[SQUEAKING]

[RUSTLING]

[CLICKING]

SARA ELLISON: So where were we last time? We had started, we had talked about, started a general conversation about estimation. And we had seen a couple of different examples of estimators that seemed to have of just-- like we just thought of them off the top of our head. And both estimators, remember we were estimating theta in a uniform 0 theta distribution. And both estimators seemed quite reasonable.

And so then, the next obvious question is, well, how do we know which one to use? How do we decide between estimators if we have more than one estimator available? And last time, we saw the first criterion that one might want to use to choose among estimators. And that was unbiased, unbiased ones. So an estimator is unbiased for theta if the expectation of the estimator is equal to the parameter we're trying to estimate for all values of the parameter, all possible values of the parameter.

And then I drew a picture. And I said that here is-- so remember, an estimator is a function of random variables. So it itself is a random variable. So it has a distribution. So we can talk about characteristics of that distribution.

Here is a particular distribution of an estimator theta hat. And this one, I've drawn so that it's unbiased. Here's a particular distribution of a different theta hat, say. And I've drawn it so it's biased. So the expectation for that. The second estimator over there is not equal to theta, so it's a biased estimator.

So let's see an example. So we've seen this particular estimator before. Remember, this is uniform. The random sample is uniform 0 theta. And we came up with the idea that we could just compute the sample mean from that random sample, multiply it by 2. And that would be a reasonable estimator for theta hat, or for theta, sorry.

So let's see if this one is, in fact, biased or unbiased. Well, we can just compute the expectation by using the properties of expectation that we saw a few lectures ago. So we don't have to go back to the definition of expectation. We just say, well, the expectation of theta hat is equal to-- by properties of expectation, we can pull out the 2 and the 1/n from the expectation and from the summation. And then we plug in the expectation of x sub i, which is just equal to theta over 2. Everyone knows where that came from.

And then we just do this summation. And we get the 2's cancel. And we get that this is equal to theta. So unbiased estimator, so that's a good thing.

How about our other estimators? So the other option that we had talked about was just using the nth order statistic. And that also seemed reasonable. So let's see what we can say about whether this one is biased or unbiased. Any guesses?

STUDENT: It becomes more unbiased as n increases.

SARA ELLISON: So you have the right intuition, but you haven't expressed it in the right way. So basically, what's happening is that it will be biased. We're going to see that. And the amount of the bias is going to get smaller as n gets larger. And do you want to do you want to share your intuition for why it's biased?

STUDENT: We talked about it last time. We could take more draws than the probability that the highest draw is closest to theta increases.

SARA ELLISON: That's right, that's right. So this estimator is sort of getting close. But it's but it's always going to be biased for a finite sample because the nth order statistic is always less than or equal to theta.

It's never greater than theta. It's always less than or equal to theta. And it's equal to theta with 0 probability. So that's sort of the way that you know this is going to be a biased estimator, the intuition I guess, that I use.

So let's verify this. We can't just use the properties of expectation like we did on the last slide to calculate the expectation of this estimator. And so we're going to do it directly, using the definition of expectation. And in order to do that, we first need the PDF of the nth order statistic.

Well, we've seen that before. So we'll just go back to previous lecture and recall that this is the general-- that's the general formula for the PDF of the nth order statistic from some distribution, F. Now, we have a specific distribution here, uniform 0 theta.

So we'll plug in that specific PDF, plug in the uniform PDF here and the uniform CDF here. And then we get this. And then it just simplifies to this.

So now we have the PDF of the nth order statistic. We'll be able to calculate its expectation with that. And so what we did is I just took the PDF from the previous slide and plugged it into the definition of expectation. Remember, the expectation of a random variable is just equal to-- that's continuous. It's just equal to the integral of x times the PDF. So that's exactly what I did there.

I solved the integral. And just a couple of steps of math later, I get that it's equal to n over n plus 1 times theta. So not only is our intuition that it's biased is confirmed here, but also our intuition that the bias is going down as the sample size is getting bigger is also confirmed. This makes sense?

So having an unbiased estimator is great. Is that the only thing we care about? I think you guessed that the answer is probably no. It's not the only thing we care about.

Oh, so yeah, I guess this is what I already said. Not surprising-- the estimator will always be less than or equal to theta. Oh, sorry. Before we get to the second criterion, I just want to give you a couple of useful results about unbiased estimators.

So the first result is that if we have a sample mean for an iid sample, then that sample mean is an unbiased estimator for the population mean, for the mean of the underlying population. So that's always going to be true. For an iid sample, the sample mean is unbiased for the population mean.

