[SQUEAKING]

[RUSTLING]

[CLICKING]

ESTHER DUFLO: Well, kind of slightly rewind since the end of yesterday was a little quick on the kernel. And assume we are finishing with histograms. And now I'm going to talk about kernel again. So given that yesterday was a little-- there was more discussion, we spent more time than I originally anticipated, I think. Most of today will be my material, looking at distributions, et cetera.

And then we'll go back to discussing conditional random variable and marginal distribution, et cetera next week. So this is what's going to happen. And then after that, with all that behind us, we'll be able to talk about expectation, variance, central limit theorem, and all that good stuff, which should place us in turn, in a very good position to start analyzing data in a way that we won't be confined to one or maybe two variables and very simple distributions.

So, that's sort of the game plan. All right, so we were there at playing with bins. And you can see that the number of bins that I put on the shape of the distribution might influence what I have to say about it. For example, with only very few bins over at the top, I might think that the distribution seems to have some skewing to the left maybe. But then when I move to a few more bins, that starts going away.

And then even more bins that start going away. And then with the most bins, it starts looking quite symmetrical and not skewed and kind of bell-shaping. So it seems that we would gain, in our understanding of the distribution of the underlying random variable, if we had a smoother, more continuous way of describing it. And that's what the kernel does.

So the kernel density is, in a sense, an extension of the histogram. Except that for the histogram, the histogram draws a bin at several points. We have to decide where the bins are. And then we estimate the value for that bin.

And the kernel is continuous. It's looking at a function. It's looking at drawing a function, which is going to be-- a histogram is a function too. But it's a discrete function, evaluated at different intervals. And the kernel is a continuous function.

So formally for a sample, we have a sample, which has a given number of observations, in this case n observation, x1, x2, xn. And we want to draw a function based on these x's. We might have very many sample points, many, many more than we could see if they are plotted slowly. Or we may have very few of sample points. Regardless, we want a continuous distribution.

So the kernel density estimator, what it does, it's a little bit like a count, like an histogram, except it's not going to assign the same weight to every observation we count. So in fact, it is not a count. It is a weighted average of frequencies.

So what it does, is that it looks-- if we want to estimate it at point x, I want to evaluate it at point x, it says, well, at point x, I'm going to apply a function, which is going to be a sort of filter. And I'm going to look at the value of this function for points that are at various distance from x within a particular bandwidth, which we are going to call the h. And I'm going to take the average of this function over the entire sample.

So in practice, how does it work? In practice, suppose that you have a sample with a few observations. How many do we have, what 1, 2, 3, 4, 5, like about a dozen, maybe. For each observation that I have, I'm going to draw the function k. So here it's looking like-- so what is the function k, I'm going to come in a minute.

It could be something flat that would be what we call a uniform kernel. Typically, we use something bell-shape, to give it a bit more nice continuity. And I'm going to draw, at each x point, I'm going to draw my kernel.

So I'm going to draw my kernel function. The width here is the bandwidth, h. So the function, the kernel function is 0 until you are at some distance from x, h/2, or h, depending on how h is defined.

And then, it's increasing as you're getting closer to x. And then, it's decreasing again, as you get further away. And it goes back to there.

There's no requirement that it can't be bell-shaped. It has to be symmetrical. It has to integrate to 1 and to have mean 0 if it's symmetric across. So it takes any negative value and positive value, it integrates to 1 and has mean 0. So that's the kernel.

So then for example, if I want to know the density for a point that is over here, I'm going to-- I have just one kernel here. So it's going to be that point. If I want to take the kernel at a larger function, I'm going to sum them up. I need the sum of the vertical distance until I have the point. And that's what it does.

So what it does in practice is some counting, except the count that it gives is-- it counts only the observation that are close by, like a histogram. And it changes the-- and it also gives less weight to the observation that-- it gives more and more weight as the observation are closer and closer from each other. Yeah?

STUDENT: Why, if you're just summing vertically, don't the right and left edges overlap?

ESTHER DUFLO: Yeah, I think they should. I think thEY should be like this. I think, as I was thinking like, there is an issue with that. I think it they should there. They should go like this, and then same thing over here as well. Yeah?

STUDENT: Is it only for continuous variables that we should--

ESTHER DUFLO: No, look here. We have very few. So we could-- oh, the function itself is not-- you're saying if the function itself, the underlying probability function is not continuous, then it would be odd to draw a kernel, I would think.

STUDENT: Do you mind explaining, or if it is at all useful or important to know the height or the altitude of these kernels? How do you actually determine? Does that matter?

ESTHER DUFLO: So the height is going to be determined by the bandwidth, basically, because they have to integrate to 1. So, they become flatter If the bandwidth is larger. Or they become higher if the bandwidth is smaller, so that they integrate to 1. Yeah?

STUDENT: So when we are determining the function of k and we're deciding between uniform, normal, or Epanech--

ESTHER DUFLO: Epanechnikov.

STUDENT: How are we deciding which one of those functional forms we want to use. Is it all with bandwidth? What rule of thumb are we using to determine?

ESTHER DUFLO: So honestly, the choice of kernel, I don't think matters a huge amount, as long as the kernel is some bell-shaped looking. A lot of people--

INSTRUCTOR: For computational reasons, people prefer the Epanechnikov over the normal kernel, because the normal kernel has positive probability everywhere.

ESTHER DUFLO: And it's a pain. So computationally, it's just more intensive than the Epanechnikov.

STUDENT: Is this weighting the closer--

ESTHER DUFLO: No, all of them, in any case, weigh the closer are more, but the normal is not bounded. It's trying to go all the way to the end. So either you need to do something like that, that is going to truncate it above your bandwidth, it still has some little mass here.

