DUANE BONING: OK, so today we're going to talk about some of the probability models associated with variation in manufacturing processes. The last couple of classes, we've tried to go and do two things. One is give you a little bit of perspective on a couple of different kinds of-- or actually, a fairly wide family of manufacturing processes, give you a little bit of exposure to semiconductor processes as well as a couple of different mechanical-- mechanically-oriented processes, and give you some ideas or ways to think about those processes, including just the physical action that's going on, as well as some of the sources of variation.

So we're going to dive in a little bit today on dealing with data taken from some manufacturing processes, and start to look at random components of that and systematic components of that. But back on the theme of trying to make sure that you have at least some exposure, given everybody's different backgrounds, to these processes, we would like to make sure you have the opportunity to at least see a couple of the kinds of tools and equipment in the facilities-- especially those associated with semiconductor manufacturing, if you haven't previously seen those. So let me turn it over to Hayden, who can check in with you well the problem is that as well, but also about arranging opportunities for you to see a little bit more, if you would like.

HAYDEN TAYLOR: So yes, on the problem set, well, if you still have questions about it, then feel free to email or call today. See me afterwards. And problem set 2 is now up on the website. That's due on Tuesday, the 26th of this month. So there's a bit longer for that, and I'm willing to start helping you with that too. So thank you.

DUANE BONING: Any questions for Hayden? Any big issues with the problem set?

AUDIENCE: Actually, I have a--

DUANE BONING: Go ahead.

HAYDEN TAYLOR: OK.

AUDIENCE: So I have a quick question. So if you can refer to the Question 3 of the Problem Set 1, and you say it's proportional to the water and the time-- the KOH. And when you talk about these [INAUDIBLE], you're talking about the concentration or you're talking about the molarity?

HAYDEN TAYLOR: What am I saying is proportional to KOH.

DUANE BONING: The square brackets around the KOH.

HAYDEN TAYLOR: Oh, I see.

AUDIENCE: The very last figure [INAUDIBLE].

HAYDEN TAYLOR: Yeah, yeah.

AUDIENCE: --proportional. The [? edge 3 ?] is proportional. So it's proportional to the concentration, or it's proportional to the molarity?

HAYDEN TAYLOR: Well, for a fixed volume, it's the same. There are two lines on that plot, and the models that are put forward by [? Seidel ?] to explain the data. The data are the open circles on that plot. And there are two models that have plotted. And what I want you to use are primarily the experimental data. So look at the curved line that is plotted on that graph.

AUDIENCE: OK and for the curved line that you want us to look at, the x-axis should be the concentration, right?

HAYDEN TAYLOR: That's right. Yes, the x-axis--

AUDIENCE: [INAUDIBLE]

HAYDEN TAYLOR: --is the weight of KOH. In other words, that's the number of grams of KOH per gram of water.

AUDIENCE: Yeah, I just want to clarify. That's it.

HAYDEN TAYLOR: Yeah. Thank you.

AUDIENCE: Thanks. Thank you.

HAYDEN TAYLOR: Anybody else? OK.

DUANE BONING: I was just also going to say, it's sort of a manufacturing process that we use for making up these problem sets. And they are not expected to be completely defect-free, so we expect part of the process is detecting errors or questions. And then we try to correct them.

Also, as a follow-up to that the schedule for the next problem set, just looking ahead for the next couple of weeks, the reason that the problem set is due not next Thursday, but rather, the Tuesday after that-- a little bit longer than a week-- is that here at MIT, next Monday, this coming Monday is President's Day, a holiday. And then Tuesday is one of these strange days.

You folks out in Singapore-- you may recall this. Tuesday is a Monday. So it's on a Monday class schedule, so we will not be meeting. OK? Go to whatever classes you would normally have on Monday. Go to those on Tuesday out here to MIT. And so then our next lecture will be a week from today, on Thursday.

OK, so also going with the material that we're starting to talk about now and the second problem set that's just been released is another reading assignment. And in particular, what we're trying to do this year is focus-- do sort of the concentrated dive of May and Spanos Chapter 4. In the past, sometimes we've assigned instead Chapters 2 and 3 of Montgomery, which are much longer.

Chapter 4 in Spanos is a shorter overview of basic statistical distributions and some very basic manipulation and use of statistics that we'll talk about today. So we're not formally assigning chapters 2 and 3 in Montgomery, but the intent here-- what I mean by this reading assignment is, if you don't have very much background on statistics, you may want to read Chapters 2 and 3.

Or you can skim them to know what's in there, and in particular, that's probably the right place to go for a longer explanation, if you need more material. If May and Spanos is too condensed for you or you'd like additional examples, there's a very nice articulation in Montgomery. Now, I have not made a formal assignment of May and Spanos Chapter 3, but that is where you should go if you want to read more or need to read more about semiconductor fabrication process physics. That chapter is a very nice, again, relatively condensed description, articulation of the key process steps used in microfabrication.

We'll, throughout the term, be popping in and out of a few of those unit processes, and I just want you to be aware that's what's in that chapter. And you will quite likely want to go in and skim that or read that, and dive in on some of those processes if you need more background on some of those things. And I think, especially before this tour that Hayden is organizing, it'd be at least really good to skim chapter 3.

