[SQUEAKING]

[RUSTLING]

[CLICKING]

ESTHER DUFLO: So let's get started. I will quickly-- I will just go straight jump in because I'm actually-- the end of last lecture. So imagine we are still Monday. And then I will start with the proper introduction of what's coming in the future and the like.

So if you remember, the formula for the density for a normal distribution with mean mu and variance sigma squared is this formula over there. This is how it looks like. You all know and like it-- bell-shaped, symmetric, thin-tailed. This is for the same mu and different variants. So you can see, of course, that as the variance becomes smaller and smaller, it's higher and higher at the mode. There is a larger fraction of the area under the curve that is towards what is both the mode and the mean of this distribution. And as the variance increases, it becomes flatter and flatter.

That's the standard normal. And very often, as of today and certainly as of after spring break, we will need to know for various reasons what is the area under the curve. What's the probability that we are below minus 2 or above 2?

And there is no closed form formula for the integral of the PDF. But a number of people have worked very hard on approximating it for us. So that's roughly how it looks like, which is you have about 68% of the area under the curve is between minus 1 and 1, 95% in between minus 2 and 2-- something we make a lot of use of in what follows. And then 97.7% is-- sorry, did I miss-- I misled you? Start again.

AUDIENCE: I think it's mislabeled a little bit.

ESTHER DUFLO: Yeah, it's just too high, no?

AUDIENCE: No, no, no.

ESTHER DUFLO: So 95-- sorry, yeah, it's mislabeled. My bad.

AUDIENCE: It's mislabelled.

ESTHER DUFLO: OK. So the little ones are correct. 2.1% is in this quadrant. 2.1% plus 2.1% is 4.2%, so that means whatever is left is 90-- it's those ones that are mislabeled. My bad. OK, I will have to fix that for posterity. These ones are correct. 2.1% is below minus 2. 2.1% is above minus 2. 13.6% is between minus 2 and 1. 13.6% is between 1 and 2. 13.4% is between minus 1 and 0. The 34.1%-- between 0 and 1.

So, in general, if you wanted to know what fraction of the curve is-- what fraction of the area under the curve is below any kind of number, you use these tables that I've distributed along. So I've distributed all of them, so I don't have one. But suppose that you wanted to know, for example, the probability that z is less than minus 1.23? You look at your tables. I think there is a bunch left. I have one and there is a bunch-- if someone doesn't have, there is a couple left here.

OK, so to use these tables, you go down the left column to minus 1.2 then you move up the top row to 0.03 and that tells you that the probability that z is less than minus 1.23 is 0.109 exactly you agree that's wrong actually 0.10.109 now suppose that you wanted to know the probability that your random variable is above minus point minus 1.68. What would you do?

AUDIENCE: 1 minus the value you find in this table,

ESTHER DUFLO: Exactly. You find 1 minus the value you find in the table for. So what I did it-- I wanted 1 greater than 1.68, 1 minus p of z minus 1 is 8, and you go down. Suppose that you wanted positive numbers, and I had not give you the positive numbers. So, in fact, that happens rarely. But what happens frequently is you want a negative number on the table only gives you the positive number. What would you do? So suppose in this case, where I gave you the positive number and I've not given you the positive numbers-- suppose you want p smaller than 1.45, and you only have the number with the size where the z is negative. Yeah?

AUDIENCE: So it's symmetrical so that you've got [INAUDIBLE] probability that is greater than 1.45?

ESTHER DUFLO: Exactly, which is 1 minus the probability that is smaller than 1.145, and then you would be in business. What if you wanted the probability that z is in between minus 1.23 and 1.45? You take the first one that we already calculated minus that.

Now what if you didn't have a standard normal? You would first standardize your variable, and then you would be in business. Now in reality, what could you do if you were not in a class with no access to computers? You could use the pnorm command in R, which is going to tell you what the probability that z is greater-- z is smaller than, say, 1.96.

If you ask for the lower tail too, which, this is asking for the lower tail, or it can give you the probability that z is greater than 1.96 if, in fact, you are writing lower-tail false and then is looking on the other end. So the probability that z is greater than 1.96 is quite high. The property that z is smaller than greater than 1.96 is much smaller.

So here, I put a little table of useful variable related to the normal distribution. So the unknown variable generates random numbers from normal distribution. If you didn't want to go through the motion of doing first a uniform and then using the inverse of the PDF using the qnorm function-- dnorm gives you the probability density function.

So, for example, dnorm of 0, 0, 5 gives you the density, the height of the normal with mean 0 and standard deviation 0.5. pnorm gives you the area under the standard normal curve. So that's what you're going to use very frequently, if you want to down the line, when it's become useful. And qnorm is the inverse of pnorm. So this is what you would use if you wanted to generate a random observation drawn from a random distribution starting with a uniform distribution. You would draw from a uniform, and you use the qnorm function. I have a code for that, actually, at the end of the text.

I'm going to spare that. I just want you to see how we use the pnorm. So this happens to be for a normal distribution that has mean 2 and standard deviation 0.5. Then the inputs are what we have below. Probability that is less than x1-- so you set your inputs. Then you ask properties that it's less than x1, probability that it's between x2 and x3, probability that it's greater than x3. You will need lower-tail false. Lower-tail true is the default, so you don't need to specify it.