Now, we actually already knew this. We didn't use the word unbiased. But remember when we introduced sample mean, we calculated the expectation of the sample mean. You guys remember that calculation? We calculated the expectation of the sample mean, and it was equal to mu, the underlying mean or the population mean. So in fact, we already knew that.

Another result that we haven't seen before because I haven't even defined what a sample variance is, but it's a useful result nonetheless, is that if we have an iid sample and we compute this estimator here, it's called the sample variance. This is going to be unbiased for the population variance.

So I am leaving out the proof of that. It's kind of a little bit of a fussy proof. And I don't think it's very instructive. But if anyone's interested, happy to show you. Yep?

STUDENT: Before we move on, I wanted to ask a question about this, the sample mean for the examples of unbiased population. Does the fact that it's unbiased mean that you're getting the actual population mean when you do the sample mean, or no? I don't think that's what's happening.

SARA ELLISON: No. So what it means is that the sample mean has a distribution. And that distribution, the expectation of that distribution is the population mean. So here, there's some distribution. Maybe it's uniform, maybe it's normal, whatever. There's some distribution that our random sample is governed by, a random sample is drawn by or drawn from.

And then, we get a random sample. We compute the sample mean. And the distribution of the sample mean-- so if we think of the sample mean as an estimator, it's going to have expectation, the sample mean.

Any particular realization of that estimator-- so like if we have the numbers, the realizations for the random sample and we plug them in to the function for the estimator, then basically, what we're doing is we're getting some realization from this distribution. So depending on how tightly distributed it is around the population mean, it could be very close to this with high probability.

Questions about this? So these are just two useful facts that we'll refer to in the future. So now, is unbiasness all we care about in an estimator? The answer is no.

So in particular, there are other aspects of distributions of the estimators that we might care about. We might care about-- I just alluded to this-- how tightly distributed it is around the unknown parameter. We could have an estimator that has a very diffuse distribution. And that estimator, even if its expectation is the parameter that we're trying to estimate, even if that's its expectation, in other words, even if it's unbiased, it might not be very useful to us if it has lots of probability far away from the real parameter.

And that's what efficiency is about. So given two unbiased estimators, theta-1 hat and theta-2 hat, theta-1 hat is more efficient than theta-2 hat if for a given sample size, the variance of theta-1 hat is less than the variance of theta-2 hat. So let me show you a picture.

So here, we have two unbiased estimators. We have an efficient one, whose distribution is pretty tightly distributed or concentrated around the unknown parameter. And we have an inefficient one, whose distribution is much more spread out. And we're going to prefer the efficient estimator.

Is that clear to everyone, why we prefer-- so note that this definition has defined efficiency here just for unbiased estimators. The notion of efficiency is broader than that. So it can exist for, sort of broader classes of estimators as well. I'm not going to give you a formal definition. But just think that if an estimator has a larger variance, it's less efficient than an estimator with a smaller variance, and we typically define that in different classes of estimators.

So fine. Suppose we have an efficient estimator and/or an estimator that has a small variance. And then we have another estimator it has a larger variance, but it's unbiased. So sorry, the first one is biased, has a small variance. The second one is unbiased and has a large variance.

How do we know which one to choose? Well, there's no right answer. I can't of tell you the theorem that says, well, you always need to choose this estimator over the other one. But there might be sort of other-- aside from just kind of picking your favorite estimator, there might be kind of a more routinized way to trade off bias and efficiency in estimation. And one way to do this is minimum or is mean squared error.

So I'm going to define mean squared error as the expectation of theta hat minus theta squared. And that can be rewritten as the variance of theta hat. So we can think of that basically, as the efficiency, a measure of the efficiency of the estimator plus the expectation of theta hat minus theta quantity squared. And that is in fact, the bias. That is, sort of this difference here is defined as the bias, so that last term is the bias squared.

So for unbiased estimators, that quantity is going to be 0. For biased estimators, it's going to be positive. So mean squared error is a way to trade off bias and efficiency. You can choose an estimator that is the minimum mean squared error estimator. And then, that sort of allows, basically allows this explicit trade off between bias and efficiency.

So I should point out, it is a reasonable way to trade off bias and efficiency. It's not the only one. And like I said, there's no right answer when it comes to choosing among estimators that are not dominated in all ways.

And I'll finally mention one additional criterion. Theta hat is a consistent estimator for theta if the following is true-- you take the absolute value, The absolute difference between theta and theta hat. And the probability of that absolute difference being less than some small number delta, goes to 1 as n goes to infinity.

So practically speaking, what does this mean? Well, first of all, let me draw you a picture, So this is what a consistent estimator looks like as n goes to infinity.

So basically, you have a consistent estimator. I've drawn this one to be unbiased for all n. It doesn't have to be unbiased. But that's just the way I drew the picture here.