Again, it's mostly on the process physics, but every now and then, it does at least show you a little bit about the equipment that's used. And it's very nice to be able-- for Hayden to be able to point to some of the equipment and components of that, and have you visualize what's going on in the process physics, or be able to ask questions about that.

OK, so what I'm going to do here at the start is just run through some example processes. And in particular, these are little not quite manufacturing processes. They are more like research scale implementations of some basic manufacturing processes that we have actually generated some data from, or students have generated data from over the years.

So I'm going to show you some actual measurements from a few simple cases. And we're going to look at that data, and what I think we're going to see is variation. And the puzzle for you and for us-- and to some extent, for me, because in many of these processes, I was not involved in the generation of the data-- our puzzle here is to look at the data and try to think about what might be going on in the data.

So here's some-- a basic turning process. So we've got a work piece. We've got under computerized numerical controls, CNC. We've got the ability to cut into that work piece, which is rotating, to try to get to diameter work pieces of a specified dimension. So here's the diameter measured on a particular location on that work piece as a function of 47-- I think it's 47 or 48-- different repetitions.

Now, I believe this may have involved more than one set of students, or more than one class. We can think of that as a shift change in a manufacturing process. We're simply gathering here a measurement of a diameter on that work piece. So talk to me. Tell me what you see. Is it uniform?

Not especially-- although you always have to be really careful looking, because on any kind of plot, you want to really look at the spread. Here it's kind of magnified, because here's 0.7 inches, and we're down to about 0.697 up to 0.702. That's not full scale.

But if we zoom in around whatever nominal is-- and in fact, it's not even clear what the target dimension was-- I might guess and think that it was 0.7 inches, because that's a nice round number, but who knows? So it's definitely not uniform. What else do you see? Yeah?

AUDIENCE: *The data is sort of drifting up over time.*

DUANE BONING: OK. So it looks like there might be a long-term trend of a drift over the number of runs. What might lead you to think that that would be a physically reasonable--

AUDIENCE: [INAUDIBLE]

DUANE BONING: Right. So if the tool wears-- and in fact, if the tool wears or some long-term change in the state of the equipment-- that's often the source of these long-term trends. Other things people see-- Yeah?

AUDIENCE: It's fairly systematic. It constantly goes up and down, and up and down. There's only two or three points before it goes back up.

DUANE BONING: Yeah, that's very interesting. Even the variation here-- it seems to be cyclical, or periodic, or something systematic going on in there. Again, I wasn't there for the generation of this data, but what might some reasons for that be?

AUDIENCE: The worker might be trying to correct with the subsequent one, and he overshoots constantly.

DUANE BONING: Interesting-- yes. That's a good idea. I guess, in fact, one can imagine, especially if you're taking a measurement after every tool, and the target was 0.7, say, then maybe, near the beginning, we're starting to see a deviation. Maybe this one was a little low, and they're trying to get back up, making an overcorrection-- because near the end here, it looks like there's a little bit more centered around 0.7.

AUDIENCE: I think that might be because of backlash [INAUDIBLE].

DUANE BONING: OK.

AUDIENCE: [INAUDIBLE]

DUANE BONING: So backlash being an operator adjustment that is made as perhaps part of this compensation strategy?

AUDIENCE: [INAUDIBLE]

DUANE BONING: Interesting--

AUDIENCE: [INAUDIBLE]

DUANE BONING: Anybody in Singapore-- any other observations or ideas?

AUDIENCE: Since measurement is measured at a different place of the things, so maybe you can see every three point will be a larger diameter. I think maybe this is measured at the other side of the thing [INAUDIBLE]. And maybe for the lower-- for the smaller diameter, it's measured in the inner side.

DUANE BONING: Yes, it's quite possible-- at least on this plot. We might learn more if we look at more data. I read this as run number, rather than measurement number. So somehow, perhaps, this is intended to be representative of the whole part, but it's possible those are multiple measurements on different part-- portions of an overall turning.

You could start to look and hypothesize on that. Is the 1, 2, 3-- is that the same pattern you see each time? And it's not clear whether we see exactly that same pattern or not. Other ideas?

AUDIENCE: [INAUDIBLE]

DUANE BONING: Right. So run number is a stand-in for time, and these are occurring. So your point here is--

AUDIENCE: [INAUDIBLE]

DUANE BONING: Sure.

AUDIENCE: [INAUDIBLE]

DUANE BONING: Right. So it may not be equipment state, or indirectly, it may be also the rest of the facility environment or the line. And over time, who knows? Maybe this thing starts heating up, or the work pieces that are coming in have a different temperature-- quite possible that there could be some other long-term state change.

We do notice there's this shift change. And referring back to this long-term trend, if we actually look at this early part, it may or may not be drifting that much. It may actually be fairly steady. And what we're actually seeing in here is an overall mean shift, but maybe not even a long-term trend.

So one might actually pose, maybe even apply some statistical methods to look and say, is there a statistically significant mean difference between this set and this set of data? So it may look like a long-term trend, or it may simply be a reflection that there is a setup change on the equipment when a new shift comes in, or there may be just inherent differences in how the operator interacts with the equipment. How they load the part might be slightly different.

