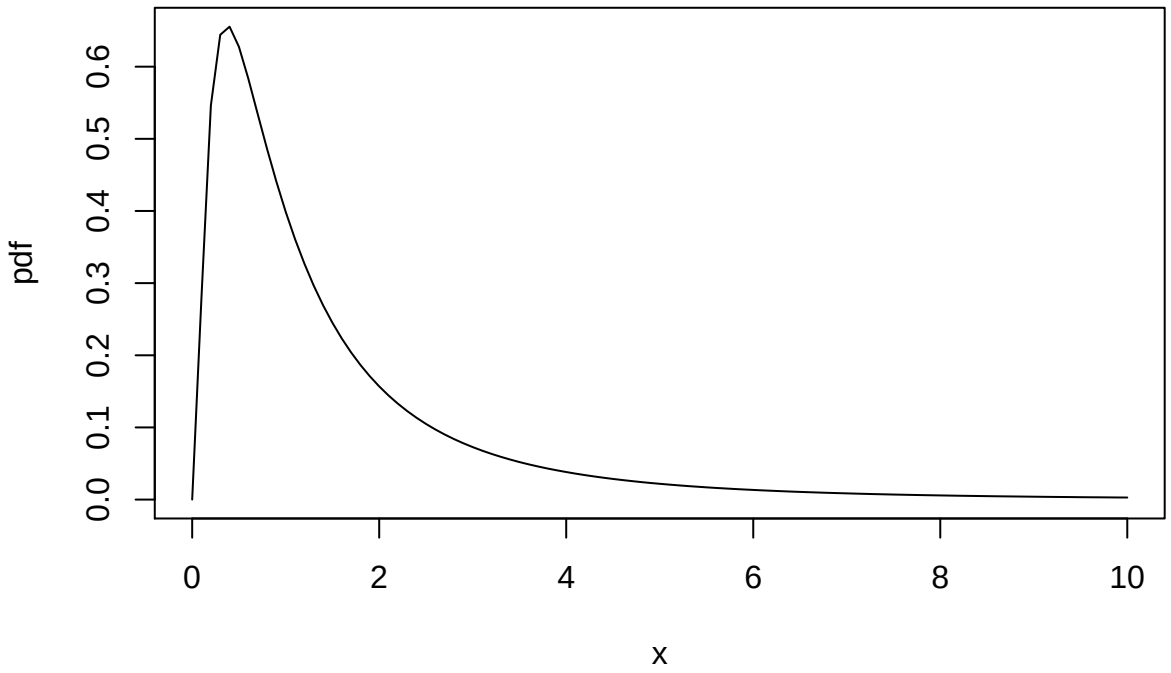
We can quickly find the exact pdf of the log-normal distribution. Remember the definitions: the CDF is $F(a) \equiv \mathbb{P}(X \leq a)$, and the pdf is the derivative of the CDF, $f(x) = \frac{dF}{da}\big|_{a=x}$. So when X is log-normal,

$$\begin{aligned}
f(x) &= \frac{d}{da}\mathbb{P}(X \leq a)\Big|_{a=x} && (1)\\
&= \frac{d}{da}\mathbb{P}(\log X \leq \log a)\Big|_{a=x} && (2)\\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(\log x - \mu)^2}{2\sigma^2}\right\}\frac{d\log x}{dx} && (3)\\
&= \frac{1}{x}\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(\log x - \mu)^2}{2\sigma^2}\right\} && (4)
\end{aligned}$$

Now that we have the pdf, we can plot it. Here's the pdf of the "standard" log-normal, $\mathcal{LN}(0,1)$:

[1] Some people write "log-normal", others write "lognormal"; no difference. I tend to use the hyphen because it makes my spell-checker happier.

[2] It really doesn't matter what "base" of logarithm we're using, whether log is really $\log_{10}$ or $\log_2$ or ln. But some of the calculus is easier if we use natural logs, so if you want to read ln for log throughout, that's fine.

