Lecture 1: Introduction the the Course, Lightning Review of Probability and Statistics Idea

36-313, Statistics of Inequality and Discrimination

31 August 2021

Contents

Read the syllabus

Seriously, go read [http://www.stat.cmu.edu/~cshalizi/ineq/21/]. It covers the goals of the course, course mechanics (including grading and the policy on acceptable collaboration and sources), and includes the detailed schedule of what topics will be covered in lectures in when. Bookmark that page; it's where homework will be posted, and where you'll find links to the readings.

Probability

Populations

This is a course about inequality, which means its about differences within groups of people, and about differences between groups of people. (Sometimes we'll be interested in differences between families or neighborhoods or the like, rather than people, but basically the same idea will apply.) We need to start with the idea of a **population**, and the **distribution** of some trait within it.

By a "population" in statistics which just mean "a collection of things we are considering together". Originally the word referred to all the people in a given area, say a city or district or country, but we've come to realize we can fruitfully consider many other kinds of population. As I said, in this course the **units** of our populations are made from will usually be individual people, but sometimes they might be, for instance, households. In other statistics courses you might consider populations of animals or plants, or earthquakes, credit card transactions, or occasions when someone using a web browser might have clicked on a link.

Every member of the population has one or more **traits**, **attributes**, or **variables** or **features** we're interested in. (These words are all more or less synonyms in data analysis, though sometimes people draw subtle distinctions between them.) You can start by thinking of things like height, weight, whether they're left or right handed, the shape of the fingerprint on their left thumb... Generally speaking, every member of the population will have some value for each variable, even if we have to make up a special "not applicable" value in some cases[1].

You could imagine recording these attributes by just creating a giant list. To be concrete, you could imagine a gigantic list of everyone alive in the US on August 31, 2021, giving their height, weight, handedness, and a picture of their left thumb print. If it's a fancy, searchable, computerized giant list, we call it a "data base". There are times when this is exactly what you want, but there are two big reasons we don't usually stop there: the giant list gives us no insight, and we don't really care about most of its details.

1. *Insight*: It's hard-to-impossible for human beings to understand a giant list, a.k.a. database[2]. We need ways of summarizing the data which we can grasp and reason with. Usually this will call for getting rid of some amount of detail, but that's OK, because—
2. *We don't really care about the details*: Somewhere in that giant list is, perhaps a record of the fact that Chuckie Johnson of Mound City, Illinois stands 6 feet 4 inches[3]. This fact matters to Chuckie and perhaps to some people around him (e.g., the Pulaski County amateur basketball league). But there are few situations where it would matter to us as statisticians or social scientists or policy analysts. We've already abstracted away almost everything about Chuckie and his life[4] (). The next step is to realize that we usually don't care about Chuckie *at all*.

What we care about, as statisticians, is usually the **distribution** of trait values across the population: how many members of the population have any given value of the trait, or any given combination of traits. In fact, we usually don't even care about the exact number of people, just the *fraction* of the population which has those traits.

For discrete traits, like handedness, we can represent the distribution by just saying how many people there are with each possible value (right-handed, left-handed, ambidextrous, NA). We typically write this as something like $\mathbb{P}(H = "left") = 0.1$, to indicate that 1/10 of the population are left-handed.

For continuous traits, there is the difficulty that usually there's only *one* person with that *exact* height. We say that Chuckie Johnson is 6 feet 4 inches and that I, Cosma Shalizi of Shadyside, PA am also 6 feet 4 inches, but that's just because of rounding. If you were to set us back to back and measure precisely, we'd have different heights. So what we usually do, as statisticians, is ask what fraction of the population is between any two given values, say $\mathbb{P}(6.3 < X \leq 6.4)$. You can convince yourself that to work these proportions out, it's enough to know $\mathbb{P}(X \leq a)$ for any given a, since

$$\mathbb{P}(a < X \leq b) = \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a) \tag{1}$$

For this reason, $\mathbb{P}(X \leq a)$ is called the **cumulative distribution function** for the variable X. We abbreviate it as $F(a)$, or sometimes as $F_X(a)$ if we need a reminder of what variable we're dealing with.

$F(a)$ is an increasing[5] function. As $a \to \infty$, $F(a) \to 1$. (If $F(a) = 1$ for some finite a, then the *smallest*

[1] Some people are ambidextrous, some people don't have left thumbs, etc.

[2] Computers also don't *understand* the contents of their databses (yet), but they don't find that upsetting (yet).

[3] Mr. Johnson is a fictional character, but Mound City is a real place.

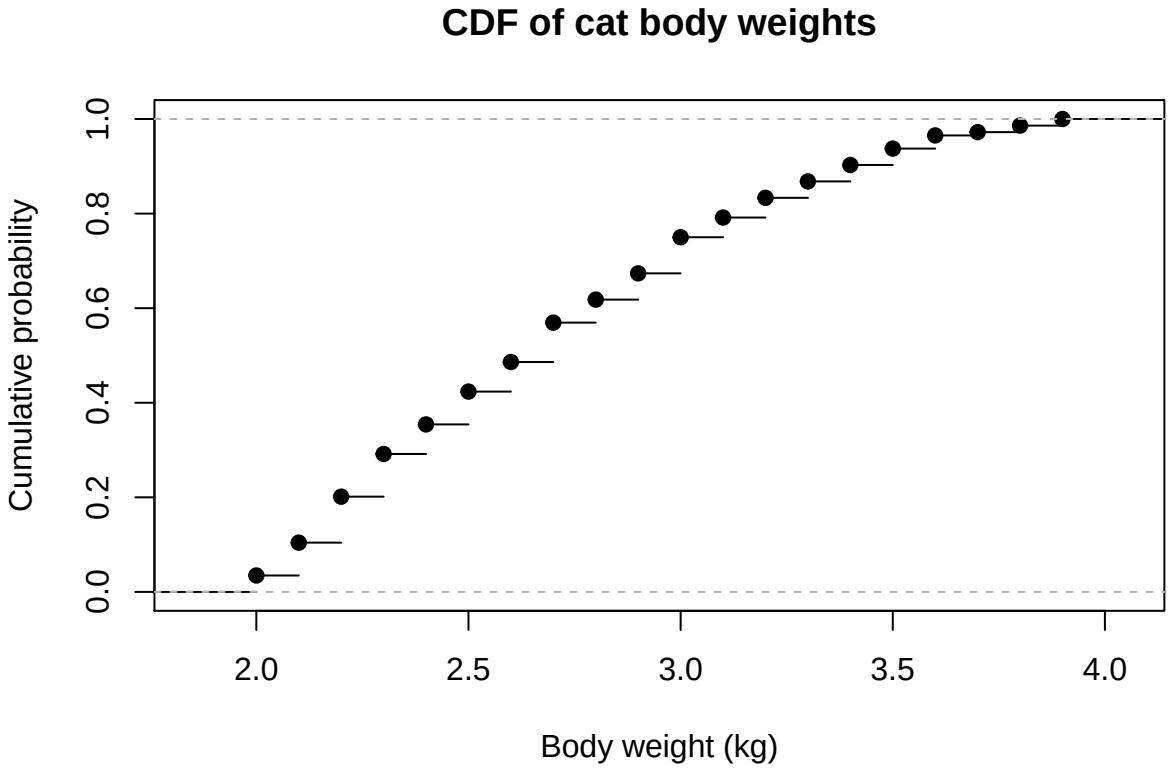
[4] his fondness for purple sneakers, his daily worry about whether his car will break down, the way he likes to put both ranch dressing and hot sauce on chicken wings, how he gets along with his boss, his memories of the pearl buttons his grandmother used to sew on to shirts, how he gets along with his exes, the way he casts his reel when fishing from the levee, what he knows about what really happened in Mermit Swamp on July 27, 1993, etc., etc.