And that's the result you are getting with these numbers-- the probability that it's less than x1 is 0.25 doesn't-- the result don't matter. But if you write the code like that, it's going to give you this kind of results. Yeah?

AUDIENCE: [INAUDIBLE] the defaults?

ESTHER DUFLO: So lower.tail=TRUE gives you what is to the left-- the area to the left of it. So its the probability that z is smaller than the threshold you put. lower.tail-- so if you don't specify, by default, its lower tail equal to. Lower tail equal false-- it's going to give you the probability that z is greater than this number, so 1 minus the probability that it's smaller. So in principle, you could-- I be very economical and only give you the lower tail, like a table, and you would figure it out. In principle, it could only accept to give you the answer for positive numbers, and you could calculate them.

But, in practice, very kindly, it will give you what you need in just one comment by putting the right-- so, for example, here, you have the lower tail false, we get the probability that z is greater than x4. Here, we didn't specify anything at the top, so it's lower tail true. It gives you the probability that the number is smaller than the threshold. So it's pretty straightforward to use.

I think I'm going to skip the last thing because we went over this many times. But the slides will stay here. It's just to tell you that in principle, you could use the inverse sampling method to sample from a normal distribution in R. So you would take a uniform distribution, and then you use the qnorm function, which is the inverse quantile function. And you would get something.

In practice, if you sample from a normal distribution, I think the algorithm they use is somewhat different. And it is much faster. It takes much less time to do it. The way they manage to do it is much faster. So, in practice, the algorithm that is used is different. But in principle, again, what we discussed still apply, and you could sample from your normal distribution using the qnorm function that's the inverse quantile function. I think I'm done discussing normal distribution for the time being, so we can leave the stage to Sara.

SARA ELLISON: OK, so as far the-- as I said, the theoretical questions on the exam are going to be similar to the practice problems that I've made available. I don't have any practice problems to make available for the sort of practical data skills questions. So I wanted to give you some idea of what to expect. Obviously, these are just kind of very loose and informal descriptions of what kinds of questions that we might think about asking.

So one thing that you've seen a lot so far this semester, one practical skill that you've developed, is creating histograms. So you should know what a histogram is. And you've always sort of used R to create histograms, but you can easily do it by hand. So perhaps you'll be asked to create a histogram by hand if we give you a table of data. We've also talked a lot about generating simulated datasets from some particular distribution. So be aware of how to do that.

We talked about web scraping, and you had that sort of substantial question on the problem set asking you about web scraping. We're not going to ask very detailed questions about web scraping, but you should know, generally, what it is, how to do it, maybe be able to give some examples of datasets that you might obtain through web scraping, et cetera. So this is the kind of flavor of the practical data skills questions that you might see. Is that OK? Any questions, clarifications? Nope.

OK, so on to the sample mean. So I'm trying to remember where I left off before [Esther's lecture on special distributions. Certainly, we were talking about functions of random variables, and I think that was the main topic of conversation. And, oh, maybe it was the--

ESTHER DUFLO: [INAUDIBLE]

SARA ELLISON: Oh, you weren't there. Yeah, in any case, we were talking about functions of random variables. And today, I want to talk about a very special function of random variables that we will encounter a lot. It's super useful, and you probably already have a notion of why it might be useful. But we'll go through that in detail.

But first, we want to define it, and we want to talk about how it behaves. And this special function is called the sample mean. So the sample mean is the arithmetic average of n random variables from a random sample of size n.

So we, at least initially, are always going to be talking about sample means from i.i.d. Random samples. You can define a sample mean from a random sample that's not i.i.d.-- no, that's not Independent Identically Distributed random variables. For now, we'll just focus on the i.i.d. Case. I just want to let you know that this notion of random-- or of sample mean can exist outside of the realm of i.i.d. data sets.

And the notation that we use to denote the sample mean is an X with a bar on top of it. And sometimes, I use an n subscript to indicate how many different random variables are in this random sample that we use to compute the sample mean.

We also call-- so right now, I've written this as a function of random variables. But these random variables could have, instead of dealing with the mathematical construct of a random variable, they could each have a realization or a value that they take on associated with them. And we also call the arithmetic average of the realizations of those n random variables the sample mean.

So it sometimes is a little confusing because we use the same terminology to discuss the function of random variables and the function of the realizations of those random variables. We don't distinguish that in our terminology. And we actually don't distinguish it in the notation either often. But it's a distinction you should keep in your mind because you have to be able to think of a sample mean in both ways. You have to be able to think of it as a function of random variables and as a function of the realizations of random variables.

OK, so arithmetic average of the n random variables or realizations from a random sample-- why would such a function be useful? So we've got this random sample. We've got n i.i.d. Random variables from some distribution. Why would we want to add them up and divide by n? What do you think? Do you have-- your hand wants to go up, but it hasn't quite--

AUDIENCE: [INAUDIBLE]

SARA ELLISON: Yes, in fact-- so what happens if we have an i.i.d. Random sample from some distribution but we don't know everything about the distribution? Maybe we don't know, for instance, what the mean of that distribution is. Well, if we have a random sample from that distribution, and we calculate the sample mean of that random sample, then maybe that sample mean can tell us something about the underlying distribution itself. That's the entire idea behind estimation. I'm going to hammer away at that over the next several minutes in examples. But that's why this quantity, why this particular function of random variables is important.

