[SQUEAKING]

[RUSTLING]

[CLICKING]

SARA ELLISON: OK, where were we? Well, so this is our summary to date that I had up right before the midterm. So let me just remind you what we had just finished up.

So we talked about a general discussion of estimation and also a discussion about the sample mean. We talked about criteria for assessing estimators. And we talked about frameworks for deriving estimators. And that's where we left off right before midterm.

So we've seen various estimators. And we've made observations about their distributions. We've discussed how we might derive them, either by just being clever and thinking of them off the top of our head, or using maximum likelihood estimation, or using method of moments estimation to derive estimators. And then we also discussed the criteria that we might use to choose among them.

That's fine. That's all useful. But when we actually have to report estimates, when we're confronted with a real data set. And we need to report an estimate of some parameter, people are going to want to have some objective measure of how good our estimate is, or how reliable, or how precise our estimate is.

So that's what I'm going to talk about today-- or one thing I'm going to talk about today.

So one way that we can do this-- so remember that the estimate is a realization from the distribution of the estimator. So we report the estimate. We report that one number that's a realization from that distribution.

And then we also might just report the variance of that distribution or, if we don't know the variance, an estimated variance. And that can give a lot of information about how reliable our estimate is and how confident we are in our estimate.

So the first thing I'm going to do is define the standard error. The standard error of an estimate is the standard deviation or, if we don't know the standard deviation, the estimated standard deviation of the estimator.

So, for example, let's go back to one of our favorite estimators, sample mean or x bar. So we have shown that the sample mean has mean mu and variance sigma squared over n. We know that. So the standard error for that estimator is just sigma over root n. It's just the square root of the variance.

Now, a lot of times-- sometimes we might know what sigma squared over n is. We always know what n is, because that's our sample size. A lot of times we might-- or sometimes we might know what sigma squared is. And so, then we just report that as the standard error.

A lot of times, we won't know it. So we have to substitute an estimate for sigma squared. And that's OK. So then we'll just plug in the estimate for sigma squared there.

So I just put this picture to remind you, estimators have distributions. And so, it's the distribution of-- this is the PDF of xn-- sorry, x bar sub n. And so, it's just the standard deviation or estimated standard deviation of this distribution that we report.

So, oftentimes, we report an estimate, some parameter, and then we also report a standard error. Oftentimes, we'll do it-- we'll put the standard error in parentheses right after the estimate. So, a lot of times, if you're reading an empirical paper, you'll see an estimate and you'll see something in parentheses after it. That oftentimes is the standard error associated with the estimate.

So, useful. Standard error certainly gives us some idea of how tightly concentrated around the unknown parameter the distribution of our estimator is. But sometimes it might be useful to report, essentially, that same information but in a different form, an interval.

So, in other words, we could think about constructing an interval using information about the distribution of the estimator. And that interval is going to be narrow if the estimator has a tight distribution around the unknown parameter. And it's going to be wide if it has a dispersed distribution around the unknown parameter. And we're going to call that a confidence interval.

So just like the standard error gives you some information about how tightly distributed around the unknown parameter the estimator, the distribution of the estimator is. The confidence interval is going to give you that same information, just in a different form. It's going to give it to you in an interval form.

So, specifically, we want to find functions of the random sample, call those functions capital A and capital B, such that the probability that A of x is less than theta is less than B of x is equal to 1 minus alpha.

So what exactly does this mean? So we've got A and B are random functions. They're functions of our random sample. And so, we can, because they're random, they're random variables, we can write probability statements involving them.

Theta is a fixed number. We don't know what it is, typically, but it's fixed. But A and B are both random functions. So we can write probabilities involving A and B and this unknown parameter theta.

And then, 1 minus alpha, that's just something we choose. So that is called the desired degree of confidence. And, basically, we just, when we're constructing a confidence interval, we just decide whether it's a 95% confidence interval, and then we'll choose alpha to be 0.05. Or a 99% confidence interval, alpha is 0.01, et cetera. That's just whatever we want-- whatever we want it to be.

So this is the problem, or this is the setup. We want functions, random functions A and B, such that a probability statement like this is true.

Now let's just go back to estimation, point estimation for a second, and see how this is different from what we've done in estimation. In point estimation, let's say we wanted an unbiased estimator. Then, what we were looking for is a function theta hat, such that the expectation of that random function was equal to the unknown parameter.

So it's kind of analogous to what we were doing with point estimation, here we're looking for two random functions such that this probability statement is true. So it's a similar exercise to what we were doing with point estimation.

Then let's suppose we can find such functions A and B. Then the interval A of little x and B of little x is said to be a 1 minus alpha confidence interval for theta. So now, we plug in the realizations. We have a data set. We plug in the realizations from that data set.

And A and B just become numbers. They're no longer random functions once we plug in the realizations. They're just numbers. And then that forms that forms an interval. And we call that interval our 1 minus alpha confidence interval for theta. Yeah.

AUDIENCE: When you say A is your random functions, does that mean functions are random variables?

SARA ELLISON: Yes.

AUDIENCE: [INAUDIBLE] figured out [INAUDIBLE].

SARA ELLISON: No, Yeah, you're right. I should have been more precise-- functions of random variables, yeah. So then, once we plug in the x's, the realizations, then those just become numbers. And so, we have two numbers, and those define a 1 minus alpha confidence interval for theta.

A couple of notes about constructing these confidence intervals. First of all, we shouldn't be surprised to find out that these functions are not unique. So you have a distribution of an estimator.

And what you're going to do is you're going to basically use information of the distribution of some estimator for theta to construct these confidence intervals. And you can imagine shifting them to the right and left and still having alpha probability outside the interval. And there are various ways in which you can do this.

So what we typically do is we choose alpha and beta such that alpha over 2 of the probability falls on either side of the interval. I'm going to draw you some pictures of this to clarify these ideas in a second. But the point I'm trying to make here is just simply that these functions are not unique. But there is a standard thing that we usually do to choose them.