So basically, you have an estimator, and it has a distribution. And as n goes to infinity, the distribution is getting more and more concentrated around the true parameter. And in fact, in the limit, it collapses to a single point at the true parameter as n goes to infinity. So this is what's known as a consistent estimator.

So let me just emphasize, I've drawn this picture for an estimator that's unbiased. A consistent estimator doesn't have to be unbiased. I could have drawn all of these distributions having bias, but the bias is going to 0.

So in the interest of time, I haven't included these examples in the lecture notes. But I will tell you, in case you're curious, that the two estimators that we looked at for the uniform 0 theta random sample, the unbiased estimator, the one that was 2 times the sample mean, was much less efficient than the biased estimator, the one that the nth order statistic. And so in that case, you might want to choose the nth order statistic, just because it's for reasonably-sized random sample. It's just its distribution is going to be much more tightly distributed near theta than 2 times the random sample or two times the sample mean.

So the criteria that I've gone through, they are probably the most important reasons for choosing one estimator over another. But they don't necessarily have to be the only considerations. So sometimes estimators can be really difficult to compute. They can be used, sort of hours of computer time or something like that. And so you might choose an estimator, just based on computational simplicity. And that's not an unreasonable thing to do.

You might also choose an estimator based on how robust it is. What do I mean by robust? Well basically, an estimator is robust if it still will do a decent job, perform well for estimating the unknown parameter, even if some of your assumptions, underlying assumptions are incorrect. So for instance, let's suppose that we want to estimate mu. And we think that the random sample is drawn from a distribution that looks like this. But in fact, it's drawn from a distribution that looks like this.

So this is still the mean. Mu is still the mean of this distribution. But we've assumed that we're drawing from this distribution, when in fact, we're drawing from that one. A robust estimator will still do a good job of estimating mu, even if our assumptions about the underlying distribution are wrong. And so sometimes we choose estimators that are robust to different assumptions if we don't have a good-- if we don't have a lot of confidence about our underlying assumptions.

And yeah, so I guess this is an example of a robust estimator that we talked about last time briefly. It turns out that the 2 times the sample median estimator from the uniform example will have less bias than the 2 times the sample mean estimator if we've misspecified the tail probabilities in that uniform example. So that's an example of an estimator that might actually not be as good as the other two estimators we saw in terms of bias and efficiency, but might have this other property of robustness that could come in handy at some point. Yes?

STUDENT: Wouldn't scale probabilities be underestimated or overestimated for that to hold? Or does it not matter?

SARA ELLISON: So robustness is basically, it's just a kind of a property of an estimator that says it still performs well if your assumptions are incorrect. And it could be that the estimator is only robust to misspecifications in one direction and not the other direction. That's possible. But just the sort of general definition of robust is that it's not sensitive to--

STUDENT: The second part of this slide is it's showing that that's just an example of what could happen. It's not necessarily always true.

SARA ELLISON: Exactly, that's right. Yes?

STUDENT: In that example, what would be the wrong distribution [INAUDIBLE]?

SARA ELLISON: So let's suppose-- so we came up with this 2 times the sample median estimator in the uniform 0 to theta example. So let's suppose instead of uniform 0 to theta, let's suppose the distribution looked like that instead. And we carried through all of the analysis of the estimator as if this were the distribution.

Then if we use the 2 times the sample median estimator, we're probably going to do a better job estimating. I would have to work it out to be sure. But in that example, we're almost certainly going to do a better job of estimating theta with the 2 times the median estimator than the 2 times the sample mean estimator.

So now, we know, at least roughly speaking, how to figure out if an estimator is good once we have one. So we have lots of different criteria to consider and some good guidelines. But how do we get an estimator in the first place? So the example that I went over in class a couple lectures ago, we just kind of dreamed them up off the top of our head, which it turns out, can sometimes result in quite reasonable estimators. But it might be a little comforting to know that you don't always just have to be able to dream things up off the top of your head.

So there are two main frameworks for deriving estimators. One is called the Method of Moments. And the second is called Maximum Likelihood Estimation. And in fact, we've seen examples of both.

So 2 times the sample mean in the uniform example is a method of moments estimator. And the nth order statistic in the uniform 0 theta distribution is a Maximum Likelihood estimator. And then as I said before, a third framework you can think of is to just come up with something clever off the top of your head. And that sometimes works.

So let me just give you, a sort of sketch out what the method of moments framework for deriving estimators is, give you an example of one, and then do the same for maximum likelihood. So the Method of Moments was developed in 1894 by Karl Pearson, that many, many people consider to be the father of mathematical statistics. There's a picture of him.