OK, so emphasizing one more time, the X's are random variables. So X bar is also a random variable since it's a function of random variables. So let's figure out how it's distributed. So a random variable-- it has a distribution. Let's figure out what it is. We're not going to be able to say precisely what it is, but we're going to be able-- at least, unless we put more structure on the random sample, but we can certainly say things about its distribution.

Oh, I should just say note that if we knew how the X's were distributed, we knew that they were all coming from a uniform distribution or a normal distribution or a Poisson distribution or whatever, we might be able to use something like the n version of the convolution formula. So it's not exactly a convolution because it's not just the sum of the X's. It's the sum of the X's divided by n. But it's similar to a convolution. For now, let's just try to-- instead of choosing a specific distribution for these X's, let's just say that they're independent and identically distributed and see how far we can get discussing the distribution of X bar with just that information.

OK, so let's see if we can calculate the expectation of this function X bar. Well, the expectation of X bar is just equal to-- I just plugged in the definition of X bar. And then I can write the expectation of 1 over n times the sum of the X's. I can write that as 1 over n times the sum of the expectations of the X's. We know that in properties of expectation. We saw that we can do that with expectation.

And then I just changed the notation at the end from expectation of X to mu, which is the notation that we often use for expectation. So I just changed it, and then we're adding up-- we're adding up mu n times and dividing by n, so it's just equal to mu.

We can do a similar calculation with variance. So let's figure out what the variance of the sample mean is. Well, it's just equal to the variance of-- we plug in the definition again. We use properties of variance to bring the 1 over n out in front. But, of course, it gets squared because of the properties of variance. And then I changed the notation to sigma squared.

And, in the end, we get sigma squared over n is the variance of the sample mean. So mu is the expectation of the X's-- each of the X's. Sigma squared is the variance of each of the X's. We take a bunch of X's, add them together, divide by n. That new random variable has the same meaning as the individual X's. And its variance is that variance of the X's divided by-- yes?

AUDIENCE: Why isn't it divided by n squared?

SARA ELLISON: Oh, because we're adding up n of these, we've got n of these divided, and then that's divided by n squared. So one note-- it's not necessarily important in this calculation, but sort of could be important later-- we use independence in the variance calculation to go from the variance of the sum of random variables to the sum of the variances. We needed independence of these random variables. You guys remember that from properties of variance? We didn't actually need it in the expectation calculation. So that can be a useful thing to note.

Now what do these calculations tell us? So we've got this sort of random sample of variables that we've taken from-- we haven't specified what distribution they're from-- from any old distribution. We've got a collection of random variables called a random sample. We add them up. We divide by n. That creates a new random variable called X bar. And we know that X bar has the same mean as all of the underlying random variables that we use to calculate it. And its variance is that variance divided by n.

So basically, we've got-- take any old distribution or even a discrete distribution instead of a continuous distribution. We get a whole bunch of realizations from these distributions, a whole bunch of realizations, and then we compute something called the sample mean. We haven't said what the shape of the distribution of the sample mean is going to look like. But we know that the sample mean is going to have the same mean as the underlying distribution, but its distribution is going to be more concentrated because its variance is sigma squared over n.

AUDIENCE: Wait, I've got a question.

ESTHER DUFLO: Yes.

AUDIENCE: Can you go back?

SARA ELLISON: Yep.

AUDIENCE: So maybe I'm getting hung up on the difference between the actual variables and the realizations, but it seems like you're averaging means. It feels like you're missing-- are the X's and i's in this equation also sets of random variables or are they--

SARA ELLISON: Each Xi is a random variable. So we have-- so this summation-- I've left off all of the subscripts on the summations. But that summation is i equals 1 to n. So we have n random variables. And so we have a set of n random variables. And what we're doing is we're asking if you consider this function of random variables, as I've specified here, how is this function of random variables going to be distributed if the X's have mean mu and variance sigma squared?

AUDIENCE: I guess where I'm hung up is if those random variables are-- [INAUDIBLE] you're calling a random variables also a distribution, and if they're not the same distribution--

SARA ELLISON: Oh, I'm assuming that everything has-- they're independent and identically distributed, yes. No, I'm glad you clarified that. So yes, we don't need to necessarily assume that when we define what the sample mean is, but here, we're assuming it.

AUDIENCE: Here, mu will be the same across all random?

SARA ELLISON: Precisely. Precisely. Yeah, absolutely.

OK, so we've got this thing called a sample mean, and we've talked a little bit about how it's distributed. And we've even talked in kind of an informal way about how it might be useful. But there's a lot more to say about the sample mean. And in particular, one of the most important and useful results in all of probability, known as the central limit theorem, deals with the distribution of the sample mean. And it really serves as the basis for statistics.

So let X sub 1 through Xn of n form a random sample of size n from a distribution with finite mean and variance. So actually, let's pause for just a second. I have one more response to your earlier question that might be somewhat helpful. So remember, I made this comment-- note that we used independence in the variance calculation down here.

But we didn't use independence here. So you were asking about identically distributed, but, in fact, this is true-- this result still holds if the random variables are not independent. So I mean, that tells you we're sort of maintaining this assumption of independence in an identically distributed, but some of these results hold even if that's not true, and that's [INAUDIBLE].

AUDIENCE: They're [INAUDIBLE] instead of i.i.d.?