[5] Strictly speaking, just a non-decreasing function. But it's usually strictly increasing.

solution to $F(a) = 1$ is the maximum value of the variable.) Similarly, as $a \to -\infty$, $F(a) \to 0$. Since $F(a)$ is increasing, the equation $F(a) = p$ has a unique solution[6] for each p. This is called the **quantile** corresponding to the probability p, often written $Q(p)$. If p is a multiple of 1/100, we talk about **percentiles**.

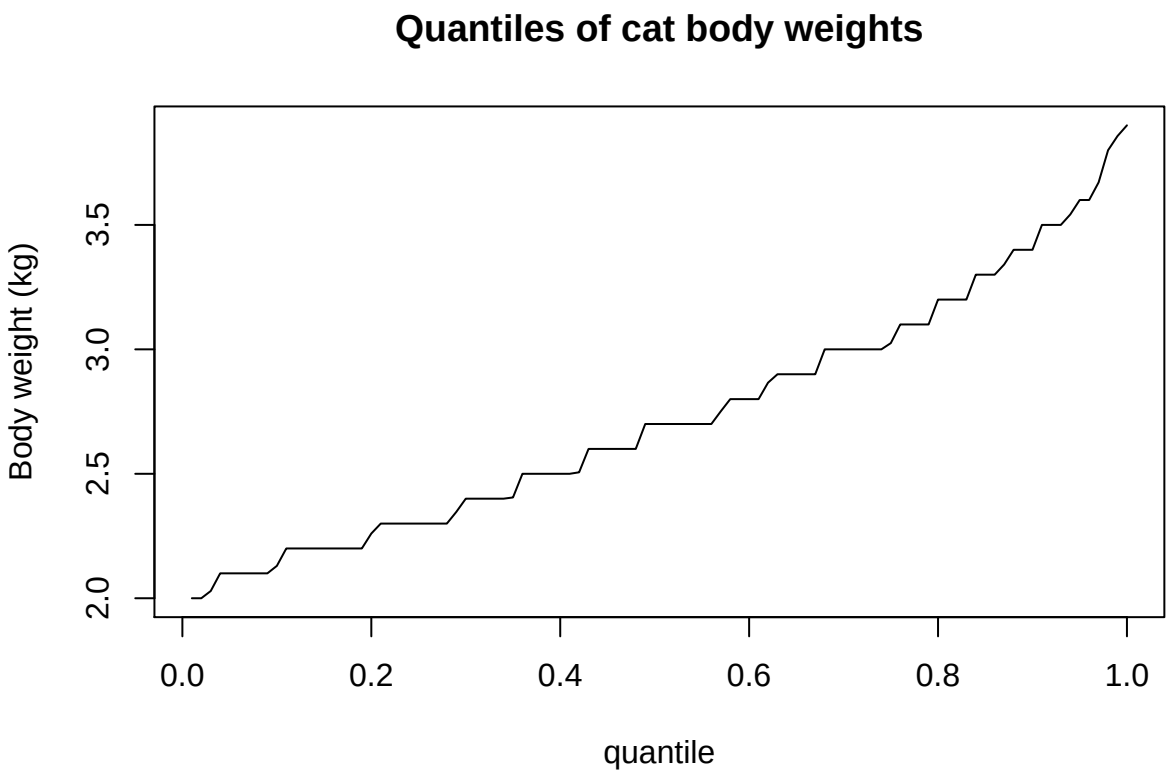
Example: Not-that-fat cats

Let's look at an example of the type of distribution you're already familiar with from other classes. For reasons lost to the mists of time, R comes with a data set about the body weight of a group of cats from the 1940s. Here's the cumulative distribution:



We can also find quantiles from this data:

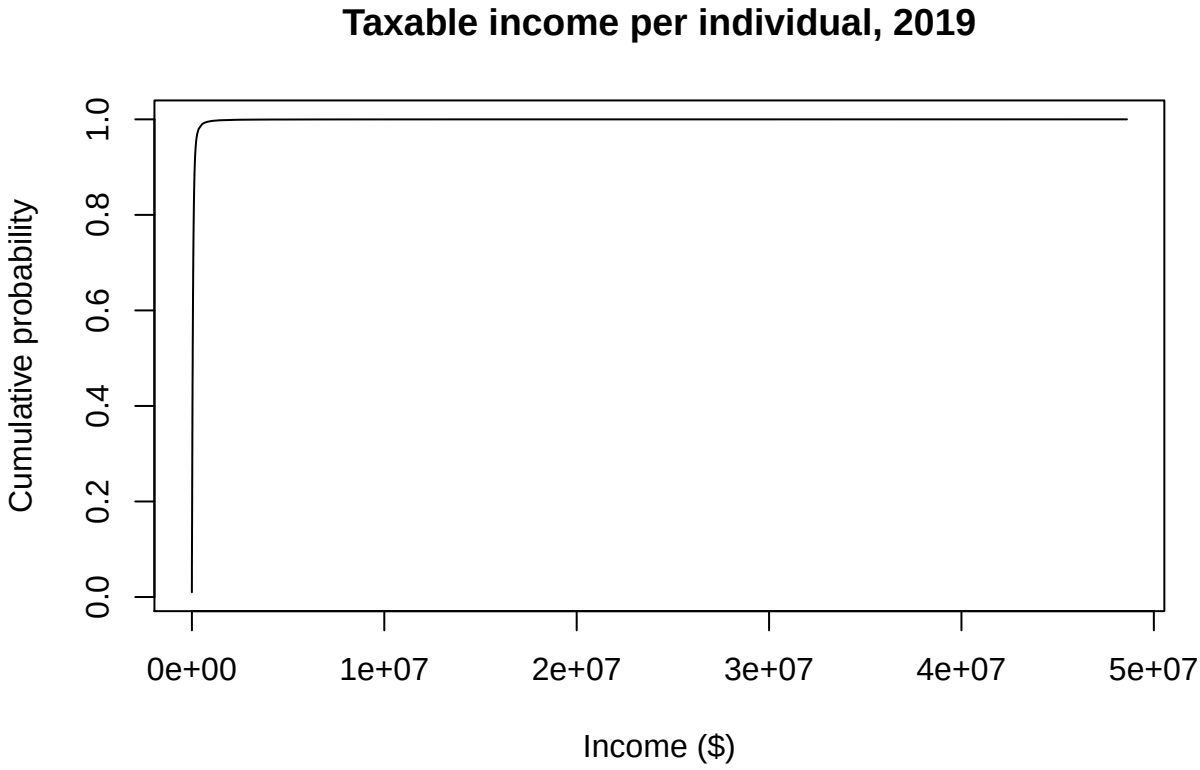
[6] If $F(a)$ is merely non-decreasing and not strictly increasing in a, then $F(a) = p$ might not have a unique solution for a given p. But it always has a unique *smallest* solution, which is usually what we pick.