Any other ideas? Let's go back to the-- because I think you had mentioned something interesting about possible oscillation in here. What might other sources or causes of oscillatory behavior, or what looks like two distinct sets of data here-- in fact, I'm not completely sure that it's purely oscillatory, because you can come down here, and it seems like sometimes you've got two things that are down low, and then one up, or three things that are low, and one up.

AUDIENCE: Well, this is a little bit farfetched, but maybe you could have [? stock ?] material of maybe 1 meter, and they always cut it into three pieces. So they always have the first piece of stock material, and they had some material properties that led to have [INAUDIBLE] maybe. And the middle piece led to middle dimensions, and the last piece led to smaller dimensions.

DUANE BONING: I like that. That might be very interesting. So if I were generalizing that, it may well be-- very often, there may be more than one set of starting material. So it may have been all one piece and cut into three, and so they're randomly picking one of the three, and therefore, you're getting a different mix of those three points in here.

But in general, it may well be in many cases that you've got two different incoming streams. If I were looking at this, instead of thinking of three, it kind of almost looks to me like this set of data here-- it's kind of randomly around some value, and this is a different set of data around some other value. So it may be that there's just two boxes apart sitting there, and they're just pulling, randomly, starting material out of one of the two. Yeah?

AUDIENCE: I also don't know if there's any control on how far from the [INAUDIBLE] the materials [INAUDIBLE] you could put it in, be a little bit farther, and then it would have a larger cantilever arm than the next user.

DUANE BONING: Right, right. So there could be lots of sources of both random and deterministic variations. So we talked in here about things like a mean shift. Maybe there's a deviation between these two shift changes. And that would be a either deterministic or-- systematic maybe is a better word of a deviation.

And then, within different sets, we may have no detailed explanation of why we have these small deviations, and a random model for some of those components might be appropriate. So that's part of the idea that we're after. So over here on the side, we're just referring back again to of our variation equation.

Maybe I'll put that up so that we can also refer to that, because I think we've touched on some of these different components. We said that our deviations in our output y you can often characterize as some component that has a sensitivity to disturbances in our process parameters, and those disturbances, and some component that relates to our controllability when we make small changes [INAUDIBLE] and then control inputs.

Among other things, we said, well, maybe some of our input feedstock, maybe delta a's-- changes in the parameters. We also-- and I thought it was very interesting, talking about overcompensation in the control scenario. It might be, in fact, that there are deviations or errors in active control, changes to the settings.

There may also be-- and we also talked about it here-- is recognition that the function y of our alpha may, in fact, be time-dependent, or that-- put it another way-- some of the delta a's have time components and time trends to them. We'll come back, and I'm sure come up with more examples as we look a little bit further.

And I think this is some of the same process. It has different run numbers on it, but this is explicitly-- it looks like there's, again, a dimension and run number along the horizontal axis, but in this case, there's the yellow, our inner; the purple, our middle; and the blue, our outer dimensions. So you can imagine here that perhaps measurements are being made at different locations on the work piece.

And I believe that this line right here is, again, a shift change. So talk to me a little bit about this. What else do we notice in this case? There's one important difference in this data compared to this data that I'm looking for in particular. What additional characteristics-- or what else could you say about this data?

AUDIENCE: [INAUDIBLE] I guess the spread on the [INAUDIBLE].

DUANE BONING: OK. So one observation is, if we were just looking-- well, right here, this is a good one, where inner, middle, and outer are all close together. But your point is, generally, we're seeing a very huge spread in the measurement within that particular part than we do over here.

AUDIENCE: [INAUDIBLE]

DUANE BONING: Much tighter band-- so now you're talking like a band here?

AUDIENCE: Yeah, that [INAUDIBLE].

DUANE BONING: Good, good-- yep?

AUDIENCE: [INAUDIBLE]

DUANE BONING: Interesting-- yeah, so it's a situation here where we don't know the specifications, so it's a little hard to actually know which is better-- because your point you were making is, in this case, the blue's always above-- or almost always above the purple, always above the yellow. And it may well be that it's meant to have a slight taper to it.

We'd need to more. And actually, that's a really important point in manufacturing processes. You can't just look at the data and always know what's going on. You also have to have information from design, know what the intended results are. So certainly, just to highlight a couple of the things [INAUDIBLE] mentioned, one here is, very often, we are also interested in within part variation.

So assuming that maybe they all really were supposed to have the same dimension, then this spread here is very different than this spread here. So characterization of within part variation may be very important in understanding that. And then the other thing here is, unlike this-- we still have all that gook on here, but in here we were basically having a similar band, and it was all just kind of moving or shifting.

But the part-to-part variation did not seem to change. The variance of the process did not seem to change. The mean shifted. So that's a very common characteristic that we're worried about. And your point here is that, in this case, this looks like a much tighter band, or the standard deviation of the invariance of the process in a run-to-run sense is very different than in this process.

And so it may be somebody with a slightly steadier hands, or a more consistent process, or a more consistent operator, or the equipment-- maybe some change was made-- improvement to the equipment saying, hm, that little screw is a little loose. Let me tighten that down. Who knows? But something clearly changed in order to apparently improve the process. So those are a couple more characteristics.

AUDIENCE: [INAUDIBLE]

DUANE BONING: Yes-- question?