SARA ELLISON: Exactly. Yep. OK, sorry-- back to the central limit theorem. OK, so let the X1 to Xn form a random sample of size n from a distribution with finite mean and variance-- so, in other words, i.i.d. Random variables. Then, for any fixed number, this is true. So let's unpack this. What does this mean?

Well, in the middle here is the sample mean, this X bar. Remember, we just calculated what the mean of the sample mean is, and we just calculated what the variance of the sample mean is. And what have I done here? I've subtracted off the mean of the sample mean, mu, and I've divided by the square root of the variance of the sample mean.

So when I divided by sigma over square root of sigma squared over n, then sigma became n or sigma-- sorry, sigma squared became sigma, n became root n, and the root n went to the numerator instead of the denominator. But that's all I've done here. I've subtracted off the mean of the sample mean, and I've divided by the square root of its variance.

So what did Esther say last time about sort of what results from doing this to a random variable? It's called standardization. You subtract off the mean. You divide by the square root of its variance. And you have a random variable that has 0 mean and variance 1.

So what I've done-- I've simply taken the sample mean, and I've standardized it. So here is something like the sample mean but with mean 0 and variance 1. And then this probability statement here, this limit as n goes to infinity of the probability of this thing is equal to that, basically is just saying that as the sample size as n gets bigger and bigger, then the probability that the standardized version of the sample mean is going to have a normal distribution is-- well, the probability that it's less than some value little x is just going to the normal CDF evaluated at X.

So I didn't say that very eloquently, but let me put some other-- oh, and I should note, I think Esther used this notation as well. This is special notation for the CDF of a standard normal random variable. So let me restate what I just said a little bit more eloquently.

So basically, we take a standardized version of the sample mean from any old distribution, let the sample size go to infinity, and this is essentially the definition of the CDF of that random variable. By definition, that's what a CDF-- the probability that random variable is less than or equal to some value X. So, essentially, the definition of the CDF of that random variable is equal to the standard normal CDF. So this is what the central limit theorem tells us.

Practically speaking, what does this mean? That means if you have a sample mean from a reasonably large distribution or-- sorry, reasonably large random sample from any distribution-- basically that isn't just a crazy distribution-- it will have an approximate normal distribution and its mean-- we already calculated this-- its mean is mu, and its variance is sigma squared over n.

So the central limit theorem tells us this sort of very powerful piece of information, which is this. We calculated this a few minutes ago. We calculated this a few minutes ago. We know that the random that the sample mean has mean, mu, variance, sigma squared over n. Central limit theorem tells us that if our sample size is large enough, it's going to be approximately normal.

So first of all, I just want to take a step back and say how kind of remarkable this result is. I didn't say that the distribution that you were drawing from had to look like this or had to look like this. The distribution could, for instance, look like that. It could be a distribution that has a big point mass here and a little point mass here. And yet if our random sample from this distribution is large enough, the sample mean of that random sample will be approximately normal. I mean, I don't know, it seems like a remarkable result to me. It sort of almost seems like magic, but it's not.

So remarkable-- perhaps. It's also useful. We don't need to know the distribution we're sampling from to know a lot about the behavior of the sample mean. So that's why it's so useful. We can be sampling from some crazy, almost degenerate distribution. We don't need to make any assumptions. We don't need to know anything about it. But we know, as long as our sample is big enough, the sample mean is going to be approximately normal. And that's just going to be super useful when we go forward into estimation and inference.

And finally, I'll just point out, we're going to rely on this theorem at least implicitly for the rest of the semester. And it also gives us a notion of why the normal distribution is so important. The normal distribution is kind of really foundational in all of statistics, and the central limit theorem is a big reason why.

And also, to, I don't know, amplify the earlier question about the nature of the random sample-- I gave you the plain vanilla version of the central limit theorem. There are lots of different central limit theorems where we relax the assumption of an i.i.d. Random sample-- lots of different versions of the central limit theorem where we can be sampling from different distributions, and the random variables don't have to be independent, et cetera. And there's still some central limit theorems that exist in many of those situations.

So are these looks of satisfaction or bafflement or does this make sense? Yep? You understand why it's important, more or less? Willing to go on?

AUDIENCE: Yes.

SARA ELLISON: OK. Ah, for the first time this semester, the title on my slide does not start with probability. Yes, we've entered a new phase. So what is statistics? Well, it's the study of estimation and inference. We'll get to inference a little bit later. So we won't talk about that for a couple of weeks. And for now, we'll focus on estimation.

I should say, just as a personal note, I was a statistics major as an undergraduate, and I took a lot of probability courses, and I took a lot of statistics courses. And my view was that these courses didn't talk to each other.

And so one reason why I like this lecture so much is that-- at least I hope. This is what I'm intending to accomplish-- I hope that this lecture is providing a bridge between probability and statistics, and telling you why the foundational material that we've been working so hard on in probability is exactly what's going to tell us how statistics behave and what we're going to be able to infer from the statistics that we calculate.

So anyhow, that's my goal. This is something that, at least in my personal experience, always frustrated me that I felt like the probabilists and the statisticians never talked to each other or something. And this is the lecture in which they get to talk to each other in some sense.

So we've actually seen examples of estimators. We saw the sample mean just a minute ago. I didn't describe it as an estimator. It is an estimator. We had this sort of notion that we talked about, that it's going to give us some information about the underlying distribution and we'll make that more formal in a minute. Also, on your problem set, you saw an estimator.

