MITOCW | Lec 5 | MIT 2.830J Control of Manufacturing Processes, S08

PROFESSOR: Last time, we started looking in more detail at some of the statistical basics. These are the basis for a lot of the tools and techniques that we're going to be learning about throughout the term, especially things like statistical process control, statistical design of experiments, robust optimization, yield modeling, and so on. And so we're going to pick up more or less where we left off.

We talked a bit about the normal distribution. And what I want to do is talk a little bit more about a few of the assumptions and why it's so common that we use it for describing some of the kinds of data that we looked at last time. So we went through a fairly substantial number of different examples and saw variation in time, variation across different parameter sets, and so on.

Just to remind us, here's the-- the standard normal is just a mean centered. So if we have x as our data, and we subtract off the mean, and then normalize to the standard deviation, we get a unit normal variable. It's another random variable z that has a distribution that is marked out in terms of numbers of standard deviation. And so this is our normal distribution.

Some nice properties that we mentioned last time are that it only has two parameters. So that completely describes the normal distribution, the mean, and the variance or standard deviation. Other properties are it's symmetric about the mean.

We actually will use that property quite a bit in terms of manipulating some of the table values that one would look up for the proportion of the distribution that's out in either of the tail. So it's perhaps obvious, but we actually do use that. We'll come back to that a little bit later.

But what I wanted to start with a little bit is talking a little bit more about this assumption, if you dive into it, that we are using a normal distribution very often. And the questions are, why? And how good of an approximation is that in most cases? When can we use it? When might we be motivated to use it?

And what we did last time is we did a couple of things where we looked at some of the data. In particular, we did histogram-- binning kinds of plots of variations. And that would often motivate, based on a general shape, that normal distribution looked appropriate. One can, I guess, do a curve fit to the histogram. Would you ever try to do that?

So imagine that you actually had, say, the tops of the bins for the distribution. So maybe I had bins like this, where sometimes I had these as values-- something like this. Now, would you actually try to do a normal distribution curve fit to that? In other words, if you said, what I'm going to try to do is minimize the errors between these points and the normal distribution, does that seem like a reasonable thing to do?

AUDIENCE: It's all driven by the size of your tails.

PROFESSOR: Yeah, there are some gotchas, certainly, with any histogram. The point was that the shape of this distribution-- if you've ever played around, especially with interactive tools, where you can bin and plot out distributions, if you were to change the size of your bins, you have this disturbing effect where the shape of your distribution sometimes changes a little bit out from under you. So if you change the bins, you may well end up with something that-- all of a sudden, this one was low, and now it's high.

And the next one is a little bit low. And this one's up here if your bins are a little bit wider. So that might be a concern, but that's actually not the point that I'm after. Would you curve fit to this distribution to fit a normal to your data?

AUDIENCE: Well, you said that normal distribution is described by a standard deviation of the mean. So you might as well just take the mean and standard deviation of your data.

PROFESSOR: Beautiful.

AUDIENCE: And use that.

PROFESSOR: Right. Right, especially-- I guess the only circumstance I can imagine where it might make sense to curve fit is you didn't have the raw data. You only had the bins. That's kind of strange. I think in most cases, you would, in fact, have the raw data. And then you simply calculate the mean and the standard deviation.

Now one thing we want to do and we'll get to a little bit today is why that's a reasonable thing to do-- to actually go in and calculate the mean and standard deviation. Why is that a good estimator for the true mean and underlying parameters for this distribution-- the true meaning of the true variance? There's other things you can do certainly to check and see.

If you had your data, and you calculated the mean and standard deviation, then you can plot perhaps your Gaussian on top of that distribution. And that, I think, is a reasonable thing to do as a quick check visually to see how well it seems to map, as well as a quick check that you had reasonable calculations, not something strange go wrong just in your numerical calculation of those parameters. Now there's a couple of other things that one can do to check quickly visually the assumption. And then there's a couple of very nice additional tools that I'll mention here for either checking assumptions or visually-- in a little bit more sophisticated way, visually or numerically check a couple of those assumptions.

But one thing you can certainly do is look at the location of your data and just do a quick comparison between the percentage of data that you would expect in different bands of this data. And we'll do a little bit more examples there so that we know what percentage of the data we expect in the plus/minus 1 sigma band, for example, or what percentage of the data we would typically expect out in the 3 sigma tails of the data. And so you can do a quick calculation and comparison of the percentage of data in each of these different bands and see, is that matching up to what we would expect from a normal distribution? This actually gets very close to the idea of confidence intervals that we'll formalize a little bit more.

Now there's a couple of additional things I've listed here. One is you can look at the kurtosis or do a quick calculation of kurtosis, which is a higher order statistical moment than the mean or the variance. In fact, if you look at the definition of the kurtosis, it's an expectation of the fourth moment. Or it is a calculation-- a normalized version of the fourth moment.

And for a perfectly normal distribution, this kurtosis value would be 1. And then as the distribution changes its shape, either gets more peaked or less peaked following other distributions, other common distributions, then it starts to deviate substantially from k equals 1. And in fact, this is a quick little tool to use sometimes if you're not sure, well, number one, if it's normal. And number two, if it's not normal, what distribution might it follow?

If you look here, this is a nice plot, although I didn't break out what all of these different distributions are. This is just a plot normalized to standard deviation of the data of a set of different distributions. And the black one here is n. So this is-- let me do the black one here, right?