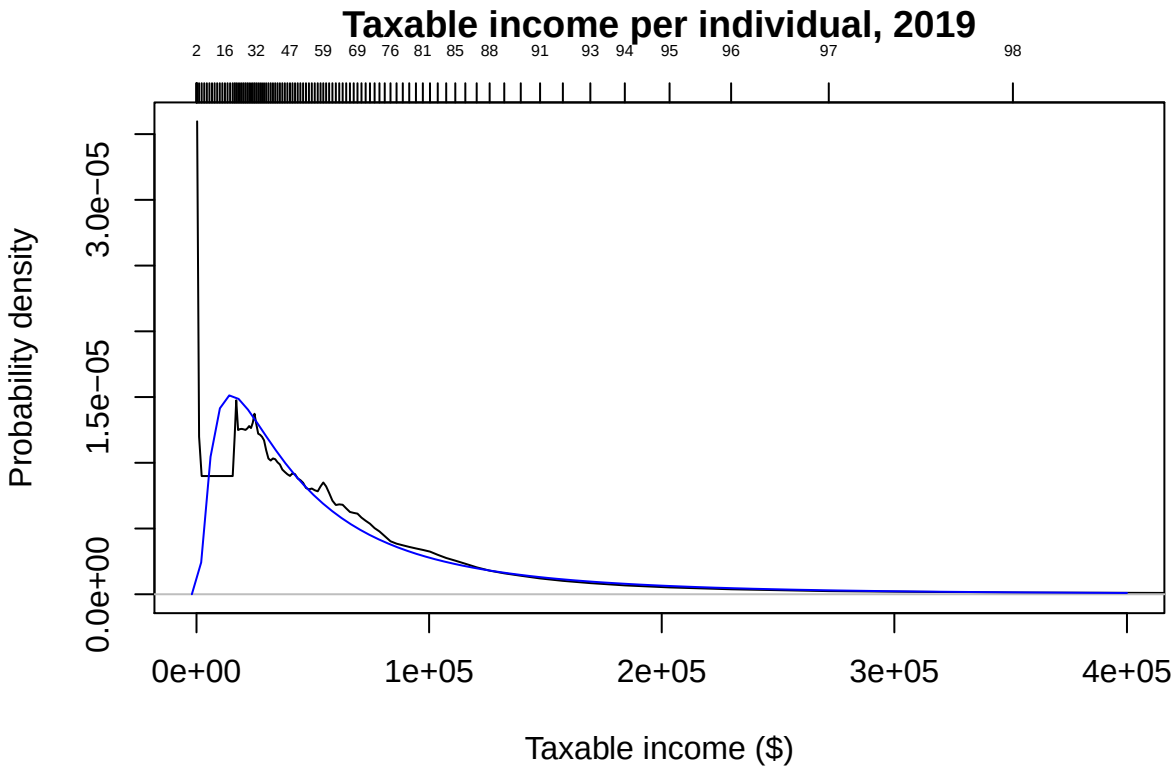


Notice that while the median is $e^0 = 1$, the mode is clearly lower than that. Since we know the pdf, we can find the mode exactly by solving $\frac{df}{dx} = 0$ (Complementary Problem 2a), with the result that the mode is exactly $e^{\mu-\sigma^2}$.

On the other hand, mean of the log-normal is higher than the median, $e^{\mu+\sigma^2/2}$ (Complementary Problem 2b). This reassures us that, as we'd expect from the plot of the pdf, the log-normal distribution is right-skewed, since $\sigma^2 > 0$.

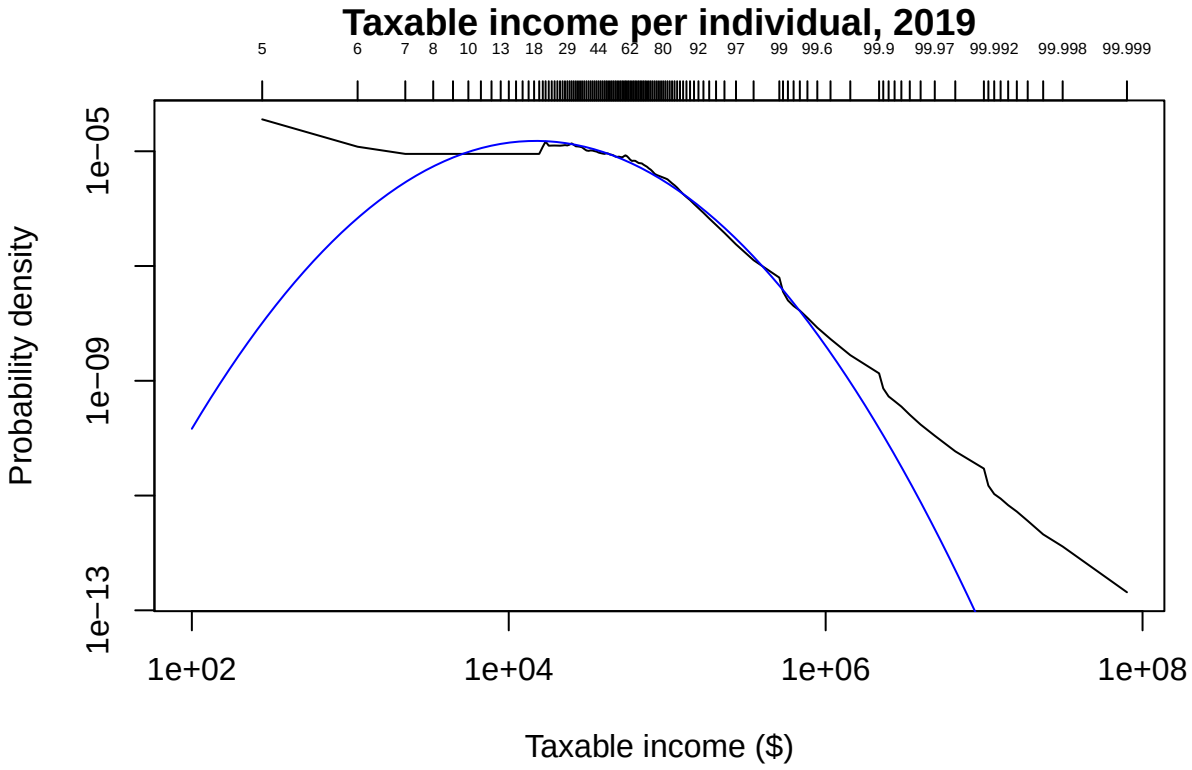
Why we care: log-normal versus income data