And again, I didn't use the terminology of estimation because we hadn't introduced it yet. But the bufflehead-- the problem with the bufflehead population, what were you doing? You were estimating the size of a population. So we're going to engage in a more general conversation about estimation. now, but keep those two examples in the back of your mind, and maybe that'll add some concreteness to what we're talking about.

So what is an estimator? Well, an estimator is a function of the random variables in a random sample. The specific function is chosen with some goal in mind. It's chosen to have properties that are useful for giving us information about the distribution of those random variables.

So the basic idea here is we have a random sample. We may know a lot about the distribution it comes from. We may relatively little about the distribution it comes from. But what we're going to do is we're going to create functions of the random sample. And those functions are chosen specifically to reveal or to tell us information about the distribution that the random sample came from. That's the whole idea behind estimation.

So before we talk a little bit more about estimation, I want to define what a parameter is. We've talked about parameters. We've seen lots of examples. I don't know that we've seen an actual definition. So I think it would be useful to say that a parameter is a contact constant indexing a family of distributions. So mu and sigma squared are the parameters that index the normal family of distributions. So every normal distribution has-- all you need to know is the mu and the sigma squared, and you know everything about that distribution if you know what's in the normal family.

The exponential family is what's known as a one-parameter distribution or one-parameter family. There's only one parameter, which is lambda-- typically, we call it lambda-- that determines the which member of the exponential family a particular distribution is in. Uniform distribution-- a and b. Those are the parameters. n and p from the binomial distribution, for instance-- often, in the discussion of estimation, I'll often use theta as a general notation for a parameter. But if you want to add some concreteness in your mind, just think about lambda from an exponential or whatever.

So I think I said this before, but the idea behind estimation is we want to determine the values of parameters that govern an observed stochastic process or phenomenon. So we have an observed stochastic process or phenomenon. We can gather a random sample from that process or phenomenon. But we might not know the values for all the parameters.

So we're going to create these functions of random variables that have the goal of estimating the unknown parameters from the distribution. The general notation we're going to use for an estimator is theta hat.

So again, I've sort of talked about this, both in this lecture and in previous lectures. But I just want to make it a little more explicit. When we're talking about statistics, when we're going from the discussion of probability to statistics, we really need to keep two notions of the random variable in our head simultaneously.

We need to think of random variables as this mathematical construct that I introduced several weeks ago-- so a function from the sample space to the real numbers. And we also have to think of it as a stochastic object that can, quote, "take on" different realizations with different probabilities.

And so we use-- and I've been sort of consistent throughout using the notation capital X to stand for the random variable and little x to stand for the realization or possible realizations of the random variables, and I'll continue using that notation. Second of all, we've got to think of a random sample. We've got to keep these two notions of a random variable in our mind at the same time. And we also have to keep two notions of a random sample in our mind at the same time.

So we think of it as an i.i.d. Collection of random variables, but we can also think of the-- we also can call the realizations of those random variables a random sample. So a random sample can refer to the random variables. It can refer to the realizations of the random variables. And sometimes, we just call that data, instead of realizations of random variables, and that's fine, too. Yes?

AUDIENCE: How did you distinguish the stochastic object [INAUDIBLE] process [INAUDIBLE].

SARA ELLISON: I'm sorry, how did--

AUDIENCE: How did you distinguish a stochastic-- how did you define the stochastic object [INAUDIBLE]?

SARA ELLISON: Yeah, I didn't-- I'm not going to give you a formal definition of it. This is just meant as-- so the question is how do I define a stochastic object or a stochastic process or something--

AUDIENCE: Process [INAUDIBLE].

SARA ELLISON: Yes. So I'm not going to give you a formal definition. And the idea here is that I just want you to have these two notions of in your mind simultaneously that there's sort of a mathematical construct called a random variable, but then the random variable-- the reason why we have the mathematical construct is there are these sort of uncertain events that happen or stochastic events that happen in the real world, and we want to have some model of them. And so we think of it as a mathematical construct. We also think of it as a model of stochastic occurrences or something like that. Does that help? Somewhat? OK, yeah?

AUDIENCE: Just to summarize, so we have a random sample, and we create a function based on those, and that acts as an estimator, which will help us to find the parameters of the--

SARA ELLISON: You've got it. You've got it. That's exactly right. And we'll see examples as we go further that I think will clarify that even more. Oh, so, actually, I think you just predicted what I was going to put on the next slide or what I put on the next slide. So we know or assume that a set of random variables a random sample is distributed-- say, i.i.d. Normal or i.i.d. Uniform or i.i.d. Exponential-- and then estimation is trying to determine the specific mu and sigma squared or the specific a and b or the specific lambda.

AUDIENCE: It will always be an error, because the sample size is not infinite as-- there will be error in the estimation?

SARA ELLISON: Yep. And the way that we quantify the uncertainty of our estimate is through-- and we'll get to this more specifically-- the way we quantify it is through the variance of the distribution of our estimator.

So I think we've covered this. You might choose a function-- maybe I should actually go through this a little bit. I think we've essentially covered it. So when I said that an estimator was a specific function of the random sample that was going to be useful for us to try to give us information about an unknown parameter, for instance, and so maybe what we do is we could choose a function whose result when applied to a random sample is a random variable that has a very tight distribution around the mean of the distribution of those random variables.