This is the n distribution. That's our Gaussian with a kurtosis. Well, I guess you got to look a little bit carefully at the definition. Actually, I think if I go back to the previous page, which is one that Dave had, this definition for sample data is essentially, as n gets very large, this subtracts off 3.

So that I believe then, in this case, this kurtosis for a normal distribution is actually more like 0. These two definitions you might look up. I'm not sure are exactly the same. Rarely would you actually use this one.

You're going to actually use this definition, which basically subtracts off a value. This goes with the plot on the next page. So they are slightly different definitions, I believe. So in that case, that's subtracting off a 3. For the normal distribution, it ends up with a value of about 0.

Now what's nice is as you get some of these distribution, such as the Laplace distribution, this very peaked one right here, the kurtosis value goes up. It's an indication of a more peaked distribution. The logistic distribution, which we might talk about a little bit later-- it's one that comes up occasionally with some quality or discrete kinds of distributions-- has a kurtosis of 1.2.

And the interesting one here also is the uniform distribution, which is less sharply peaked than a Gaussian. And it actually has a negative kurtosis with that subtraction of the 3 off it at the end. So you might find that as a useful tool.

I've rarely used kurtosis actually as an indicator. But I want to mention it to you because it is out there as at least a hint at looking at some different distributions. A more useful tool-- yeah, question?

AUDIENCE: So there's two different formulas? Because--

PROFESSOR: Well--

AUDIENCE: What you said or--

PROFESSOR: Yeah, so this is for sample data. And I think if you were to actually go in-- I mean, essentially this-- I have not checked this. This was some definitions from previous class notes.

I do believe when I did a quick lookup on what kurtosis is, I believe this is a better definition in terms of actual calculation formulas that you can use for calculating it. This is to give you the sense. I mean, it's sort of lurking in here.

You can see the expectation operation down in here, and then the normalization to the standard deviation. In this case, this has to be your calculated standard deviation. This is the abstracted one.

So if you actually poke around, you will find in the literature more than one definition of kurtosis. My point was that this is what I would use if you want to use the plot on the next page in terms of coming up with a number that might also indicate if there's a different distribution that you might look at. So it's related to the fourth moment.

A more useful tool-- and this is one that I actually do use-- is probability or quantile-quantile plots. And there's a section in Montgomery on that, as well as different toolboxes. We'll be able to generate these things. And so here's an example for a quantile-quantile plot.

What I've started doing on the lecture notes on the web is put up an early draft as early as I can for the next couple of weeks of lecture notes. But then as I'm editing and adding in, I'll have the most up-to-date one. So if you've got slides, you may be missing a couple of these. If you printed them out before 9:00 or 10:00 PM last night, I think these got updated about that time. So this plot, for example, was not in the early draft of the slides. And I'll try to indicate that with a little "draft" on the web page if they're still early drafts of the slides.

So what are quantile-quantile plots? These are a little bit subtle in terms of explaining. So let me try to give it a shot at explaining. And then if you have questions, let me know. And then normally, it's going to be generated by your statistics package. There are hand ways to do it, and I'll refer you to Montgomery for practice with actually trying to generate them if you had to by hand.

But here's the basic idea. What we're plotting is the actual data that you've got. And in the y-axis, you'll be plotting your data in terms of normalized distribution. So you would normalize to the mean or center to the mean and then scale it to your standard deviation. So think of these as unit standard deviations. So you simply find that as your y location for your data.

Then what you're plotting on the x-axis is the normal theoretical-- I'm not sure I'd use quantiles here-- but your normal theoretical standard deviation for that number of data points that you would have had and the location for each of those data points. So imagine this is 50 data points. I'm not sure exactly how many data points this is.

If you were to take 50 data points and draw 50 data points from a normal distribution or order them and put them where you would expect on a normal distribution, what you would have is many more data points near 0. And as you get further and further out, 1 out of 50 times or 1 out of 25 times, you would expect to find a data point about whatever it is-- 2, 2.1 standard deviations away. In other words, if I were to compare the actual location of that data point in terms of its value within my sample distribution of 50, compared to if I just drew randomly 50 data points, that would be its location.

Then what I can do is plot that coordinate for that data. So what you end up with is taking all of your data, if you will. You sort it from low to high.

And then starting at the center, in some sense, you start working outward from the center, ordering the data from the location of its index in your sorted data from number of standard deviations away from the mean that it would be, compared to how far that data point actually was away from your sample mean. And what that gives you, if it were perfect, and there was not any sort of noise in your data, that would give you this perfect matching line. Every data point falls where you would expect it to.

Now in your actual data, you're going to see some deviations from that. But what this is basically doing is a compression of your data or an expansion of your data out in the tails, but a compression of your data near the center to be able to basically tell you how closely your data here is following the assumed distribution. And for this case here, we plotted the location of 50 data points, assuming it was a normal distribution. So that's where my x values were coming from.

And as you can see here, the data pretty much nicely follows this distribution. You get a few little things that look like it's wandering or trailing off a little bit. And then you also often look out here in the tails.

And you find even out here for over two standard deviations away, it looks like I've got pretty good fidelity to those tails. I might have values that are a little bit further away from the mean than I might expect from a normal distribution, but it's pretty close. So this is the kind of plot that you would expect to see for data that, in fact, followed a normal distribution.

All right, so I know that's confusing. Are there questions that people have on what this--

AUDIENCE: Yes.

PROFESSOR: Yeah?

AUDIENCE: I have a question. So for each point, you get the y-axis by the sampling value from your data.

PROFESSOR: Right.