And then, the second thing I want you to do is keep in mind that these functions, A of little x and B of little x, they're just numbers. So once you plug in the realizations, the probability statement that I had up a couple of slides ago doesn't make sense.

So, for instance, once we have numbers, this is 1.3 and this is 7.9, it doesn't really make sense to say, oh, theta is between 1.3 and 7.9 with probability 0.95. So maybe this is a slightly pedantic point, but I think it's an important point, that once we plug in the realizations, these are numbers, and we can't make probability statements involving them and the unknown parameter.

So where do we find these functions? How do we even get started? Well, there's going to be a problem on problem set five where we walk you through finding confidence intervals from scratch based on a estimator that you're familiar with.

In most cases, that's not what we have to do. So, in most cases that you're going to encounter in your everyday life, people have-- others have found these functions, A and B, and the resulting formula for, say, a 95% confidence interval for the mean of an unknown distribution with sample size greater than 30, or whatever it is, other people have found those functions and found the formulas. And then, what you do is you just plug into them to create the 95% confidence interval or whatever percent confidence interval.

So you will have one problem on the problem set where you have to construct it from scratch. But, typically, when we're constructing confidence intervals, people have already done most of the hard work for us, and we just plug in to the formulas they provide.

So I'll go through an example in just a second. But in order-- I just want to emphasize one thing. In order to find these functions and derive the formulas, one needs to know how an estimator for the unknown parameter is distributed.

So, typically, just having the mean and the variance isn't enough, if you just know the mean and the variance of the estimator. You really have to know the whole distribution of the estimator to construct a confidence interval.

Ah, so now we're going to take a little detour. As I said, I'm going to go through an example in a couple of minutes. But we have to take a little detour into learning about three additional special distributions.

So I just said that it's important to know how estimators are distributed in order to construct confidence intervals. And this is precisely where our friends the chi squared t and F distributions come in. So you've probably encountered them.

You've probably seen things like t-tests, and F-tests, and confidence intervals, or t confidence intervals, and so forth. And so, I'm going to tell you today why those distributions are important and where they come in.

So these are special distributions. But, unlike the ones that Esther covered before, they don't really appear in nature. They don't really describe phenomena that we encounter. They're really just invented distributions. They were invented because estimators or functions of estimators had distributions that needed to be described and tabulated. And these were invented for that purpose.

So, first, the chi squared distribution. So do you guys remember, a few weeks ago, very briefly, I had a slide up talking about an estimator for the variance called the sample variance, denoted s squared. I don't know if you guys remember that? So that was up on a slide earlier.

And s squared, just to remind you, is equal to 1 over n times the sum of xi minus the sample mean squared. And when I talked about s squared earlier, I said that it was an unbiased estimator for the variance of a distribution.

Well, it turns out that a particular function of s squared, n minus 1 times s squared over sigma squared, has a chi squared distribution. So when we construct a confidence interval for the variance, we're going to base it on that particular function. And we know that that function has a chi squared distribution. So that's going to allow us to construct the confidence interval.

And I should also point out, there's, in addition to the fact that the chi squared and the t and the F distributions don't really occur in real life and they're kind of invented for the purpose of constructing confidence intervals and performing tests, there's also something else different about them. And that's that their parameters have a special name.

So the chi squared distribution is a one-parameter family. So there's just, there's one parameter that tells you which member of the chi squared family a particular distribution is. But instead of just calling it a parameter, we call it the degrees of freedom.

So you can just-- you don't have to think about it any differently. It's just simply a parameter. But just so you know the terminology, it's called the degrees of freedom. So that's the chi squared.

The t distribution-- and we'll see an example using the t distribution in a few minutes. Well, why is the t distribution useful? So let's suppose you have a random variable x with a standard normal distribution and another random variable z with a chi squared distribution. And they're independent.

Then it turns out that x divided by z over its number of degrees of freedom, the square root of that, has a t distribution. Why do we care? Well, so I'll give you some indication without going through all the formal calculations. I'll give you some indication of why this is useful.

Well, suppose we're sampling from a normal distribution, mu sigma squared distribution. We already know that x minus mu-- oh, I think that's supposed to be x bar. Oops. x bar, right here, x bar minus mu over sigma squared-- over the square root of sigma squared over n has a standard normal distribution. We're all in agreement of that.

And then, we also know that n minus 1 times s squared over sigma squared has a chi squared n minus 1 distribution. I just told you that the last slide. We actually don't know that they're independent but, in fact, they are. You'll just have to take my word for that.

Well, then what we can do is we can form this unwieldy looking product. And if we do a little bit of algebra, cancel a few things, we get something-- we get this. So we get x bar minus mu-- for some reason I put the root in up here instead of putting it in the denominator down here. But we get x bar minus mu divided by s over root n.

Look familiar? Yeah. So this is exactly the function of x bar that had a normal 0, 1 distribution, except, instead of having sigma here, we have s. So, basically, our estimator for sigma is here instead of the true sigma.

So, basically, what this tells us is, if we form this standardized version of x bar, but instead of using sigma we use s, then instead of having a normal 0, 1 distribution, it will have a t distribution. Does this seem perplexing? Yes. [LAUGHS]

Any specific questions you'd like to ask? No?

AUDIENCE: Yes.

SARA ELLISON: Yes, go ahead.

AUDIENCE: Could you make that a little more concrete with maybe a context in which you would use this and these [INAUDIBLE]?

SARA ELLISON: So I will do that. If you want to wait for a couple of minutes, I will do that. But maybe I'll try to give you some indication verbally.

So, basically, what happens is, we have a random sample. If we know what the variance-- let's say the random sample is from a normal distribution, just to make things simple. So we have a random sample from a normal distribution.

If we know the variance and the mean of that normal distribution, then what we can do is we can form this function of x bar. And we know that this thing has a normal 0, 1 distribution.