Here's our first plot of the pdf of income (as calculated from the income percentiles) again, with a log-normal distribution added in blue.



(This log-normal is fitted in exactly the way you're learning to do in HW 1. We'll cover fitting methods in more detail next lecture.) What you can see is that while the log-normal distribution isn't quite right at very low incomes, and in particular messes up how many people have incomes close to 0, from about the 10th percentile of the income distribution onwards it's a close match to the data, at least visually, through the 98th percentile and perhaps above.

Let's look at this same comparison in a log-log plot, which will emphasize the right tail.



Here we can see that there's a visually good match over most of the range of incomes, from about the 10th percentile (or maybe a little lower?) to above the 99th percentile. The fact that the log-normal distribution systematically under-predicts how many people have (or report) very low incomes shows up again. What's new is being able to see what's happening in the right or upper tail: the log-normal also under-predicts how many people have extremely high incomes. (Said another way, it predicts that the highest percentiles of income should be much lower than they actually are.) So while the log-normal is a pretty good approximation to the distribution of income for most of the population, "the rich are different than you and me".

The Power-Law or Pareto Distribution

A good (but not perfect) model of the upper tail of income and wealth distributions has actually come down to us from the first social scientist to seriously investigate the issue, the pioneering economist and sociologist Vilfredo Pareto, who first published his findings in the 1890s. He used tax records to examine the number of people with very high incomes or net worths in a number of countries and cities for which he could find the data. What he did visually was plot the number of tax-payers whose income met or exceeded any given level (e.g., \$1000, \$2000, ...) against that threshold level. In other words, he plotted $\mathbb{P}(X \geq x)$ against x. In modern terminology, we'd call this the **complementary cumulative distribution function** (CCDF) or **upper CDF** or **survival function**, often written $\overline{F}(x)$. When Pareto made these plots with log scales on both the vertical and horizontal axes, he found that the curve became a straight line, at least above some minimum income or wealth. This implied that $\overline{F}(x)$ was proportional to some power of x, at least above that minimum level.

What we now call the **continuous power law distribution** or **Pareto distribution** has the following pdf:

$$f(x) = \begin{cases} 0 & x < x_{\min} \\ \frac{\alpha-1}{x_{\min}} \left(\frac{x}{x_{\min}}\right)^{-\alpha} & x \geq x_{\min} \end{cases}$$

In words, the pdf is a power of x, hence the name "power law".

If you integrate this, you will find that, for $x \geq x_{\min}$,

$$\mathbb{P}(X \geq x) = \left(\frac{x}{x_{\min}}\right)^{-\alpha+1}$$

so the CCDF is *also* a power of x.

At this point I need to say that there is a small but annoying discrepancy in the literature. Basically everyone writes α for the Pareto exponent, but for some people, that's the exponent in the pdf, and for other people, that's the exponent in the CCDF. The two exponents always differ by exactly one, so this isn't a big deal, but if you just want to borrow a formula, or compare empirical results, it's annoying to have to constantly check which convention is being used by a particular author.