AUDIENCE: And how do you get x? Do you get it based on the probability of that, simply pulling from your sample distribution that you referred to the theoretical normal distribution with the same probability, then you get the y-- x-axis?

PROFESSOR: Yes, very, very close. So for the y-axis, you've got it exactly right. For the x-axis, what's interesting is you don't actually use the values of your data. You just use its index location in a sample of the size that you've got.

In other words, if I had a million points, I would look at the lowest. And I would expect that to be-- in a normal distribution, I would look at where the probability, the number of standard deviations where 1 out of 500,000 points is that far away from the mean. So I would look up the inverse probability on a normal distribution of being-- of where 1 in 5-- 500,000-- point-- what? 0.02 to the whatever. So I basically look on a tabulated normal probability plot, going backwards from where that index-- my smallest point was. And then I could do that for every point in my sample to figure out what the probability for its location should be on the x-axis.

So here's another example. Maybe this gives you a feel because these q-q plots-- the quantile-quantile plots-- can actually be used with other distributions as well. They are not always q-q norm plots. They can be applied to whatever assumed probability distribution you might want to investigate.

So here's an example where we again took the data. But in this case, the theoretical quantiles are actually lining up. I'm assuming a normal distribution. But in this example that I'm showing here, the data actually came from-- let me erase. Let me get rid of all this. The data actually came from an exponential distribution.

So this is an example where I would have assumed things were coming from a Gaussian. So this is still for the normal quantiles. But with an exponential and e to the minus x or an e to the x kind of distribution, what you end up with are a lot of data values that are much larger, much further away from the mean, than you would expect from a Gaussian. And you also get a bunch of data that's much larger than you would expect from a Gaussian. So this would be an example here, where the normal q-q plot doesn't seem to match up. It's telling me my data really is not following along the normal distribution line.

Now, I didn't pull a plot. But one could then ask the question-- maybe you'd look up kurtosis. Or maybe you look back at your data and say, I think maybe an exponential distribution is really what this is following. How would you plot that on a q-q norm?

AUDIENCE: Question.

PROFESSOR: Yeah?

AUDIENCE: Why doesn't the line go through 0, 0?

PROFESSOR: This is a good question. These don't appear to be mean-centered to me. So there's something weird on the plot.

AUDIENCE: So the line should be for a normal distribution, not fitting.

PROFESSOR: Yeah, this does not-- I think what has happened here is these are not quite mean-centered and normalized because-- well, so in terms of 0, 0 following on the plot, that that's not happening. So I'm a little-- I'm not sure exactly what's going on there.

AUDIENCE: We need the closing function taking the mean of the data et cetera. It's a conceptual normal on the data mean.

PROFESSOR: Yes, it should. And that's what I'm saying, is this plot I don't think is correctly mean-centered because it should then-- 0, 0, by definition, has to fall.

AUDIENCE: Right.

PROFESSOR: Oh, that's what you're saying.

AUDIENCE: No, I was saying you could take the mean of the data you send to the normal that you're plotting is aligned with that data.

PROFESSOR: Right. But I'm saying here's my y data, and my 0 mean is not-- I don't have any negative-- I don't have any data lower than the mean. And therefore, that doesn't make any sense. So this is not mean-centered correctly.

AUDIENCE: It looks to me like the mean of the data does give us a slightly less than 1 in the point data, so that coincides with the mean.

PROFESSOR: But if I mean-center and scale to 0, then the mean of my data have-- by definition, that ought to be at 0, right?

AUDIENCE: Oh, I see. I don't think you're shifting the data, though.

PROFESSOR: When you mean-center, yeah, you're shifting.

AUDIENCE: Oh, I think you're shifting, but I think conceptually, you're not shifting the data. You're shifting the normal, that you're saying my might correspond to the data.

PROFESSOR: No. In a normal-- in the standard q-q norm plot, you mean-center. You actually take your data, you mean-center it, you normalize it to the calculated sample standard deviation and plot that. And that does not look like quite what they've done here. I think these are still normalized to standard deviation, but I think it's not quite mean-centered.

But in some sense that doesn't actually matter in terms of the data following along the line. It's still indicating. That would just be a shift. That would be a shift.

AUDIENCE: You said the data hasn't been normalized or hasn't been mean-centered. But if it's an exponential distribution, can you still normalize a bit?

PROFESSOR: In this first use of such a plot, you would be testing the question. Did your data-- you don't know yet that it's exponential. You just have data, and you're testing. Does it fall on the normal line? So you would still follow that procedure. We'll look at an exponential distribution in a minute.

And of course, every distribution has a mean. So you can always mean-center. Similarly, every distribution has a variance that you can calculate. The neat thing about the exponential is the mean and the variance are the same.

But that's not entering in here. There's something else weird. So there's the risk of pulling a plot off at 9:50 at night. I hadn't noticed that the-- it doesn't look correctly mean-centered.

But the additional point I wanted to make is I could actually take this same data. I could produce a different plot, not a normal q-q plot, but an exponential q-q plot. And if I were doing that, in that case, what I would do is take my data, still plot it hopefully mean-centered, and then number of standard deviations away.

But then along this axis, I would calculate the location in numbers of standard deviations based on the probability of an exponential distribution, not based on the probability of that index location in a normal distribution. So I would basically say, for my 50 data points, I expect the 25th data point larger than the mean to occur in that distribution. I have to go 2.1 normalized standard deviations away in order to get to that probability. So that it takes my same y data, but then it plots along the line, where if it really is exponential, my data should follow along a 1 to 1 correspondence line.