And first, to tell you what Method of Moments estimation is, I've got to define a couple of things. So I've got to define population moments. And I've got to define sample moments.

So population moment, we haven't used the terminology population moment. But we've already seen the first population moment about the origin, which is just the expectation. The second population moment about the origin is the expectation of x squared, third population moment, expectation of x cubed.

So for any distribution, we can just know we can calculate these as functions of the parameters. Or we can look them up on Wikipedia. Or we can look them up in the back of our statistics book or whatever. But we can get the population moments as functions of the parameters of the distributions. So as a function of parameters of the distribution, what's the first population moment of a normal distribution?

STUDENT: Mu.

SARA ELLISON: Mu, how about of a Poisson? Does anyone remember that?

STUDENT: Lambda.

SARA ELLISON: Lambda. How about of a uniform 0 to theta?

STUDENT: Theta over 2

SARA ELLISON: Theta over 2, right. So you guys get the idea. The sample moments, we've seen obviously, the first sample moment. It's just the sample mean. The second sample moment is the same thing. But we're going to square all of the observations instead and so forth. You have a question?

STUDENT: Going back to the sample variance that you showed us, that was divided by n minus 1. So is there any inconsistency? Or how do you reconcile the fact that this sample moment is odd?

SARA ELLISON: So the sample moment that I showed you was not a sample moment. Well, first of all, it was not about the origin. It was about the mean. So I'm defining sample moments here as-- I left it out. But I should have said sample moments about the origin.

So there are two different types of moments, sample moments for a distribution. One is about the mean, and one is about the origin. These ones are about the origin.

And the other thing is that to be honest, I'm not sure the sample variance may not-- I believe that actually, the definition of the second sample moment about the mean is 1/n. It's not 1 over n minus 1. So the sample variance is defined differently than the second sample moment about the mean. And the reason why is that the second sample moment about the mean is actually a biased estimator of the variance. Yeah, thanks for asking that question, but yeah, there's no inconsistency.

So we've got these population moments. You can just think of these as like functions of the parameters. And then, we've got the sample moments. And we're going to use those together to derive estimators.

So to estimate a parameter, we equate the first population moment, which is, as I said, a function of the parameter, to the first sample moment. And then we just solve for the parameter. So let's do a quick example.

So this is an example, we've already seen. The first population moment, expectation of x of a uniform 0 to theta, is just theta over 2. And the first sample moment is just 1 over n times the sum of the x's. So we equate those two things. We have to now stick a hat on the theta, because once we equate the sample moment and the population moment, now we're solving for an estimator, not the parameter. So we stick a hat on it, and then we solve for theta hat.

So that's exactly the estimator we saw before. And this is just exhibiting that it is, in fact, a method of moments estimator. What if you have more than one parameter estimate? So for instance, what if you have a normal distribution, and you want to estimate of both the mean and the variance from the normal distribution? Or you have a binomial distribution, you want n and p or something like that.

Well, no problem. So the Method of Moments can accommodate that. You just use as many sample and population moments as necessary. So each pair of a sample moment and a population moment, each one of those is called a moment condition. And then, if you have k parameters to estimate, you just have k moment conditions.

And then that basically means you'll have k equations in k unknowns, the k unknowns being all of the parameters with hats on them. You're solving for estimators for all those parameters. And so then, you just solve those k moment conditions for the k estimators.

So the second framework-- the idea for Maximum Likelihood Estimation has been around for a very long time. And if you try to figure out what its origin is, it's really a little unclear. I think that a lot of historians of statistics would attribute the idea to Lagrange in the 1770s. But the basic idea probably predated that. And sort of the analytics like proving that Maximum Likelihood Estimation was a reasonable thing to do, that sort of came much later around 1930 with RA Fisher.

And there's those guys. You can probably figure out which one is from 1770s and which one is from the 1930s. So what is a Maximum Likelihood Estimator. Well, it's the value theta hat, which most likely would have generated the observed sample. That's why it's called maximum likelihood.

So what exactly does this mean? How do we find it? So let's think of it this way. So this is not-- this is just a motivation. This is not exactly. This is not, in fact, how you find a maximum likelihood estimate, but this is going to serve as motivation.

So let's suppose we have a random sample. And we know that our random sample is from some particular distribution. But we don't know which member or some particular family of distributions. But we don't know which member it is. So in other words, we know it has a beta distribution. But we don't know what the parameter of the beta distribution is, or we know it's a normal distribution, but we don't know what mu and sigma squared are.

So you can think of being able to create a histogram with our random sample. And this is sort of the empirical counterpart of the PDF. On the other side, you can think of taking this family of distributions that we're assuming or that we know that our random sample has come from, and varying the parameter or parameters in that distribution, that family of distributions. And we get all kinds of different possibilities here. And presumably, we get a continuum of possibilities.