Some properties of the Pareto distribution

Mode

The mode of the Pareto distribution is $x_{\min}$: the highest probability density is *always* at the smallest possible value.

Median and other quantiles

If we assume that there's no probability mass below $x_{\min}$, then it's easy to find upper quantiles:

$$\begin{aligned}
\mathbb{P}(X \geq x_q) &= q & (5)\\
\left(\frac{x_q}{x_{\min}}\right)^{-\alpha+1} &= q & (6)\\
x_q &= x_{\min} q^{1/(1-\alpha)} & (7)
\end{aligned}$$

Thus the ordinary, lower quantiles are also easy:

$$\begin{aligned}
\mathbb{P}(X \leq x_p) &= p & (8)\\
x_p &= x_{\min}(1-p)^{1/(1-\alpha)} & (9)
\end{aligned}$$

Matters are a little more complicated if we're just using the Pareto as a model for the tail above $x_{\min}$. *If* $x_q \geq x_{\min}$, then

$$\begin{aligned}
\mathbb{P}(X \geq x_q) &= \mathbb{P}(X \geq x_q | X \geq x_{\min})\, \mathbb{P}(X \geq x_{\min}) & (10)\\
&= \left(\frac{x_q}{x_{\min}}\right)^{-\alpha+1} \mathbb{P}(X \geq x_{\min}) & (11)
\end{aligned}$$

Setting this equal to q and solving for x_q gives

$$x_q = x_{\min}(q/\mathbb{P}(X \geq x_{\min}))^{1/(1-\alpha)} \quad (12)$$

Expectation and higher moments

The expected value of X is easily calculated:

$$\begin{aligned}
\mathbb{E}\left(X|X \geq x_{\min}\right) &= \int_{x_{\min}}^{\infty} x \frac{\alpha-1}{x_{\min}} \left(\frac{x}{x_{\min}}\right)^{-\alpha} dx && (13)\\
&= \frac{\alpha-1}{x_{\min}} x_{\min}^{\alpha} \int_{x_{\min}}^{\infty} x^{-\alpha+1} dx && (14)\\
&= (\alpha-1) x_{\min}^{\alpha-1} \frac{1}{\alpha-2} x_{\min}^{-\alpha+2} && (15)\\
&= x_{\min} \frac{\alpha-1}{\alpha-2} && (16)
\end{aligned}$$

Notice that as $\alpha \to 2$, this goes to infinity. The practical meaning of this is that when $\alpha \leq 2$, as we get larger and larger samples, the sample mean will keep getting larger and larger without any limit. This is *not* how most random variables work, but it is the way of very heavy-tailed distributions.

Higher moments are even less well-behaved:

$$\begin{aligned}
\mathbb{E}\left(X^k|X \geq x_{\min}\right) &= \int_{x_{\min}}^{\infty} x^k \frac{\alpha-1}{x_{\min}} \left(\frac{x}{x_{\min}}\right)^{-\alpha} dx && (17)\\
&= (\alpha-1) x_{\min}^{\alpha-1} \int_{x_{\min}}^{\infty} x^{k-\alpha} dx && (18)\\
&= \frac{\alpha-1}{\alpha-k-1} x_{\min}^{\alpha-1} x_{\min}^{k+1-\alpha} && (19)\\
&= x_{\min}^{k} \frac{\alpha-1}{\alpha-k-1} && (20)
\end{aligned}$$

Since the variance is $\mathrm{Var}\left(X\right) = \mathbb{E}\left(X^2\right) - \left(\mathbb{E}\left(X\right)\right)^2$, it follows that the variance is infinite if $\alpha \leq 3$. Again, the practical meaning is that if you keep drawing larger and larger samples from such a distribution, the sample variance will grow without any limit.

Why we care: Pareto distributions versus tail data

You will explore this in detail in Homework 2. I will just say here that while Pareto distributions aren't a *perfect* model fit to the upper tail, they are pretty good approximations. You can, for instance, see this in the log-log plots of the pdf I gave above — the right tail is very close to a straight line.

Calculating measures of inequality from theoretical distributions

One use of these models is to simplify or "regularize" the calculation of Lorenz curves, Gini indices, etc.

Lorenz curves and Gini indices from the Pareto distribution

Let's do an example of calculating Lorenz curves and Gini coefficients from a Pareto distribution. To simplify matters, I will assume that there is *no* probability mass below $x_{\min}$, but you can re-do the calculations without this assumption (Complementary Problem 3).