So you don't often see the use of these q-q plots from the perspective of different distributions, but you can use them. What you often will see is really this. You'll see q-q norm plots. And they're lovely plots. It's a wonderful tool to do-- use-- because you're actually seeing all of your data. It's got all of your actual data.

It's showing you that it corresponds roughly to a normal distribution. It's also giving you very nice information about essentially your variance or standard deviation. And there are variants of these plots that you will often see in the literature, especially the semiconductor literature, dealing with large numbers of samples coming from different kinds of measurement.

So for example, if you want to make contact resistance measurements for literally thousands of contacts and very succinctly present that data, you will see families of q-q norm plots. So for example, maybe you did a bunch of contacts at a particular size. You would plot them like this.

And then maybe you had another data set, where you had attempted to pattern those contacts slightly larger, slightly smaller. And you would often see then another-- oops, that's not very straight, is it? It's meant to be another underlying set of data. But you might find your data looking something like this.

And that kind of plot is really useful for showing that there is a mean shift, a mean difference between your data. But also, the variance is different in the two cases. Now, exactly what you're plotting here might be a little bit different.

You might actually not plot quite normalized data. You might actually use it in an unnormalized fashion. Here, you might plot this not in terms of standard deviations, but rather actual-- keep it in the quantiles or the probability of being that far away-- probability of that x-- that's weird. The probability of that x value.

So for example, you will often see these kinds of plots which would show things like 0.001, 0.01, 0.1, 1, or something like that, getting up to-- I guess 0.5 would be the equivalent for the mean. And then you start going larger-- 0.9, 0.99, 0.999. In other words, you might actually plot as-- I should have put these on the x value-- the probability that you would find a data point that far away as opposed to implied probabilities in terms of number of standard deviations.

So there are some really cool variants of these plots that are very useful. And I think we'll see some of these when we talk a little bit about yield and some other distributions.

AUDIENCE: I have a question?

PROFESSOR: Yeah.

AUDIENCE: Yeah, after I have the q-q plots, how can I tell the confidence level that I have to say whether or not my data is normally distributed?

PROFESSOR: So the q-q plot does not actually tell you confidence intervals on either the hypothesis that it's normally distributed or confidence intervals on the parameter estimate. There are some formal statistical tests where you can test that hypothesis of normality. And essentially, you can use those from your-- never going to hand-calculate some of those statistics, and then the probability associated with a derived statistic associated with normality.

You'll use your statistics package for that. This gives you a good visual indication. But to actually test, is it normal? Or what is the probability that the data is non-normal? That's a different question. And then today, we will start talking about confidence intervals on the mean and the variance, which you also would not use the q-q norm plot to generate. So in fact, let's get to that because that's-- yeah?

AUDIENCE: For that plot, can you use regression to see how far it is from the normal?

PROFESSOR: Well, first off, again, if you were actually trying to estimate the parameters of normality, you would just use the data and calculate the sample mean and sample standard deviation. I think if you are-- essentially what you are posing here is, could I go in and look at these deviations and do some, I don't know, sum of squared values of those deviations? That's actually getting really close to calculating a statistic.

Call it a W or some number, a W statistic that I would form based on sum of squared deviations on one of these plots or some other-- maybe it's a sum of absolute distance deviations. Now I've got a statistic W, and that's getting really close to the kinds of statistical tests that one would run to ask the question of normality. I don't actually know what the formula is used in coming up with a W value and then what the normality tests are.

But that's the kernel of the idea, is to actually look at your data, form an aggregate value for that statistic, that W statistic. So for example, if it was sum of absolute values, and it-- for a sample of size 50, and that W is very near 0, then you have high confidence that it's a normal distribution. But as W gets bigger, that would seem to indicate more and more likelihood that it's not normal. And that's exactly the kind of thing that's going on in the formal statistical test for normality.

So here, we've given you a few tools for being able to look at the data, get a feel for is it normal or not. But it hasn't answered the question, how come so often we're using a normal distribution when we're actually looking at manufacturing data or other kinds of experimental data? And a really important thing is the following observation-- the following fact-- that if we are forming a sum of independent observations of a random variable-- so x has some underlying distribution.

And it doesn't actually matter what the underlying distribution is. But I form n independent observations of that random variable. And then I look at the distribution of the sum of x1 plus x2 plus all n random variables. The fascinating fact is that the sum of independent random variables tends towards a normal distribution. This is the central limit theorem.

So here's a neat little example. If my underlying distribution is in fact something like a uniform distribution, and if I'm, say, pulling off 20 samples of x1 and 20 samples of x2 from a different uniform distribution, and I form, say, 100 samples of, I guess, 100 sets-- each one of these is, I guess, 1,000 points in this example. But I essentially take the sum of all of these random variables and form a new random variable. The new random variable tends towards a normal distribution with some mean and variance.

Some of you I saw in 2853. And I had a nice link to a website. And I'll actually dig that up and post it for this class. It's the SticiGui website. It's a statistics-- interactive statistic package out of UC Berkeley. And it's really fun. You can actually form these kinds of sums of random variables out of different underlying distributions and plot them and start to see how close the sum or the normalized sum of these distributions are to a normal. So there's some very, very nice interactive tools that you can play with.

Now, an important point here is if I'm calculating the mean-- so I'm calculating an x bar across my data. And I've got 100 samples, each drawn-- and I'm assuming I'm drawing it from the same underlying distribution, whatever that may be. What is the distribution of the sample mean?