A lot of times, even if we're sampling from a normal distribution, we actually might not know mu and we might not know sigma. But let's put aside the fact that we don't know mu for a second because it turns out that won't be important sometimes. But let's suppose we don't know sigma and we only have an estimate of sigma.

So what we want to do is we want to form this function-- we want to form something that's close to this function as we can. And this is what we come up with. Well, it no longer has a normal 0, 1 distribution, it has this other distribution.

And so, basically, if we can form this function of the estimator x bar, and we know it has this distribution, then we can use that to build a confidence interval around. And I'll do an example both when we know sigma and we use the real sigma and when we don't know sigma and we use s instead. And we'll compare them. We'll talk about a comparison of those two. Does that help a little bit?

AUDIENCE: Very helpful.

SARA ELLISON: OK, good. Other questions?

AUDIENCE: I have a quick question.

SARA ELLISON: Yep.

AUDIENCE: So, a few slides ago, you said that in order to-- if you can just go back a few slides. You said-- wait, before, it said, rather than-- before, one slide.

SARA ELLISON: Oh.

AUDIENCE: Oh, actually, so you say, we need to know how the estimator for the unknown parameter is distributed. But I thought the central limit theorem told us that all estimators were--

[INTERPOSING VOICES]

SARA ELLISON: If n is large enough.

AUDIENCE: If n is normal, OK.

SARA ELLISON: If n is large enough.

AUDIENCE: OK.

SARA ELLISON: Yeah. So we'll talk about that. So, basically, what happens is, a lot of times we don't know the variance of-- well, let's suppose that we're sampling from a normal distribution. And we know that the sample mean is going to have a normal distribution, but we don't know its variance.

So, in that case, then we have to appeal to things like this. Even though we know x bar has a normal distribution, this function of x bar doesn't have a normal distribution. It has a t distribution. So now, as n gets very large, the t distribution is going to converge to the normal distribution. And so, for a large n, it doesn't matter what we do.

Then but your question was a little bit different. Your question was, well, what happens if we're not sampling from a normal distribution? What happens if we're sampling from some crazy other distribution? The central limit theorem tells us that our estimator-- our sample mean is normally distributed. But, again, that's only true when n is sufficiently large.

So if we have-- yeah, the hardest case is when we're sampling from a crazy distribution and our number of observations is very small. Then we can either appeal to the central limit theorem or to this distributional fact. But I'll talk about that a little bit more.

AUDIENCE: Thank you.

SARA ELLISON: Yep. Let's see, where was I. So we have-- this is just a summary of what I said. So we have the square root of n times x bar minus mu over s has this t distribution.

And that's when we're sampling from a normal mu sigma squared distribution but we don't know-- or, well, if we don't know sigma squared-- I mean, even if we know sigma squared, this is true-- the statement is true. But if we don't know sigma squared, that's why we would create a function like that. Yeah.

AUDIENCE: [INAUDIBLE], so, in the case where I cannot sample a large n, and I'm forced to sample a small n, a small sample size--

SARA ELLISON: Yeah, if you could speak up just a little bit.

AUDIENCE: If I'm forced to sample a small size--

SARA ELLISON: Yeah, you have a small sample.

AUDIENCE: That means that the smaller my sample, the-- let's say, the smaller my confidence-- I mean, the larger my confidence interval would be, obviously, right?

SARA ELLISON: Yeah.

AUDIENCE: So my question is, I know sometimes a small sample is not representative of the population. But sometimes it's OK. What is-- my question, what is the cut-off confidence interval in everyday research, in experiments, where it's OK for me-- is it OK for me to say, I'm 20% confidence interval-- 20% sure that my sample is representative when it's obviously not? But is 80% OK [INAUDIBLE]?

SARA ELLISON: So you've asked a question that has a number of pretty subtle answers to it. So there's a lot of things going on in your question. So in case people didn't hear it, the question was, how large does your sample have to be before you have a reasonable amount of confidence that the estimates that you're deriving from the sample are reasonable. Is that a fair statement?

So there's a lot of things going on. So, first of all, we're going to assume in this class, for the most part, that your sample is randomly drawn. And sometimes you might have a situation where you have a sample but it hasn't been randomly drawn. And that complicates the analysis.

So we're only-- so when we talk about small samples, they're still going to be, quote, representative. They're still going to be randomly drawn from the population-- we're assuming that. But, when they're small, it just means that our estimates-- we don't have very much confidence in our estimates. And that's going to be exhibited in how large our confidence intervals and how big the standard error of our estimate is.

And there's nothing wrong with that. We're just-- you just say, look, I'm estimating this unknown parameter, but I don't have very many observations. So my estimates aren't very precise. I'm not very sure of them.

And that's why we want to quantify how confident we are in our estimates, by doing things like creating confidence intervals.

AUDIENCE: I guess my question comes from a publication point of view. Is it OK for me to publish something where I'm 80% sure, or [INAUDIBLE].

SARA ELLISON: So your referees would be the ones to tell you [LAUGHS] whether it's OK to publish that. And it just depends. I mean, there are some cases where it's relatively easy to get much larger samples, and then referees would typically tell an author, you should go out and get a much bigger sample and produce estimates you're more confident in.

And there are some cases where maybe it's impossible to get a larger sample, and the evidence that you have is the best out there, even if it's not that good. So it's, yeah. So there are different ways to answer your question. There's no right answer. There's no-- I can't tell you an n, a magic n, that you always need to have.

Does anyone know the origin of the t distribution? The t distribution was actually formulated by William Sealy Gosset in his job as chief brewer in the Guinness Brewery in Dublin. So I went to the Guinness Brewery with my family a couple of years ago, took a photo of this plaque that's right on the wall of the brewery.