Income shares and the Lorenz curve

Let's begin by working out the quantiles of the Pareto distribution. The starting point will be the CCDF:

$$\begin{aligned} \overline{F}(x_q) &= q && (21)\\ \left(\frac{x_q}{x_{\min}}\right)^{-\alpha+1} &= q && (22)\\ \frac{x_q}{x_{\min}} &= q^{1/(1-\alpha)} && (23)\\ x_q &= x_{\min} q^{1/(1-\alpha)} && (24) \end{aligned}$$

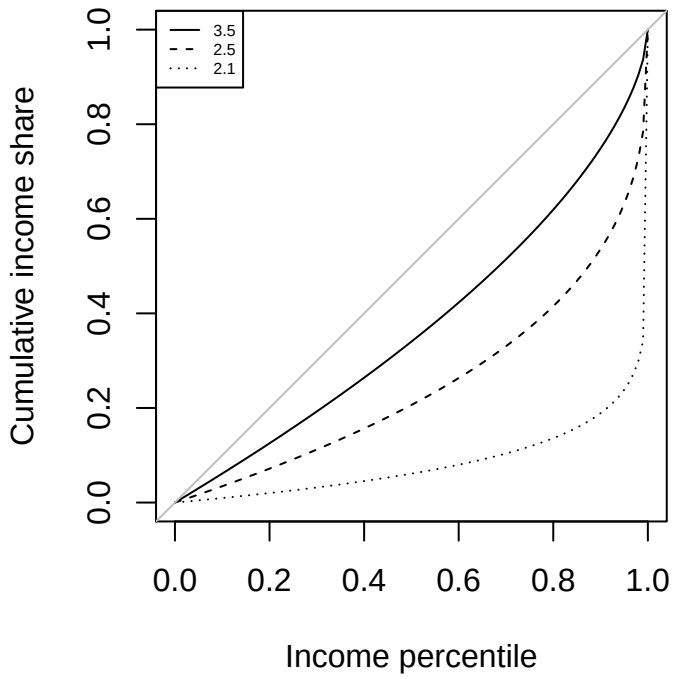
With this in hand, we can work out what fraction of total income goes to the upper q fraction of the population, call this $S(q)$:

$$\begin{aligned} S(q) &= \frac{\int_{x_q}^{\infty} x \frac{\alpha-1}{x_{\min}} \left(\frac{x}{x_{\min}}\right)^{-\alpha} x}{\mathbb{E}(X)} && (25)\\ &= \frac{\frac{\alpha-1}{x_{\min}}^{\alpha-1} \int_{x_{\min} q^{1/(1-\alpha)}}^{\infty} x^{-\alpha+1} dx}{x_{\min} \frac{\alpha-1}{\alpha-2}} && (26)\\ &= x_{\min}^{\alpha-2} \frac{\frac{1}{\alpha-2} x_{\min}^{-\alpha+2} q^{(\alpha-2)/(\alpha-1)}}{\frac{1}{\alpha-2}} && (27)\\ &= q^{(\alpha-2)/(\alpha-1)} && (28) \end{aligned}$$

Now by convention the Lorenz curve is plotted using *lower* percentiles, say $p = 1 - q$, so we can say

$$s(p) = 1 - (1-p)^{(\alpha-2)/(\alpha-1)} \quad (29)$$

Here are some pictures:



We can see that as α decreases, the curve gets further and further away from the diagonal, indicating increasing inequality.

(Note that this result will break down once $\alpha < 2$ — why?)

The Gini index

Recall, from Lecture 2, that the Gini index is determined by the area under the Lorenz curve:

$$G = 1 - 2 \times \text{area} \tag{30}$$

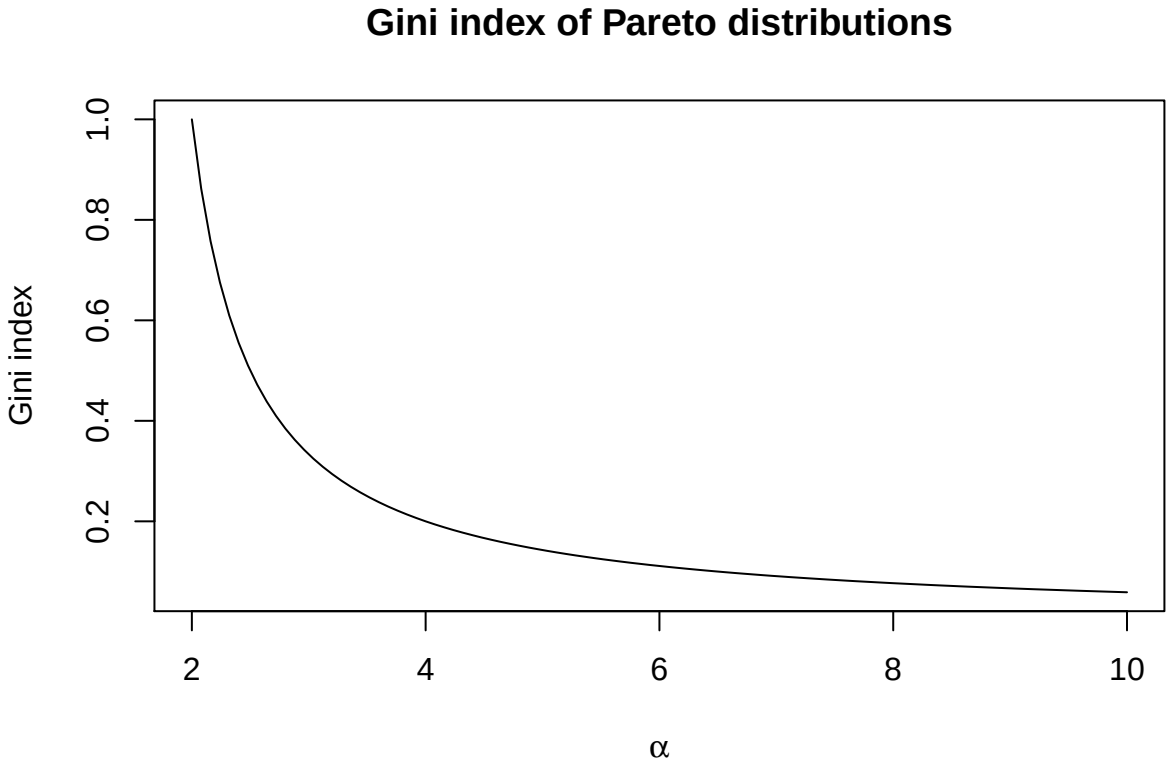
The area under the Lorenz curve is straightforward to find from the income shares:

$$\begin{aligned}
\text{area under Lorenz curve} &= \int_0^1 s(p)dp && (31)\\
&= \int_0^1 \left(1 - (1-p)^{(\alpha-2)/(\alpha-1)}\right) dp && (32)\\
&= 1 - \int_0^1 (1-p)^{(\alpha-2)/(\alpha-1)} dp && (33)\\
&= 1 - \int_1^0 q^{(\alpha-2)/(\alpha-1)}(-1)dq && (34)\\
&= 1 - \int_0^1 q^{(\alpha-2)/(\alpha-1)} dq && (35)\\
&= 1 - \frac{1}{\frac{\alpha-2}{\alpha-1}+1}[1-0] && (36)\\
&= 1 - \frac{1}{\frac{\alpha-2+\alpha-1}{\alpha-1}} && (37)\\
&= 1 - \frac{\alpha-1}{2\alpha-3} && (38)\\
&= \frac{2\alpha-3-\alpha+1}{2\alpha-3} = \frac{\alpha-2}{2\alpha-3} && (39)
\end{aligned}$$

so

$$\begin{aligned}
G &= 1 - 2\frac{\alpha-2}{2\alpha-3} && (40)\\
&= \frac{2\alpha-3-2\alpha+4}{2\alpha-3} = \frac{1}{2\alpha-3} && (41)
\end{aligned}$$

This leads to the following relationship between α and the Gini index:



Lorenz curves and Gini indices for the log-normal distribution

Homework 1 quoted some results about the shape of the Lorenz curve and the Gini index for the log-normal distribution, but without asking you to prove them. Here are quick sketches of the proofs.

Income shares and the Lorenz curve

Let's start by working out the (lower) quantiles of the $\mathcal{LN}(\mu, \sigma^2)$ distribution:

$$\begin{aligned}
\mathbb{P}(X \leq x_p) &= p && (42)\\
\mathbb{P}(X \leq \log x_p) &= p && (43)\\
\Phi((\log x_p - \mu)/\sigma)) &= p && (44)\\
\log x_p &= \mu + \sigma\Phi^{-1}(p) && (45)\\
x_p &= e^{\mu}\exp\left\{\sigma\Phi^{-1}(p)\right\} && (46)
\end{aligned}$$

Now what's the share of income going to the lower p fraction of the population?