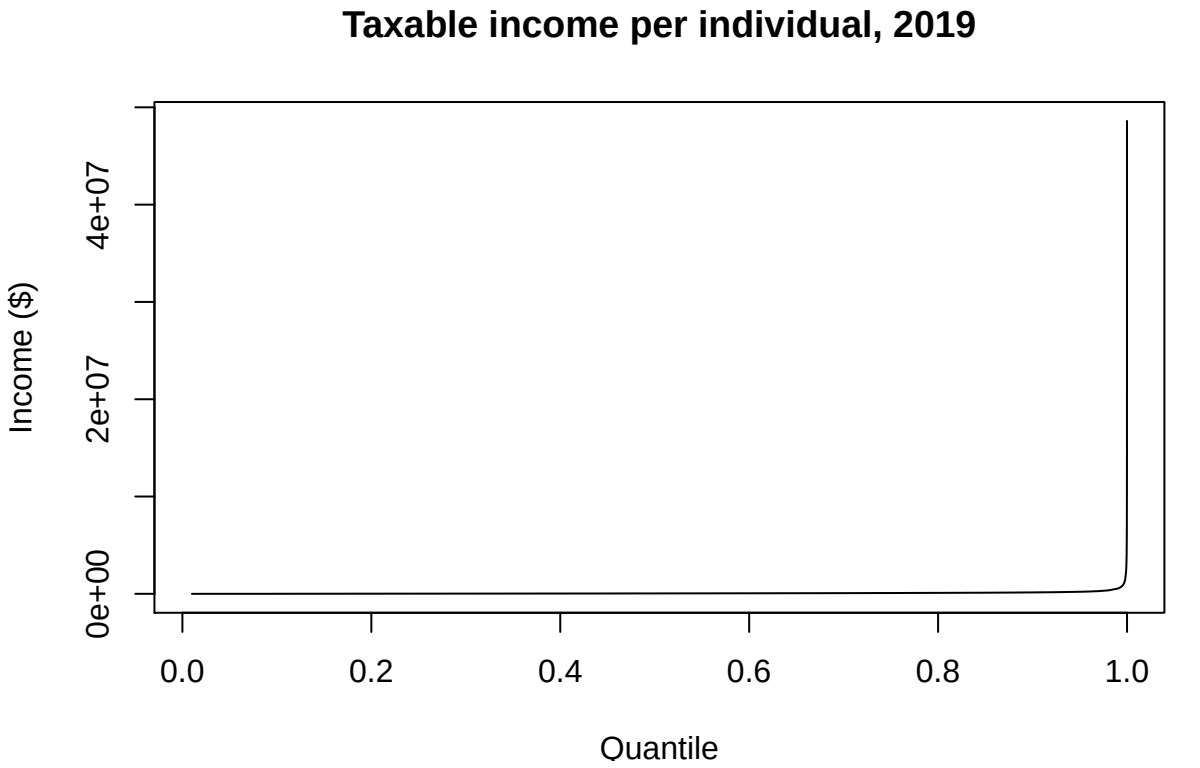
This distribution has a lot of features you're probably used to from other classes: there's a spread of values for body weight, but it's fairly small, there aren't too many very small or very large cats, even the biggest cat isn't that much bigger than a typical cat, etc.

Example: Income distribution

Now let's look instead at data of the kind we'll be dealing with in the rest of the course, which (for better or worse) is not about cats. Here is a plot of the cumulative distribution function for the taxable income of each individual person in the US in 2019, courtesy of the World Inequality Database) (which in turn gets its data from various countries' tax authorities, in this case the IRS).



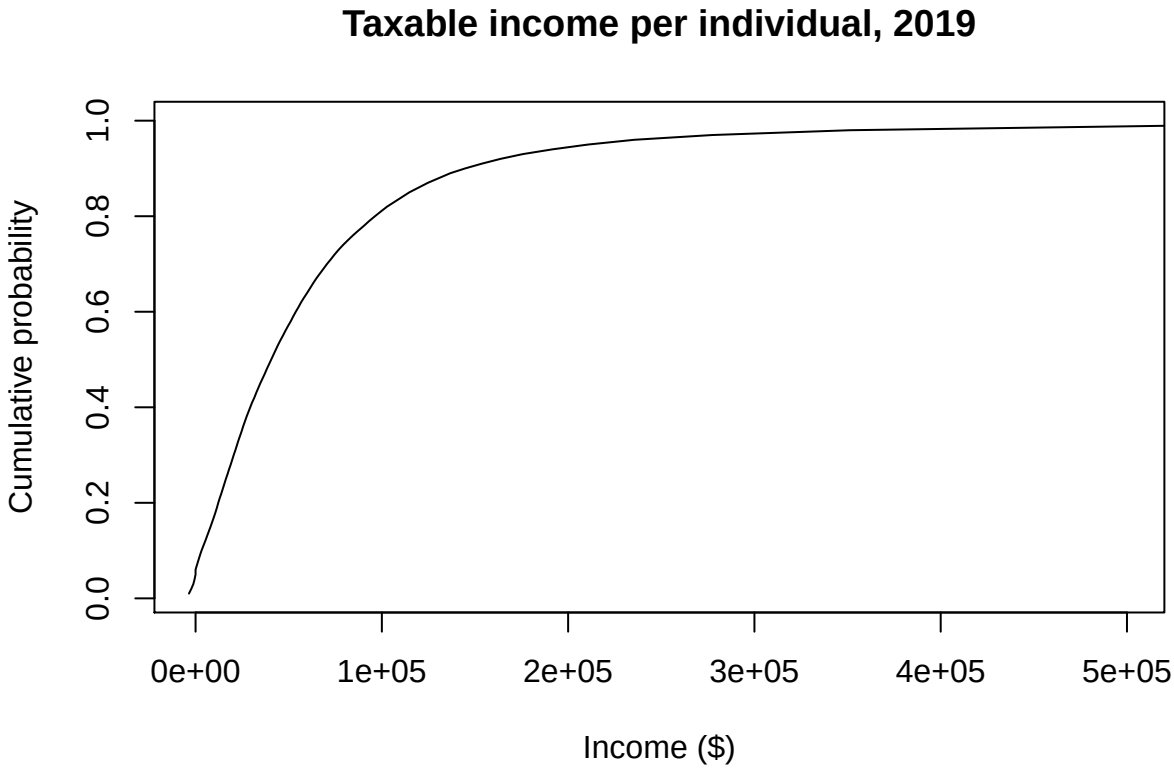
The quantile plot is basically the same, just turned on its side.

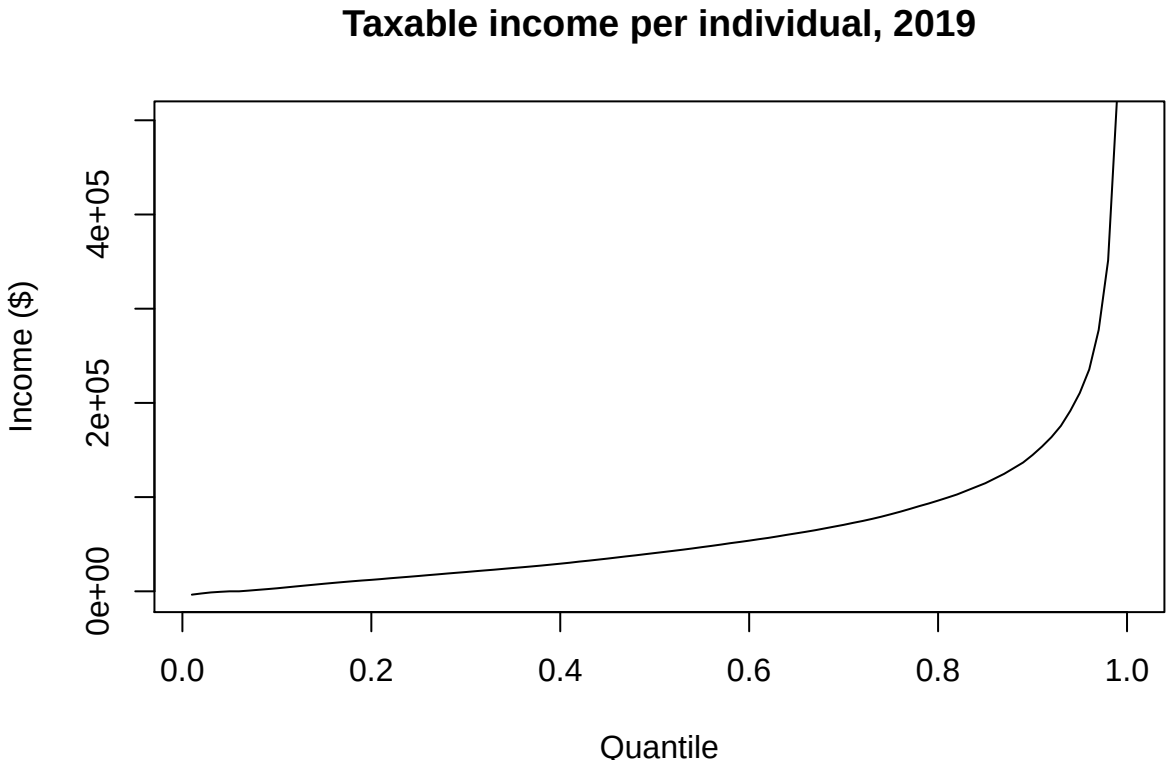


Let's zoom in, by limiting ourselves to those who make no more \$500,000 a year[7] which is a nice round