So it's going to be like that. It's still going to have some little mass here that's left. Whereas, the Epanechnikov is exactly 0 at the kernel. And then it's bell-shaped. It's also optimal in some sort of statistical sense that's too involved for me, to be fair.

So the kernel, you will typically go with Epanechnikov. The bandwidth is much more relevant choice in a sense, much more sensitive and relevant. Let me show you what happened with bandwidth. In fact, I'll skip something.

I'll come back. I'll show you a picture, where it matters a lot. I'm not telling you what the picture is for the moment. We'll get back to the nature of the picture.

But so here is an histogram. This is our friend Steph Curry. And this is a short distance from the midline of course, he's taking. So it doesn't matter why this is like that. But it turns out that's the shape of the histogram.

You see it has this slight peak at the end and a big peak in the bottom. And that's the first band, which you could use, which turns out to be the default bandwidth. I'll explain to you in a minute what the default is trying to do.

And here is what you would get if you tried to use a bandwidth that is much wider. So you would completely lose. If you had not plotted the data and you start with a big bandwidth, you would completely lose what this distribution is looking like. It might be looking to you kind of almost bell-shaped, when in fact, just plotting the histogram would have convinced you that it's not bell shaped whatsoever.

So that illustrates the-- ideally, I would have plotted one more, where we have a very small bandwidth, much shorter than that. And what you would get is like it goes squiggle, squiggle, squiggle, squiggle. And then top squiggle, squiggle, squiggle, squiggle. So it'd be very variable and not very nice to look at. And that illustrate the issue.

If we go with a bandwidth that very wide, we might get the shape altogether wrong. A technical way to say that and much more accurate way to say that is that we might get bias. Because basically, our kernel starts getting wrong and not estimating the right, it's not converging towards the right function at the end of the day. As if I start with a very large bandwidth, as I add more and more and more observation, this is still never going to go to the right shape. So this is bias.

If I have too small a bandwidth and I don't have that many observations, it's going to look much more bumpy than it really is. But in this situation with a very small bandwidth, if I continue to add observations, it will eventually smooth itself out, so for a given sample size that is variance. So the choice of bandwidth in kernel density approximation-- and by the way, in kernel regression, which we are going to see a little bit further along the way, is trying to balance these two sources of errors, the bias and the variance.

So, what it does is basically trying to optimize so you can choose. You can think of it as a minimization problem, where you're trying to minimize some function of to a natural object that summarizes the two is mean squared error, which is how far away are you, how far away is each of your observation from the truth, which is a combination of variance and bias. So the optimal bandwidth is a solution to this mean squared error minimization problem.

There are various solutions to the mean squared error minimization problem that people have suggested. The one that is implemented in R is one of the most common one. So in a lot of cases, you can I think, go with the R default, which if you type nothing, this is what's going to come. Or if you type if you type, and now the 0 is going to be the default. And then you can play with making it smaller, making it bigger with respect to 0 and see what happens.

There is no harm in trying it out. You can see what-- there is good reasons to revert to the default, not only because someone has shown that it was optimal in some ways, but also because it's a number that at least, you've not manipulated in some ways to get the shape you wanted. For example, someone who didn't like this shape, say no, I really think it has to be bell-shaped, would eventually be able to get it by just manipulating the sample a bit more. Or alternatively, someone who had a prior that the high distribution in Bihar is really-- we can go back to very few kernel-- is vary shape.

It's really, in fact, skewed towards the left. Might have said, well, I'm going to try and espouse it in some way. There's going to be difficult. But you could, for example, try to-- someone might show this histogram and convince themselves that there is bunching at a meter 50, which is plausible, simply because people might round there. The people who are measuring the height might be rounding at a meter 50.

And with enough squiggliness, you might actually be able to find it even with the continuous kernel, which of course, might not in fact, be there in the data. So for all of these reasons, if we're a little bit worried about specification searching, default is a good thing to go for over and above the fact that it is, in fact, optimal in some ways. So I would always show some default bandwidth. And then I would also play going back and forth to see whether you're actually not missing something. If you play with the kernels, you will see that with the exception of the uniform kernel that can give you slightly different shapes, whatever kernel you use is going to look prettier.

So now that we've plotted it, let's talk a bit about its substance. Let's talk a bit about the shape of this. Let me show you one where we don't have just one. What do you think about the shape of this distribution? How can we characterize this distribution in terms of its physical shape? Yeah?

STUDENT: You look at its mean and standard deviation.

ESTHER DUFLO: Yeah, we could do that. We could look at it mean and standard deviation. I was just asking you just, to look at the shape it has.

STUDENT: It's Gaussian

ESTHER DUFLO: Sorry? It looks Gaussian. What makes you think it looks Gaussian?

STUDENT: It's symmetrical.

ESTHER DUFLO: It's symmetrical, it's bell-shaped, it doesn't seem to have thick tails. So that reminds him of a Gaussian curve or what we also call a normal distribution. What does it remind you of things we've seen actually, before? We've already plotted curves looking very much like that, a couple of lectures ago.

STUDENT: Binomial?

ESTHER DUFLO: Exactly, it looks very much like a binomial as you increase the number of attempts. And that makes sense that it should look like both. Because the binomial distribution is approximately normal with mean np and variance np times 1 minus p for large n and p not too close to 0 or 1. So this curve is actually a quite normal-looking.

I don't want to say it is normal, because how would I know? I haven't estimated it. I haven't tried to reject it. But it looks quite normal-looking.