AUDIENCE: Professor--

DUANE BONING: Yes?

AUDIENCE: Since you mentioned that maybe with the shift change there's some improvement in the [INAUDIBLE] of data, and that's why it gives us a smaller band right now. But I noticed that the outer, middle, and inner data-- they are pretty close to each other, as in the difference between these three sets of data are much wider in the previous shift than in this right now. So can we still say that maybe this-- the object we are measuring is tapered? Because it doesn't look like it's tapered right now.

DUANE BONING: Absolutely-- well, but the point was made that we don't actually know the design intent, so this was a hypothesis in here. I think it's probably unlikely that there was an intended taper, and I'll agree with you. But in here, it looks like there is somewhat closer within part correspondence in these dimensions.

OK, so let me go to one more. Here's another one. This is brake bending of a metal sheet. The basic idea in this process is we're trying to change the geometry under the application of force in a bending process. So we might have some tooling. We have some piece of metal, say, and we press down, and the intent is to create some permanent deformation in the part.

An important component here is that, when we remove that force, remove that pressure, there can be spring back in that part-- so little bit of the inherent characteristics of this. And so here's some data looking at using the same tooling, and I believe the same set of downforce, with-- I believe this is a measurement of the resulting angle.

Alpha is not a good variable to use here, is it? We call that just angle of the resulting component. And what else is changing here is that the source material is changing. So in some cases, we're using aluminum, some cases steel-- and also with different thicknesses.

So I won't even ask you to share-- the most obvious observation is, depending on the material type, we certainly get different angles, different degrees of spring back. That's the most obvious thing. What else do you see in this? Yeah?

AUDIENCE: It depends on the thickness of the material.

DUANE BONING: OK.

AUDIENCE: So for the same material [INAUDIBLE] the springback, it will be more.

DUANE BONING: Right. OK, and one could certainly imagine good physical reasons for that, that thicker material is going to perhaps spring back more readily. And one could even go off and imagine trying to build a first principle physical model that would tell you what the resulting angle or degree of spring back would be as a function of that.

And in fact, it's probably worth it to go and consult some literature and see if there is physical insight. But if I were to ask you to actually tell me what you think the result might be if we had a 0.45-inch steel part, what would your first inclination be? Would it be to go to the technical literature and look up a physical model for this?

AUDIENCE: [INAUDIBLE] averaging between the [INAUDIBLE].

DUANE BONING: Yeah. You would be basically using some averaging to deal with noise and manufacturing variation, but you would be building a very simple empirical model based on the data that you have, and then interpolating in some way. And here we've only got two data points, so you might pick the middle.

We don't know. You might want to actually go and design some experiment where you add a third point so you know if there's some non-linear dependence, and so on. So that's certainly an important thing. There is a very clear input-output, a deterministic effect that, again, you would like to be able to understand enough to be able to deal with that.

And the same may also be true of other discrete design choices-- discrete choice being the material, whether it's aluminum or steel. You would like to know, is it a similar trend? One thing that's very interesting here is it appears that the delta as a function of those thicknesses is about the same-- very nearly the same in the two material cases. And then there's a material delta.

So I'm already starting to think, OK, I can have a very simple additive empirical model that has a delta effect due to the material type. It's a binary choice-- or a binary coefficient, depending on that. And then an additional component is a function of thickness, and maybe that function of the thickness is independent. That delta from amine is independent [INAUDIBLE] material type. OK, anything else going on? Yes?

AUDIENCE: [INAUDIBLE] there is kind of a drop in the middle of the set. [INAUDIBLE].

DUANE BONING: Yes-- very interesting, and maybe a hint almost of something, but then it-- yeah. So there might be something systematic going on there. The human eye is wonderful for looking for patterns, isn't it? Actually, one of the challenges often is using both engineering judgment and when the data is simple enough to detect trends, but also have statistical methods when the volume of data is huge or there are many things changing at the same time to detect possible shifts and trends. Anything else? Yeah, you had more.

AUDIENCE: OK. [INAUDIBLE].

DUANE BONING: Yes.

AUDIENCE: [INAUDIBLE]

DUANE BONING: So you're worried about this, and maybe this also.

AUDIENCE: Yeah.

DUANE BONING: In fact, we don't even know if that's a still component or an aluminum. Yeah. So in some sense here, if I were looking at this data, I almost start to think these are relatively-- well, you'd have to ask as a function in comparison to specifications to actually say how well-controlled that is, but there is relatively tight bands with a couple of these shifts, and then almost outlier points.

Now, would you think that these points are coming from the same source of deviation that might be at work in this band? Might be-- but it's such a large deviation that it starts to feel unlikely. And another use of statistical methods will actually be to quantify, how likely is it that we would observe by chance alone a deviation within the natural variation band of such a point?

And the key reason is, if you see something big like that, there might be a [? point cause. ?] And it's quite possible in here that the operator is making a measurement on each part and said, whoops-- something just happened here. Something got misadjusted on the tool. I need to readjust it and get it back in. Maybe there's even a little bit of something going on there.

Or it might have been a single event that wasn't-- didn't require adjustment. Maybe you go in, you investigate, and you go, that one piece of source material coming in, that piece of metal coming in was much thicker than it was intended to be.