I've just drawn-- what is this? I've drawn five different possibilities here. But we have sort of an infinite number of possibilities as we vary the parameter or parameters of these distributions. So where did we get these? Well, as I said, we're assuming a particular family and then just varying the parameters.

So how do we choose which one of those? Well, just intuitively or graphically, we want to choose the one that sort of fits our histogram the best, so that's what we do. We choose the member of the family of distributions that's most likely to have produced our data. So if we chose a member of the family whose shape was much different than our histogram, then it's unlikely that that parameter would have produced the data that we saw. The one that fits our histogram most closely, that's the one that's most likely to have given rise to our data.

So theta hat, our maximum likelihood estimator, is the value of the parameter associated with this particular best fit member of the family of distributions. So is that, sort of concept clear, even if you don't know how to find that yet? So conceptually, it makes sense, or at least I hope it does. Operationally, how do we go about this? How do we find one of a bunch of PDFs that's most likely to have produced our data?

Well, the key is that we have to think about the joint PDF of our data in a somewhat different way. So we know how to-- if we have a random sample, we know how to get the joint PDF of that random sample. We're just it's just the product of the individual PDFs of all of the random variables in our random sample.

Now we think of this joint PDF as a function of the parameters now. We'd always sort of taken the parameters as given. They were just constants in our function that we sort of took as given. Now we think of that joint PDF as a function of those parameters. And we're going to maximize that joint PDF over the possible values of the parameters.

Yes, of course. And we'll see a couple of examples as well. But I'm happy to go over this, because it's a little subtle. It's a little confusing.

So basically, what we do-- actually, let me go to the next slide. So in other words, what we do is we take the joint PDF of our random sample. Instead of calling it a joint PDF, we're now going to call it a likelihood function. That's just suggestive to us that we have to think of it in a different way. It's exactly the same thing. The likelihood function is the joint PDF of a random sample.

How do we get this likelihood function? Well, we just take the product of the individual PDFs of all of the guys in our random sample. Do you guys remember that? What's a random sample? It's independent identically distributed random variables. To get the joint distribution of independent identically distributed random variables, you just take their distribution, and you multiply them together because they're independent.

So this is the joint PDF of our random sample. We're going to relabel that the likelihood function. And we're going to think of the likelihood function as being a function of theta, given the data, as opposed to the other way around. And then we're going to maximize that likelihood function with respect to these parameters theta.

So like I said, I have a couple of examples. And this actually just emphasizes what we said before. So we've got this likelihood function. We just maximize it over all possible values of theta.

And one thing to keep in mind, for this class, you don't really have to worry about this. But if you go out of this class and you encounter a maximum likelihood in the larger world, then you're going to have to realize that a lot of times what we do, instead of maximizing the likelihood function, is we maximize the log likelihood function.

So that's a perfectly fine thing to do. Whatever maximizes the likelihood function is going to maximize any monotonic transformation of the likelihood function. So it's fine to take a log. The reason we do it is computationally, it just is much easier, typically, to maximize the log likelihood function So you go out into the broader world, you will encounter log likelihood functions, not likelihood functions, and that's why.

So it's good practice. I'm not necessarily saying you guys have to do it. But it's good practice if you're sort of interested in these kinds of techniques, to write down some joint PDFs, maybe take the logs to make computation easier. Take the derivative with respect to theta. Set the derivatives equal to 0. Solve for the maximum likelihood estimators.

You certainly can do that. We're not going to do it here. What I'm going to do instead, is I'm going to do a couple of examples that are both unusual in that they can't be solved this way because their maxima are not at a point where the likelihood function is differentiable. But then also, we don't have to get bogged down with all the computation. So I think they're kind of easier examples to show you, I think.

STUDENT: So do we guess what we think the PDFs are going to be, and then we--

SARA ELLISON: Good question. So the question was, do we guess what the PDF is? So sometimes we may know. Why do we know?

I don't know. Maybe we're engineers, and we've sort of worked in a particular field for our entire careers. And we know that the random draws from this particular-- from some random event that happens with this particular machine always has an exponential distribution. And we just don't know what the parameter is. But we just know from vast years of experience, it's going to be exponential.

Or maybe we have a theoretical reason to believe that there is a particular distribution that's giving rise to the random sample. Or sometimes we just guess. So it just depends a lot on the problem, whether how confident we are about the distribution, the underlying distribution.

So I said, we'll do a couple of examples. But I'm going to do ones that don't involve the standard, sort of taking the log of the likelihood function, taking derivative with respect to theta, setting the derivatives equal to 0, et cetera. I'm going to do some kind of non-standard examples instead.