And so if we have an estimator function of random variables that's very tightly distributed around the mean, and we don't know the mean, well, if we get a realization from that estimator, then that's going to give us important information about where the mean is. So--

AUDIENCE: When we don't know the mean, then how can we say that the estimator is tightly around the mean?

SARA ELLISON: No, we can't say the estimate-- we can say, analytically, mathematically, we know it has to be tightly distributed around the mean. And so what we do is we take a specific random sample, we plug it into that function that's the estimator, the estimator gives us a value, and we know that value is a realization from that distribution that's tightly distributed around the mean.

So let me-- maybe drawing a picture would help. So let's suppose we have a random sample from this distribution. This distribution has a mean, mu. We don't know what it is. We want to know what it is. So what do we do? We gather a random sample from this distribution. So it's just a bunch-- it's a bunch of realizations from this distribution. So they're both-- well, I should say they're both random variables that have this distribution, and the data is the sort of realizations associated with that.

So we gather that random sample. We plug that random sample into this particular function, X bar, which is equal to 1 over n times the sum of the X's. We know, analytically, that this thing here has a distribution that looks like this.

Maybe that was a-- maybe that's a little exaggerated. But basically, this thing here has a distribution. We know, if this sample size is big enough, this is going to be approximately normal-- I'll put an A here for approximate-- and it's going to have mean mu, and it's going to have variance sigma squared over n. Central limit theorem-- well, we knew this part and this part before the central limit theorem told us this part. So this is now how this distribution-- how the sample mean is distributed.

So if we get a realization from this distribution, then we're pretty sure that realization is somewhere close to this unknown mean because this distribution is tight around this unknown mean. And that's the fundamental idea.

So one thing I do want to point out on here is that just like we use the term random sample to refer both to the set of random variables and to the set of the realizations associated with them, we use theta hat to stand for both an estimator, which is a function of random variables, and the estimate coming out of that estimator, which is what you get when you plug in the realizations into the estimator. So again, statistics does not make a distinction between the estimate and the estimator, typically, in the notation.

An example, maybe, that will add a little concreteness-- so let's suppose that we have a random variable X, and it has a uniform zero theta distribution. So we don't know. It has a distribution that looks like this. We don't know what theta is. We want to figure out what theta is.

So what could we do? Let's suppose we have access to a random sample, the realizations, from this distribution. What could we do? Any ideas?

AUDIENCE: [INAUDIBLE]

SARA ELLISON: We want to do something with those-- we want to plug those realizations into some function that's going to give us some information about this parameter theta. Yes?

AUDIENCE: I think if you take the mean [INAUDIBLE].

SARA ELLISON: Yes, so what would you do to the mean, then? Sample mean.

AUDIENCE: I would estimate the sample mean is [INAUDIBLE].

SARA ELLISON: Well, OK, so I wanted you to say that you take the sample mean and then you multiply it by 2. So we were saying the same thing, but you can-- so one thing you can do is you can get a random sample. You calculate the sample mean. We know that the sample mean is going to have a distribution that's sort of centered around here and is going to be more concentrated around the mean of this distribution than the distribution itself. So you plug them into to the sample mean function and you multiply by 2, and that's going to give you an estimate for theta. Seem reasonable? Other ideas? Are you yawning or-- oh, sorry.

AUDIENCE: [INAUDIBLE] take the [INAUDIBLE].

SARA ELLISON: That's right. So how about if we have a random sample and, instead of computing the sample mean, we just take the n-th order statistic from that random sample and use that as an estimate? Seem reasonable, right? So we'll study these two examples. We'll figure out how both of them are distributed in the time to come, like this lecture next. But they both seem like reasonable ways to take a random sample from this distribution, plug that random sample into a function, and gain information about theta from-- choose the function that's going to allow us to gain information about theta.

So this is just what we said verbally. Two reasonable procedures come to mind-- gather a random sample, compute the sample mean, and multiply by 2, and use that as our estimator. Or gather a random sample, compute the max, the n-th order statistic of the random sample, and use that as our estimator.

And here they are n equations instead of words. So theta hat sub 1 is just equal to the n-th order statistic. Theta hat sub 2 is equal to 2 times the sample mean.

So let's think-- I have a couple of slides that help you picture what's going on. So here is the n-th order statistic from a random sample of size 5 from this distribution. Here's the n-th order statistic from a random sample of size. I don't know what that is-- 12 or something like that. And here's the n-th order statistic from a random sample of size a lot. I don't know how many little marks I put on there.

So as n is getting bigger, as our sample size is getting bigger, our estimator seems to be getting better. That's not just-- I mean, obviously, I could have drawn many different pictures here. That's not just a mistake. In fact, the estimator does get better. It becomes a more accurate estimate of theta as n gets bigger.

Here's 2 times the sample mean. So the sample mean was, I don't know, somewhere around in here, and I multiplied it by 2. I just kind of did this visually but, I think this is probably close. So here, the sample mean is out here somewhere. This particular random sample that I dreamed up-- it, by chance, had more observations in the upper half. So 2 times the sample mean would be somewhere out there. And then this random sample-- seems like the sample mean was pretty close to mu. And so 2 times the sample mean would be just about theta.

So again, in these examples, the estimator is getting better as n is getting larger, but it does sort of bounce around a little bit. And again, that's kind of not-- that's not a mistake. That's actually a feature of these two estimators that we can prove. Yeah?

AUDIENCE: I'm just wondering if we have a way of evaluating the estimators?