AUDIENCE: [INAUDIBLE]

DUANE BONING: Yeah, maybe that's a piece of aluminum right in there, or this was-- exactly-- this was a steel [? 0.3 ?] in the aluminum set, or steel [? 0.3 ?] in the first one of those. Absolutely-- that's a cool observation. Yeah, that point seems to fit down in that distribution, doesn't it? Very interesting-- so a couple of the points here-- the kinds of things we might be looking for here are deterministic and systematic kinds of effects that we might want to use modeling for.

And then we also are starting to get close to the statistical process control thinking, which is to say we would like to be able to know what the inherent or natural variation of the process is having to do with these sorts of bands, and then be able to detect with high degree of confidence that something strange has happened, and I better take action. And that SPC, or statistical process control.

Maybe we're going to see many of the same characteristics here. This is just a little observation from injection molding. So this is a plastic molded part. We've got a wit, the part along this dimension, and down here we've got a number of the run. And also notice-- I'll just draw your attention here-- there are clear intentional bands here, where this first segment here says holding time was five seconds, injection press 40%.

And we're changing here-- this is a holding time of 10 seconds. Here's another holding [INAUDIBLE] of five seconds and 60%-- and whatever is lurking under there. I think it's a holding time of 10 seconds. We could find out, but-- talk to me about this one. Oh, I guess also here, notice, we're plotting two pieces of data. We've got a width and then some average.

So it's quite possible that what we're doing here is this is a point width at a well-known location, and than the average is we take that measurement on many, many locations around the part. Or it's possible we also actually run multiple parts, little mini batches of-- who knows-- maybe five parts, and I'm plotting the average across that little batch-- that five-part batch, which is another very common thing to do in SPC. So tell me a little bit about this. What do you see? Yeah?

AUDIENCE: [INAUDIBLE]

DUANE BONING: Yeah. So you're talking about this band here to here. Right. In fact, this might be the result of a designed experiment, where somebody is trying to explore the effective different process conditions, these four different cases. And part of the goal of that may be to identify not only things that are closer to a mean target, but are also more robust-- inherently robust, meaning they have a smaller delta a, or are the process itself has a smaller sensitivity among at that operating point to whatever inherent disturbance, whether it be temperature deviations or what have you.

So that's a good observation. It's also kind of amazing to me here that the average looks rock steady in these cases, but these individual width measurements have a substantial deviation.

AUDIENCE: Is that the average of all [INAUDIBLE]?

DUANE BONING: We don't really know. I don't know. Yes, it's quite possible that this is just a plot. So that's your point. This may be [INAUDIBLE] suspiciously so, right? This may be the average across that ensemble, and all that is doing it's not saying it's an average of multiple measurements or multiple samples, but that's just highlighting that there's this delta. And that's got to be it, because-- exactly. There's zero deviation in those average points.

OK, but these are some of the same kinds of themes. We're, again, looking at offsets from deterministic process conditions. We're looking at inherent variation, ranges. We may be doing design experiments. You could start to think of empirical models that would help to guide us to an optimal point, the point being there that maybe these four conditions give you a little bit of an exploration across two different process parameters.

One approach would be maybe your target is 40.85 or 40.87. You might pick the one that was closest to it. But if you actually built a model, including a model with continuous parameters, you could then interpolate and drive even closer to a target. Yeah-- another question or comment?

AUDIENCE: [INAUDIBLE]

DUANE BONING: Right, one would guess.

AUDIENCE: [INAUDIBLE]

DUANE BONING: Right.

AUDIENCE: [INAUDIBLE]

DUANE BONING: Yes. So that's actually a very interesting point. When we start talking a little about more about design of experiments, about the synergistic effects, it may not be a simple additive model. You're saying, as we go here, from 5 to 10, we got an increase. Here, when we went to 5 to 10, we got a decrease. So the effect of holding time may be completely in the opposite direction at two different pressures-- very good point.

So that means there is some strong interaction between those two-- or mutual dependence between those two parameters. Man, we're getting very abstract in this data. We've got measurements on the output and we've got-- some number of run number on the side. This is actually looking very similar to some other things. We see definite shift effects. We see mean effects. Again, we see perhaps an outlier.

The point here was that-- and this didn't come up earlier-- we keep looking and trying to put blame back on the manufacturing equipment or the source material. You also have to look carefully at your measurement apparatus. There's inherent measurement error. Even if the parts were almost perfect, very often, you're going to have an inherent measurement limitation and deviations due to measurement error.

And you may have loading problems into the measurement apparatus. So it's not always an indication that there's part deviations. In some sense, the measurement equipment is manufacturing process, tool, or equipment itself that is also subject to deviations. And in fact, part of the applications of some of the statistical methods we'll talk about seek to characterize or do gauge studies of the accuracy and capability of your measurement equipment and the components of variation that affect that.

So for example, there's a pure reproducibility and there's a repeatability component to typical gauge studies. One component, just real quickly, would be I've loaded the part on the measurement apparatus, and then I just fire whatever the measurement sequence is multiple times, and think of that as a pure repeatability-- somehow the physical-- maybe there's-- it's an optical measurement. There's scatter from extraneous light sources nearby off of that.