[7]These plots might suggest to you that a few percent of individuals had negative incomes in 2019. This is in fact correct. You might well wonder how anyone could possibly have an income of less than zero. This comes primarily from people who

number that's just a bit below the $\$\ 5.34 \times 10^5$ cut-off to be in the top 1%.





operated businesses that made net losses (rather than profits) during the year. These businesses typically took in *some* money during the year, they just had even higher expenses. There are also some other situations where US tax law can end up giving people negative incomes for tax purposes — and remember these figures are derived from tax records. In homework 1, we'll work with a different data set on *gross* income, which is necessarily ≥ 0.

The fact that the curve of $F(a)$ for income is so smooth suggests that it should have a derivative. This is, strictly speaking, false. There are only a finite number of individual tax-payers in the US (about 247 million adults in 2019), so even if we used the complete database, rather than the WID's summary of it, the curve of $F(a)$ would really just make 247 million (or so) small steps. The derivative of a step function is 0 in between steps, and infinite[8] at the steps. But it's often convenient to **idealize** the population as infinite, with some fraction lying between any two values (no matter how close), so we can talk about the **probability density** at a point, as the derivative of the CDF. In symbols,

$$f(x) = \left.\frac{dF(a)}{da}\right|_{a=x} \tag{2}$$

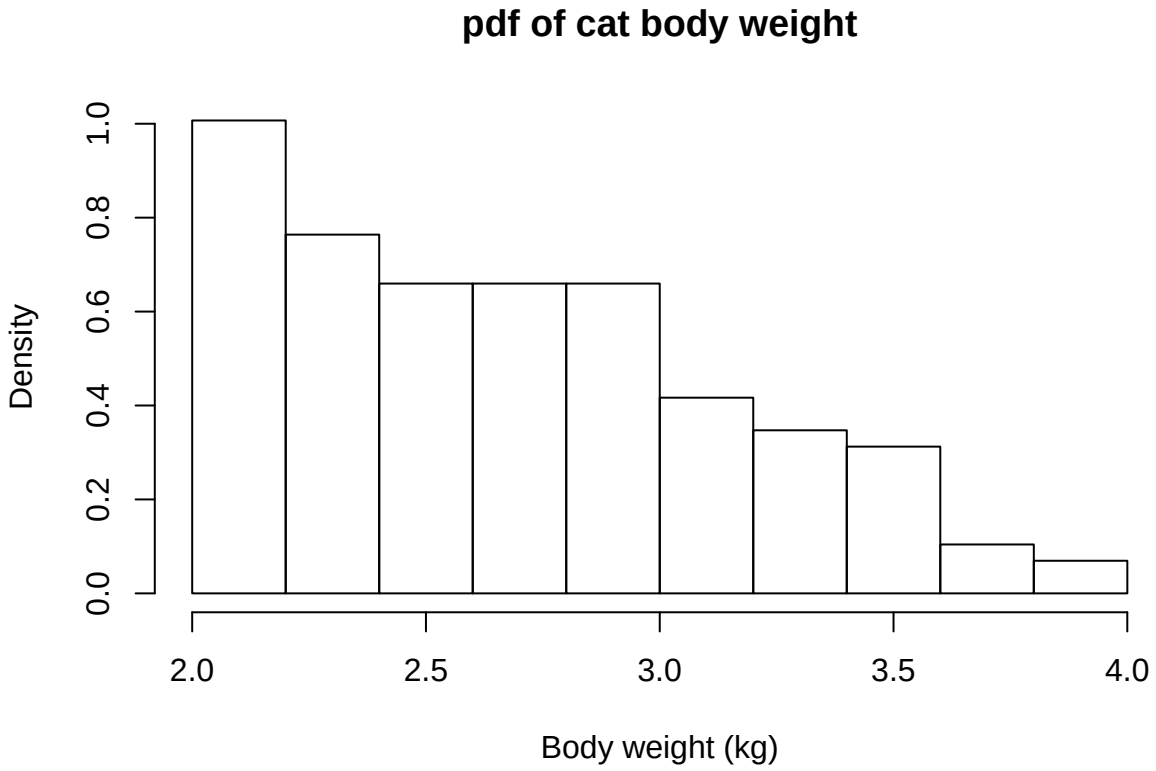
$$= \lim_{h \to 0} \frac{F(x+h) - F(x)}{h} \tag{3}$$

Going the other way,

$$F(a) = \int_{-\infty}^{a} f(x)dx \tag{4}$$

$$\mathbb{P}(a < X \leq b) = \int_{a}^{b} f(x)dx \tag{5}$$

If we have the underlying data, a simple way to estimate or approximate the pdf is to create a histogram: divide the range of X up into equal-length bins, count what proportion of individuals fall into each range, and divide proportion by length. Here's how it looks for the cats:

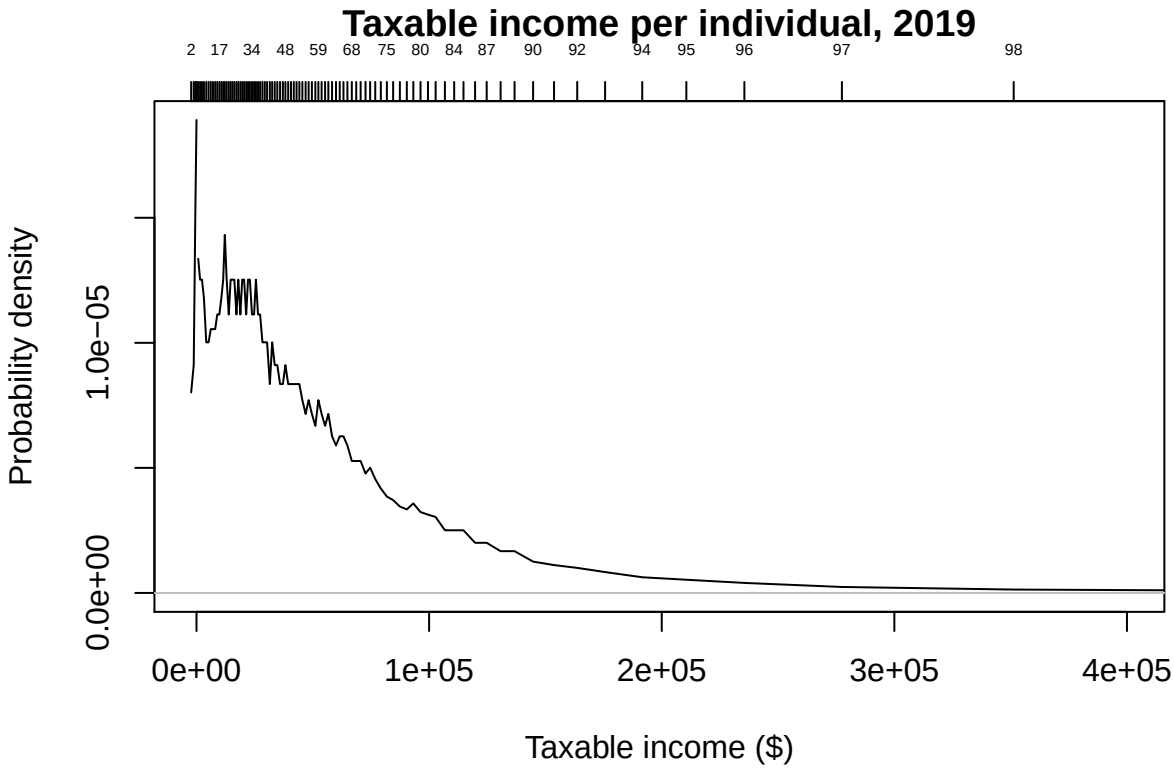


Since the figures I downloaded from the WID give $Q(0.01), Q(0.02), \ldots Q(0.99), Q(0.991), \ldots Q(0.999), \ldots Q(0.9999)$, I can again calculate (approximate) values for the derivatives of $F(a)$, and plot them[9]. Again, I'll limit

[8] If you're a mathematical purist, undefined.

