MITOCW | Lec 6 | MIT 2.830J Control of Manufacturing Processes, S08

PROFESSOR: So as I was saying, what we want to do is get up through use of some of these statistical distributions for making hypotheses tests and understanding the relationship of probabilities associated with hypotheses such as a point belongs to this distribution or that distribution. And that will set the ground for talking about statistical process control and SPC charting, where you're asking the question of a new piece of data off of the manufacturing line, does that piece of data come from the in control distribution? Or does it come from some out of control distribution? So it's all about probabilities on SPC charts.

And we want to build up the rest of the machinery that we need for that today. To do that, one of the subtle things that we have to understand a bit more about is sampling and sampling distributions. And really what we're dealing with here is issues of the use of statistics dealing with observed data.

And I have this philosophical picture of what I think of as statistics meaning. And that's our real goal and statistics is to reason about, think about, and be able to argue about processes-- in our case, real manufacturing processes-- when there's uncertainty in those processes. There is noise. There's other things we don't know.

But the key idea in statistics is we are getting some evidence. We're getting some data. And what we want to be able to do is use that data to start to infer things back about the underlying population, the underlying process or distribution. So there are some preconditions in here.

A lot of what I said here is we're reasoning based on evidence from observed data. But that really means we are taking fundamentally a probability model of what's going on. And we talked last time, for example, about assumptions with normal distributions and parameters of normal distributions.

And what we're going to do today is focus a little bit more on evidence coming from finite sets of observations, drawn from that population, and then calculations we do on that-- simple calculations, like calculating the sample mean. And then we have this number, this sample mean. What's it really telling us? What can we infer back about the underlying distribution-- what the true mean of the underlying population is?

And then a little bit later, we'll flesh this out more. But already, even as we start building these simple arguments based on our data, we have an underlying implicit model of the process. It may be a purely probabilistic model, saying it has a certain mean and a Gaussian distribution or a certain mean in a normal-- or a uniform or a Poisson.

There is a model there. And so we have to keep in mind that it is only a model. A little bit later, we'll also build up other kinds of functional relationships when we get to things like response surface modeling. But for now, these are relatively simple models, mostly focused on the probabilistic or stochastic nature of that.

So here's the plan for today. What we're going to do is talk a little bit about sampling distributions. We touched on this a little bit last time when we talked about the distribution of the sum of random variables and the central limit theorem, where the sum or the average always tends towards the normal.

In some of the cases, we're going to be calculating things, like the sample s squared, the sample variants that are not going to be normally distributed. They will have other statistical shapes or statistical distributions such as the chi-squared. There will be other cases where the student t-distribution is operable.

So we want to get a sense of these sampling distributions and understand how to use those to make not only point estimates-- that is our best guess of things like the underlying population mean-- but also confidence intervals-- where, with some probability, we think the true mean lies or where, with some probability, we think the true variance lies based on one set of observations. So that's where the sampling distributions come into play.

And we'll talk about the effects of sample size on that as well as things like what kind of inferences, these point and confidence interval inferences we can make on those. And then, again, leading up towards hypothesis testing. And then really, this will be for next time. We'll dive into SPC charts.

So here's how we typically are using sampling. We have some underlying-- I'll refer to it as the population distribution or sometimes the parent distribution. It's the set or universe of all possible parts, say, coming off your manufacturing line or all possible observations. What we're typically going to do is just draw some number, some finite number, of samples, some n samples of the process output-- so some x sub i drawn from a parent distribution with some PDF p. And what we're going to be doing is calculating these sample mean, sample variants, other sorts of sample statistics.

A key point here is the underlying process. That basic variable x has a probability distribution function associated with it. This new variable x bar that we calculate, this statistic that we calculate, also has a probability density function associated with it. And it's a different one than the parent one. And so what we'll need to understand is what those probability distributions are that arise from sampling, and then how to work backwards from those to make inferences about the parent.

Now, a quick thing. I guess there's both definitions on this slide, but also a quick thing about definitions or terminology or notation that I like to use. And in particular, I'm, again, distinguishing between the population or parent distribution, and then these sample statistics. And typically when I talk about "truth" or the population as a whole, we're using Greek variables like mu, sigma, rho, xy, for the correlation coefficient. And those expectations, those different moments, are calculated over the entire population. Typically we're doing those analytically if we have a closed form description of what the population is.

In contrast, I'm going to typically use Roman characters-- x, s, r, xy, for example-- to indicate the finite sample statistics calculated from some n number of observations. And so that's when we have a finite discrete number of observations. And we have simple formulas for the calculation of those statistics.

A little bit later in the term, we will come back and start to look in particular at covariance and correlation between two different random variables, some x and y. Those are especially important when we're looking for functional dependencies. Right now, we're simply looking at one set of data or one population, one random variable x. So we'll focus on univariate stuff today.

There is a term, "random sampling," that actually has a technical definition that I want to point out that's very close to the intuitive notion here. But it actually is a little bit stronger in requirements for its definition. We said sampling is this act of taking some finite observations out of a population. Random sampling is when every observation that we pull is identically distributed, has the same PDF associated with it, and is independent from any other sample that we pull from that population.

And this would not always naturally be the case-- if you had, for example, finite populations, and you pulled out a sample, held it in your hand, recorded it, pulled out another sample, for example. Imagine that you've got a bag of 17 blue and red marbles in it. And I pull a marble out, and it's red. I hold it in my hand, and I pull another marble out.

Do you think I'm sampling from the same underlying distribution? No, because I did not replace that original marble. So now the mix of blue and red marbles is different within that bag, and the probability is different. It is not identical and independent anymore. The observation that I made first, based on the first draw, changes the probability for later draws, changes-- there is dependence as well as no longer an identical distribution.