And then there might be a separate component, if I pulled that same part out, took it out, and put it back on, and did that sequence of operations, including the operator operations. Then you have a reproducibility component that might be, in fact, much larger. So you'd need to understand, and maybe decompose and break down to those different components. And then, if you started to see that one was a source of deviation, then you might look at strategies to control or compensate for that. Yeah?

AUDIENCE: [INAUDIBLE]

DUANE BONING: Yeah.

AUDIENCE: At one of my internships, the company was getting a lot of [INAUDIBLE]. They were really trying to figure out [INAUDIBLE], and when it went bad, at the supplier [INAUDIBLE].

DUANE BONING: Absolutely, absolutely-- in some sense, it's often the analogy of it's not working-- oh, check the plug on the wall. Always check your measurement equipment and apparatus before you start diving and tearing your hair out, going too deep into debugging of the manufacturing process. Check each step along the way-- glad they discovered that pin problem back in the-- at the case.

Oh, here's more. What do we see here? Again, some output as a function of run number [INAUDIBLE] again, maybe this is another sheet bending kind of component. We see effects of shift change multiple times here. These are different shifts, maybe different shifts or different parts.

What's very interesting here is that, under some conditions, we have a nice stable, not drifting in time-- fairly nice, stable, or stationary mean, with some deviation-- every now and then, maybe something that looks outside of the distribution. But in some cases, holy cow-- that's a very different effect than what's going on in here. And it's back to this is a very clear drift that is definitely not just a one-time shift. Yeah?

AUDIENCE: When I looked at the [INAUDIBLE], and the first [INAUDIBLE]. I'm wondering if maybe the [INAUDIBLE].

DUANE BONING: Perhaps-- or maybe this worker finds a way that corrects for it in a comparatively effective fashion, whereas the strategy used in this case over here is not nearly as effective or quick at trying to get back, if that was the target. And so in fact, one of the challenges often in SPC or other control algorithms-- or not even algorithms, but control practices, especially involving people-- is one shift or one set of operators or engineers in charge of a process may learn things that don't necessarily get captured and conveyed to other shifts.

OK, so what is going on in lots of these examples is perhaps the most important point here-- is that, very often, there are two highly different characteristics that keep coming up. There's systematic or deterministic effects having to do with the type of the material, settings on the equipment, choice of process conditions. And then there's also random components. There's the inherent spread.

And a key point here is inherent spread you can almost never completely get rid of. I guess, philosophically, you could ultimately maybe appeal to quantum mechanics, that there would always be some deviation. I think very rarely-- perhaps in some semiconductor processes, we do approach truly inherent physical randomness. But generally, it's that the understanding is limited in what is contributing to that small degree of spread, or it's impossible to completely control those-- that inherent randomness.

But then we also have big disturbances. They may still be random events, but they are not this little band, but they are big events with a clear cause. We may not always be able to discern the cause, or may not always be able to discover it, but in many cases, we've been hypothesizing about many of these big shift changes or a big drift.

And a bit of investigation and more knowledge about the process would readily show us what's going on in those cases. And in those, when there's something systematic or deterministic going on, the whole idea is you make some changes to the equipment to stamp those out. And what we're often going to be dealing with is designing experiments and whatnot to try to understand systematic components and eliminate them, maybe as part of the initial design implementation of a process.

Or in an ongoing process, when we see a deviation, we've got to go and debug it, find out what caused that, and eliminate it. And the natural occurrence then is that, over time, as you have learning about your process and manufacturing system, you're generally squeezing out these deterministic sources. It's something-- was repeatable, systematic, and therefore, you can come up with strategies either to eliminate it or to compensate for it.

Maybe you have to have a control strategy that recognizes that source of variation and keep squeezing it out. And what you generally are left with are these random components, hopefully in a very nice, narrow band. But we still need to understand and model those, and so that's actually where-- another place where the statistical analysis tools are really crucial, because then we want to be able to understand, what is the natural or reducible-- irreducible component that we don't know how to stamp out one at a time-- and use that information to be able to detect in that small scatter when something unusual happens.

And that's the idea of statistical process control. So what we want to do is characterize the process-- particularly just the random component, assuming that we've been able to squeeze out many of those other systematic components. Why do we want to characterize that process? There's some of the things that I've mentioned, and that you've mentioned. If I see one of these points that looks a little bit different than the distribution, we said that [? by i, ?] but very often, the question is, maybe there's a point that's a little bit above some of the other points.

Then you would often want to ask the question, do I really believe the output changed, or is it part of that natural variation? And if we start to think about some natural variation and models for that, including distributions-- which we'll get to in just a moment-- then we can start to ask that question in a very systematic way, in terms of actual probabilities.

Other reasons is we may go back to, did the input actually cause the change? Did I make some output change? A very important point here is, how confident are we have these answers? Very often here, we were appealing to it looks like that's a bigger change, or this is-- it was a mean offset. Well, that's OK to try to convince your management or convince your operators that, look, something changed, but in many cases, you'd actually like to be able to quantify how confident you are that that was either a deviation point or that some specific decision that you made had a true effect.