[9] The calculation I'm doing here is almost the opposite of a histogram — I know each of my intervals contains 1% of the population, and I'm calculating how far apart the two ends of the interval are. If there's a standard name for this, I don't know it, but it's a crude version of a way people sometimes estimate densities in high-dimensional spaces (Gershenfeld 1999, 170).

the plot to those earning $\leq$ \$500,000. (The small tick marks on the upper boundary of the plot show the percentiles, so "98" indicates where the 98th percentile begins, etc.)



Central tendencies

The **median** income is the 50th percentile, so exactly as many units are above the median as below it. Here, the median is \$ 41 thousand.

Now, the **mean** with a finite population is just the ordinary, arithmetic average: add up everyone's incomes and divide by the number of people in the population. With our imaginary infinite population, represented by the pdf, we use an integral, which is the limit of averaging. We write means with respect to this infinite theoretical population with a special symbol, $\mathbb{E}()$, and talk about **expected values** or **expectation values**.

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) dx \tag{6}$$

The same rule applies if we're thinking about some function or transformation of the variable, say $h(X)$:

$$\mathbb{E}(h(X)) = \int_{-\infty}^{\infty} h(x) f(x) dx \tag{7}$$

In this case, the mean income is \$ 75 thousand, a bit below twice the median.

Skew and heavy tails

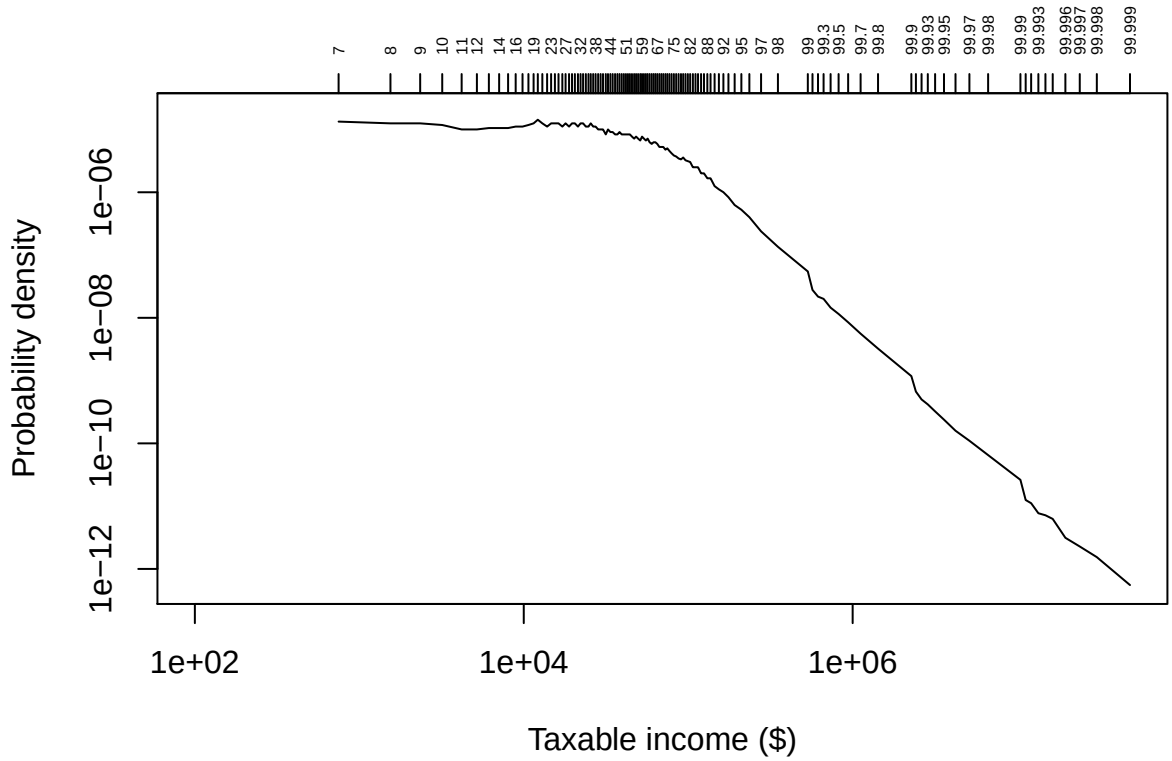
Because the mean is bigger than the median, we say that the distribution is **right-skewed**. (If it was the other way around we'd say it was **left-skewed**.) Distributions of variables like income, wealth, etc., are typically quite right-skewed. In part this is just because there are lower limits to these variables, but no real upper limit, and there are *some* people out there in the right tail. But really most of the skew comes from

the fact that there are usually substantial numbers of people with literally orders of magnitude more money than the median person. Remember the median income was $ 41 thousand in 2019, but that more than 1% of the population had incomes over $500,000. In fact, 0.1% of the population, or about 247 thousand people, had incomes over $ 2.28×10^6. To un-skew the distribution and make it symmetric around the median, we'd need a balancing number of people whose incomes were equally negative, which just doesn't happen, not even with the most creative accounting.

Variables like income (and, as we'll see in latter lectures, wealth) thus have *very* different distributions than what you're probably used to from examples in other statistics classes, or from thinking of biological variables like height and weight. It's true that there's a minimum weight for a cat, while the upper limit on weight is a lot more open-ended, so the distribution of cat weights is a little bit right-skewed. (The median cat in the dataset weighs 2.7 kg, while the mean cat weight is 2.72 kg.) The same is true for, say, height in human beings. But there is *no* cat which is 56 times heavier than the median cat, just as there is no human being who is 56 times taller than the median person, let alone a quarter of a million such people in the US. But there *are* that many people who make (at least) that multiple of the median income. (For their part, they look with similar astonishment on the 2500 people in the top 0.001%, which begins at $ 49 million[10].)

Distributions like this, where going out to high quantiles keeps taking us to orders-of-magnitude larger values, are said to be **heavy tailed**, or to have **fat tails**. There are some situations where both the right (upper) and left (lower) tails are heavy[11], but in social phenomena there's usually only one heavy tail, and it's usually the one on the right. A somewhat more precise definition would be that the pdf $f(x)$ goes to zero as $x \to \infty$, but does so more slowly than any exponential function[12].

When plotting heavy-tailed distributions, it can be helpful to use logarithmic scales on both the vertical and horizontal axes. Here for instance is the pdf plot again (with negative values of income omitted):



The long straight-ish segment on the right is one signature of a heavy tail.

[10] Cf. the now-classic movie *Wall Street*, where a character is taunted for being an under-achiever who's content being "a $400,000 a year working Wall Street stiff flying first class and being comfortable".

[11] Many of these occur in the physical sciences, e.g., the branches of physics which study turbulent motion in fluids. Schroeder (1991) provides a wonderfully-readable introduction.