Well, if you look at the formula for the sample mean, it's not exactly a sum of your data. It's the sum of your data, then divided by 1, right? It's summed from i equals 1 to whatever n is of your individual samples. So it is a sum, and then with a constant out front.

But the point is, by appealing to the central limit theorem, the sample mean distribution, the PDF associated with a sample mean, always tends towards the normal distribution. So we're going to come back to this idea of sampling and what the distribution is for sample statistics a little bit later. But more generally, very often what we're doing is pulling data out of a process that in itself is already, by the physics of the process, highly averaged. And therefore, it's averaging lots of perhaps other underlying strange physics or difficult physics.

But in aggregate, that averaging nature of the data itself-- not the operation that we perform, but each individual underlying data point-- each individual x sub i-- underneath of that may have some averaging by the physics going on that will help to drive it towards itself being a normal distribution. So just to remind you, the central limit theorem is probably the most used and perhaps often abused appeal to why we're using normal distributions very often. It is still good to test it. But there is a good reason why very often, our data does come up as normal distributions.

So I want to talk a little bit now about sampling because we are very often using actual measurements and data to try to get estimates for, or more generally, build a model of our random process and estimate parameters of that random process. And we've said in general, p sub x is unknown. The data-- always plot your raw data first and foremost.

And very often, the raw data will suggest a distribution. Or then histograms may provide some insight. So for example, a very quick histogram will very often give you the difference between a normal distribution and a uniform distribution. If it's evenly falling, and I don't have this falloff in the tails, that's very important. And then we can also use things like the q-q norm plot to test some of those things.

So the first job is to come up with what likely [COUGH] distribution you want to use. Nine times out of 10, normal distribution will be appropriate. And then the second thing is to estimate parameters of the distribution.

And the normal distribution, again, to remind you, just has these two parameters, mean and variance. And now what we want to do is estimate them. Now, everybody is used to the formulas. We've got the formulas right here for calculating from your sample, your limited number of pieces of data, what things are-- what a few important statistics are or characteristics are of that data, like the sample mean or the average, and the sample variance.

But what I want to give you a feel for today, perhaps the most subtle idea, an important idea for interpretation, for establishment of confidence intervals, for actually being able to say where you think the real values lie-- the subtle idea is that these themselves are statistics that have their own PDF, their own Probability Density Function. They have a sample statistics that fall that tell you the likelihood of observing particular values of them-- that establish bounds for where, if I had a different sample, how close you think the new sample, still drawn from the underlying parent distribution, would actually lie to the particular sample that I just drew.

So I'm going to explain that in a few more slides or several slides here. But the key idea is it's really easy to calculate a couple of these moments-- the mean and the variance. For the normal distribution, that tells you everything for an estimate of your raw data. But then I want to get to the more subtle idea so that we can start talking about things like confidence intervals.

And a simple example to give you a little bit of a feel for this here is if I were to ask you what distribution applies to the sample mean, where does that come from? Where does this notion of a distribution associated with the sample mean arise? So if we look at the formula for the sample mean and expand it out, in some sense we've got just a sum of independent random variables, like we were talking about with the central limit theorem.

There are different constants in here. And in this case, for the sample mean statistic, all of the constants are the same, which is just 1 over the total number of data points or sample points that I've got. Now, you can go back to the definition of expectation that we talked about earlier and do the expectation operator across this and do expectation math.

So the expectation of ax is equal to just that constant times the expectation of the underlying random variable. So the 1 over n simply comes out to the left. And if I were to ask, what is the mean of the PDF associated with x bar, it is going to be 1 over n-- the same mean.

Now what else is going on here is if you look at the standard deviation of x bar-- I hope you guys can see that. There's a variance of x bar in here. So that's an x and a bar, which I just-- the pen doesn't line up exactly with the screen.

You can also do the expectation operator for-- oops, not the expectation, but the variance operator. And if you do the mathematics on variance of some ax, that's equal to a squared, the variance of the underlying variable. And if you follow that math through for the definition of x bar and relate that to the variance of each of these x sub i's, what you find is that the variance-- I get an n times-- I'm summing n of these random variables.

So I've got n times-- 1 over n is the constant in here. So I get a 1 over n squared times the underlying variance of my x. So that I get a cancellation, and the variance then of my x bar is just equal to what I've shown here, a 1 over n of the variance of the underlying distribution.

So what's interesting here is if I start to ask about the distributions associated with what are the mean and the variance of the normal distribution associated with x bar-- what is the mean of an x bar that I would typically observe from lots of samples of my underlying distribution? What is a variance I would observe? It's related to the underlying distribution, but it's not exactly the same.

I've got a new random variable, an x bar, that has a different mean and variance. It's got the same mean in this case, but the variance is actually scaled. And this is extremely useful because the variance of my averaging means that I'm getting a tighter distribution-- a narrower or smaller variance compared to the underlying distribution. I'm going to show you that in a little bit more of a graphical fashion a little bit later because this is-- that's a preview to this whole idea of sampling, which is really critical. We've already talked about this.

So the key thing here is to get to this notion of sampling distributions, what are the key distributions arising from the fact that I'm drawing multiple pieces of data from a parent distribution, and then calculating things about that? So we'll get to some of these key distributions besides the normal distribution. We'll actually talk about these next class.