And in fact, you'll be able to show very soon that it makes sense that it also looks like the binomial, since the binomial is quite normal looking after for large ends. We will come back to our friend, the normal distribution in great detail. Because it will turn out, for a very good reason, that not only it's the limit of the binomial as n becomes large, but many random variables, we will play with in economic application and other applications can conveniently be assumed to have a normal distribution.

And in fact, if you open any statistics textbook, height is a canonical example of something that's normally distributed. So some exercise that I've been doing is asking myself, why on Earth should height be normally distributed? It's actually not completely obvious to me that height should be normally distributed, even now. So when we know a bit more about where the normal distribution is coming from, we will come back to whether there is a good reason for height to be normally distributed or not in a particular population.

So I want to get back to that. At some point, we will do it. So here is one reason why it should not be normally distributed, is for some reason, there was a bunch of Peace Corps volunteers in your data set. And and they all came from America. Then you would have a very different shape of-- if by mistakes, we would also have taken the height of all of the Peace Corps volunteers in the same area Bihar, the distribution might not have been looked so much the same way.

The reason being that here, I've not plotted the distribution of a sample, which has both. I've plotted two distributions separately. And the distribution of height in the United States look quite different from the distribution of height in Bihar. So if I had put both in the data set, what would have been the distribution in the overall sample? Which shape it have had?

It might not have had a hole because I need to add them up. But it's going to have an elephant-looking shape a little bit, the sum of the two. So while looking at these two. It's sort of interesting. What can we see from-- I told you the big fact. But what can we see from this? Why can I see almost intuitively, that very quickly, just by plotting these two, that women in the US tend to be taller than women in India? Yeah?

STUDENT: The nutrition?

ESTHER DUFLO: Oh no, that is certainly a part of the explanation. I'll get to the explanation or some explanations in a minute, if you want. But just as a matter of looking at the graphs, why can we say immediately that women-- Yes?

STUDENT: The mean for US women is shifted towards the right.

ESTHER DUFLO: Exactly. So in this distribution that are bell-shaped like that, the mean, the mode, and the median are all in the same place. So that's kind of the peak of the distribution. It's very easy to see. And we can see that it clearly has moved to the right.

What might be a little bit harder to see from this picture is whether this is true at all level of the distribution or maybe, or not. Yes?

STUDENT: I mean, in every quarter of the quantile of the distribution, it gets taller.

ESTHER DUFLO: Yes. What makes you think that? And how do you to make this statement? What did you need to look at?

STUDENT: Just with the distribution, about how everyone is away from the average. It's just goes to the right.

ESTHER DUFLO: Exactly. So to make this statement, what you had to do was to look at surfaces. It's a little bit hard. But it's possible. You look at for every possible quantile, the probability that a woman is smaller than that quantile is given by the surface here. And it is true, I think, and one can, given that the way the distribution are, you can probably convince yourself here, that all the triangles here are different than all the triangles on the red line.

But comparing triangles is not very easy. So there would be another-- if I had given you the data in another form, it might have been a little bit easier for you to make this exact point. Which way would you like to have seen the data? Which form would you like to see the data to be able to make this point that you made, which is at every quantile, we seem to have fewer American women that are shorter to a particular level than others?

STUDENT: With the data, does it use the z-score for standardization?

ESTHER DUFLO: Well, then if you standardize, you're going to miss everything about levels. So I think you're going to get-- if we were to standardize the data for that, they have the same mean and the same standard deviation, I think you might find them to be similar. Because you would have removed what is different about them.

STUDENT: What if you use a box plot, box and whiskers?

ESTHER DUFLO: So if we did a box and whisker plots, what would we get from a box and whisker plot, just to summarize for people who haven't? We would see-- what does in the box and whisker plot tell us?

STUDENT: We see for the US women, that it's shifted towards the right. And there are more-- so yeah, I think it will be shifted towards the right.

ESTHER DUFLO: Exactly. So that summary is going to be true, because the mean is shifted towards the right. But with a box and whisker plot, what would we have the mean? Yes?

STUDENT: It would also, the mean would be shifted also to the right for the United States women. But also, the box would be larger.

ESTHER DUFLO: And what is the box in the box and whisker plot?

STUDENT: The box is the middle 50%.

ESTHER DUFLO: Yes, and what are the whiskers?

STUDENT: They are the outliers, 25% each.

ESTHER DUFLO: Or depending on your whiskers, they're sometimes the 95%, et cetera. So it's already some data, but it's not all the data. So by plotting a box and whisker plots, compared to that, I'm still losing some of the-- I still cannot answer this question that is true at every quantile. Yeah?

STUDENT: The PDN.

ESTHER DUFLO: Yes. We can also decide that, OK, from the PDF, I can instead, I could plot the PDF. It's no different information. But it's a way of plotting it differently that will directly bring us to this fact. So you can do it in an histogram, which instead of plotting instead of plotting a histogram, you can plot a cumulative histogram, which is at every bin, you're not counting the number of cases that falls in that bin. But you're counting the number of cases that falls in that bin or below. So the histogram is going to be non-decreasing until it gets to 1. Or you can--

INSTRUCTOR: Just to clarify-- so if we think about a histogram, for instance, as an empirical counterpart for of a probability density function, then what Esther is going to talk about is the empirical counterpart for a cumulative distribution.

ESTHER DUFLO: Exactly, it's the CDF is a cumulative distribution function. So the histogram will tell us-- the cumulative histogram will give us something that will look like a CDF, but bigger. Or we can do it in a similar technique to a kernel, using again-- R gives that to us for free. And we get a descriptive representation. Or we get estimation of the cumulative distribution function.