So when we do random sampling, as I'm defining it here, and random sampling for calculation of some of these sampling distributions, we're assuming if it's coming from a finite population, you would always put the observation back in and do another sample from the same pool. Typically what you're often doing is assuming there's no connection from one to the other, and the same process, physics, is operable from one point in time to the next. So we are typically making this IID, this Independent and Identically Distributed assumption. And then we're going to, again, as I said, calculate some statistics from those.

Ultimately, when you have a sample-- sample of size 14, drawn from a big population. You calculate x bar. What do you get? A number. You get an actual number because I observed those 14 things, measured length or whatever it was that I was measuring on those.

And so a key point here is that the statistic is a function of the sample and the sample data. And so it's actually a value that you can compute. If I do that, I grab one sample, I calculate that x bar, I've got one number.

If I were to go back and draw another sample from that distribution, I get a different number. And so if I keep going back and drawing multiple, multiple samples, that's how you build up a distribution function associated with that statistic, that calculation. So that's where this notion of statistics, x bar or whatever, as a random variable also comes into play. For any one sample it's a number. But when I go and take multiple samples, multiple sets, of n, now I build up a distribution function associated with those.

I'm going to switch here to-- or do I want to? No. I'm going to switch to-- this is here on the web. I mentioned last time this very nice website. I don't even know what the acronym stands for-- this SticiGui. It's out of the Department of Statistics at Berkeley. It's got a lot of different-- I guess sort of an online course kind of thing.

But what I really like in this is the Tools tab. So if I go to that Tools tab-- let me do that-- it's got a number of these little Java utilities online. And one that I want to look at here first is sampling distributions. So let's see. Let this load. Loading up Java, here.

So here's an example of sampling from some a priori distribution. And this is actually drawing from a uniform distribution with discrete values, 0, 1, 2, 3, and 4. So that's our underlying true population, and they all have equal probabilities.

And what I'm going to do is calculate a-- I'm going to draw a sample down here at the bottom. It's a sample of size 5. So I'm going to do random sampling with replacement. So I'm going to draw five independent and identically distributed samples out of that underlying parent distribution.

And then I'm going to calculate some statistic. What I want to do is to actually calculate the sample mean. So there in blue is our underlying population. Let me take one sample of size 5, calculate the mean, and plot it. There it is. It's a mean of 1.4.

Let me take another sample. I take another sample. Do you think the value is going to be 1.4 again? It might be.

AUDIENCE: Might be.

PROFESSOR: But probably not, right? Let's see what happens. There it is-- 2.4. Let me do a few more. So the green bars are popping up, as I think I've done something like 1, 2, 3, 4, 5, 6-- something like 8 different samples, each of size 5, plotted the mean. Now to speed things up, I can keep taking more and more samples. What distribution do you think this is trending to?

AUDIENCE: Normal.

PROFESSOR: Normal. Down here at the bottom, I can take samples that are a little bit larger. Or let me take-- excuse me, we take-- the thing tells me how many samples I'm taking, so I don't have to just take one sample of five, plot it. I can take 10 samples of 5, each of 5, and plot it. So it's just speeding up my button clicks so that we can get a little bit better shape on that.

So there's the point. That's a very fascinating point. I find it fascinating that I can sample from a non-normal distribution, take the average, the sample average, x bar, and over lots and lots of sampling, I get a normal distribution.

What else? What other observations or what other points might you make about that green distribution? What do you think is true about that green distribution? There's a really important fact which motivates why we can't calculate x bars all the time and believe the numbers that come out of an x bar calculation.

AUDIENCE: It's centered around 2.

PROFESSOR: It's centered around 2. Out of the numbers 0, 1, 2, 3, and 4, what do you think the average is-- the true average? 2. So one thing that's very nice about the sample mean is that it trends toward the true population mean. It's unbiased.

That if I were to take enough samples, the average of or the mean of all of these sample averages is equal to the true underlying population mean. It's unbiased. Doesn't have a bias or delta, a fixed delta, a fixed offset error in it. It is an unbiased estimator. So I can take lots and build that up.

Turns out there's another thing that's true which I don't want to go into and don't want to try to prove. But it turns out that the sample mean is also not only unbiased, but it's also the minimum error estimator. So on average, it's the best estimator of the mean that you can use as a statistic, meaning its distribution in some sense is the narrowest. The x bar distribution is the narrowest estimator you can have for trying to calculate the sample mean based on your distributions.

Now another important thing that comes up here is at least a few of the times, I got a sample mean that was 0.6. Is it wrong? If you do just one sample, it's quite possible, out of this set of four, I drew a sample of size 5. I might have gotten a value of 0.6. That's all the data you have.

What's your best guess for the true mean of the underlying population? That 0.6, whatever that value was. But now there is some spread on it. And so if you're wise, you would also start to want to hedge your bets a little bit here, right? You want to be able to say, my best guess is 0.5.

But I think I'm only drawing a sample of size 5. So I know there is, in fact, this kind of Gaussian spread. And I think the true mean probably lies within some range of that. And so you would like to have this confidence interval idea. We'll get back to that a little bit later. In fact, there's another very nice little tool in here for illustrating confidence intervals that we'll use at that point.

I want to do one more thing, and then we'll go back to the lecture slides. One of the neat things you can do with this tool, and it's lots of fun for you guys to connect up with and play with, is you can change the sample size. Let's say you wanted a better or a tighter estimate for the x bar.

You're not happy with the idea that sometimes, with fairly substantial probability, you might be off by plus or minus 1. You have a substantial probability of estimating, say, the-- or guessing the sample mean to be more than one value away from the true population mean. What might you do to try to improve your likelihood of being closer to the true mean when you're doing sampling?