[12] If you crave even more mathematical detail, Adler, Feldman, and Taqqu (1998) and Resnick (2006) are good places to start.

Modeling

We've already gone from the concrete population, which consists of certain particular individuals, each unique and potentially infinitely complex, to a more abstract idea, a distribution of traits across a population. This abstraction makes it easier to grasp the patterns in those traits and *that* aspect of the population, at the cost of ignoring almost everything about the individuals.

It's often useful to abstract even further, and to try to approximate the actual distribution, which still has a lot of fine-grained detail (all the little bumps and jags in those last few plots), by a simpler model, which fills in all those details by applying some mathematical rules. Typically those models have some adjustable settings which go into the rules, the **parameters** of the model. There are (at least) three reasons for using probability models like this.

1. *Economy of thought*: Thinking is painful and expensive and we want to do as little of it as possible. For human beings, remembering details is an especially expensive form of thinking. If instead of having to remember hundreds or thousands of little curve segments, you could remember just two or three numbers, you should be tempted.
2. *We don't care about the details*: There are very few conclusions we'd draw from the last figure that would be different if the little jags and wiggles were slightly differently angles and spaced, or slightly smoother or slightly rougher. Since those details don't matter to us, why *not* replace them with something simpler?
   Said slightly differently, we have every reason to think that some of the details of that distribution curve are just accidents. How much money people made depended on things like the weather (think of farmers, rain-coat sellers, etc.), the random ups and downs of the stock market, literally gambling and lotteries, and innumerable other tiny factors which could just as easily have turned out differently. We don't, for these purposes[13], *care* about those accidents, but rather about the more stable or recurring patterns that would emerge no matter how the weather or the cards turned out. Using models is a way to (try to) separate the stable pattern (signal) from the accidents (the noise)[14].
3. *We care about the parameters*: In many situations, the parameters of probability models are meaningful, because they're also part of a scientific model, a story about how some part of the world works. Much of statistics was originally invented in the 1700s and 1800s to estimate physical parameters, like "How strong is gravity?" or "What is the mass of the Earth?" or "How much does the Earth deviate from being perfectly spherical?" or "Where is that asteroid and which way is it headed?" (and consequently "Will it hit Earth?") (Farebrother 1999). In modern applications physical or biological[15] parameters might answer questions like "How old is this species?" or "How quickly does this radioactive waste decay?" We will see examples later in the course of *social* parameters which answer meaningful questions like "How concentrated is wealth in this population?", or "How much more likely is a job application to be rejected if it includes a criminal record, all else being equal?" or "How much less likely is someone to die within a year of drug overdose if they finished college, all else being equal?"

With all that throat-clearing and motivation, let me now briefly present our first probability model. (We'll have a lot more to say about this model over the next few lectures.) You remember the Gaussian or "normal"[16]

[13] Remember our friend Chuckie Johnson from above. Obviously, *Chuckie* would care about what I'm calling the accidents which affect him and those around him. If we were telling the story of his life, those would be *incidents*, not accidents. But, again, doing statistics, or social science, involves zooming out from that human scale, not because it doesn't matter but because it's the only way to see the larger patterns in which Chuckie's is enmeshed. The double vision is not always easy to maintain.

[14] If you worry that this leads to very difficult questions about how we dcide what's pattern/signal and what's accident/noise, you're right. A good place to start attacking those difficulties is Ruelle (1991).

[15] The original point of gathering the data on the cats was to work out the relationship between total body weight and the weight of the cats' hearts. If, like me, you are a sentimental cat person, it is not pleasant to think about how they got the data about heart weights, but when I had a cat who needed a heart medicine, I was glad that there was a rule which let the vet gauge how much of the drug my pet needed based on her body weight.

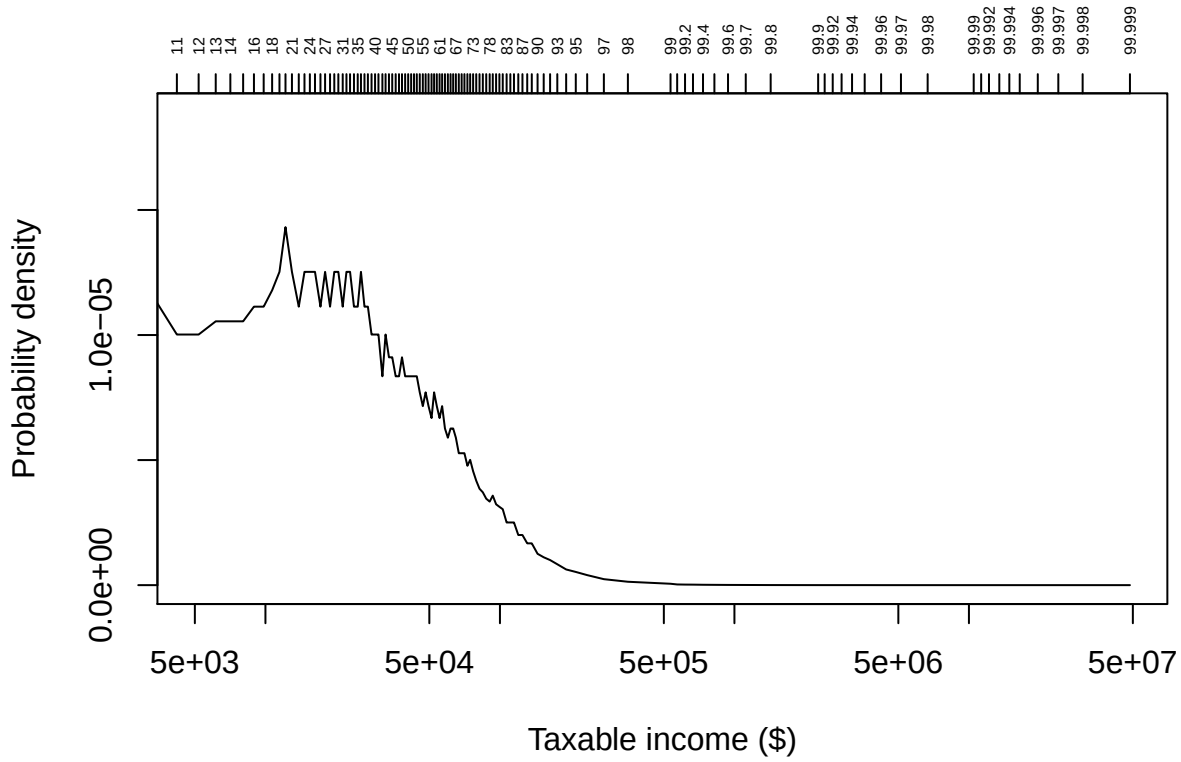
[16] I dislike using the word "normal" here, both because "normal" is used for so many different, unrelated things in mathematics, and because these distributions are not actually all that common that you *should* take them as the default, so I will almost always write "Gaussian" instead. But if I tried to write $\mathcal{G}(\mu, \sigma^2)$ I'd just be making you confused about the notation you'd encounter in every other statistics reference.

family of distributions from your earlier classes; it's conveniently defined by its pdf,