And so the probability and statistical tools are really, really important for that. So what are some ways that we can actually characterize and model the random components? Well, that's where we're getting into random processes and random variables. And I hope you're seeing this is the basis for statistical process control I've referred to. And we'll dive into some of the tools and techniques, as well.

Design of experiments-- so if we were making that pressure change or that time change, how big of an effect is it? How certain are we within the random natural variation that the deviation-- or the controlled decision I made actually had a big effect? And then finally, even with compensation strategies-- like this-- one of the very first things that we said-- in that oscillating case, there might be inherent noise. And it's quite possible that somebody could be making process control feedback changes to try to compensate.

But if they're trying to compensate for something that doesn't have, in fact, a connection between the control decision and the true source of noise, they may, in fact, be injecting more noise in the process. So understanding actually how feedback control works when we've got random, and not just the systematic component, is very important.

What are some ways of describing randomness? Actually, philosophically, there's a couple of very different ways of thinking-- if you open up a statistic book, in the first few pages, we'll talk about this. And from a manufacturing perspective, well, we're really dealing with data. In many ways, the data tells us a lot.

And one approach here is to look at actual large ensembles or collections of data, and look at histograms, frequency distributions. And we'll look at some examples of that. A very different approach is to start with a pure probability model that assumes there is a universe of and an infinite number of samples, and there are characteristics of that infinite number of samples. And that gives us the tools to put empirical observations, like frequency histograms, on a mathematical basis that we can reason with.

So for example, a way of-- let me jump forward here-- this is some thermoforming data. Again, this is some dimension, and then, over this collection of parts, we have a-- this bar chart, this frequency histogram. And what that's really saying is, for each of the bends in that histogram, I'm simply plotting the frequency-- or relative frequency that some output was between the bounds of that bin.

And so what we are doing is converting from historical data into bins. This is still discrete bins, but then we're associating that and interpolating forward to say, well, the probability in most applications-- if I were to do some more part, we think that this histogram, in some sense, represents a larger set of data, and I can use it to talk about the probabilities of occurrences of new data in here.

So what do you see in here? Whoops-- there's a little insert here with the raw data. Now, if I hid that from you-- it's hard for me-- I'm trying to put my hand over this. If I hid that from you here, and I hid that from you here, and you just looked at this data, typical things you would say about it-- it looks normal, or Gaussian distributed. Maybe you would think almost it's completely random. It's got to mean and the deviation. It's got a normal distribution.

Now look back at the data. It's kind of drifting around. Just a peek forward to other things to be cautious about-- don't always believe or just look at your statistics. Always go back to the raw data, because if this were a characteristic of my process, I would say there is a random component, I think, here, but then I'm also seeing maybe some systematic drift that-- I'd like to go in and squeeze that out.

And then, if I were able to get that out of my process, then I might have a random component that doesn't have that deterministic or systematic component. And it might have a distribution, but it might be much, much tighter. So an important point here is purely empirical data doesn't always-- it may look random, but it may have embedded in it-- if you were to deconvolve all of the sources of variation, this big random-- or this big normal distribution may have mixed into it lots of different systematic components, as well as underlying natural variation.

Imagine, for example, that I actually went in and I got rid of this long-term drift. Would you believe that the remaining variation is all random? I bet I could go and I could plot it, and I bet it'd look-- the remaining deviations would look mostly normal as well. You think so? Well, who knows?

You might go in and do some of the other kinds of hypotheses about systematic sources that we were talking about in the other cases. For example, it looks like there may be a bit of an up, down-- there might be something cyclical or periodic going on. That might be another systematic source. It could masquerade as a random normal distribution. So there's always that mix in.

However, in many cases, maybe you also want to live with-- it's not worth it to you, or too expensive to go in and eliminate, or you tried and we're not able to eliminate also some of these wandering drifts. And in that case, maybe the best you can do is at least characterize that remaining variation, which has a systematic component. But you would still like to have a probability model for it so that you could detect deviations that were outside of that variation source.

So if I saw a big shift up, I would say I've got a point way out here in the tail of the distribution that's highly unlikely. And you'd still want to use that. So now, we might want more compact ways of describing this distribution. Actually, in the age of computers, I think, with lots of big disks and lots of great data collection, one of the most underutilized opportunities is keeping around all the raw data, and actually using empirical distribution.

So in some sense, it's kind of a holdover from the days when it was really tough to keep track and use vast amounts of data. And we almost always would start to say, OK, I'd like a compact way of describing this, and I want to describe it as something that only has maybe a couple of parameters. It has a mean and some variance component.

But that's a different philosophical point. There's, actually, I think, a huge amount of statistical methods that are underutilized that deal with sampling or resampling from raw data that actually can capture subtle things going on in the data. We're, in this course, not going to be using those things. We're pretty much going to seek compact descriptions of randomness in ways that we can reason about.

There are data-intensive ways of doing the same kinds of reasoning. So one important point here is we're often going to take this discrete binned kind of information-- we want to characterize it with perhaps a continuous parameter variable and a compact distribution, such as a normal or Gaussian distribution. And this is especially good when the outputs are themselves truly continuous parameters, like a width, a dimension of the part.

And one point here, just to remind you-- especially when we go to continuous parameters, you've got to be a little bit careful about talking about the probability of your part or your variable taking on any specific value in that continuous distribution. Back here when we had things binned, I could talk about the probability of measuring something in that bin, within that range, 1.485 to 1.487.