AUDIENCE: More samples.

PROFESSOR: More samples. More samples? I guess you could do more samples. But in some sense, really, that taking one sample of size 5 and another sample of size 5, that's like one sample of size 10. Larger samples.

AUDIENCE: Oh, yeah, larger samples.

PROFESSOR: Larger samples. So if I do that here, let's take-- instead of samples of size 5, let's do a modest increase first and take samples of size 10. See what happens now. Oops, let me just do-- OK, that's good.

I'm taking a lot of samples here. I've taken several hundred samples, each of size 10. And sure enough, that distribution is a little bit tighter. Let's say if I took a really big sample, sample of size 100. Yeah, looking a lot tighter.

So one question is, we know as I take a larger samples, the distribution gets tighter. One of the things we want to do is understand how much tighter do they get as a function of the sample size? So it turns out-- let me go back now to-- it turns out that if I'm sampling from a parent distribution, the variance in the estimate of that x bar, or the PDF, the variance of x bar itself, shrinks with size n. And the variance in fact scales as 1 over n. It scales inversely proportional to the size of the sample. That's true always as you take larger numbers of samples.

For this special case, if my underlying population is in fact really a true-- has a true probability distribution function that was normal, then it turns out that x bar is not just trending towards the normal, but is itself, even for very small numbers of samples, also a normal distribution. So in that little demo I showed you, drawing from a uniform distribution, for large enough n's, large enough samples, large enough numbers of samples, the mean does trend towards a Gaussian. But it's even a stronger statement, a stronger relationship, if the underlying population is itself normal.

So let's say we start with an underlying random variable, an underlying process x, that has some mean and some variance. Now if I take samples of size 1 and plot out the distribution, what do you think it looks like?

AUDIENCE: [INAUDIBLE]

PROFESSOR: Yeah. I'm just repeating. I'm replicating my underlying distribution, right? So part of the special case of a sample of size 1, if I do that long enough, I build up the same distribution. But now, if I take larger numbers of samples, even a little bit with n equals 2, again, we get that effect that we saw with the SticiGui of the narrowing of the distribution, PDF associated with the x bar.

And in particular, the PDF or the Probability Distribution Function associated with x bar is exactly normal with the same mean-- it's unbiased-- and with reduced variance. So the variance goes as 1 over n. So we start with the population distribution here, and we end up with a sample mean distribution that is a different PDF. Everybody clear on this? So key points-- statistic itself is the random variable has its own probability distribution function.

Now what we want to do is reason about the underlying population based on those observed statistics. Somebody's cell phone is going crazy. Not mine. Everybody hear that click? Can you even hear that click in Singapore? Yeah? All right. Hopefully that will go away in a second.

So once we know the sampling distribution, say, for x bar, now we can argue about the probabilities associated with observing particular values of x bar. We can make observations or arguments about how much probability's out in the tails of these things. And then we can invert backwards and reason about the actual population mean.

And again, we're after not only the point estimates, our best guess, but also interval estimates-- confidence intervals where we think the actual value is going to lie. And these are critically dependent on probability calculations of the sampling distribution. So here's an example.

So suppose that we start out with some assumptions. We start out with some a priori beliefs about the distribution of some parameter. In particular, we're interested in the thickness of some part. We don't know the mean of it.

But based on maybe lots and lots of historical data, we do believe we do know a couple of things. We know its variance. The standard deviation was 10. So let's just assume that we know the standard deviation. And we also know-- the second thing is that the thickness of these parts is normally distributed. Those are our starting assumptions. our a priori assumptions.

Now what we do is we go, and we draw 50 different random parts with the IID assumption. And we calculate the average thickness from those. And I'll tell you, of those n equals 50 samples, the actual sample mean that comes out from that one sample of size 50 is 113.5. There you go. You're blessed with that piece of data.

Now the first question here, based on what we've seen, is what is the distribution of the mean of the thickness? What is the PDF associated with t bar? Everybody should know this. What's t bar distributed as?

AUDIENCE: It's normal.

PROFESSOR: It's normal, right.

AUDIENCE: Centered around the mean.

PROFESSOR: Centered around the mean, so it would have the same mu unknown. And what would its variance be?

AUDIENCE: 2.

AUDIENCE: 2.

PROFESSOR: 2, very good. So it has the same mean, and the variance scales as 1 over n. So we had 50 samples, so the variance goes down by that factor.

One quick notation point here is when we use this notation of normal with mu and sigma squared, I try to be very consistent and put the mean and the variance in there. You will sometimes find different texts and different writers or whatever putting the mean and the standard deviation.

So you always want to confirm that, because one's a square, and one's the square root of the other. So be a little bit careful-- a little bit careful on that. I try to be consistent and have that be the variance.

So that was a first easy question. We know that based on sampling theory. We know the distribution function for the sample mean. Now the key question is, how do we use that to reason about the actual population mean? Well, it's really easy already-- the best guess.

But the more subtle question that we've been talking about is, where do we think the true mean of the population lies based on this one observation? What range do we think the true mean has with some degree of confidence? Do you think it's plus or minus 2 around that mean? Do you think it's plus or minus 20 around that mean? If I were to ask you to bet your life on what the true mean is, you would want to be able to say with some degree of confidence, it's actually within this amount of distance.

I have to say one more thing, because if I said it's within some amount of distance of that, well, with non-zero probability, that thickness could take on values all the way from plus infinity, if it's truly normally distributed, all the way to not quite negative infinity, because this is a thickness to 0. So it's still an approximate model. So if I just asked you, bet your life. Tell me where you think the true mean is, if you wanted 100% chance of saving your life, you'd say, it could be anything.