And, basically, he derived and tabulated this distribution to aid in his analysis of data for quality control across batches of beer. So he had a relatively small number of batches of beer. He didn't know what the variance that he-- of the distribution that he was sampling from was.

And so, he was the one who actually formulated and tabulator derived and tabulated this distribution. And, if you go to the Guinness factory, seek out this plaque, and you could have your picture taken by it and email it to me or something.

[LAUGHTER]

Yeah.

AUDIENCE: [INAUDIBLE] students' t distribution, I always thought this was just a simplified version of the t distribution [INAUDIBLE].

[LAUGHTER]

SARA ELLISON: Nope. Nope. Totally not. That's the t distribution. And it was-- yeah. So he published his paper-- I think it was in *Biometrika,* under the name Student. Because, I guess, even though the Guinness Brewery was supportive of his efforts, they didn't want-- I don't know. They didn't want it known that their chief brewer was publishing papers in mathematical statistics or something.

[LAUGHTER]

So I don't know.

Final distribution, special distribution I'll talk about now, is the F. And, again, I'll just give you an indication of why the F might be useful or important.

If we have a random variable x that has a chi squared distribution with n degrees of freedom and z chi squared with m degrees of freedom, and they're independent, then the ratio of those two random variables divided by their respective degrees of freedom has an F distribution with nm degrees of freedom.

So those-- so n and m, again, are just the-- it's a two-parameter family. Those are just the parameters that characterize the distribution, the F distribution. And, again, they're called degrees of freedom instead of parameters.

So why is this a useful fact? Well, suppose that we have samples from two different populations. And we might want to know whether the distributions in the two populations are the same. And, if they are, that means their variances are the same.

That means if we form the ratio of the sample variances, the-- or, sorry. I said the ratio of the sample variances. What I meant was actually the ratio of the function of the sample variances that has a chi squared distribution. So let me just go back to that quickly.

So the ratio of this, for both of-- for the two different populations. If we form that ratio, then the true variances, if the population have the same distribution, the two variances will cancel out because they'll be the same. And then the ratio has that above distribution.

So that's why the F distribution was invented and tabulated. No good stories about the F distribution that I know of.

So now, let's proceed to construct some confidence intervals. So we're going to focus on, initially, on two cases. These are not the only cases that you'll ever encounter, but they are, by far, the most important cases. And having these under your belt is important.

So case one, we're sampling from a normal distribution with a known variance, and we want a confidence interval for the mean. So sampling from a normal distribution and we know the variance. Case two, same situation but we don't know the variance.

So let's consider case one. So what we're going to do is, we want to construct a confidence interval. And we're going to construct the confidence interval based on a good estimator we have for the unknown parameter, the mean. And that, of course, is x bar.

Now, we know, since we're sampling from a normal distribution, we know not only the mean and the variance of this thing, we also know its distribution. Does everyone remember that? So we're sampling from a normal distribution. So the sample mean will have mean, mu, variance, sigma squared over n, and be normal itself.

So what do we know about this function here, x bar minus mu over sigma divided by root n? What do we know about that?

AUDIENCE: [INAUDIBLE]

SARA ELLISON: Yeah, it's a standard normal distribution, normal 0, 1 distribution. Exactly right. So never mind the fact that we don't know what mu is. That doesn't matter. We're just doing calculations. So we just use mu here.

So we have something in the middle that has a standard normal distribution. And what we want to do is we want to construct an interval. And, remember, what is the definition of a confidence interval? We're looking for functions A and B such that the probability that A of x is less than or equal to mu is less than or equal to B of x is 1 minus alpha.

So let's start with this step first. We've got something whose distribution we know perfectly-- I mean, up to mu-- whose distribution we know. And then, what are these things on the outside?

Well, this on the left, this capital phi inverse of alpha over 2, that's simply this number right here. It's just the inverse CDF of the standard normal evaluated at alpha over 2. So it's the number such that alpha over 2 of the probability in a standard normal is to the left. Got it?

Now, over here, what's this number? Well, we could have done 1 minus capital phi inverse of alpha over 2. But, because of symmetry, we just put a minus sign in front of that one, which is fine.

So does everyone believe this probability statement and see where it came from?

So now what do we do? It's not in the form of a confidence interval yet. Remember, we have to have mu, the unknown parameter, isolated in the middle. Well, we just do a little rearranging, that's all.

So we do a little rearranging. And then, on the left, we have x bar plus capital phi inverse of alpha over 2 times sigma over root n. And then, on the right, we have something similar but with a minus sign.

This is still a probability statement until we actually plug in the realizations. And so, then our confidence interval is just simply x bar plus this quantity and x bar minus this quantity.

So, here, this is still a probability statement. So you can think of x bar as being a function of random variables. Here, you can think of plugging in the realizations for x bar there.

So a couple of things I need to point out. One is that everything in this confidence interval is either known, or we can look it up in a book, or we can calculate it from our random sample.

So we assume that sigma is known. That's what case one was. We know the size of the sample. We look this up in a book. And this we just calculate from our sample, from a random sample. So this is a number. And this is a number. And that's our 1 minus alpha confidence interval for mu. Make sense?

So how about case two? Yep.

AUDIENCE: Before we move on. So once we have that interval, what does it tell us again? Say alpha was 0.05, so then 95%. So then you say-- and then, let's say, the interval was 1 to 3. So you would say we know that the mean is between 1 to 3 with 95%?

SARA ELLISON: No. [LAUGHS] No. So that was the thing that I said. It's kind of a pedantic point, but it's no longer-- we can no longer make a probability statement once we plug in the realizations. Because there's an unknown parameter mu, but it's fixed. It's not a stochastic object.

So you can't-- nothing is stochastic once we plug in the realizations. We just have two numbers in an unknown quantity. So it doesn't make sense to talk about probabilities.

So, to be perfectly honest, that's how most people think about a confidence interval, because that's the probability statement is how we construct it. But then, once we plug in the realizations, it's not really correct to make probability statements.