And then, we can immediately see the point that we saw with surfaces, which is what can you say about the US curve, again, compared to the Bihar curve? So the US curve looks shifted to the right, regardless compared to the Bihar curve. So what does it tell us?

It tells us that whatever the threshold that we pick, there are more women that are below this threshold in Bihar than they are in the US. For example, at a meter 50, we have about, let's say, what would you say, 0.22? I'm making that up. 0.22 women who are below in Bihar, and we have less than 10% women who are below in the US.

And we can repeat that exercise many times. So it looks like, of course, eventually we will want to have a formal test of that. But visually, it does look like the Bihar distribution is pushed to the right-- sorry, the US distribution is pushed to the right, compared to the Bihar distribution. And hence, that what we might want to test is whether the the US distribution, first order stochastically dominate the Bihar distribution. That is formally, we will want to test. And we'll give you a test statistics for that when we come to that point, that at every point in the distribution and for every value of height, there is more people in Bihar that are similar to this value than they are in the US.

So, when might we prefer to plot CDF versus PDF? The underlying information is similar. But when you're interested in probabilities, for example, when you're interested in making this type of statement, representing the distribution, plotting the CDF or plotting an estimate of the CDF of the distribution is more conventional than plotting PDF. Often, although people do both. But you will see more often when people try to make the argument, try to argue exactly your point. Although you were able to do it with a PDF, most people will do it with CDF.

There is a reason, statistical reason, which is the test that we are going to talk about, is expressed in term of cumulative distribution function. But there is also, I think, a more, sort of something that has more to do with cognition and how we perceive things, which is the PDF represents probabilities to be in a particular interval, to be smaller than a number or in a particular range or whatever you're interested with areas. We have to look at triangles.

Whereas the PDF does it with vertical distance. And it's a bit easier for the human eye to compare distances than to compare areas. We don't have to do the counting. We don't have to integrate in our brain. So if we're interested in areas, then we might as well just-- it's much easier to do it with the CDF. We can see that it is true everywhere that the CDF of the US is below CDF of the Bihar.

So I think that's kind of a reason why I, at least personally, much prefer to see CDF than PDF when this type of question are being asked. On the other hand, sometimes you might be interested in densities, not probabilities. Both the CDF and the PDF shows density. But to get the density from a CDF, you need to take the slope.

So you need to take the derivative in your head, which you can do. But it's a little bit unintuitive. Whereas the PDF gives it to you by distance, which is easier for us to look at. So if you're interested, for example, in knowing where the mode of the distribution is, whether the distribution looks symmetrical or not, whether it looks skewed or not or stuff like that, it is possible to do it from a PDF. But it's much less good from a CDF. But it's much less good, so typically, you're going to plot a PDF.

So that's the-- and then generally, of course, at least for your own eyes, you need to be able to go from one to the other. And and then you're going to choose, depending on the argument you're trying to make, what is going to be visually represented in your presentations a the end. So I hope you are super excited because I've given you a very new novel fact that American women are taller than Indian women. So it was worth three lectures of probability and statistics to get to that point.

We even have an explanation for it. It might be nutrition. Although some people have argued it's about genes. Little aside, while we are at it, what would be an empirical strategy to distinguish between the two?

STUDENT: How would somebody, who's genetically Indian, but grew up in terms of the same nutrition as the United States.

ESTHER DUFLO: Yes, so for example?

STUDENT: Myself.

ESTHER DUFLO: Yourself. You're pretty tall. So that would be a good. So if we did that, if we plotted that, we would have the first generation, the children of Indian immigrant. So take Indian immigrants. Where your parents born in India?

STUDENT: Yes.

ESTHER DUFLO: And your name is?

STUDENT: Kutan

ESTHER DUFLO: Kutan. Say it again?

STUDENT: Kutan.

ESTHER DUFLO: Kutan-- so Kutan's parents were born and brought up in India. And he was born here. And of course, he's a male, which sort of screws a bit, the example. But if he had been a female, the distribution actually, of people like him, who are female instead, is exactly in between the two. So its first order stochastically dominates the Indian women. And the US kids, first order stochastically dominant first generation immigrant.

Does that tell us we are about half-half? Half gene, half nutrition?

STUDENT: No.

ESTHER DUFLO: Why?

STUDENT: Because it still didn't give us any cause. We just changed one. So we don't know that specifically, genetics that causes that difference.

ESTHER DUFLO: Yes. So that's one thing. Yes?

STUDENT: There's also other omitted variables. Altitude could be one, the amount of oxygen you get as you're growing up.

ESTHER DUFLO: Yes, that is all the other stuff going on. And there's another reason. Yes?

STUDENT: Maybe it would also be worth looking at more than just one gender.

ESTHER DUFLO: Exactly. So what if we looked at grandchildren? Do we have grandchildren here? Yes. So if we looked at grandchildren, would you have a guess where we would be? Yeah?

STUDENT: Could that be that born in the US?

ESTHER DUFLO: Yes, they are pretty much in the same place. So the grandchildren are pretty much in the same place. So the problem with children is that-- I don't know if it's a problem. But the fact with children is that, so it turns out that when you grow up with poor nutrition, your entire body gets affected, including your reproductive cells, such that your parents probably got, assuming they grew up or the parents of the average immigrant person-- I know nothing about your parent.

But if your parents grew up in a place where they didn't get very good nutrition, there is what they call epigenetic modification of the reproductive cells, which means that the children of people, who are themselves not properly nourished, will not reach their-- will also be smaller. But meaning they will not reach their genetic potential. But the children of these people, because there is no epigenetic modification of these guys because they were perfectly well nourished from the beginning, the children of these people, so the grandchildren of the immigrant will be back to full possibility.