So I have to give you, when we're talking about confidence intervals, another piece of bounding information. I want the range. How far away from that one observation of the mean do I need to be with some probability? 95% confidence or 95% of the time, where do we think the true mean would lie?

What that means is if I were to go and calculate another 50 samples and calculate the mean, again, we have that distribution. And what we're looking for is that 95% central region of the PDF associated with x bar, which is where 95% of the time, the mean is actually going to lie. So that gets us pictorially and formulaically here to this notion of the confidence interval and how we actually go about calculating that.

What we're asking-- what we've got in this situation is the variance is known, so I'm not trying to estimate the variance. I'm just trying to reason about the mean. And I want to estimate it to some percent, some confidence interval. You always have this chance of being wrong when you talk confidence intervals.

You've got some alpha probability that the true meaning is even further away than you think in your interval. But you're trying to quantify that and bound that. So we typically talk about, say, an alpha of 5% or maybe 1% probability of being outside of your interval. So there's this alpha probability of error associated with any confidence interval. So that's that second piece of data I had to give you.

The first is we want to know this range-- what the size is. So the way this works is we're wanting to know, based on our calculated x bar from our sample of size n, where the true mean actually lies. So we know what we're doing is saying that the true mean, mu, is going to be bounded on the left by the x bar, but then going some portion of the distribution to the left and some portion of the distribution to the right until we get the 1 minus alpha.

So this area in here is the 1 minus alpha-- the 95%, say, central component of that distribution. And then we're evenly spreading the error part, the alpha, into 2 alpha over 2's on each side, saying I've got for a 95% confidence interval, a 2.5% chance that the true mean is a little bit further off to the left and a 2.5% chance that it's a little further off to the right. I guess in this picture here I'm doing an 80% confidence interval with a total alpha error risk, error probability, of 0.2.

And so the question then becomes, how far do I have to go out? And we know that from the basic probability manipulations from a normal distribution you guys have been dealing with already. The whole question is, how many unit standard deviations of a unit normal do I have to go? How many z's out do I have to go until I have exactly alpha over 2 out here in the tail?

So for example, we might know what this is going to do is I've got to go out 1.28 standard deviations to the left in order to be able to have just that alpha over 2 to the left of that tail, and similarly to the right. Now, notice that we're also unnormalizing. The z is the normal-- how many z's you get to, out of the unit Gaussian, the probability out in the tails.

But what we wanted to do is reason about the location of the true population. We want to know the true population mean. And so we have to do a little bit of unnormalization and say, z alpha gave me the number of unit normals. Now, in terms of my actual population variance or population standard deviation, what does that correspond to?

And this is where the sample size also comes into play. We were reasoning about the distribution associated with a x bar. And the x bar is scaled. It shrunk by that square root of n in terms of the standard deviation.

So when I expand it back out, I'm counting number of-- first off, together, this is number of standard deviations in my x bar. And then when I expand that further out to the number of standard deviations in my population, I have to divide back out by that root n. So what we've got is the rationale for being able to use the PDF associated with x bar calculate, probabilities off of the details, and get finally to this nice-- this is my fast way to erase everything-- get back to my nice distribution here or a nice formula, which you'll see in Montgomery, you'll see in all of the textbooks.

It's a wonderful note to have on your one page set of notes or cheat sheet for taking quizzes in this class and elsewhere. This is the interval, the confidence interval formula, for the location of the true mean when the variance was known. So any questions on that?

We actually want to return to our example and see what numbers pop out because I want to know-- we knew x bar was 113.5. But I actually want to know, what is the 95% confidence interval for that? And so we can simply go back to our second question.

Use the fact that we had-- you guys told me what the distribution was of t bar was our unknown mu. And the variance was scaled, 100 over 50. So now for a 95% confidence interval, what is the true mean?

So I've pictured it here. And what we're saying is we want-- we've got this red curve which, again, goes with this PDF associated with t bar. And I want the plus/minus z alpha over 2, the alpha being 0.05. That's my probability of being wrong to get to a 0.95 confidence interval.

So how many z's do I have to go out to have 95% in the center? We actually showed some examples. If you remember, last time we looked at plus/minus 1 sigma, plus/minus 2 sigma, plus/minus 3 sigma for a Gaussian. And it's actually a very close approximation.

That plus/minus 2 sigma is 95% of a distribution. That's a good rule of thumb to remember. It's actually 1.96, not quite 2. But about plus/minus 2 sigma has 95%. So you'll often see 95% confidence intervals graphically shown.

So we need about 1.96 standard deviations. Now that translates to a confidence interval that tells us, as a function of n, the distribution for where we think the true population is, based on the sample size that we had. The compression that we got because of sampling gets us that tighter standard deviation. And I've got a symmetric plus/minus 2.77 for my 95% confidence interval.

Now, notice that all you had to do here was be told what the actual calculated t bar was and what the underlying variance was and the size of your sample. I didn't even have to actually give you a list of all those values, right? But I did have to tell you the sample size. If sample size changed, that PDF would narrow or widen, and your confidence interval would narrow or widen, right? So any questions to where we are now? It's all seeming pretty clear?

So this is the relatively easy part because it's dealing with normal distributions. This notion of sampling is a little bit subtle because there is a different PDF, and you got to know how that scales with the sample size. Now I'm going to throw a few different curves at you, the different curves being different probability distribution functions than normal distributions. And I'm going to briefly cover three of them, and all three of them are ones that we actually will be using in multiple scenarios in statistical analysis and statistical techniques and tools that we're using.

The first one is a relatively easy step, and that's to look at the student t distribution. I'll come back to this. But basically, if we go back to the example I gave you. I said, we assumed we knew, based on, I don't know, lots of past history what the underlying variance was on the thickness of our parts.