AUDIENCE: You may have already answered this, but what do the two numbers tell you that? Sorry if you already answered it.

SARA ELLISON: No, no, no. I haven't answered it. They just give you information about the distribution of x bar. How--

AUDIENCE: If the interval is wide, then you know it has a large variance.

SARA ELLISON: Yep.

AUDIENCE: [INAUDIBLE]

SARA ELLISON: That's it. Yeah.

AUDIENCE: Couldn't you just take the variance instead of the [INAUDIBLE]?

SARA ELLISON: Sure, absolutely. Yeah. So this is only-- so it's important for you guys to see confidence intervals because they're everywhere and to know where they come from. But it's really-- I mean, it's essentially the same information that you would be presenting more or less if you just reported a standard error. But it's in a different form. That's all. That's all.

ESTHER DUFLO: [INAUDIBLE] in economics, we often see a [? lot of ?] tests [INAUDIBLE]. And now they [INAUDIBLE] to be [? able ?] to test and to come in a a minute. So, for example, a very popular test, you reject this expressions a confidence that is zero. [INAUDIBLE], for example, in medicine, in all of the medical journal, thinking about what the report is [INAUDIBLE].

It's a little bit a matter of convention. It's very important. [INAUDIBLE] a bit of a richer sense of-- because, after all, you might not care about [INAUDIBLE] going [INAUDIBLE], we have a focus on zero. But we might not always-- we might care about any other number.

For example, you might wonder [INAUDIBLE] whether an effect is large or small. Suppose you see always inside the confidence interval, that's one useful piece of information. [INAUDIBLE] come back to the discussion that's [INAUDIBLE] earlier, it's not a very useful piece of information [INAUDIBLE].

And it becomes a more useful information if you can [INAUDIBLE] haven't told you what the unit was. What I'm saying is kind of senseless. But it would-- if zero is a confidence interval, but 0.001 is outside, [INAUDIBLE] the effect of what [INAUDIBLE] confidence interval [INAUDIBLE] readily available information. So [INAUDIBLE] really have to multiply by this magic number that's going to come in a minute.

[LAUGHTER]

[INAUDIBLE]

SARA ELLISON: Good. Questions before we go to case two? So case two, remember, is exactly the same as case one. We're sampling from a normal distribution. But here we don't know the variance of the distribution.

So, because we don't know the variance of the distribution that we're sampling from, we also don't know the variance of x bar. Because, remember, the variance of x bar is sigma squared over n. So we don't know its variance.

So the only thing that we have is the ability to estimate it. And, remember, if we form this function that looks very much like the function we formed in case one, but put an sn instead of a sigma n, then we-- actually, that's going to let us construct a confidence interval because we know how that thing is distributed. We know it's distributed t with n minus 1 degrees of freedom.

So we're going to do exactly the same thing we did last time. But instead of having the inverse CDFs of the standard normal distribution in the probability statement, we're going to have the inverse CDFs of the t with n minus 1 degrees of freedom in the probability statements. So it's the same exercise. So I could draw-- basically, I would draw almost the identical picture that I had on the previous slide.

And then, again, what we do is we just rearrange. And we get this on the left and this, oh, I guess, starting up here, on the right. And, again, everything on the left, everything to the left of the less than sign, is something that we can compute or that we know from our random sample. So we just plug that in.

So we don't know sigma. So sigma is not here. But we can compute s. So we plug that in. And then we look up the inverse CDF of the t in a table, et cetera.

And then, once we plug in the realizations for x-- the realizations from our random sample for x bar and for s, then we have two numbers. And this is called a t confidence interval.

So what's the difference? Well, the t distribution, if I drew a picture of the t distribution, it would look identical to my picture of the normal distribution just because maybe my drawing is not quite precise enough.

But, basically, the difference is that the t distribution is symmetric, just like the normal. It's centered at zero, like the normal. But it has, quote, fatter tails. It just doesn't-- its tails don't fall off quite as quickly as the normal. So it does in fact converge to the normal as n goes to infinity.

But for small values of n, like n equals 10, for instance, the t is going to be more spread out than the normal distribution. Well-- oh, and so, I should say, actually, because it has thicker tails, this number here for the t distribution is going to be further out than for the normal, the inverse CDF evaluated at alpha over 2.

So what does this mean for the confidence interval? Well, it means that the t gives you a wider confidence interval than the normal for finite n. The intuition and-- well, I should say, I say finite n, which is true.

But, I mean, basically, practically speaking for n less than 50 or something like that, it's going to give you a somewhat wider confidence interval. And, above that, I think it's not even tabulated separately from the normal distribution.

So what's the intuition behind this? Well, the intuition is that even though you're drawing-- in both cases, you're sampling from this normal distribution that's identical. One case you know what the variance is. The other case you don't know what the variance is.

And so, when you construct your confidence interval for the mean, you're kind of getting penalized for the fact that you don't know what the variance is. And so, that decreases the confidence you have in your estimate, and it results in a slightly wider confidence interval.

So the t essentially penalizes your confidence interval by making it wider, reflecting that you have greater uncertainty. As n goes to infinity, this uncertainty becomes irrelevant. And then the t confidence interval converges to the normal.

So I said before, case one and case two don't cover everything. But they're really important. And what makes them so important? Well, first of all-- oh, well, I'll get to what makes them important in a second. But, first of all, you might ask the question, if we don't fall into case one or case two, what do we do? Is there something else we can do?

Well, you can construct a confidence interval from scratch on your own. Like I said, we'll see one of those on the problem set. In practice, what we normally do is we appeal to central limit theorem-like results to argue that the estimator that we're basing our confidence interval on has an approximate normal distribution. And, most of the time, it will.

And then, what we do is we just use the t confidence interval with an estimated variance. So, basically, we're assuming that we're in case two even if we're not sampling from a normal distribution. And that's because the time that we use the fact that we were sampling from the normal distribution was to figure out how this function of x bar was distributed.