SARA ELLISON: Yes. Excellent question. So the question is do we have a way of evaluating these estimators, deciding how accurate they are and which one to choose? And we will get to that. So yes, those are important questions.

Here's another procedure. Gather a random sample, compute the sample median instead of the sample mean, and then the number-- the median, I'll remind you, is the number above and below which half of the sample falls. And then multiply that by 2 and use that as theta hat. Seem reasonable? Yeah. So I can tell you that that also is a reasonable estimator and it has sort of particular properties that we can talk about later.

Here's another procedure. Gather a random sample, throw the whole thing away and have R generate a random value for you and use that as theta hat. What do you think? No. So we can guess that this procedure does not have very good properties. And, in fact, we can prove that it doesn't have good properties. I think we probably won't bother to, but we can, in fact, prove it doesn't have very good properties.

So these are some of the questions you just asked and one other. So first of all, how did we come up with these functions? We just kind of dreamed them up off the top of our head. Is there a more systematic way that we can generate these estimators? And the answer is yes. We'll see that next time. How do we know if they're reasonable? How do we choose among them? So we have two estimators here-- 2 times the sample mean and the n-th order statistic, and they both seem like reasonable estimators. How do we choose among them or between them, in this case?

So for the rest of this lecture and some of the next, we're going to talk about these topics-- criteria for assessing estimators and the frameworks for choosing estimators. And these two topics are going to answer those questions that were just posed.

So how can we even think about criteria for assessing estimators? Well, remember, an estimator is a random variable. So everything that you need to know about a random variable is embodied in its distribution. I mean, that's essentially what a random variable is. It is its distribution. So our criteria for estimating or for assessing these estimators will be based on characteristics of their distributions. So what characteristics of the distributions matter? What do we care about?

Well, first one-- I'm going to define a criterion called biasness or define a sort of a characteristic called unbiased. An estimator is unbiased for theta, which is the sort of parameter that we're interested in estimating, if the expectation of theta hat is equal to theta for all theta in-- this is a capital theta, by the way-- so in-- this just means for all possible values of theta.

AUDIENCE: [INAUDIBLE] estimator of [INAUDIBLE].

SARA ELLISON: So theta hat is our estimator, and it can serve two-- that notation serves two roles. It both stands for the function of the random sample, and it stands for a particular number, the estimate, once we plug in the realizations into the function of the random variables. So theta is two things. In this case, it's serving the role as the estimator. So the expectation-- and remember, the estimator is a random variable. So it has a distribution. So here, we're saying if the distribution of that estimator, theta hat, is equal to the parameter we're trying to estimate, theta-- and if that's true for all possible values of theta, then we're going to call this estimator unbiased.

So here's a picture. Yep?

AUDIENCE: So when you take the [INAUDIBLE], that means [INAUDIBLE].

SARA ELLISON: Theta.

AUDIENCE: Theta is [INAUDIBLE].

SARA ELLISON: No, it's a parameter. It's a parameter. So, for instance-- so here's an example. The random variable-- we have sort of a random sample that we're drawing from this distribution, and we can call those X's or Y's or whatever we want to call them. And then this has-- distribute each X has a distribution uniform 0 to theta. Theta is just-- it's a number. We don't know what it is, but it's a number. It's just a number governing this distribution.

AUDIENCE: [INAUDIBLE] symbol [INAUDIBLE]?

SARA ELLISON: That's a capital theta. So it's just the space of all possible values of little theta. So for instance, in this case, capital theta has to be the sort of positive, real numbers. So for instance, you can't have a uniform 0, negative 4 distribution. And so the capital theta is just the set of all possible values of theta.

So here on the left is a picture of an unbiased estimator. So what is this curve here? That's the distribution of the estimator. And that distribution has a mean or an expectation. You compute the expectation. If that expectation is equal to this parameter theta, then it's unbiased. And if this distribution is not equal-- or sorry, if the expectation of this distribution is not equal to the parameter theta, then it's a biased estimator for theta. Yes?

AUDIENCE: So this normal that you've drawn, that's the normal that you get when you take sample-- when you take a lot of different-- what's the word-- realizations from this uniform distribution, and then you get this normal--

SARA ELLISON: If your sample size is big enough, it's going to be approximately normal. Yeah, that's right. That's right.

AUDIENCE: Can we say it's biased when theta-- when the mean of that random draw is exactly equal--

SARA ELLISON: Unbiased if the X-- so this is the PDF of theta hat, the distribution of theta hat. If that distribution has an expectation that's equal to theta, then we call that estimator unbiased. Yes?

AUDIENCE: The parameter theta that we are talking about here represents the mean of the actual distribution?

SARA ELLISON: Well, it might be, and it might not be. So here, it's not the mean.

AUDIENCE: But then how would I say expected value of theta hat is theta in this case because expected value of theta hat will be the mean?

SARA ELLISON: No. So I'm glad you asked this question, so let me try to clarify. So here, theta is a parameter governing the distribution that we're drawing the random sample from. It might be the mean of that distribution. Here, it's not the mean of the distribution. Here, it's the upper limit of the distribution. It could be any other parameter governing the distribution.

What we do is we come up with an estimator that we think-- we maybe have good reason to believe is going to be sort of a good estimator for that unknown parameter. And so in the case of this uniform zero theta, we came up with 2 times the sample mean or we came up with the n-th order statistic. And both of those were functions that we thought might do a pretty good job of telling us something about this unknown theta. And, in both cases, those functions have distributions.