What if you don't know that? What if you have to estimate that, too? Well, if you had to estimate it, you'd probably use sample standard deviation, that formula, and come up with an estimate.

It turns out when you do that, that additional uncertainty on what the underlying variance is means that the right distribution for arguing about the mean when you didn't know the underlying variance is no longer a normal distribution. It's actually a t-distribution, and we'll talk about that. It's a slightly different-- it's very close to or looks qualitatively close to a normal distribution, but we do want to cover that.

And then more have to do with not the mean, but arguing about the variance. If I calculate sample variance from a distribution, I calculate s squared using the formula for a sample of size 50, I get a number. I do that lots and lots of times. I trace out a PDF. The PDF associated with the values of sample variance calculated from that sample is a chi-squared distribution. So we'll talk about what that shape looks like.

And then we've got a variance that we've calculated from a sample. And a very strange distribution is the F distribution, which is the distribution of the ratio of two normally distributed variances or two variances drawn from normally distributed sample data. Good heavens. Why would you ever be calculating ratios of variances?

What a weird distribution. Why would you ever calculate ratios of variances? Where might that come up? There's at least a couple of cases-- one that's kind of subtle, but one that's pretty obvious.

AUDIENCE: I think it's you're thinking about the variation of the actual population, which varies from your sample.

PROFESSOR: Certainly, the variance associated with a sample of smaller size than your true population. So that's exactly true. That's one important area. The fact of sample size entering into spread and things is very important. That actually will come up more in the chi-squared.

But I think a second very obvious place is I make a change to a process. And I'm maybe not trying to mean center it. I'm trying to get a reduced variance process. I want to know, is this process better or not? Is its variance smaller?

So the ratio of those two variances are something I might be very, very interested in. I want to look at those and see, well, I did get a smaller variance. It's half as small. Do I have confidence that the true population variance is really smaller or not? And so that's where the F distribution is going to come into play. So we want to be able to manipulate and deal with that one as well.

Let me do the student t-distribution first. Actually, I can't do that. Let me do the chi-squared distribution first. For the formal definition of the t, I need the chi-squared, even though conceptually, it doesn't really matter. So let's talk about the chi-squared distribution first.

If I start out with truly normally distributed data and unit normal, mean 0, variance 1. And now, I take a sum of n of these unit normals, each one of which is squared. So each x sub i is normally distributed.

I do this weird operation where I take that sample. I square it, I take another draw or another random variable, also from the same distribution, square that, and then take the sum of n of those squared random variables to create a new random variable y. y is the sum of squared unit normal random variables.

Then I get this chi-squared distribution. The distribution of this new random variable y is chi-squared with n degrees of freedom. Good heavens, what a weird thing to be doing. Why would you be taking random variables, squaring them, and taking sums of them?

Well, think back to the formula. Let's see if I can do this. What page is that? Anybody got it there? 8? There we go, page 5. Look back at this formula for sample standard deviation.

First off, I'm subtracting the mean off of some sample. So now I've got a 0 mean variable. Now I'm taking squares of them. Well, that sounds kind of like this squaring operation. And then I'm taking a big sum of them.

That sounds a lot like this operation I was just describing for chi-squared. So this creation of a new random variable, this F squared here, is very closely related to-- that didn't work. There we go-- very closely related to the definition of chi-squared.

Now the chi-squared, the PDF associated with the chi-squared, looks kind of funky. It's clearly not normally distributed, right? It's kind of skewed. Notice it's got a long tail out here to the right for large values. Because it's a sum of squared values, it can't be negative. So it's truncated. There's nothing-- can't be smaller than 0.

Another really weird thing is that the maximal probability value is not equal to the mean of the distribution. That's kind of interesting. And there's another really interesting fact that is truly useful and occasionally comes up on problem sets and that sort of thing. The mean, the expected value of the chi-squared distribution with degrees of freedom n, is n. So as I have larger numbers of variables, the sum of that larger number keeps getting bigger. So that makes sense when you think about it.

So the point here is when we actually do that calculation of a sample standard or a sample variance or a sample standard deviation, the PDF associated with that is actually related to this chi-squared distribution. Now there were some other constants in there. They're scaling factors.

So for example, we did a mean shift x bar, but we didn't normalize to the true variance, because we didn't know it. So there is this relationship or a scaling factor before we get to the chi-squared distribution. We also had this other n minus 1 factor back on the calculation of the sample-- sample standard or sample variance. So we have tc do a little bit of moving variables around to get to a chi-squared distribution.

Another important point is that the-- let me clean up some of this-- is that the sample variance is actually related to a chi-squared with n minus 1 degrees of freedom. And I really don't want to go into a whole discussion of degrees of freedom because it's a little bit subtle. But this reminds me of another point that I didn't make back on slide 8.

Get me to 8, please. There we go. Oops, not 48, 8. Oh, it wasn't 8. Where was it? 4, 5. There we go. Back here on this, notice that when we calculate sample mean, we used 1 over n. But when we calculate sample variance, we always use 1 over n minus 1. Why do we do that?

It turns out that if you need or want an unbiased estimator for a sample variance, you need to divide by 1 over n minus 1 or divide by n minus 1, not n. Now, as n gets very large, the difference doesn't really matter. But you can go through some statistical proofs to show that the best unbiased estimator needs that n minus 1.

Now the other thing that's going on in this formula is we were subtracting off the mean. And in this case, we were also estimating the mean. So we're using up essentially one degree of freedom out of our data to calculate the sample mean, leaving us only n minus 1 degrees of freedom really in the remaining data to allow variance around the mean.