So back to this sort of uniform 0 theta example-- remember, what the maximum likelihood estimate from what the-- remember, I told you. I didn't show you, obviously. But remember what the maximum likelihood estimate for this is? The nth order statistic.

STUDENT: Over n plus 1?

SARA ELLISON: That's the expectation of that estimator. So the maximum likelihood estimate, I just said it verbally, it wasn't on a slide. But I said that in fact, it's the nth order statistic. So I think I did.

So let's figure that out. Let's verify that, in fact, that is the maximum likelihood estimator. So what do we do? We first have to write down the likelihood function.

How do we get the likelihood function? It's just the joint PDF of our random sample. So we have each of the individual members of the random sample has this distribution, 1 over theta for x in 0 to theta and 0 otherwise. Everyone agrees to that.

So let's think back to the definition of the maximum likelihood estimator, it being the parameter that's most likely to have given rise to the sample that we observed. Why? So I say here, I claim here that we obviously would not pick any theta hat less than the nth order statistic. Why is that, given the definition of the maximum likelihood estimator?

STUDENT: Well I guess, theta is supposed to be the maximum. So if there's a greater value that shows up in the random sample.

SARA ELLISON: It cannot have happened. It's a probability 0 event. So let me just draw a quick picture to illustrate.

So here's our distribution. Theta is unknown. If we were to choose-- so we have sort of a random sample chosen from this. The entire random sample has to be less than theta, just has to be. I mean, that's how the problem is set up.

So let's suppose we have-- actually, maybe what I'll do is erase this part. And we have an nth order statistic. We would never choose a guess for theta below the nth order statistic, because that would suggest a situation like this that we know can't happen Does everyone understand that? So theta always has to be at least as big as the nth order statistic or theta hat.

So now, we've got the PDF of each individual guy in the random sample from that. Since we know that they're independent, we can write down the likelihood function, which is also the joint PDF of the random sample. So remember, it's the product of all of the individual PDFs. So it's just 1 over theta raised to the n for all x or for x sub i's in this interval. And it's going to be 0 otherwise.

Now, we can actually restate this condition here. Instead of saying x sub i in 0 to theta for i equals 1 to n, we can just say instead, that's the same as the nth order statistic being less than theta, just equivalent. It's an equivalent statement. You can say all of the x's are in the interval 0 to theta. Or the biggest x is less than theta.

So we can say we can say it like that instead. And so what is the value that maximizes this likelihood function? It is, in fact, the nth order statistic. So graphically, let me show you.

It might not be 100% clear when you see it in this form. I hope when I show you graphically, it will become more clear. Why is it that this function is maximized at the nth order statistic?

So this likelihood function is equal to 0 for all values up until the nth order statistic. Think of the likelihood function. Think of this now, instead of being a joint PDF of the random sample, this is a function of theta. So basically what it's saying is the function is equal to this if theta is greater than the nth order statistic, and it's equal to 0 otherwise.

So for any values of theta less than the nth order statistic, this function is equal to 0. So off to the left, it's equal to 0. Then starting at the nth order statistic-- let me flip back again-- it's just equal to 1 over theta to the n. So when we graph it in theta space like this, we get that it's equal to 0 until you get to the nth order statistic. Then it jumps up to 1 over theta to the n.

And then it declines as theta gets larger. So this is maximized at the nth order statistic. Questions? This make sense?

STUDENT: N has to be greater than 1?

SARA ELLISON: N? Yeah, exactly. Because it's just the sample size. Oh, I guess the other thing I want to say about is, you can probably see now by looking at the picture why we didn't solve this by writing down the likelihood function, taking the derivative with respect to theta, setting the derivative equal to 0, et cetera. Because in fact, it's not differentiable at the max. So if we had tried to do that, we would have run into some trouble.

Second example-- let x sub i be iid uniform 0, or sorry, theta minus 1/2 to theta plus 1/2. So here, we again have a uniform distribution. It's not the length of the interval that's unknown.

We know that the interval is length 1. It's just the location of the interval that's unknown. So the parameter theta is determining the location of this interval. so this is theta minus 1/2 up to theta plus 1/2.

So this is our uniform distribution. We're interested in estimating the unknown parameter theta. The unknown parameter theta just determines the location of this interval.

And since the interval is length 1, then the PDF is always equal to 1 in the interval. It's 0 otherwise. So let's write down the likelihood function. Well, this one we have to be a little clever about. We have to think about it when we write it down.

So obviously, the likelihood function-- so 1 raised to the nth power is just equal to 1. But then what we have to be clever about is these bounds here. And so let me go through the explanation of how we get those. So we're writing them in terms of order statistics.