But what we want to do is go back and get a little bit more feel for not only the normal distribution, but a few other distributions that often arise in manufacturing, and then also start talking about these notions of where the data actually lies. What are the probabilities of data falling out in the tails? And using that then to start to get towards the idea of building confidence intervals and where we think the real mean of our underlying parent distribution sits. Next class, we'll also get to hypotheses tests, which arise naturally and actually start to get really close to statistical process control charting, which is one of the fundamental tools of manufacturing control.

So what I'm going to do here is go back-- this is the plan for the next-- the rest of today and starting into tomorrow. We're going to go back, just remind you of some of the discrete variable distributions, then talk about some of the-- which are more applicable to attribute modeling or yield modeling, sort of discrete things. Then we'll come back and talk a little bit about the continuous distributions, and then also touch on how you manipulate some of these distributions.

Discrete distributions-- people seen the Bernoulli distribution before? Good. This is like the simplest distribution-- the very simplest. You do a trial. You do an experiment. Can only have two outcomes, success or failure.

You get to label what success is. We'll label a success with the random variable taking on the value of 1 and failure taking on 0. I could flip that. You can start to see already a little bit of inkling of yield in here. Does the thing work or not?

The very simplest, coarsest, crudest kind of model for functionality, and the probability or statistics associated with that is, does the thing work or not? And often, we talk about what is the probability that the thing is functioning at the end of the line? Maybe that's 0.95. So 95% of the time, I think I've got yielding parts out.

For any one experiment, one outcome, I've simply got a p and 1 minus p probability associated with that. And the PDF can be expressed as shown here. Now we can go in and use our expectation operations for discrete random variables and calculate what the mean and the variance are. And those have nice, closed form functions for those two outcomes. So that's the Bernoulli.

Now the second easiest-- although it can actually look a little confusing at first glance. But the second easiest distribution is the binomial distribution because it's saying that I'm simply taking that success or failure with a fixed probability p and running repeated trials of that. So now I'm flipping my coin, say, which has-- perhaps it's a weighted coin, and it comes up heads with probability p that's not 0.5. Maybe it's 0.9.

But now I'm doing that repeated times. I'm doing that n times. Now what's the probability of having n successes? Or let me state that again. What's the probability of having x successes when I ran n repeated trials? So n is the number of trials.

So if I ran 100 trials, the probability that I had exactly x equals to 7 successes is given by this formula, here. And you can actually see this lurking in here. How do I have 7 successes? Well, that meant p, the probability of having a success, had to come up exactly 7 times. And the rest of the times-- if I was running 100 trials, the other 93 trials all had to be failures.

So I've simply got the product of all of those probabilities. And then we've got the combinatorics, the n choose x, which tells me how many different orderings could have occurred by which I would get the 7 successes and 93 failures for n equals 100. So that's simply the different numbers of combinations that can come up with that.

So the notation here, by the way, that we would often use-- and I already snuck it in some other places-- is this little tilde symbol here we're using to read as "is distributed as some distribution." And I'm using the big B to indicate the binomial distribution, which has associated with it the underlying Bernoulli probability-- success for any one trial-- and then the number of repeated trials. So this is a discrete probability. What's the probability that x could take on 0.7? 0, right? It's the number of successes out of this.

And here are some examples that just give you a little bit of a feel for what the Bernoulli distribution looks like. This is the number of successes plotted as a histogram for some values. I think that this is-- if you try it, I think this is a live spreadsheet. So actually, if you double-click on this from your PowerPoint, it may bring up the underlying Excel spreadsheet.

So you can actually play with some of the parameters in this. I don't remember what either p or n was for this. But you can start to see, it's really-- it does not look quite normal because you can never have negative numbers of successes. It's always truncated.

And you get these very non-normal kinds of distributions. This is a binomial distribution. But its location and its shape can change somewhat as you play with p and n. By the way, up here-- this is just the cumulative probability function, just saying the probability that I've got x less than or equal to some value. So that's also shown.

So then this is also in this histogram, normalized to the fraction of products. And so now, you can start to look at calculating. If this were my data, and I simply-- it was actually coming from a line where I was looking at the probability of any one part succeeding or not, I could start to ask questions about the probability of seeing, out of 1,000 products coming off the line, some number of defects or some number of failed products. You can appeal to the binomial distribution for that. Now this is all still pretty coarse, right? It's just a very simplified model-- failure or success for yield.

Now another discrete distribution is a Poisson distribution or also sometimes referred to as an exponential distribution, although terminology there sometimes varies, depending on whether people are including this component or not. But the formal definition for the Poisson distribution is shown here. Now it continues to be a discrete distribution.

So I'm asking, what is the probability associated with observing x taking on actual discrete integer values? But this is a very nice distribution associated with kinds of operations that many of you saw in 2.850 or 2.8-- yeah, 2.853. The arrival times in queuing networks will often be Poisson distributed.

But it also can come up when we are dealing with very large numbers associated with the binomial distribution as a very good approximation to the binomial. And this turns out to be really nice, because if you actually go back to the binomial formula and try to calculate it for situations where, say, n or x are very, very large, or p or 1 minus p is very, very small or very large, very close to either 0 or 1, you end up with some problems, some numerical problems. Because if you actually try to calculate it for, let's say, p is equal to 0.0001, or maybe 1 minus p is equal to that.

Let's say you had really, really high yield. And I take that, so if that's 1 minus p-- and I'm doing this for a sample of size a million. I've got 0.0001 to the one millionth power. And numerically, you start losing the digits. You can't hardly keep track of that.