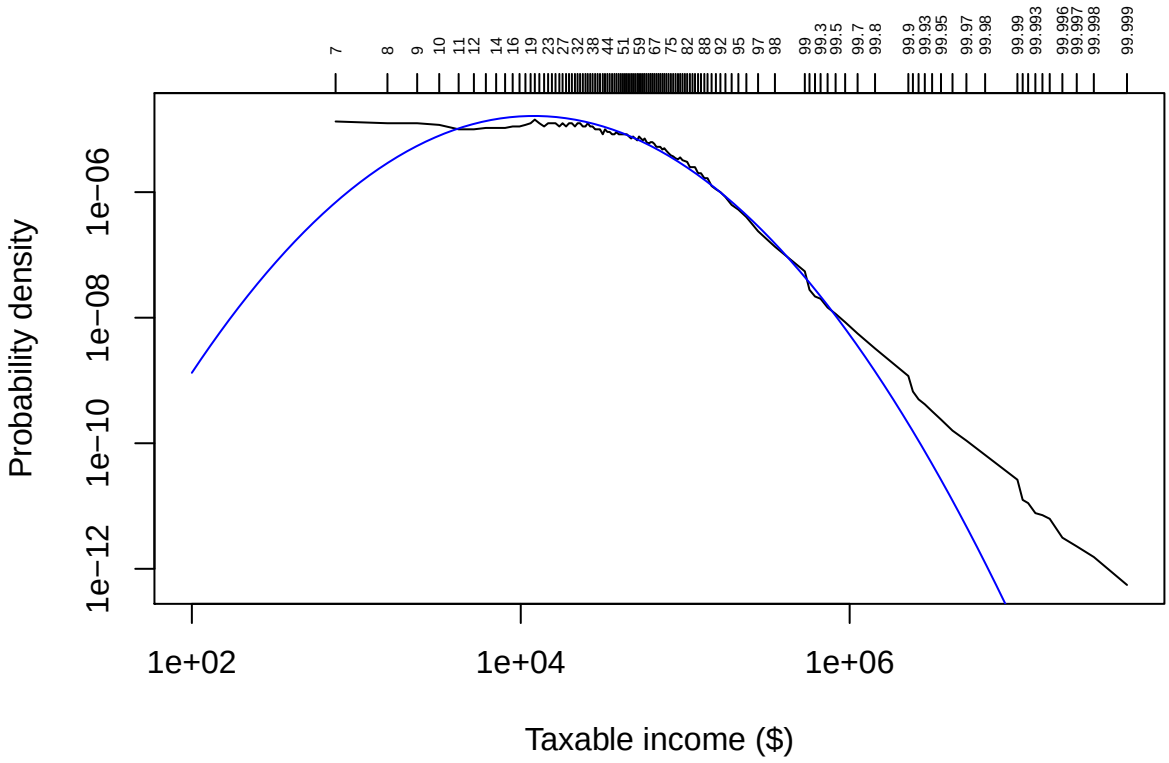
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{8}$$

We write this particular distribution $\mathcal{N}(\mu, \sigma^2)$. Here μ is the mean of the distribution, and σ^2 is its variance. The Gaussian distribution is a *horrible* fit to the data on income, because Gaussians are always symmetric about their means, which are equal to their medians, and this data plainly looks nothing like that.

What is remarkable, however, is that a simple modification of the Gaussian distribution is actually not bad here. Let's define $Y = \log X$ as the *logarithm* of people's income. (We ignore people who show up has having incomes ≤ 0, and we'll use natural logs, though that won't matter much.) If we go back the very last plot, and re-do it using only a log scale on the horizontal axis, we're getting an approximation to the pdf of Y, and that looks a lot more like a bell curve (particularly if we leave off some very small incomes):



Let's suppose that $Y\mathcal{N}(\mu, \sigma^2)$. The corresponding distribution for X is called the **log-normal** distribution, written $\mathcal{LN}(\mu, \sigma^2)$ or $LN(\mu, \sigma^2)$. Be aware that now $\mu = \mathbb{E}(Y) = \mathbb{E}(\log X)$, not $\mathbb{E}(X)$, and similarly σ^2 is the variance of $\log X$, not of X. This is actually a pretty good fit to a very large part of the data:



Here the blue curve is the pdf of a log-normal distribution with parameters chosen to match the data (in a way you will learn to do in homework 1). You can see that the log-normal provides a pretty good match to the pdf we approximated from the data, from about the 10th percentile of income to about the 99.5th; outside that range it underestimates the number of very poor people[17], and the number of very rich people.

This last figure helps illustrate some of the points I made about models above. The blue curve is simpler than the black one, and definitely easier to remember; most calculations we'd care to make using the two curves would come out very similarly, especially if they're just about that 90 percent of the population; and the fact that the model fits suggests ways we might go about starting to explain, rather than just describe, the income distribution[18].

On the other hand, the model is also definitely imperfect — it makes systematic mistakes for some parts of the income distribution. In particular, if we're interested in some question where the very rich matter, then the log-normal model is going to be misleading, because it only works up to about the 99.5th percentile; it thinks the top 0.1% have much less money than they really do, to say nothing of the top 0.01%. One way to deal with this is to try to come up with different, and usually more complicated, models that capture more of the distribution. Another is to be clear that we'll use different models as approximations in different circumstances, and to be clear about the "scope" of each model, about when it's a good approximation. Thus in a few lectures, we'll see a different model, the **power-law** or **Pareto** distribution, which is a good fit to the right tail of very high incomes, but works poorly for the "body" containing most of the distribution.

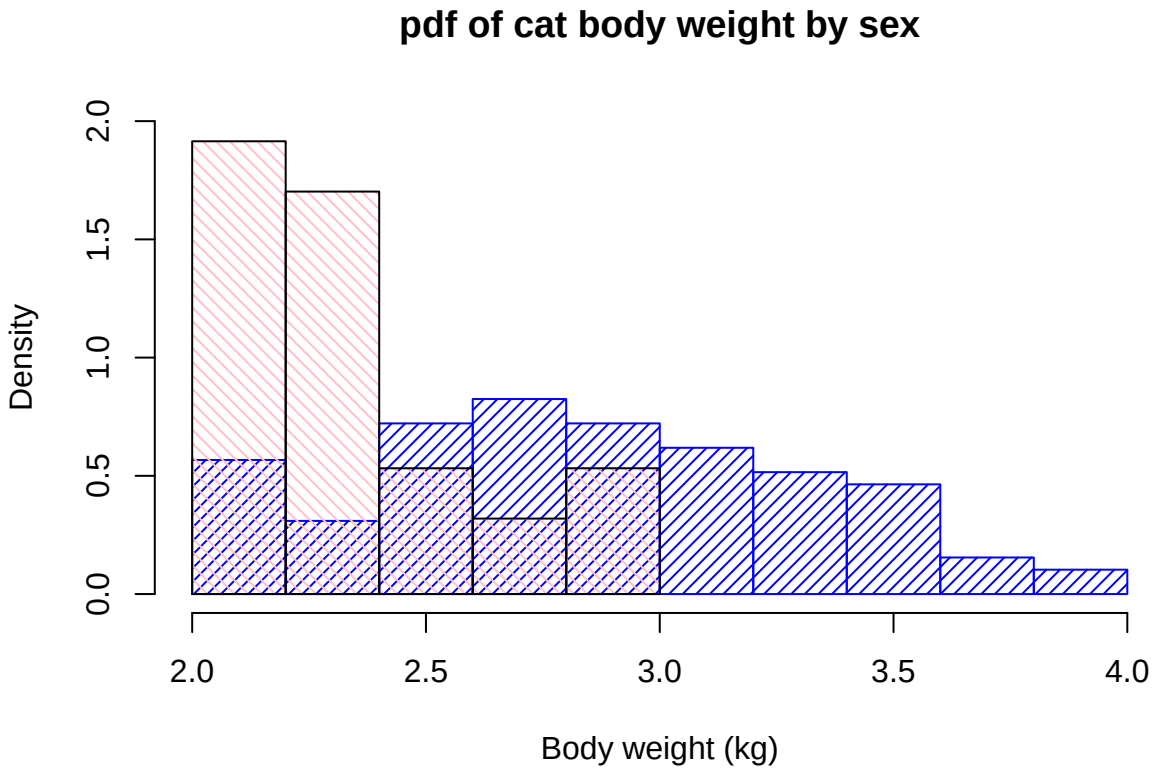
[17] More cautiously, the number of people who, under the tax rules, get to report extremely small incomes. There *are* people in the US who survive, somehow, on $2 a day or less (Edin and Shaefer 2015) (or about $700 a year), and even people who survive without any money at all. But it's not clear how many of the people showing up in this data at these extremely low incomes are in those situations, and how many are people with large tax write-offs. Again, we'll look later in the course at other data which measures gross income more directly, though it misses a lot of what's going on with very high incomes.