And let me just point out that actually once I convinced you that these bounds are correct, which I hope to do in just a moment, then the maximum likelihood estimate is going to be any value in that interval. Because any value in this interval maximizes the likelihood function. It's just flat over that interval. It's equal to 1. So any value in that interval is going to maximize the likelihood function.

So now I just have to convince you that this sort of condition is correct. So let's look at this one graphically. So let's suppose we've got a random sample. The interval that the underlying distribution has a uniform, where the interval is length 1 and it's centered at theta. And it's got to be here somewhere. This is our random sample, so all of these observations have to live in that interval.

It's got to encompass all the data. So what are the possibilities Well, the interval could be there, it could be there, could be there, could be there. And in fact, there's a whole continuum. And all of those possibilities are equally likely.

Can it be there? No, no, way. Because if it's over here, then these happened, these occurred with probability 0. Or they couldn't occur. Likewise, it can't be there.

So I think I've convinced you that there's sort of a range of different possibilities, that for the location of this interval of length 1, and the only question is, how do we express what that range is? Well, the theta can be, at most, 1/2 above the first order statistic. Let me just flip back.

So over here is the first order statistic. We know that the left edge of the interval can't be any higher than the first order statistic. So therefore, theta can't be any higher than 1/2 above that.

Then the same reasoning goes with the nth order statistic. Theta can be at most, 1/2 below the nth order statistic. It can't be any further down. Or we're going to start missing the observations at the top.

And so that gives us sort of a little region in which theta can live. And all the values in theta, in that little region in that window are equally likely, because we wrote down the likelihood functions. It's just equal to 1 in that region.

So in fact, I can flip back if you want. But this is, in fact, the condition that I had on the likelihood function. So let me just-- that theta is in the nth order statistic minus 1/2, up to the first order statistic plus 1/2. There should be an overlap, and that the overlap of those two sort of conditions, gives us the region where theta can live. Does that make sense?

So maximum likelihood estimators, sort of it's a very popular framework for deriving estimators. And part of it is because they have some very favorable properties. Well, part of it is because they're kind of always available. If you're willing to take a stand, make an assumption about what the underlying distribution is, then you can write down a maximum likelihood estimator.

Of course, that's also true for method of moments estimator. But sort of maximum likelihood estimators are sort of available if you're sort of willing to write down a distribution of the underlying data. But they also have some very favorable properties.

So if your assumptions are correct, so this is all assuming that you haven't made a mistake in assuming the underlying distribution, then the maximum likelihood estimator is going to have the following properties. If there is an efficient estimator in a class of consistent estimators, maximum likelihood estimation will produce it.

So that tells you right there, you don't have to check for the efficiency of your estimator. If you're using maximum likelihood, it's going to have that a good property in that respect. Also, there is a central limit theorem-type result having to do with maximum likelihood estimators.

So under certain regularity conditions, if the maximum likelihood estimator is not on some boundary of some space or something like that, then asymptotically, sort of as your sample size is getting bigger and bigger, maximum likelihood estimate are going to have a normal or approximately normal distribution. So just like we had the central limit theorem for the sample mean, there is a central limit theorem-like result for maximum likelihood estimators. And that can be very useful for inference, which we'll see after spring break.

So does this mean that maximum likelihood estimation is always the right thing to do? Well, no, in fact. So we saw an example where even given that our assumptions were correct, the maximum likelihood estimator was biased.

So that was the example, where we were estimating theta from the uniform 0 theta distribution. And our maximum likelihood estimate for any finite sample size was going to be biased. Now, in fact, in that particular example, we could fix the bias, because the bias is a function of n. And we know how big our sample size is, so we can actually undo that bias.

But there are lots of examples where you don't know the size of the bias. And so you can't always undo it. So they can be biased. And we saw an example. And sometimes, we're not interested in biased estimators.

They might be difficult to compute. And then finally, they can be sensitive to incorrect assumptions about the underlying distribution and more so than other estimators. So the method of moments estimators tend to be more robust to the underlying assumptions. Because they rely less. So the maximum likelihood estimators really rely on the entire shape of the distribution of the random sample. The method of moments estimators don't. They just rely on the moments.

And so you can imagine just kind of intuitively that something that relies on the entire shape of a distribution, if you're wrong about that shape of a distribution, it might be sensitive that something can go wrong. And that's, in general, true. Maximum likelihood estimators can be sensitive to incorrect assumptions about the underlying distributions. Questions?

No? So I would like to just go through a summary of what we've done so far this semester. And feel free to chime in with questions and ask about if you have questions about the exam, relative to this, what you might be responsible for in the exam. I'm happy to answer those questions.