But I might be asking, what is the probability of some substantial number of failures? And this, the combinatorics, end up being a really, really large number. So overall, the overall probability of seeing 10 failures out of a million parts might be substantial.

But to calculate it, you can't do it numerically, because I've got a huge number times a really small number. I get overflow or underflow. And I can't actually calculate it. What's useful is in those kinds of situations, where, say, n and p together-- the product of those things-- are reasonable-size numbers, then the Poisson distribution is a very, very good approximation.

And this applies to things where you have very, say, low probability. So p might be very small. But I'm asking-- or I have many, many opportunities to observe that very low-likelihood event.

So an example here that comes up in semiconductor manufacturing are things like the probability of observing some number of defects on a wafer. The likelihood of seeing a point defect on any one location is very, very, very small. But I've got lots and lots of area on the wafer-- lots and lots of opportunity for the appearance of that small defect. And so you can start to talk about the product of those things or a rate per unit area that starts to become reasonable.

Another example is the number of misprints on a page of a book. You don't expect for any one character on a book for that to actually be a misprint. But over the entire aggregate number of pages in your book, you expect some number of misprints. And the statistics that go with that are typically Poisson distributed.

And I already mentioned that the mean and the variance, if you actually apply those formulas to this distribution, come out to the fascinating fact that they are numerically the same value. By the way, units-wise, they're not. But x is an integer and-- oops. That should be x, by the way. Come on. Cut that out. There we go.

So here are some example Poisson distributions. You can start to see one here for a mean of 5. It looks close to the binomial distribution that I showed you earlier. And then as the mean here is increasing, and the lambda parameter, you can start to see this distribution shifting to the right.

We said lambda is the mean. It's also a characteristic of the variance. The variance is also equal to lambda. So that will also broaden out for larger numbers of-- or larger values of lambda.

There's another observation in here which is useful. What are they starting to look like for large lambdas?

AUDIENCE: Normal.

PROFESSOR: Normal, right. If you looked at that, it doesn't look very normal distributed. It's truncated. It's a little bit skewed. But another approximation is for large lambda, that also tends towards a normal distribution. So very often, you've got this success or succession of approximations, where you might take a binomial, approximate it as a Poisson. But then for large numbers, a normal distribution also can be a useful approximation.

So let's go back to the continuous distributions, the normal and the uniform. And here, I want to start getting to actually how you use or calculate probabilities of observations in certain ranges and in particular things, like the probabilities of observing things out in the tails. So here's a continuous distribution that has a probability density function.

This is the normal-- excuse me, the uniform distribution that has the same probability density for values in some range. And then I've also indicated with a capital F our cumulative density function for that. So this is just reminding you of a little bit of the terminology there.

But I'm highlighting the uniform distribution because there's a couple of very standard questions, that if you have a known PDF or CDF, these are the kinds of questions that you're going to be asking again and again and again. And they're nice and intuitive off of the uniform distribution. When we get to the normal and other distributions, they're not quite as intuitive. But seeing them here for the uniform first, I think, helps.

One of the typical kinds of questions is I want to know, what is the probability that some x is less than or equal to some value if I were to draw it from this underlying distribution-- from a normal distribution? And so one could ask that using either the PDF or the Cumulative Density Function. And sometimes, one or the other, if they're tabulated or available to you, is easier to use.

Clearly, if this is a Probability Density Function here, I can ask it in terms of the interval question. Oops, excuse me-- the interval question right here, and say, well, the probability that x is less than or equal to that x1 is simply the integration up of that probability. And you can do that numerically or just by hand on such a simple distribution.

But the point that is actually exactly the value that is tabulated on the Cumulative Density Function. That's the definition of the Cumulative Density Function. So if you've got the CDF, you simply look it up and say, what is f of x1 equal to whatever your value is for that probability function?

Now similarly, you can also ask the question, what is the probability that x sits within some range, say, between x1 and x2? And again, you can do that either off of the underlying density function, just integrating and saying, yes, x has to lie between those values, and integrate up the density. Or you can recognize that the probability that x is less than x2 is simply that value and subtract off that the probability that x was less than x1 is that. And so therefore, the difference between those two corresponds to the integration on the underlying Probability Density Function. So that's pretty easy. That should be pretty clear.

Let's talk about that also for the normal distribution because some of those values are not as easy to integrate up by hand. In fact, there exist no closed-form formulas. But they are tabulated for you. And that's where going to the table on the normal distribution for things like f of x are going to-- is an operation that you will actually perform quite a bit when you're manipulating normal distributions.

So here's another plot. We've already talked, or I've shown other examples here of the normal distribution. I've tagged off on this plot for us a few useful little numbers to have as rules of thumb. This is actually, I think, a moderately useful page to print out and have off on the side for your use.

In particular, what I'm showing here is for the normal distribution you've got a formula. You're hardly ever going to actually plug in values for the formula. But if you look out plus 1 standard deviation, plus 2 standard deviation, on the PDF, I've tried to indicate here how rapidly the value of that probability density falls off. So for example, one standard deviation, I'm about 60% the peak. Two standard deviations, I'm down to about 13.5% of the peak.

Now more often than asking what is the relative probabilities of these things, you're actually more often asking, what is-- how much-- what is the integrated probability density of the random variable out in some tail or in some central region? And that's where the Cumulative Density Function is really the one that you want to use. And so what I'm showing here is out for some number of standard deviations-- this is mu minus 3 standard deviation.