So in order to look at the genetic things, we need to look at more than one generation. In the same way-- so we hac one question. Yeah?

STUDENT: I just wanted to understand, how do you use the word first order?

ESTHER DUFLO: First order stochastic dominance?

STUDENT: Why first order?

ESTHER DUFLO: First order stochastic dominance is about the fraction of people who are below a certain height. So first order distribution, first order stochastic dominance. Another one-- if the probability to be below a certain height is smaller, regardless of the threshold. Second order stochastic dominance is if maybe it doesn't, But the variance is smaller at each of these points. So we'll get back and write down formally those distribution, as we move to statistics. But that's the reason.

OK, this was an aside in order to make that graph interesting for 5 minutes but much more interesting, we can go back to Mr. Steph Curry and plot CDFs, cumulative distribution of-- so someone makes-- I think it's Kutan actually, made the claim. Someone made the claim, the bold claim that he might have been the best 3-point shooter? Is that the point?

Shooter.

Shooter, absolute shooter in the history of basketball. So in order to check this claim, we should have given you more data than that. But these are at least him compared to a couple of other contenders for that title, according to knowledge. So you don't ask questions on that.

[LAUGHTER]

Or if you ask them, you direct them to Sarah over here. It looks, in fact, that the colors are all a little bit blue. But Steph Curry is the darker here. And it looks, in fact, that the distribution of the-- the distribution of his shots in term of where he took the shot from, distance to the basket, stochastically dominates LeBron James and Kevin Durant.

So you might say, well, it's easy because you can try. m in fact, when you look at successful shots, it's true as well. So Steph Curry is, in fact, maybe the best shooter in NBA history. There is more data to prove that point, if you're interested.

But that's not the only thing we might be interested about with him. In fact, one thing we might be interested in is where he shoots from. And there are two places to consider, since the basketball court is a plane, is the distance from-- the horizontal distance from the basket-- the distance from where he is to the basket in this way and then if he's on the side or not on the side, so distance to the baseline and distance from the sideline. I'm going to do it with respect to the midline of the court.

So if I plot each of them separately, here is the basketball court. So I'm going to call the midline a line that's not this in here. But it goes from 0 to minus 25. And then the distance from baseline is-- suppose this is 0 to 94. Principle, you can shoot from anywhere you want. But you'll get more points if you shoot from here. And of course it's going to be much, much easier if you hide below the basket.

So if I were to plot either distributions, I could be interested in distance from the baseline of court. What's interesting here, I've plotted the histogram, and they can help with the optimal boundaries. What comes out of this?

Yeah.

STUDENT: That huge peak could be due to free throw shots, maybe.

SARA ELLISON: those shots are not in the data set.

ESTHER DUFLO: So what's the [INAUDIBLE] even I would know that? Yeah.

STUDENT: Would it be that, say, like, 5 feet away is where they start jumping to dunk?

ESTHER DUFLO: Yeah, you can just dunk, exactly. So we have a peak there. And then we seem to have something over here. So as we go, they seem to be-- the further he is, then there is not too many shots within a close distance. Then they start kind of going up, but in a smooth way, maybe nothing very-- I mean, it's kind of hard to interpret exactly what's going on here. So the peak we understand quite well. But then the [INAUDIBLE] and why it goes slowly is not fully clear.

And if I did that, we already knew. I already showed this picture. From the midline of court, likewise, we have a big peak at 0 because they are right under the basket. Well, then it kind of slows. Maybe it's trying to go up again. We might miss it if we have to smooth the boundaries, as we discussed. But it seems to be a bit--

So what's going on here? What am I missing by plotting the distance to midline and distance to baseline symmetrically?

Yeah.

STUDENT: Since it's radial, it might be that you can interpret that, oh, he's better at shooting 3 from [INAUDIBLE] symmetrically? Or you can interpret it as, oh, he's better when he's closer because [INAUDIBLE] nearer to the basket. So you cannot really tell the difference because that he prefers to shoot from the middle or if he's actually [INAUDIBLE].

ESTHER DUFLO: Yeah. And in particular because it's radial, just showing the distance to the baseline is going to change, depending on where on the-- suppose that what he was trying to do is to try to not go past the 3-point line, but yet be as close as possible from m because closer is easier, then that's going to depend on-- the distance from the midline is going to depend-- that he's going to shoot from is going to depend on where he is with respect to the baseline, and vice versa.

Which means that in order to understand whether it is, in fact, a pattern to the way that he's throwing, in particular to test my hypothesis that he's going to try to be as close as possible to the 3-point line. What do I need to show you, instead of showing you this data separately?

STUDENT: Could you make the 3-point line a 0 and then just somehow come up with a new coordinate to give you distance?

ESTHER DUFLO: Yes. So we could renormalize the data, everything with respect to the 3-point, line And in order to even know whether it's worth doing.

STUDENT: One possibility would be to draw it on the x-axis and the y-axis to sideline the midline--

ESTHER DUFLO: Exactly.

STUDENT: --and then that points somehow to another-- the radius of the circle or something that then determines the number of successes at that coordinate.

ESTHER DUFLO: Tada!

We can, in fact, do that. It looks better on my screen. This is a three-dimensional graph that does exactly that. So now, I am drawing something which-- basically what you were saying is that, look, the distance to the baseline, distance to the midline are not independent distributions. They are not independent. I don't want to be drawing them separately.

So either I force the functional form and transform it back into one variable. We could do that. Or I could-- and then we should be able to see bunching at the 3-point line. Or I draw it in three dimensions, which is something like we get here. So it looks better on my screen.