[18] You remember that the Gaussian distribution is what happens when we add up lots of small, random terms; this is the "central limit theorem". Apply this origin myth to $Y = \log X$, so $Y = \epsilon_1 + \epsilon_2 + \ldots + \epsilon_m$, for some large number m of small, random terms ϵ_i. It follows that $X = e^Y = e^{\epsilon_1+\epsilon_2+\ldots+\epsilon_m} = e^{\epsilon_1} e^{\epsilon_2} \ldots e^{\epsilon_m}$. This would mean that income X is the outcome of *multiplying* a lot of little independent random "shocks". Whether that's a good scientific model of income determination is something social scientists can investigate.

Different Populations; Sub-Populations

When we have multiple traits or attributes, we can use them to split up a population into sub-populations, defined by one or more of the traits. Alternately, if we have multiple populations, we can imagine combining them into one bigger population. However we get sub-populations, we can compare them.

With the cats, for instance, we have their sex as well as their body weight. What I showed you before was the over-all distributed across the whole population, but we can look at (for instance) the histogram across the two sub-populations, defined by sex.



Here I've plotted the female pdf in pink and the male in blue. There are two things which are notable about the comparison between these distributions, just from this plot:

1. *The two distributions are different*: It seems pretty plain that the two distributions are different. They have different ranges, different means, different medians. Female cats are *typically* smaller than male cats. (This is true of most species of mammals.) We are going to develop a lot of techniques for verifying the plain impression of our eyes that two distributions are different, and for describing the differences.
2. *The two distributions aren't that different*: There are plenty of male cats who weigh less than some female cats. If, in ordinary language, we say "male cats are heavier than female cats", that sort of all-or-none statement is at best a rough summary of the actual state of affairs.

(Incidentally, both the male and the female body-weight distributions are well-fit by "gamma" distributions, where $f(x) \propto x^{\alpha} e^{-\beta x}$, but with different parameters for the two sexes.)

(Incidentally, the WID data doesn't allow us to split up taxable personal income in the US by the sex of the tax-payer, though WID does have that information for some countries. We will look later, using other data sources, at differences in income distributions across different social categories, such as sex, race, ethnicity and education, and look at how we might, as statisticians, begin to explain those differences.)

Samples and inference

The income data set we've been working with here is somewhat unusual, because it includes every member of the relevant population, namely US taxpayers in 2019. This sort of **complete enumeration** of a population, also called a **census** of the population, is a very valuable sort of data (nothing's missing!), but it's also rare. Getting everyone in a country to tell the truth, or even something like the truth, about their income is a big undertaking, which can really only be done by powerful, well-organized, honest and intrusive governments[19].

Much more typically in data analysis we deal with a much-smaller **sample** of the population of interest. This is the situation with the cats: these `r nrow(cats)` felines were not the whole of the relevant population, but just a small part of it, selected in the hope that this part would tell us about the whole.

We want to make guesses about the actual population from the sample. Because "make guesses" sounds crude, we call this "drawing inferences", or "drawing statistical inferences" if we want to emphasize that we're using partial, noisy, incomplete data and so might be wrong, as opposed to drawing *deductive* inferences, which have to be true so long as we're right about the data and don't botch any calculations[^deducclarify]

For instance, with the cat data, concluding that the *sample mean* is 2.7236111 kg is a *deductive* inference. It's a necessary consequence of the data and can't really be wrong. (I guess that I could have made exactly the same arithmetic mistake as R did, but that's about it as far as possibilities for error go.) But this is just a fact about this sample of cats. It says nothing about any other cat or group of cats. If I go on to say that the mean adult cat in British laboratories in the 1940s weighed 2.72 kg, that is *not* a deduction from the data, but a statistical inference. Whether it's a *reliable*[20] inference depends on whether this was a **representative sample** of adult British lab cats in the 1940s. Even if this sample was representative of that population, whether I can reliable extrapolate that conclusion to all adult British cats in the 1940s, all adult cats in the 1940s, or all adult domestic cats ever, would in turn depend on how representative those various sub-populations were of the broader groups. (It'd obviously be foolish to extrapolate to all cats of all ages, because lots of cats are very small kittens [proof: Internet].)

In your previous statistics classes, you'll haver learned about different types of samples, such as convenience, purpose, random, etc., and learned that the easiest way to ensure that a sample *is* representative is for it to be a simple random sample. (If you need refreshers, I strongly recommend the well-written little book by Cox and Donnelly (2011), which you can download from the university library.) This ideal is often unattainable, and it's sometimes undesirable. For instance, in a simple random sample, where every member of the population has the same probability of being included in the sample, it's often hard to say much about the properties of small sub-groups[21]. Organizations which gather social data often deliberately **over-sample** small sub-groups of scientific or policy interest, which gives us more information about them, but also makes analyzing the survey more complicated.

Read the syllabus

In conclusion, read the syllabus and go do homework 0 (about course policies) and after-class exercise 1 (which is just a survey about why you're taking the course and your previous statistics classes).

[19] In fact, the discipline of statistics has (one of) its origins in analyzing information like tax records collected by governments, to help them figure out how much money they could expect to raise, how many men they could draft for the army, and so forth; the very word "statistics" comes from the word "state" (Hacking 1990).

[20] The qualifier "reliable" is important here. It's a free country and if you *want* to make wild and foolish extrapolations, no one is going to stop you, but this is a class in statistics rather than journalism or activism, so let's suppose you'd rather be right than grab attention.

[21] In a simple random sample, the margin of error for most quantities we'll calculate will be $\propto 1/\sqrt{n}$ where n is the sample size. Say we sample $10,000$ individuals. If we're interested in a sub-group which is only 1% of the population, we'd expect only ≈ 100 members of the sub-group in our sample. The margin of error for the sub-group will then be $\sqrt{10000}/\sqrt{100} = \sqrt{100} = 10$ times larger than the margin of error for the over-all population.

References

Adler, Robert J., Raisa E. Feldman, and Murad S. Taqqu, eds. 1998. *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*. Boston: Birkhäuser.

Cox, D. R., and Christl A. Donnelly. 2011. *Principles of Applied Statistics*. Cambridge, England: Cambridge University Press. https://doi.org/10.1017/CBO9781139005036.

Edin, Kathryn J., and H. Luke Shaefer. 2015. *$2.00 a Day: Living on Almost Nothing in America*. Boston: Houghton Mifflin Harcourt.

Farebrother, Richard William. 1999. *Fitting Linear Relationships: A History of the Calculus of Observations 1750–1900*. New York: Springer-Verlag. https://doi.org/10.1007/978-1-4612-0545-6.

Gershenfeld, Neil. 1999. *The Nature of Mathematical Modeling*. Cambridge, England: Cambridge University Press.

Hacking, Ian. 1990. *The Taming of Chance*. Cambridge, England: Cambridge University Press. https://doi.org/10.1017/CBO9780511819766.

Resnick, Sidney I. 2006. *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. New York: Springer-Verlag. https://doi.org/10.1007/978-0-387-45024-7.

Ruelle, David. 1991. *Chance and Chaos*. Princeton, New Jersey: Princeton University Press. https://doi.org/10.2307/j.ctv10crf7w.

Schroeder, Manfred. 1991. *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise*. San Francisco: W. H. Freeman.