This is saying the probability that x is less than mu minus 3 sigma is exactly that value. That equals f of mi minus 3 sigma. And I simply look that up. And that's about 0.00135, or less than 0.1% of your data should fall less than 3 sigma off the left side of your distribution.

And then I've tabulated that for two standard deviations, one standard deviation. By the way, what's the probability, now that I've marked it up, that your data falls less than your mean? 50%. It's a symmetric distribution.

And so, in fact, you could then ask also the question, what's the probability that my data is all the way from my left tail up to two standard deviations above the mean? And that's 97.7%. But I want to also point out these-- this distribution itself is also anti-symmetric around the mean. So this value and this value sum to 1. So in other words, 1 minus whatever is out in the upper tail is equal to the probability of being below the lower tail.

So what's tabulated is not mu minus numbers of standard deviations. But what will often-- what is actually tabulated are the standardized or unit normal distribution-- again, the mean-centered version, where I subtract off the mean and divide by the standard deviation. And that gives a PDF and a CDF that is universal.

And that is what will often be then tabulated as the unit normal Cumulative Density Function. In some sense, that's what I actually showed on this plot, by just labeling it as a function of mu and standard deviations. But now when you normalize, that becomes in units of z as 0 and the numbers of standard deviations off on the side.

Now, if you look at the back of Montgomery, there is a whole bunch of these tables. And you'll be using these tables in some of the problem sets and so on. And there is a table for the unit normal.

And in particular, what's tabulated is this Cumulative Density Function for the unit normal. And we have a little bit of terminology here that I want to alert you to, because we often talk about percentage points off of some distribution or percentage points of the unit normal, as pictured here. And what we're talking about is relating percentages of my distribution that are in some location, usually the tails, to numbers of standard deviations that I have to go in order to apportion that amount over in the tails or in the central regions.

So a very typical question I might ask is, how many z's-- how many "unit standard deviations," how many z's-- do I have to go away from the mean in order to get some alpha or some percentage of the distribution located out in those tails? So for example, I might say I want the 20%, 20th percentile percentage point, the 0.2 probability that my data sits in the two tails.

So for a total probability that all my data or the remain-- the portion of my data is on either of the tails, some further away than some z, that means 10% is in each of the tails. And I'm asking the question, how far-- how many standard deviations do I have to go to get 10% in the left tail and 10% out in the right tail? So I'm essentially asking the question, what is the probability on the cumulative unit normal Probability Distribution Function to get to-- how many z's do I have to go to get to half of that alpha probability being in each of the tails?

One observation here is that these things are, again, anti-symmetric. So I can also ask the question either looking just the right tail or the left tail. And then you can do the inverse operation using the table.

So I'm actually asking the question, what is the z associated with that? And I'm looking up on this plot. So I might ask, OK, I need 10% there in the tail. How many z's does that correspond to?

And to get 10% out in that left tail, I got to go out 1.28 standard deviations off to the left. That's the operation that one would look up in the table. So very often, you would get to these kinds of lookups, where you're relating the probability alpha of your data lying below that number of standard deviations and what that corresponding standard deviation is.

So I didn't copy one of the tables out of Montgomery, but you'll get some practice with that on the problem sets. Now, there's other related operations you can do once you have that. So for example, now I can ask, what is the probability not just that data lies out in the tail, but what are the probabilities that it also or instead lies in the middle region? They're all the same kinds of operations.

And so for example, here's a quick tabulation for three different kinds of examples, where I'm asking not what is out in the tails, but I'm asking what is within the center plus/minus 1 sigma region of the data? And if you look very carefully, I'm using exactly these Cumulative Probability Density functions for the unit normal. This is for a unit normal.

And looking out, what's the cumulative probability over in the left tail? The right tail? Doing those observations. But these are also very nice rules of thumb to have ready for you, which is saying within plus/minus 1 standard deviation in the normal, 68% of your data is going to fall in that 1 sigma region. In the case of if I expand out to 2 sigma, now I've got 95% of my data should fall roughly in there. And if I expand out even further to the 3 sigma, that's the 99.7% of your data would be falling-- should fall within those center three standard deviations.

So the percentage points out there, the part that falls outside of that, is about 3 and a-- 3 and 1,000. We'll come back to this when we see statistical process control and control charts because you may have run into these control charts. We're often plotting the 3 sigma control limits.

And essentially what we're saying is only a very small fraction of my data-- 3 out of 1,000, if I'm using plus/minus 3 sigma control limits. 3 out of 1,000 points of my data, by random chance alone, should be falling outside of those 3 sigma bounds. So that starts to get as close to statistical process control.

So what we're going to do next time is start to look a little bit more closely at statistics. When I do form, again, things like the sample mean, or I form the sample standard deviation or sample variance from my data, those themselves have these probability densities associated with them. And from that, we're going to be able to go backwards and essentially work to try to understand things about the underlying process distribution, the parent probability distribution function, associated with that.

So we're going to have to understand more complicated PDFs than the normal distribution because things like the sample variance is not going to be normally distributed. It's going to have its own bizarre distribution-- in this case, the chi-square distribution. So we'll return to looking at some additional distributions, but these same manipulations will come up again.

And what we're ultimately going to want to be able to do is make inferences about the underlying distribution-- the parent process-- what its mean is, what its variance is, based on the calculated sample mean and sample variance that we might be using, and then also make inferences about the likelihood that the true mean lies in certain ranges. Or to put it another way, next time, we'll also be talking about confidence intervals. So we'll see you on Thursday. Watch for the message from Hayden about tours and enjoy.