$$\begin{aligned}
s(p) &= \frac{\int_0^{x_p} xf(x)dx}{\mathbb{E}(X)} && (47)\\
&= \frac{\int_0^{e^{\mu}\exp\left\{\sigma\Phi^{-1}(p)\right\}} \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2}\frac{(\mu-\log x)^2}{\sigma^2}dx\right\}}{e^{\mu}e^{\sigma^2/2}} && (48)\\
& && (49)
\end{aligned}$$

Set $z = \log x$, so $dz/dx = 1/x$ but $dx = xdz = e^z dz$; make z the new variable of integration and change the

limits accordingly:

$$\int_{-\infty}^{e^{\mu}\exp\left\{\sigma\Phi^{-1}(p)\right\}} \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2}\frac{(\mu-\log x)^2}{\sigma^2}\right\}dx \tag{50}$$

$$= \int_{-\infty}^{\mu+\sigma\Phi^{-1}(p)} \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2}\frac{(\mu-z)^2}{\sigma^2}\right\}e^z dz \tag{51}$$

$$= \int_{-\infty}^{\mu+\sigma\Phi^{-1}(p)} \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(\mu^2-2\mu z-2z\sigma^2+z^2)\right\}dz \tag{52}$$

$$= \int_{-\infty}^{\mu+\sigma\Phi^{-1}(p)} \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(\mu^2-2(\mu+\sigma^2)z+z^2)\right\}dz \tag{53}$$

$$= \int_{-\infty}^{\mu+\sigma\Phi^{-1}(p)} \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}((\mu+\sigma^2)^2-2(\mu+\sigma^2)z+z^2-\sigma^4-2\mu\sigma^2)\right\}dz \tag{54}$$

$$= \int_{-\infty}^{\mu+\sigma\Phi^{-1}(p)} \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{\sigma^2/2+\mu\right\}\exp\left\{-\frac{1}{2\sigma^2}((\mu+\sigma^2-z)^2\right\}dz \tag{55}$$

$$= e^{\sigma^2/2+\mu}\int_{-\infty}^{\mu+\sigma\Phi^{-1}(p)} \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}((\mu+\sigma^2-z)^2\right\}dz \tag{56}$$

At the end of all this manipulation ("completing the square"), what we're left with as the integrand is the pdf of a Gaussian with mean $\mu+\sigma^2$ and variance σ^2, so this becomes

$$e^{\sigma^2/2+\mu}\Phi\left(\frac{(\mu+\sigma\Phi^{-1}(p))-(\mu+\sigma^2)}{\sigma}\right) \tag{57}$$

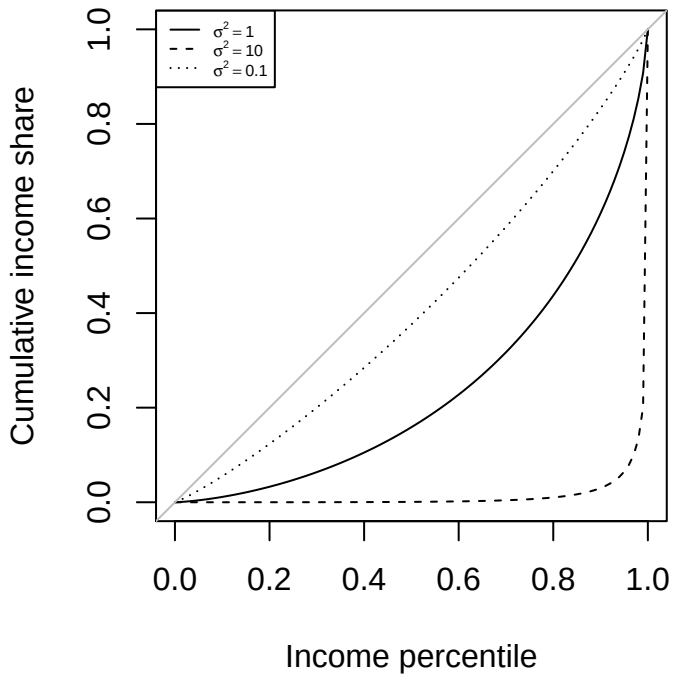
$$= e^{\sigma^2/2+\mu}\Phi(\Phi^{-1}(p)-\sigma) \tag{58}$$