But this is kind of a heat map with where we both-- there is both a heat map and the height of the graph. This is done in R. You have various ways to represent three-dimensional objects in R. One of them is this that gives you both the height of the bar and the colors. Another one is just a heat map, which will just use the color, the color. And otherwise, you can have just [INAUDIBLE] whatever you prefer. I thought this would be the best-looking one, but it's actually not that great.

So what you see now is that clearly, at something, it looks like the map-- it looks like the basketball court viewed from above, where the height of each of the bars, in fact draws very nicely the 3-point line. And then we also have a huge peak at where they can dunk-- where he can dunk. And then now we have something that looks interesting.

So from then on we could say, well, clearly, there is something we could now have a functional form that expresses everything with distance with two kind of summary variables is distance to the basket and distance to the 3-point line. And that could be put in a two-dimensional graph.

OK. So these are kind the type of things you can do as soon as you get your hands on some data, plotting them up. But something that-- I want to give you another very specific example of why it's really useful to understand the underlying distribution of our variables. We're going to see plenty of examples of

That. But before we do any statistics or econometrics, [INAUDIBLE] the power of understanding probability. And this is about top income distribution. You might be interested in the distribution of income. So first let's look at distribution of wages. Here is California, CPS, year 2014, data that you could download off the internet.

Does it look Gaussian? Just when I was telling you that lots of stuff in economics look Gaussian. And there, it doesn't look Gaussian whatsoever. So forget that little guy for a while. We'll come back to him.

[LAUGHTER]

It does look pretty skewed to the left. In fact, it turns out wages are not at all normal. But the log is a little better. So the log is-- actually, before we plotted that, I would have thought it would be more bell-shaped once it's logged. Wages are frequently assumed to be log normal. It's a little surprise that in the log, it, still doesn't look very symmetrical.

So it is useful to plot the data. When you get some data, it is very useful to plot the histogram to see how it looks like. When you see a shape like that, you might think that it's worth taking the logs and that the log would look a little bit better.

But even once we've done that-- I've done that kind of getting rid of the little guy to the end. But what is that picture? What's that-- it looks like there is a mass of wages at outside of the range. But what is going on here? Lots of people with exactly the same income seems unusual because we might think that wages are actually continuous, random variables. So we shouldn't expect a mass point.

STUDENT: CEOs?

ESTHER DUFLO: Yes. But you think CEOs earn all exactly the same amount?

STUDENT: It depends on how big [INAUDIBLE], Right? So what's the range of the [INAUDIBLE]?

ESTHER DUFLO: I'll tell you, you make it as little as you can. And it's still going to be this. Yeah

STUDENT: It could have a legal basis. There may be an accounting reason why it only has so much income at a specified value for tax reasons. And then the [INAUDIBLE]

ESTHER DUFLO: OK, so it could be that. So if not, we know that CEOs make all sorts of different wages, Yes,

STUDENT: [INAUDIBLE] that is sort of constant to that, because if you can identify [INAUDIBLE].

ESTHER DUFLO: Exactly. Wages are top coded in the CPS.

So

[INAUDIBLE]

Wages are top coded in the CPS. I think we have it in a slide. Ta-da-da-da. Wages are top coded in the census and the CPS to protect privacy. What does top coded mean? Anyone who makes more than that is going to be brought back to that. So anyone who makes-- I don't know what that is, but it's not that high. And anyone who makes more than that amount is just [SNAPS FINGER] brought back to that level, The top codes vary from year to year, et cetera.

And this is exactly for privacy reason. You don't want to identify-- there are not that many people who make-- if you wrote down every single wages of people who are interviewed in the CPS, it would not be very difficult to find them. Ah, so that's fine if you're interested in the impact of minimum wage on people's lives. Because they are not very many people who are the-- top coded people are unlikely to be affected by that.

But what if we're interested in CEO pay in Silicon Valley? Then that's not going to be the right data. And the right data might be a little bit difficult to get hold of. In fact, there are sources of data. Forbes publishes some data, et cetera. You can look for it. But it's generally not that easy to find data, certainly not if you are interested in the history of very top wages.

You will find some data that is published on CEO pay. But for recent years, and if you wanted a continuous time series going back a long time to see how things changed over time, you would not be able to do it.

And we could be-- yeah.

STUDENT: I'm curious, if one wants to get the real wages of CEOs in that quantum corner, would it still be having [INAUDIBLE]? Because I'm just thinking of people who had some of the big technology platform companies. They're basically sort of consolidating a lot of the wealth. So would it still be a nice degradation? Or would it be--

ESTHER DUFLO: That's a very good-- that's a very good question. And we are going to answer it right now.

[LAUGHTER]

So some people are interested in top 1%, particular the rest of us.

[LAUGHTER]

There was a lot of discussion, this 1% captured the imagination, the one person, how much income does the one person [INAUDIBLE] captured/ it got embodied in this Idea of we are the 99% that now you people discuss, 1%, 99%. It's just part of the language, thanks to the Occupy Wall Street movement.

But where they get it from? We've been talking a lot about the Republican presidential candidates. So let's even the balance and talk a bit about the Democrats. So I think it turns out-- so I was trying to trace the 99% history. My best guess, my best answer at the moment is that it comes from this one speech our friend Bernie Sanders.

December 10, 2010 made himself-- well, he was kind of famous already, but made the 1% famous with his famous Filibernie, which was an eight-hour speech on these issues, opposing further tax credit, et cetera for the rich,

And here is what he said. "We cannot give tax breaks to the rich when we already have the most unequal distribution of income of any major country on Earth. The percentage of income going to the top 1% nearly tripled since the 1970s. The top 1% now owns more wealth than the bottom 90%. That's not the foundation of a Democratic society. The fact is 80% of all new income earned from 1980 to 2005 has gone to the top 1%. People should be mindful of this fact. The last time that type of income disparity took place was 1928. I think we all know what happened in 1929."