Well, if we appeal to central limit theorem-type results, we know that that function of x bar is going to be approximately normal. So then, that puts us in case two, even when we're not strictly speaking in case two.

And then, furthermore, as I said before, for large N, case one and-- case two converges to case one. So, for large N, It doesn't even matter. We just, not only can we apply central limit theorem-type results, it's as if we knew the variance of the underlying distribution because our estimator is so good because n is so big. Does that make sense? Questions?

So now we know what an estimator is. We know how to estimate unknown parameters. We know a couple of different ways to express how confident we are in our estimates. So, in particular, we know about standard errors and we know how to construct a confidence interval.

And that goes a long way and gives us a really good foundation for studying all kinds of estimation going forward. But there is one additional piece of foundational material that's quite important. And that's hypothesis testing.

And, as Esther indicated a couple of minutes ago, it's particularly important to economists and social scientists. And so, I'm going to introduce the idea of hypothesis testing and give you some indication of how we do it.

So, in social science, we often encounter questions that we want to know the answer to. And some of you have started formulating these questions for your empirical project in this class.

So what are some of the questions that we care about? Do the lifespans of popes follow a log normal distribution? Does the income tax rate affect the number of hours employees are willing to work? Do used books cost more on the internet than in brick and mortar stores? Has NAFTA hurt US manufacturing workers?

So these are the kinds of things we-- we not only want to get estimates of unknown parameters, but we also want to use those estimates to answer questions like this. And this is very common and very important in social science.

So the tool that statisticians have invented to help answer such questions and quantify how confident we are in the answers is the hypothesis test. So another way to describe the hypothesis test is that it has the following purpose. Given a random sample from a population, is there enough evidence to contradict some assertion about the population?

So, in other words, maybe my assertion is the tax rate has no effect on workers' willingness to supply labor or something like that. That's my assertion. I have a random sample from a population. Does that random sample give me enough evidence to contradict that assertion and say, no, in fact, the income tax rate does matter in workers' willingness to supply labor.

So let's build the structure underlying the hypothesis test. So there's lots of definitions. So I'll try to go through these pretty quickly because there's nothing very deep about them. But you'll hear these terms a lot, and we'll use these terms. So you need to be familiar with them.

So a hypothesis is an assumption about the distribution of a random variable in a population. A maintained hypothesis is one that cannot or will not be tested. And I'll show you an example in a second. A testable hypothesis is one that can be tested using evidence from a sample.

The null hypothesis is, in fact, the testable hypothesis that we will test. And the alternative hypothesis-- and I should say, the null hypothesis typically denoted H sub 0. The alternative hypothesis typically denoted H sub a, sometimes you see it denoted H sub 1. The alternative hypothesis is a possibility or a series of possibilities other than the null hypothesis.

So, for instance, let's say we want to perform a test concerning an unknown parameter theta, where x sub i-- we have a random sample x sub i from this distribution characterized by theta. So I've got this random sample. Theta is some parameter in the-- governing the distribution of the random sample.

So our null hypothesis we might specify as theta being in some space, capital Theta 0. And the alternative hypothesis we specify as theta being in some other parameter space capital Theta sub a, where Theta 0 and Theta a-- capital Theta 0 and capital Theta a, are disjoint. So you can't have overlap in the null hypothesis and the alternative hypothesis. Yeah.

AUDIENCE: [INAUDIBLE] the null hypothesis and the alternative hypothesis kind of compliment each other?

SARA ELLISON: Yes, that's what that means. They're complimentary. They're just-- so H-- sorry, Theta 0 and Theta a being disjoint means that they're-- oh, sorry. You mean-- sorry.

You're absolutely right. You mean not just mutually exclusive but also complimentary? They may or they may not be. We'll actually see an example where they're not complimentary. Typically, they will be complimentary.

AUDIENCE: So there's no chance of them being in any other-- it's either or. There's no other hypothesis possible?

SARA ELLISON: We'll talk about this in the example I do. I'm going to do an example where each is a-- known as a simple hypothesis. There's one point associated with each hypothesis.

And then we'll talk about-- we'll analyze that case because it's very simple to analyze. And then we'll talk about, well, what happens if, actually, the parameter is not one of those two values. So we'll get to that. But, yeah, thanks for pointing that out.

A simple hypothesis, as I just said, is one characterized by a single point. So this where Theta 0 lives is just equal to one point. A composite hypothesis is one characterized by multiple points or a range of values.

Usual setup is that we have a simple null and we have a composite alternative. And, again, the usual setup is that those are complementary. So they cover the entire parameter space. Doesn't have to be that way, but that's usual setup.

So let's see just some examples, and I'll throw in a few more definitions as we go along. So let's suppose x sub i, our IID normal mu sigma squared, where sigma squared is unknown-- or, sorry, where sigma squared is known. And we're interested in testing whether mu is equal to 0.

So we could set up the hypothesis test this way where mu is equal to 0 and the alternative is that mu is equal to 1. And, in this case, the stuff at the top are the maintained hypotheses. The stuff in the second line is the testable hypothesis, or one of the testable hypotheses, I guess.

And then, the hypothesis that we're actually going to test is, in fact, the testable-- the one labeled as the testable hypothesis. We're going to call that the null hypothesis. And note that it is simple and the alternative hypothesis is also simple.

What happens if we have a hypothesis, the alternative hypothesis, which now becomes complementary? So the union between the null and the alternative covers the entire parameter space. Well, that's a very common setup.

Again, this is called the alternative hypothesis. In this case, it is a composite hypothesis. And we also call it a two-sided hypothesis. And that's basically because we're going to reject the null for evidence that mu is greater than 0 or evidence that mu is less than 0. And so, we're basically rejecting the null on two sides. So it's called a two-sided hypothesis.