So I'm not going to go into much more detail, other than to simply say the fact is, when we're calculating sample standard deviation, we're actually calculating two random variables or two statistics, x bar and variance. And so you would need-- you essentially don't have complete independence between those two things. You use up one degree of freedom for one of those.

Let's use this. Before we use this, just to give you a qualitative feel, here's-- again, plotted a few different chi-squared distributions. When n is very small, it becomes very skewed. It's quite interesting.

Again, the mean you can see for n equals 3 here is 3. It's this blue curve. And as n increases, the distribution shifts to the right. The mean shift to the right.

But it also spreads out, which kind of makes sense. If I've got more and more random variables, and I'm looking at the variance and estimating that sum of random variables, its spread is going to get large. And another observation is that as n gets larger and larger, this also trends towards a normal distribution, which for very large n can be a useful fact.

I want to actually go in and use-- not that one-- use this chi-squared distribution to ask another question on that thickness example. I'd actually want to know, what's the best guess for the variance of my thickness of parts? And better than that, what's a confidence interval for where I think the true variance lies, based on just this one number for sample variance, based on my sample of size n equals 50.

And this is where we do the same kind of a formula for the range where we think the true variance lies, based on our observation from one sample of sample standard deviation. And this is using that relationship between the chi-squared distribution and F squared and the true underlying variance. So if you go back to one of those formulas, what I did was took-- sigma squared was lying out here. I moved it up here and divided the chi-squared down here. So this is essentially right in here that equivalence that we said before about how F squared was distributed as a chi-squared with n minus 1 degrees of freedom.

So what we've got is a bound-- let me get rid of all this gook-- a bound, upper and lower bound, on where we think, again, the true variance is, based on our calculated F squareds. And what we're doing again is putting some alpha probability of being wrong in each of the tails. I want the central part. I want the 95% central part of where we think the true variance lies.

Now an interesting point here is chi-squared is asymmetric. So if you ever see somebody going off and writing, I think the true variance is equal to F squared plus or minus 14.2, that should be a great, big red flag. It's somebody who doesn't know what they're talking about. Well, maybe they have a huge sample size, and they're appealing to a normal distribution.

But what they're probably doing here is something very wrong. Because the chi-squared distribution is not symmetric, I have my best point estimate of F squared. And then I'm going to go a different distance to the left and a different distance to the right.

So here's, still for our same example, the chi-squared distribution for n, a sample size of 50. So this is a chi-squared with 49 degrees of freedom. And again, I want 2.5% in the left tail and 2.5% in the right tail.

And so if I apply that formula, and I have to look up chi-squared with 0.025 and 49 degrees of freedom, and then the chi-squared where I need to know-- I want 97.5, everything, leaving except just alpha over 2 out to the right. The s squareds are the same in both cases. My n minus 1 is the same.

But because these values, the chi-squareds, are not equal-- whoops. I guess I got these flipped. Actually, when you look at the tables at the back of Montgomery or Mayo and Spanos, be careful on the definition. They often show you a little plot that looks a lot like this. And they shade in what their percentage points are. And sometimes they go from the right, sometimes from the left.

But the point was when you actually look that up, you get different values for the left and the right. And when you divide those out, you get a range-- get that out of the way. You get a range finally for where your true variance lies.

AUDIENCE: So is that through a [INAUDIBLE] or estimates of variance or from chi-square distribution, or is that--

PROFESSOR: The point is that all estimates-- well, it's strictly true if I'm drawing from a population that is normally distributed. But an approximation is no matter what, any time I'm calculating a variance, the variance tends to be chi-squared distributed. So it's always going to be these kinds of chi-squared calculations.

So it's not that the chi-squared was a special case. It's the PDF that you should always associate it with s squared. And notice here, we had 102.3. That's our best guess. And we had 71.4 and 158.1 for the range and variance.

I always find this a little bit shocking. A sample size of 50? I took 50 samples, right? And I had-- my underlying variance, I guess, was 100. But I took a lot of samples.

And it always shocks me a little bit how big the range is on the estimate of variance coming out of this. Here, my estimate of variance is 102.3. Well, that's at least reassuring, because that's close to the example that I gave here, where a priori, I thought it was 100. I just basically popped that out.

What's shocking is I can go down to 71. That's like 30% lower than that, or 158, which is 68% higher than my point estimate. And a really important thing just to know qualitatively is that estimating a mean is pretty easy. And actually, as sample size grows, you can get pretty good tight estimates of mean.

But the estimates of variance are hard. You need a lot of data to estimate that second-order statistic. And so we get big spreads in variance. So you've got to be really careful in your reasoning about variances. And that'll bring us back to the F-statistic a little bit later.

So let me go back now to the student t-distribution. And it has a formula and a formal definition here, which is if I start out with a random variable z, that is the unit normal. And then I divide it by a random variable that is chi-squared with k degrees of freedom, divided by k, I get a new distribution, a new variable t, that is a t-distribution with k degrees of freedom.

And it's the same question. My god, why would you do such a cruel thing to a random variable-- divide it by a chi-squared random variable and some constant k? And the answer is that's essentially what we're doing when we are normalizing data like this, when instead of normalizing to the true underlying population variance or the true underlying sample variance, I'm also having to estimate not only the mean, but also estimate the population standard deviation.

We already said, what is s? s squared is chi-squared distributed. So s is a square root of a chi-squared distribution. So buried in this unit normalization that we like to do to get to a probability distribution function-- we can talk about confidence intervals on the mean.

We subtract off some mean, and then we normalize to s over root n. But s itself is this chi-squared. So it's really closely related to the operations that we do when we are normalizing our sample data, when we also had to estimate the standard deviation.