You have to agree that compared to most political speeches, this is chock full with facts.

[LAUGHTER] Even

Better if you wanted to fact check this fact. You could. They come back to these guys. These facts come verbatim from research that Thomas Piketty and Emmanuel Saez wrote some years before. Their first article was for the late 1990s for the US. building on a lot of Thomas Piketty's work for France done in 1998.

SARA ELLISON: By the way, Thomas was a faculty member here at MIT. And Emmanuel was a PhD student here [INAUDIBLE].

ESTHER DUFLO: And how did they do it? They were interested exactly in your question. What's the shape of the income distribution past the line? Does it still look very bunched, or what does it look Like? And the issue-- and they calculated it for many years. And you can go back to Mr. Sanders' speech.

And I assume-- and I hope you will be impressed by the fact that everything that's said actually is faithfully taken from that one graph, which tells us the top 1% income share. In fact, Emmanuel recently extended the graph. The original paper I think went till 1998. And Emmanuel extended-- keeps extending every year. He adds more data. So it goes now till 2014.

And the black line is the share of income that goes to the top 1% of the US population. And you can see that he's right. Which is, in 1928, it took a huge beating with '29 crises. And then it recovered a little bit. Then there was the war. And then it was flat for a long time.

And since the 1980s, really, it's been steadily increasing. You can see the impact of the dotcom bubble and the 2008 crisis, where it falls down. And then it went back up again, In fact, between 2009 and 2014.

So this is a lot of data. And you might ask yourself, where is that data coming from? How did they find out since there are no places-- there is no data set that tells you line by line. In fact, now, for recent years, the IRS makes available to the Piketty [INAUDIBLE] team and people who work with them in very, very controlled situation access to the individual tax data, but this is recent, and this is not available for a longer time period. So how could they do these graphs?

So the only source of data on the distribution of income that's consistently available over a long period of time is the tax data. And the tax data doesn't tell you which person pays what taxes, thank goodness. And there would be a lot of private information devolved on us. But it's available in the form of tabulations.

So what is a tabulation? Basically, it tells you for-- for bracket of income, it tells you how many people are in this bracket, correspond to a tax bracket? Or how many people in this income bracket-- obviously, a bracket is defined by a minimum and a maximum amount that defines the bracket. And the tax data also tells you the average income of people in the bracket.

So this data comes from combining this data with population census data and aggregate income source to calculate the share of total personal income accruing to various top income groups that you would be interested, the top 10%, the top 1%, the top 0.0001%, if that's what you're interested in.

How do they do that? They exploit that-- they do that by exploiting a known fact about distribution of top incomes, which we can actually verify in the data, which is that it tends to be well approximated by a Pareto distribution. So it's not a bunch. It's a Pareto distribution.

What's a Pareto distribution? In technical terms, that's the density-- or, 1 minus the density. 1 minus the density, that's the probability that X is bigger than any number. X is X-m over X to the power lambda for any X bigger than Xm and 1 over, otherwise.

So in order to draw the PDF, the PDF looks something like that. So say if Xm is 1, it looks like this and then like this. We're typically are not interested in income close to Xm. But this gives us the scale parameter. And then lambda is called the shape parameter.

So what it reflects is something-- it's more intuitive to see that where it's the average income of those whose income is greater than some line, Y, is Y times the constant lambda over lambda minus 1.

So for example, suppose lambda is equal to 2. What's the average income of people whose income is larger than $100,000? $200,000. If lambda is equal to 4?

STUDENT: [INAUDIBLE] times.

ESTHER DUFLO: Exactly. Which is bigger or smaller than 200?

STUDENT: Smaller.

ESTHER DUFLO: Smaller. So the smaller lambda, the more inequality there is at the top of the distribution, in the sense that the bigger the share of the income that goes-- the larger is the average income that goes above any given line. This, of course, doesn't-- this is for lambda greater than 1. Otherwise, that's not true. If lambda is greater than 1, the average distribution-- the average above any given threshold is infinity.

So this is the-- that's kind of what's the known property of the Pareto distribution is-- and it's a very useful property. Because once you know that, then it's very easy to calculate the coefficient of the Pareto distribution and the Xm parameter. I'll show you how they do it in a minute. And then once you have that, then you can do it for any quantile you're interested in.

Before we get that, a little aside. The Zipf distribution is special. [INAUDIBLE] whose income are not distributed. But it's a special case with lambda equal 1. And one way that Zipf distribution is described is in terms of the rank. And it says that the log-- the way you can write the Zipf distribution is that the relationship between the log rank and the log size is 1.

So I'll show it to you for city. The log population plotted against the log of the rank is exactly minus 1. That means the-- say, for example, the word that-- it's also works for words, in the English language anyway. So the word that is spoken the most in the English language is, what you think?

It is the, the word the. And it's spoken about 7% of the time. The next most frequent is of. And it follows as it flow. So you can tell me how often number two, given that I anchored it at 7% for how frequently is the of spoken. It's around 2.

STUDENT: 3.

ESTHER DUFLO: 3.5, exactly. And it continues like that. So many phenomenon are approximated by Zipf's law. And people like to give explanations for these things. Maybe if we have time down the line, we'll give you the one for cities. But it's more economics than statistics. So we might not have time to get there. But you can see that Zipf's law, cities-- the farm size also is quite well approximated by something like Zipf's law.