So here, I've drawn the PDF of theta hat as kind of normal because that's often what it is. It doesn't have to be normal. If our sample size is small, it may not be normal. But here, I've sort of drawn it normal just because I had to pick some shape to draw. But the theta hat, which is the function that we're using to estimate this unknown parameter, has a distribution. If that distribution-- if the expectation that distribution is equal to theta, the thing that we're trying to estimate, then it's an unbiased estimator. Does that help? Yes?

AUDIENCE: I'm a little bit confused because it seems a little counterintuitive to me. So the PDF of the estimator-- in order for you to generate it, you need to have a lot of samples and multiple estimators, right?

SARA ELLISON: No. If you have a very large sample-- so if you have a very-- I think, actually, you might be confusing the two different senses of an estimator, or you might be confusing an estimate with an estimator. So an estimator is a function of random variables. It's a sort of mathematical construct that we can talk about the distribution of. In the real world, what we typically have is we have one random sample, which is multiple draws from some distribution, and that random sample that we have is sort of a series of realizations from that distribution, a series of numbers. So what we do is we choose a function that has properties that are going to be useful for us, and then we plug those realizations into the function. And basically, an estimate is one realization from this distribution. It's just one realization. And that doesn't mean it's going to equal theta. I mean, we could have a realization here, a realization here. If our estimate is from the distribution that's centered around theta and is pretty concentrated around theta, then we have some confidence that our estimate is close to theta.

AUDIENCE: So theta is a true value for [INAUDIBLE].

SARA ELLISON: True value for--

AUDIENCE: The estimator.

SARA ELLISON: No, it's a parameter from the underlying distribution.

AUDIENCE: So it's the actual value that we're trying to estimate?

SARA ELLISON: Yes, exactly.

AUDIENCE: So if I'm trying to estimate the actual value, that means I don't know it. In order for me to prove that it's unbiased, I actually have to know it.

SARA ELLISON: No. No, you don't. So mathematically, we can do the proof that an estimator is unbiased without-- we just know that, under the assumptions that we lay out, we can prove that an estimator is going to be unbiased. And then, when we have a particular random sample, obviously, we don't know where in this distribution we're drawing-- the random sample that we use to-- we plug that into the estimator, and we get a particular value from that distribution. We don't know where in the distribution that value is coming from.

So maybe this helps. I hope this helps. We actually have proved unbiasedness just a few minutes ago. So remember the sample mean? We said here is a function of random variables. We're going to assume that they're i.i.d. Random variables. We're going to add up a bunch of i.i.d. Random variables and divide by n.

Let's see what the expectation of that new random variable is, that X bar-- the new random variable we're going to call X bar. We computed-- we mathematically showed that its expectation was equal to mu. In a practical situation, we don't know what mu is. But we can show mathematically that that function of a random variable has a distribution and its expectation is equal to mu. So that's how to prove unbiasness. Does that help? Yes?

AUDIENCE: I think the definition [INAUDIBLE], it seems as if it would be unbiased, yet we could still not always-- so it's easy to satisfy the definition, but depending on the sample, you wouldn't always find the actual theta, so what [INAUDIBLE] bias is actually representing?

SARA ELLISON: So let me see if I can answer this question without muddying-- it's a good question. I'm going to just try to answer the question without sort of muddying the waters more. So basically, unbiasness just tells us-- let's consider this thought experiment. So let's suppose that we gather a random sample from, say, this distribution. And the random sample is actual numbers, realizations. We compute the sample mean, and then we multiply it by 2, and that gives us some number.

And then let's suppose we did that 100 times-- a new random sample. We compute the sample mean. We multiply by 2. That gives us another number, a brand-new random sample. We do this 1,000 times or a million times or whatever. The size of our random sample is going to determine how this thing that we're computing, this 2 times the sample mean, how it's distributed. So let's say our random sample was size 15 each time, and we did this 100 times or 1,000 times or whatever. So what we're going to do is basically, then, we'll have something that looks like-- I'll erase this. 2 over n.

So let's say I said 15, but let's say n is equal to 20. Then we can mathematically figure out what the distribution of this 2 times the sort of sample mean from a random sample of size 20 is. And maybe we're going to get something that looks a little bit like this. I'm not sure. And so, basically, if we think about the thought experiment where we do this 1,000 times and then create a histogram of the values of the estimate-- those 1,000 different values of the estimate-- the histogram is going to look something like that. Does that help? OK, great.

So it's a lot to keep straight. But these are sort of the fundamental ideas underlying estimation. So I'm glad that you guys are asking questions and trying to get these concepts straight in your minds.

AUDIENCE: So I just want to [INAUDIBLE] real quick. So n is 20, so we're taking 20 samples from this distribution. Then you've got this kind of for loop, where you're doing that 1,000 times.

SARA ELLISON: Yes. And then you're creating a histogram of the 1,000 estimates that you get from each of these random samples.

AUDIENCE: So that second number, 1,000-- clearly you just came up with that on the spot, that's not a number that we're considering here?

SARA ELLISON: No.

AUDIENCE: n is the number.

SARA ELLISON: Exactly. So I only said 1,000 because I wanted you to picture in your mind that this histogram is going to look a lot like this distribution. Oh, so we're done. Time is up.