So the way to think about the t-distribution is it's really close to the normal distribution, except it's perturbed a little bit, because we didn't really know the underlying variance. We're having to estimate it also. So here's some pictures, some examples.

The red is the unit normal distribution. And now for different sizes of sample, so for an n equals 3, you have this little blue distribution. That's the t-distribution with degrees of freedom 3.

Notice that it's a little bit wider than the normal distribution, reflecting a little bit less certainty on really the location of that random variable. Now as n gets bigger, so we've got an n equals 10 example in here in the green, the chi-square-- or the t-distribution gets a little bit tighter. And for n equals 100, it's basically almost lying right on top of the normal distribution.

So what the t is reflecting is a little additional uncertainty because we didn't know sigma squared. I had to calculate s squared from that same sample distribution. So that's all that's really going on there.

If we then say, OK, I want to get back to a confidence interval. But now, I don't know the variance, and I have to estimate that also from my data. We have essentially the same confidence interval formula, the only difference being instead of z related to the unit normal distribution, we have numbers of standard deviations on the t-distribution that we're arguing about, again, reflecting that that t is a little bit wider.

But it's essentially exactly the same thinking, just recognizing that now, the sampling distribution for x bar when variance is unknown-- is not a normal. It's a t-distribution. But all the other operations are exactly the same. We look for what alpha error we're willing to accept, what our chance of being wrong on our bounding of the interval is, and then allocating that to the left and the right; figuring out how many units normal over we go on not the underlying population distribution, but our sampling distribution. So we still get the benefits of increasing n getting tighter. But we just do that all on the t-distribution.

AUDIENCE: So this is-- will be necessary for small sample sizes.

PROFESSOR: Exactly. So the point or the question was this is only necessary for small sample sizes. And that's exactly right because of the effect that we see back with the t-distribution getting very close in approximation to the normal distribution for n becoming appreciable.

I've heard different kinds of rules of thumb. Some people like to say for n about 25, you're pretty close to a normal distribution. Some people like to draw it at n equals 40. It really depends on what kind of accuracy you're after. But you can be substantially wrong for very small sample sizes-- of sample size 5, which is a natural sample size you would often use in some manufacturing scenarios. So you do have to be aware for very small n to use the t-distribution.

This was an example where we had n equals 50 in our part thickness example. Let's see how different things pop out if we use the t-distribution or the normal distribution. So let's go back to our example. But now, let's say we don't know either the variance or the mean. Both of them are unknown.

We already calculated the sample mean. We had 113.5. And now I'll tell you-- I guess I already gave you this number previously. But I'll tell you that we apply the sample variance formula to the data, and out pops the number 102.3. So again, that's your best estimate of the sample variance. So these are your point estimates.

But now, I want to go back to the question, where's the confidence interval on where we think the true mean would be 95% of the time? Well, now we have to use the t-distribution. When we do that with 49 degrees of freedom, again, k minus 1, because we're using up 1 for calculation of the sample mean. Now we have this slightly different formula.

Here, we can use the plus/minus, because the t-distribution, like the normal distribution, is symmetric. So I've got plus or minus some number of unit, z's. In this case, it's unit t's because the operative distribution is the t-distribution. I plug that in.

Notice that for 2.5% in each of the tail, the t-distribution is slightly wider. Remember, back with the unit normal, we said 1.96 plus or minus standard deviations is 95%. For the t, you got to go a little bit further-- 2.01. Not a big difference-- 2.01.

And when you come out with that, you get a slightly wider confidence interval. I'm less confident. I got to go further to get to my 95% confidence on the range because I'm also estimating. So in this case, the difference is pretty much negligible.

And if I had a sample of size 50, I would probably just use the normal distribution. And that's a good example, showing that difference is 5 parts out of 200. It's really quite small.

One more distribution I want to mention-- we're not going to use it much here. I think I've already described it briefly-- is this F distribution. And this arises if I have one random variable that is chi-squared distributed. I take another random variable that's chi-squared distributed.

And I form a new random variable R that is the ratio of those two, each normalized to the degrees of freedom or the number of variables that went into each of those chi-squared distributed variables. And that is an F with u and v degrees of freedom.

Again, this comes up when we're looking at things like ratios and want to reason about ratios of true population variances, based on observations of sample variances. And the key place where that might come up that I mentioned is experimental design cases. So this is an injection molding example, where you might be looking at two different process conditions-- a low hold time and a high hold time.

And there may be other things varying, maybe even other variables varying, that cause there to be a spread. Or there's just natural variation in the two populations. And you might ask questions like, are these two variances different? Did I improve the variance with that process condition change?

Maybe-- maybe not. Certainly not obvious here, so you might have a very low confidence. So we're going to go and use the F distribution a little bit later when we do analysis of experiments, especially where you're looking to try to make inferences about whether there is differences between a couple of populations.

And again, because we're dealing with variances, there's a huge spread that arise naturally in these distributions, purely by chance. This is a good place to re-emphasize that a lot of what's going on here in random sampling is they're spread in the observations that you get. So here's a very simple numerical example.

If I start with a variable that is unit normal, and I'm just going to take two samples, sets of size n equals 20. So I'm taking two different samples, same underlying population. I'm not making a process change, say. I'm just taking two samples, each of size 20.

By chance, when I take that first sample size, I calculate a particular sample variance, s squared. And by chance, I calculate another one for the second sample. And if I form the ratio of those two, what typical range am I going to observe in the ratio of those two variances? For example, what ratio might I observe 95% of the time or what range? And that's the F distribution.

In fact, if I look at the upper and lower bound on the range of that ratio for a 95% confidence interval for this ratio of two samples of size 20, I can go anywhere from 2.5 to 0.4 in that ratio. That's with samples of size 20. That's a huge range, right?