What happens if we just consider the case where mu is greater than 0? That's our alternative. So we could set up a hypothesis test this way. Why would we want to do that? Well, maybe we have some kind of a theoretical reason to know that mu can't be negative perhaps.

AUDIENCE: [INAUDIBLE]

SARA ELLISON: That could-- yes, exactly. That could go up in the maintained hypotheses that mu has to be non-negative. Yeah. So, again, this is the alternative hypothesis. Again, it's composite. But, this time, it's one-sided instead of being two-sided. So then, what we do is we can either accept or reject the null hypothesis based on evidence we have from our sample.

But, obviously, mistakes can be made. So the sample is not the entire population. It's just, there can be random variation in the sample that can lead us to the wrong conclusion.

And so, in order to-- what we want to do is we want to set up the hypothesis testing procedure to control these various types of errors. And so, what we have to do first is we need a taxonomy of the errors. We need to define what the errors are so that we can control them.

So this is just a little 2 by 2 grid. If the null hypothesis is true and we accept the null, obviously, we haven't made an error. But, if we reject the null, that's called a type I error.

And, likewise, if the null hypothesis is false and we accept it, that's called a type II error. And so, we'll want to calculate the probabilities of these errors and control them when we set up a particular structure for a hypothesis test.

So a couple more definitions. The significance level of the test is alpha, the probability of type I error. So probability of type I error is called alpha. And it's also known as the significance level.

The probability of type II error is we denote as beta. And that's known as the operating characteristic of the test. And then we call 1 minus alpha the confidence level. And we call 1 minus beta the power of the test. So just definitions.

AUDIENCE: 1 minus alpha minus beta would be no error?

SARA ELLISON: Would be what?

AUDIENCE: No error?

SARA ELLISON: So alpha and beta are probabilities.

AUDIENCE: Because then you are in either the-- you're in the no error zone, right? 1 minus alpha minus beta.

SARA ELLISON: Oh. You're asking what the probability of falling in one of those-- yes. Yeah.

And, finally, one more definition. We define the critical region of the test-- and we're going to denote it C or C sub x-- is the region of the support of the random sample for which we reject the null.

So we're going to think about, when we're constructing a hypothesis test, we're going to think about what are the possible values of our entire random sample that we might get. We'll simplify it, actually, and we'll basically be talking about the values of a test statistic.

But you can think of it more broadly as the values of the entire random sample. And what is the support of the random sample for which we're going to reject the null, where we are going to decide that that's enough evidence that we want to reject the null. And that's called the critical region.

So let's do a specific example of a constructing a hypothesis test. Let's suppose that the x's are IID normal mu sigma squared. And sigma squared is known. The hypothesis, our null hypothesis that we want to test, is that mu is equal to 0. The alternative is that mu is equal to 1.

So, as was pointed out earlier, this is maybe-- this is a little non-standard and maybe a little non-realistic because we're only allowing for the possibility that mu is equal to 0 or 1. We're not allowing for any other possibilities. But, as I said before, this is a much easier case to analyze and for us to understand what's going on.

So let's first think about the case where our sample size is small. We only have two observations. What kind of sample would lead you to believe the null or doubt the null in favor of the alternative? What do you think?

AUDIENCE: If both are zeros, then it is null. And [INAUDIBLE].

SARA ELLISON: Keep in mind that these are random draws from a normal mu sigma squared distribution. So they don't have to both be exactly equal to whatever the mean is. They're from a normal mu sigma squared distribution. Yep.

AUDIENCE: [INAUDIBLE] less than 0.5 [INAUDIBLE]?

SARA ELLISON: Yeah, so you're one step ahead of me. You're thinking about characterizing the critical region in terms of the sample mean. And that's exactly the right direction to go into. But let's think of it a little bit more broadly first. Let's just think about the region of the x1, x2 plane for which we'd want to reject the null in favor of the alternative.

Basically, it's going to look like this. We're going to draw a diagonal line, which is equal to x1-- the line is x1 plus x2 is equal to some value k. We haven't specified what k is yet.

And, everything, if our sample falls above that line or to the northeast of that line, we're going to want to reject the null. That just means that our two observations are kind of large. So that's going to tell us, they probably didn't come from a mu equals 0 distribution.

And anything below that line we're going to want to accept the null because, well, those are pretty small values and, yeah, it probably didn't come from a mu equals 1 distribution.

So we'll get back how to choose k in a second. That's an important piece. But, for now, let's just say that there's some constant k. And that's going to divide the-- where our sample lives into two regions.

So we, equivalently, we can do this and, actually, well, you're two steps ahead of me. Because I'm first going to say that, equivalently, we could, instead of thinking about the x1, x2 space, we can just think about forming the sum of x1 and x2. That's going to be exactly equivalent to what I just put up.

But maybe it's easier to think about just forming this sum and rejecting for large values of the sum and not rejecting for small values. It's exactly equivalent to the picture above.

And also equivalent to the picture is, instead of forming the sum, form the sample mean. So, in this case, if we want to keep the same procedure, we would-- this k value would get divided by 2 because the sample mean is divided by n. And n is equal to 2.

So does everyone understand how these three-- so all three of these ways of thinking about defining the critical region for a hypothesis test are equivalent. And they're going to result in identical procedures.

And does this make sense? We get large values of our sample, we want to reject the null, small values, we want to accept the null.

So do we prefer one of these three ways of thinking about a hypothesis test over the others? Not really. But as n gets big, then my ability to draw an n-dimensional sample space deteriorates very quickly.

So it just makes it-- obviously if-- well, I'll just say that we'll probably want to just focus on one of these two. And, typically, we'll focus on this one. But I just want to make the point that these three ways of thinking about defining the critical region are all equivalent.

So we'll base our testing procedure on the test statistic. We'll call this sample mean now the test statistic. And sometimes we'll denote it t. And, in this case, our test statistic is just equal to x bar.