Thus

$$s(p) = \frac{e^{\sigma^2/2+\mu}\Phi(\Phi^{-1}(p)-\sigma)}{e^{\mu+\sigma^2/2}} \tag{59}$$

$$= \Phi(\Phi^{-1}(p)-\sigma) \tag{60}$$

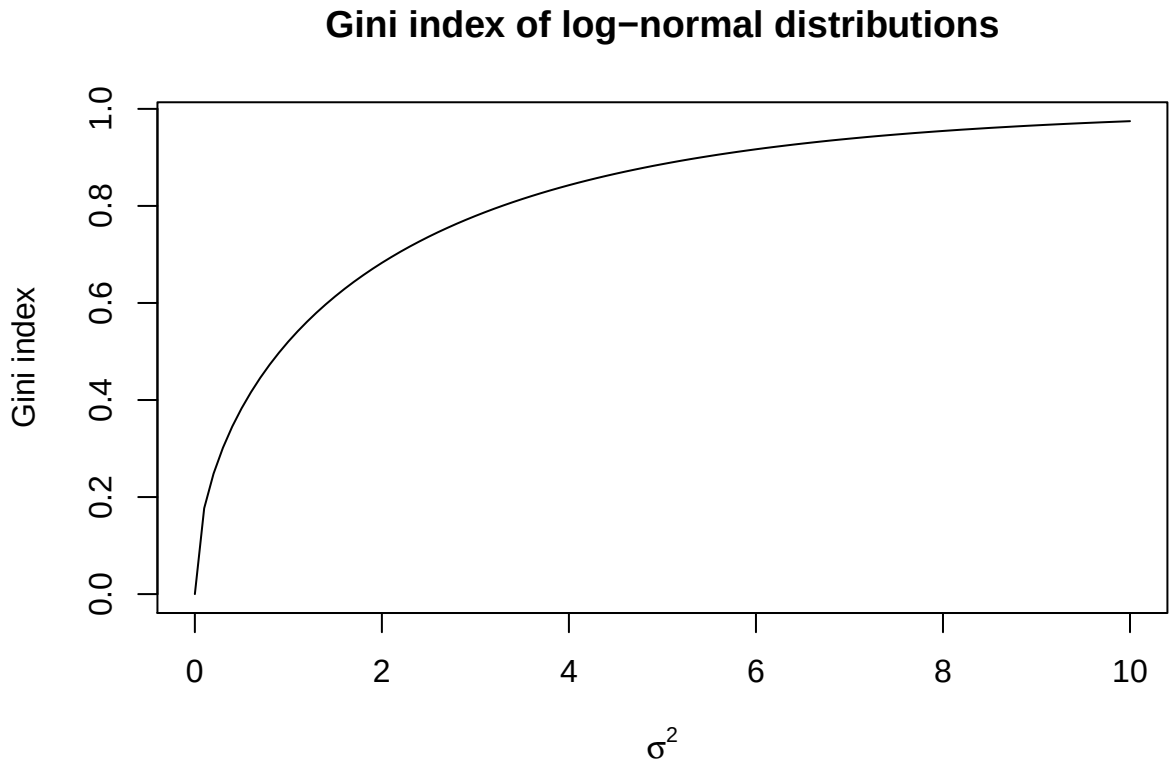
as claimed (without proof) in Homework 1.



Notice that μ doesn't appear in $s(p)$ at all, so two log-normal distributions with different μs but the same σ will have the same Lorenz curve. This makes some sense: changing μ will change everyone's income by the same factor, and that's precisely the kind of change which the Lorenz curve ignores.

Gini index

This $2\Phi(\sigma/\sqrt{2}) - 1$, but you're tired of reading me do integrals by now, so I'll make this Complementary Problem 3. Notice that since the Lorenz curve for a log-normal only involves σ and not μ, the Gini index couldn't involve μ.



Complementary Problems

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

1. The code for this lecture downloaded income data for multiple countries and years (as in Lecture 2), but I've only shown the pdf of the income distribution for the US in 2019 (as in Lecture 1). Pick another country and year, say Finland in 2005, and make similar plots of the pdf of income. Do this again for a bunch of other countries and years.
2. *Some calculus with the log-normal distribution*
   a. Show that the mode of the log-normal distribution is $e^{\mu-\sigma^2}$. There are a couple of ways to do this, but the must straightforward is to solve $\frac{df}{dx} = 0$ for x.
   b. Show that $\mathbb{E}[X] = e^{\mu+\sigma^2/2}$. Again, there are a couple of ways to do this, including brute-force integration.
3. Show that the Gini index of a log-normal distribution is $2\Phi(\sigma/\sqrt{2}) - 1$. One way to do this is to find the area under the Lorenz curve.