Imagine, 2 and 1/2 times bigger variance over here, compared to over here. And that occurs purely by chance. So in 95% of the time, I might have ratios within that. But 5% of the time, I'll even observe ratios that are bigger or even smaller than those extremo points. So you've got to be really careful in reasoning about variances.

So we're mostly there. The last thing I want to do here is draw the relationship of some of these two hypotheses tests. And that gets us very close to some of the Shewhart hypotheses that are the basis for control charts that we'll talk about in the next lecture.

But I do want to get the basic idea in the last five, 10 minutes on what statistical hypothesis is and how that relates to some of these confidence intervals that we've been talking about. So the basic idea we've been doing with these means is we've been hypothesizing that the mean has some distribution, say a normal distribution. And then when we talked about this confidence interval, I would say, accept or reject the hypothesis that the mean was within some range with some probability.

We can extend that to asking other questions or other hypotheses, and then looking at the probabilities associated with it, and saying, with some degree of confidence, I believe the hypothesis. Or I have enough evidence to counter it. And a typical example might be a null hypothesis, often referred to as H0, that the mean is some a priori mean, some phi 0.

The null hypothesis is based on this sample, this sample that I'm drawing from the population. I have this alternative hypothesis that the mean has changed. It's no longer the same mean. Do I have enough evidence to say with some degree of confidence that the mean has changed?

And it's a little tricky because there's all these probabilities associated with random sampling. So I observe a particular value with some deviation. How do I know to what degree there's actual shift, say, in the mean or not? So let's look at this.

What we do is we form the hypothesis. We then look at the probabilities associated with the two cases, and then based on those probabilities, say with some degree of confidence, I choose one or the other. And what's important is there's always the chance of being wrong, making an error-- those alpha errors out in the tails, for example-- with that decision. So that's where this confidence level comes in.

So let's say we're looking at this test. We're asking-- the null hypothesis is I have a normal distribution with some a priori mean and some a priori variance. I'm going to draw a new sample. And based on that, I want to either decide that a shift has occurred or that the data-- or not-- that the data comes from that distribution or not.

And so what we're going to do is use essentially this same confidence interval idea and say, say to 95% confidence, 95% of the time, if my value lies in the central part of that distribution, I'm going to accept the-- well, in this case, the null hypothesis that my new sample still comes from that same distribution. So that would be my 95%, my 1 minus alpha, if alpha is a 5% error.

But if I observe a sample mean, say, or I observe a piece of data that lies out here, I'm going to reject the null hypothesis. I'm going to say instead, I think I've got an unlikely event by chance that I think instead indicates something has changed. Something has changed in the process. And we'll call that the region of rejection.

So again, already you can see one kind of error that's likely to pop up. There is a confidence interval, this alpha. There is a significance level to the test, very similar to the confidence interval idea and the alpha error associated with that.

So right away, you see there's one kind of error-- it's referred to as a type I error-- on these kinds of hypothesis tests. We're rejecting the null hypothesis out here in the tails with some probability alpha. If I observed a point out there in the tails, even if that population or that distribution is still operative, it is, in fact, true. My samples are still coming from that distribution.

But I happened to draw a sample way out in the tail. And I said, well, that was unlikely. That was unlikely in this picture. I'm rejecting the null hypothesis. I'm claiming this is evidence that something changed when, in fact, nothing changed. I just got unlucky, right?

So the first type of error that you can make is this type I error. It's also sometimes referred to as producer error, producer risk. You're the manufacturer. You reject your part because your-- or you reject a batch, say, because your sample was way out here in the tail.

You're taking the risk of rejecting and throwing away good product, even though it really was good. If I took more samples, it would go back and really indicate what was going on-- that the product was still good. So it's also sometimes referred to as producer risk.

But there's another possible error. There is an error associated with the distribution shifted or changed. I still accepted it based on a random sample from the different distribution that happened to fall in my other distribution. And that's referred to as type II error-- has a probability associated with that called beta. We've been talking all about these alphas. Well, there's also a beta. It's also sometimes referred to as a consumers' risk.

The manufacturer did a little inspection. The mean happened to fall in the region of acceptance. He shipped it. Turns out, it was actually by bad chance just happened to fall in the good region. It really is coming from a bad distribution. So let's look at that. What is this beta?

Well, for the type II errors, we essentially have to hypothesize a shift of some size, some little delta. And then we assess the probabilities that I'm drawing from the tail of that shifted distribution and just happen to fall over here in this region of acceptance for our good distribution. So this is the probability associated with our null hypothesis. This is our starting distribution.

Our alternative hypothesis here is that I had a plus delta shift in the mean. So this is our possible new operative. And in fact, for a type II error, this is actually at work. Remember, this is the region of acceptance. So I'm claiming this is good.

But if the population actually shifted over there to the right, notice off on the left here we've got this whole tail, where if I drew from the shifted distribution, I've got that tail, that lightly shaded blue tail, falling in the region of acceptance, where I would say it's a good distribution and erroneously except. And one can simply apply the same probabilities to basically go in and calculate-- just integrate up and do the cumulative normal distribution function to calculate what that tail is. So it's all the same probabilities.

So the applications of this are really going to be on-- of hypothesis testing. This would be shifts of the mean. You can start to see worrying about monitoring your process and seeing if something changed in your process, a shift occurred, and being able to detect that. And that gets us to control charting that we'll do next time. So this is all pretty much the same stuff.

And now this is a peek ahead. You'll see process control. And we'll talk about repeated samples in time coming from the same distribution next time. So we will see you on Thursday. And we will dive into Shewhart control charts.