And what we're going to do is reject for large values of x bar. We're going to reject the null for large values of x bar. So, in other words, the critical region is going to take the form x bar greater than k for some k that we have yet to determine.

How do we choose k? Well, I said a couple of minutes ago, where we care about the probability of these two different kinds of errors. And, intuitively, it makes sense that our choice of k is going to determine our probability of rejecting a true null and our probability of accepting a false null.

So what have I done here? I have drawn the distribution of our test statistic under the null. So this test statistic has a normal distribution with a mean-- with a variance sigma squared over 2 and mean 0. Because, under the null, the mean is equal to 0.

And then, this is the distribution of the test statistic under the alternative. So it looks exactly the same, but it's shifted over 1. It has mean 1.

So this is an important point. Is everyone on board with these two different distributions and where they came from?

So I've drawn the distribution of the test statistic under the null and under the alternative. So the probability-- so we set some k here. And we accept the null for values of our test statistic less than k. And so, this red shaded part here is the probability of accepting a false null.

How did we get that? This means, this is how the test statistic is distributed if the null is false. And so, this is the probability that our test statistic is less than k under that presumption.

What's the probability that we reject a true null? Well, the true null means that our test statistic is distributed this way. And we reject the critical region means that-- the shape of the critical region is we reject the null for large values of x bar. So anything above k, this is going to be the probability that we're rejecting a true null.

So you can see then, what we do is, as we move k, we increase beta and decrease alpha, or vice versa. So, basically, the choice of any one of alpha, beta, or k determines the other two.

Oh, wait, that didn't-- yes, that is true. Choice of any one of those determines the other two. So if we set alpha to some particular value, that's going to tell us what beta is, given a sample size, et cetera. And it's also going to tell us where to set k.

And then, the other important point to make is that choosing them involves an explicit trade-off between the probability of type I and type II errors. So we choose k, we move it, type I error goes way down, type II error goes way up, or vice versa. Is that clear? Yep.

AUDIENCE: So when you say H0 and Ha are disjoint when they have [INAUDIBLE]?

SARA ELLISON: So the parameter spaces that characterize H0 and Ha are disjoint. That doesn't mean the distributions of the test statistics are disjoint. Yeah, good question.

So let's compute alpha and beta using some specific numbers. So let's suppose-- remember, I don't know if you recall this, but we started this example by assuming that we knew sigma squared. We've just been-- sigma squared has been floating around. But we know sigma squared. Let's say that's equal to 4. And let's let n equal 25 instead of 2.

And then we have that our test statistic has a normal 0, 4 over 25 distribution under the null, and a normal 1 comma 4 over 25 distribution under the alternative. How did I get that? How did I know that? Anyone care to explain?

AUDIENCE: You're taking a [INAUDIBLE] from your null hypothesis and you're [INAUDIBLE] is distributed [INAUDIBLE].

SARA ELLISON: Yeah, so, basically, the x bar is our test statistic. So now I'm using the notation t. But I could use x bar instead. So x bar is our test statistic. So, under the null, meaning the null is correct, we know exactly how that test statistic is distributed. It has mean 0 and variance sigma squared over n.

Under the alternative, we know exactly how that test statistic is distributed. It has mean 1 and variance sigma squared over n. So that's where these came from.

And so, then, what I can do is, I can just compute alpha and beta explicitly if I know all of these things. So the probability of type I error is just equal to 1 minus capital Phi of the critical value, whatever that is, minus 0 over the standard deviation of that distribution. And beta is something similar.

So, basically, if you plugged in different values of k into these two formulas, then you would get-- you could trace out the trade-off between alpha and beta in this particular example.

So if you have-- so if you plug in different values of k, let's suppose that you wanted to have very small probability of rejecting a true null, a very small alpha. Then what kind of k would you have? You could go back and look at this picture. Very small alpha. You'd have a large k. If you wanted a very small beta, you'd have a small k.

And so, this point here corresponds to k-- I hope I get this straight-- k being infinity-- no, negative infinity. Is that what I just said? And this point corresponds to k being positive infinity. I hope I didn't get that mixed up. [LAUGHS]

Anyhow, you can look back at the picture just to make sure that that's correct. And then, as I said, you vary k and you get this-- that's the shape of the trade-off between alpha and beta.

So what happens if, let's say, you're not comfortable with the values of alpha and beta that you have. There's just too-- the probability of a type I error and a type II error are just both too high for you. What can you do?

Well, there's only, really, one thing you can do, assuming that you've constructed your test in an intelligent way, and that's get a larger sample size. So if you increase n, you get a larger sample size, what happens is this trade-off becomes more favorable for both alpha and beta. So you can get, if you have a larger n, what happens is the denominators there change.

And you can-- the trade-off between alpha and beta comes closer to the origin. And you can do better in terms of both alpha and beta. And this is just a picture of what the alternative and the null distributions of the test statistic will look like as n gets bigger. They'll both get more concentrated around the null and the alternative values.

So a couple of notes about this particular example. Let me just say a couple of quick things, and then I'll let you go.

So what if mu is neither 0 or 1? Well, a lot of the times, we set up the hypothesis test so that the union of the two-- the alternative and the null space is the entire parameter space. So either the null or the alternative must be true. That means that one or both of the hypotheses are composite.

When we have a composite hypothesis, the test becomes more difficult to analyze because, basically, alpha and beta are no longer just functions of k, but they could be functions of also the unknown parameter theta. So that's why we start with this simple example so you can really see the mechanics of the hypothesis test. But it is not a standard setup.

So what if our hypotheses looked like this? Mu is equal to 0, and the alternative is mu is not equal to 0. We could use the same test statistic. But I will just leave you with the question, what should the critical region look like in this case?

So, remember, go back a few slides, think about how we thought about what the critical region should look like in the one-sided test in the example I just did. But what if we have a two-sided test. What should the critical region look like? So I will leave you with that thought.