So we started out the semester with some probability basics. So we introduced the concept of probability and talked about simple sample spaces, independent events, conditional probabilities, and Bayes' rule. And in some sense, that was kind of foundation for sort of going on and studying probability with the kind of mathematical construct of the random variable. But do note that we didn't study Bayes' rule in the context of random variables.

You can actually formulate a Bayes' rule using random variables, which we didn't do. But Bayes' rules are very important concept, that aside from, I guess, sort of the one concept that we covered in the beginning that wasn't sort of just foundation for the rest of what we're doing. Then we introduced random variables. And we defined what a random variable was.

We discussed ways to represent distributions. So a probability function, a probability density function, so those were kind of analogous, sort of things. But one was for discrete random variables. And one was for continuous random variables.

And then we also talked about how the same information in PFs and PDFs can be embodied in a CDF. And that's the definition, of a CDF is the same for both continuous and discrete random variables. And then we also covered sort of the random variable analogs to the sort of stuff we covered in probability basics.

So we had talked about independent events. Then we talked about independent random variables. We talked about conditional probabilities. We then talked about conditional PDFs.

And then, Esther covered some, sort of I don't know, empirical counterparts to some of the things that we had been talking about theoretically in class. So she talked about both, histogram, defined what a histogram was, showed us some nice examples, and then also talked about kernel density estimates and what the difference between a histogram and a kernel density estimate is and showed us some pictures of those as well.

And we saw lots of examples that we hope gave you a sense for how these theoretical objects function and what information they tell us and how to interpret them. Then we started talking about functions of random variables. So we saw some basic strategies, or I guess, actually one basic strategy for figuring out the distribution of a function of random variables. And then I also did-- I went through several important examples. And so I think it's useful to go back and be comfortable with the example.

So what were some of the examples that we went through? Well, we did sort of the probability integral transformation. And we did a convolution. We did linear transformations of random variables. We did order statistics.

Maybe that's it. I'm not sure. Maybe there was one more. I don't know. Those are sort of, those are all important examples and all things that we sort of have drawn on since and will continue to draw on.

Then we sort of introduced moments of distributions. So we defined mean variance. We defined a couple other moments. But we mostly focused on the mean and variance.

And then in addition to learning how to-- or expectation if you'd rather, expectation and variance. In addition to learning how to directly compute the expectation and variance, given the distribution of a random variable, we also learned a lot of techniques and properties to help compute moments of functions of random variables. So you have a random variable, maybe you know its moments. But what you're really concerned with is a function of that random variable. Do you have to figure out how that function is distributed and then compute the moments directly?

You can do that. Usually you don't. Oftentimes, we can just rely on properties of expectation and properties of variance, properties of covariance, et cetera, to figure out the characteristics of distributions of functions of random variables, Then Esther went through a whole lecture discussing special distributions. And these are all distributions that are useful. They're distributions that will re-enter your life at some point in high probability, I would say.

They're distributions that we will continue to discuss throughout the semester. And so it's good if you have some level of comfort with them. We're not interested in having you memorize PDFs of special distributions. We're not interested in having you memorize the expectation and the variance of special distributions.

Sometimes that just happens because you use them so much. So on the exam, if you feel comfortable, you can write them out on your page of notes that you bring in. But we'll also provide things like that on the exam.

And then finally, in the last couple of lectures, we started talking about estimation. So we started our segue into estimation, in some sense, was the central limit theorem. So we've been talking about functions of random variables. And we introduced a special particularly useful function of random variables, called the sample mean.

And we computed what the expectation of that function was, what the expectation of the sample mean is. We computed what the variance of that sample mean was. But we wanted to know even more about its characteristics and how it was distributed. And the central limit theorem told us that. The central limit theorem had this sort of remarkable result, that no matter what kind of distribution you're drawing from and computing the sample mean of, if your sample size is big enough from that crazy distribution, then the sample mean is going to have approximately a normal distribution. And that's just super useful.

The full usefulness of it has not occurred to you yet because we haven't gotten to inference. But we will after spring break, get to inference. And then you will know.

So we talked about the central limit. We talked about the sample mean. So that's kind of the most important estimator that we talked about. We talked about the central limit theorem, which told us some important facts about how the sample mean behaved. We had sort of a general discussion about estimation.

And then, the last two topics we covered, these are topics that we will not test you on the exam. The last two topics we covered were criteria for assessing estimators. So whether an estimator is unbiased, whether it's efficient, whether it's a minimum mean squared error estimator, whether it's consistent, and then we also talked about frameworks for deriving estimators, method of moments estimation and maximum likelihood estimation.