And income, it turns out, follow Pareto distribution. And what you can do is estimate them. So income, it's at the top. Top income tend to be Pareto distribution, which is consistent with the fact that the rest of the income tend to be log normal. Because at the tail, it's very difficult to distinguish a log normal distribution from a Pareto distribution.

So if income distribution is Pareto at the top, then you can derive simple expressions for the share of the top 1%. So I'm going to let you do this calculation in your homes, very simple integration of the-- you take the derivative of the-- from this, you can get, of course, the PDF, take the derivative of the PDF, and integrate over whatever you're interested in.

You can convince yourself very easily that the top-- the q percentile share of income can be derived as q over 100 to the power of lambda minus 1 over lambda. So for example, with lambda equal 2, the top 1% share is 10%. With lambda equal 3, it's 4%. So the smaller the lambda, the more inequality there is at the top.

So all we have to do now is to use available data to estimate the shape parameter lambda and the scale parameter Xm. And that's what Piketty started doing for France. By the way, not he's not the first to have observed that. Jim Poterba, who is here, is one of the first persons who have done it. And he started doing it for-- Piketty first wrote a big book on France and started doing, systematically, a little bit differently than Jim Poterba had done, but very much same principle. And then Piketty and Saez did it for the US, and then did it again for wealth distribution, et cetera.

So while it is true that the top distribution is approximately Pareto-- and you can check that because you can calculate the b for various brackets that you would get at various brackets. And you can calculate the b above 100,000, above 200,000, above 300,000, whatever you're interested in. And you can see that the b does look approximately constant as you move up. But it's not exactly constant. So it is best to use the b that is the closest to the quantile that you're interested in.

So how do you do that? Well, the table gives you, for each tax bracket, the number of people in the bracket and the sum of the income of people in this bracket. So if you divide the sum of the income of people in this bracket by the overall income of the economy, you get the average income of people in this bracket.

And if you remember, the average income of people in a bracket is the multiplicative factor times the minimum income of people in the line. So that allows us to calculate the Pareto coefficient, b. b is b over b minus 1, as the average people of income in the bracket divided by the lower bound of the bracket-- average people, sorry, above a particular line divided by the line.

From b, we can calculate lambda. And from lambda, we can calculate the scale parameter. And then we have everything now that describes-- we have an entire description of the distribution at the top. We have everything we need.

We take the share going to the relevant percentile using the closest tax bracket that gives us the closest line, S. And we calculate the top income of everybody above divided by total income to obtain the average income, obtain the coefficient b, obtain lambda, and then applies this to the quantile we're interested in.

So that's what they do in a systematic way. And then once you know that, you can just do whatever you're interested in. And that's how we get this nice graph, which have had, I think, a pretty altering consequences, at least on the political debates of the last few years. Yes.

ESTHER DUFLO: I was just curious if the slope of the Pareto is changing as the years have progressed.

Well--

STUDENT: Especially, it seems to be getting more and more intense. Because my assumption would be not coming [INAUDIBLE]. Technology is allowing more and more platforms to consolidate things, in a way the inequality stretch, also keep on increasing.

ESTHER DUFLO: So there is a lot in your question. The first point is an empirical fact. You're asking if the Pareto coefficient changes. You can almost see there. By definition, it does. This is why this is telling us, in a sense. The top 1% is getting a larger and larger share of the income. That means b is becoming smaller and smaller, at least since the '80s. And it was higher before.

The second thing is you're providing an explanation for this fact, which is technology. There is actually a very interesting paper by [INAUDIBLE] about CEO pay, which very much makes this assumption, in the sense that the technology means that CEO pay will be-- sort of derived the reason why CEO pay would be Pareto distributed and that the coefficient might change as a response of how much leverage there is in the economy. Yeah.

STUDENT: Is this all before or after taxes?

ESTHER DUFLO: So this is before taxes. And then after taxes is the same ship. (LAUGHING) Because--

[LAUGHTER]

--because our tax rates are not hugely-- of course it's flattened a bit. But our tax rates are not extraordinarily progressive, not progressive enough to undo that. So the shape is similar and the levels are different.

And you're asking a very good question, which is, for some problems we're interested in before. For example, if we're interested in testing his hypothesis that the technology-- development of technology will lead to a concentration of the income towards few people. To a changing of the Pareto distribution we are interested in pre-taxes.

If we we're interested in how much inequality there is in actual spending power, we are interested in post-taxes. So depending on what we're interested in, we're going to be interested in both.

I should say that the big sort of competing explanation for this type of phenomenon, one is what you're talking about, technology-led explanation. The other is institution-led explanation, in a sense how COP are set by boards and stuff like that.

There was a very interesting *NPR Planet Money* episode a few days ago about CEO pay that was very much leaning in the institutional direction, which is-- and the Piketty says-- do favor the institutional direction. But there is lots of good works on both sense. Once you have the facts, you can start playing a lot to try and explain them. Yep. No, you. You had your hand--

STUDENT: [INAUDIBLE] about assuming the Pareto distribution? Or is this sort of generally accepted?

ESTHER DUFLO: Well, you can verify it. So first, it's quite generally accepted. Second, you can verify it. It's very easy to check because you have several tax brackets at the top. So what you can do is to calculate the Pareto coefficient for several lines, and you can see if, in fact, they look roughly constant. And what you do is that-- what you see is that for a given year, they look roughly constant, but then they change, after a certain income, after-- it's only true at the top tied to the [INAUDIBLE].

All right, so I think we can stop here. No point starting something new. And we will start again on Monday with first deepening our knowledge of joint distribution, and how we play with them, and then function of random variable, and the like.