But now, if I were to ask, what is the likelihood that I measure exactly 1.486, be careful. That is a very, very discrete value that-- when we talk about continuous probabilities, the probability of getting any one value-- and it's truly continuous-- is 0. You're never going to see exactly that.

Now, with real measurement tools, we never have perfect continuous measurements either. There is always a precision and a discreetness capable from the tool. So this is an abstraction dealing with the mathematical niceties. I'd just point out it's not exactly true and measured reality and data.

So very often, when we talk about probabilities, we got to be careful. The probability of any one value is 0 what we talk about really is the probability of our measurement or whatever being in some range that's inherent in the distribution. But we can also talk about, for example, the probability that the value be in the range from minus infinity up to some value y star, and that is a reasonable question to ask in terms of probabilities.

And that is embedded in this notion of a cumulative probability function, or cumulative density function, or CDF. Just to illustrate that a little bit more, we can do the same thing, have a cumulative frequency with discrete data. So if I had that underlying histogram here, I could also build a cumulative frequency histogram that simply was giving me the probability that a part or measured value was up to, but less than some particular value.

And that's simply the integration across the PDF up to that cumulative density function. And we can do the same thing that we did with the discrete probability frequency relationship. So we can approximate the frequency with a PDF. We can also look at the cumulative frequency and approximate that with a cumulative density function, or a CDF.

So we can have that kind of an equivalent. And those equivalents also, again, help us interpolate and ask probability questions also on values that may not have been exactly lined up with our bins. So very often, we're going to be, I guess, talking either about the density function-- the PDF of some variable-- or the cumulative probability.

And there is this inherent relationship between the two. I describe the cumulative probability as an integration up. The probability that some value is less than or equal to would be an integral up to some x star. And so that value right there corresponds to that. And then the inverse is-- relationship is true as well.

If I had a cumulative density function, I can differentiate that to get down to the probability density function. So this is some basic statistical concepts that you probably have seen before. Many of you have seen it in much more detail in 2.853.

OK, so very often, the histogram does suggest an underlying approximate or convenient probability density function. Very often, we've talked about normal distributions, but a key thing to look at is always look at your raw data, because there may be other perfectly reasonable continuous or discrete probability density or probability mass functions that apply in the discrete case, such as a uniform distribution.

And so one of the things that we'll see is the family of some typically emerging or typically observed kinds of distribution models. So very often, what we're doing is we're gathering data, looking at these histograms, and looking to see, is there a consistent pattern, such as we saw with the underlying normal distribution?

And we're often then trying to intuit what that reasonable underlying distribution may be. There are ways-- and we may touch on them a little bit-- to be able to test whether or how good a particular distribution model may fit with that data. And on the problem set, in fact, you'll explore things like a qqnorm plot.

And I'm not going to go into detail here, but essentially, that-- assume a model-- for example, what the probability density function probabilities would be associated with certain kinds of observations, if it were a normal distribution-- and then you plot your data against that and start to see how well the data fits that particular model.

Let's go back to that original set of data, the [? c and c turning. ?] If I were to plot that as a distribution, I might get something like this. I might argue that my bins may be too coarse here, but what do you see? This is a little tricky, right? If I were to plot that data as a distribution, I might be tempted to say, OK, that's one big Gaussian distribution.

You could try that. But again, if we look back at the data, another interpretation here is, really-- you can start to see it-- it's bimodal. It's a mix of two distributions. So I guess the simple point here is don't always assume it's going to be just a single distribution that's underlying it. There may be a mixture.

Similarly, here's another set. And very often, again, plotting your empirical data in time order tells you a lot more, but also be sensitive to those kinds of questions. I could have started the whole class off and never shown you the raw data, and just started plotting distributions. And in some cases, things might be more observable there. Usually it's easier to see perhaps lurking up here, when there's such a very clear time shift.

But here, this is the histogram associated with these, and it sure looks like one normal distribution here and another normal distribution here. Now, by the way, if these things were ping ponging back and forth, or they were randomly picking and it was one of those source material deviations, it might actually be hard to detect that, whereas the bimodality-- the fact that I've actually got what looks like two mixtures-- may actually appear more easily in a histogram kind of form.

There's another example. Here we start to see, again, these two shifts, and now you can start to convert these perhaps to two statistical distributions, two Gaussians with different means. And now you can also estimate parameters, like what the mean shift is. So once you've got distributions, now you've got mathematical approaches for being able to estimate those parameters. So this is just going back to some of that other data.

OK, so some of the key points here out of this portion is, even if there's no strong input effect, there may still be disturbances. And we'll often see a consistent histogram pattern emerge. And what we often want to do is use that either to detect deterministic deviations and squeeze them out, or once we've done all those process improvements and gotten those all squeezed out, we may still have an underlying inherent distribution. Maybe it's random variation associated with a Gaussian distribution.

And the key idea is we can even use, and importantly, we want to use knowledge of that underlying pattern. It's not a systematic pattern. It's a random pattern associated with that distribution-- so that we can predict behavior, and in particular, set limits on normal behavior. This is a nice case. Maybe this is what we'll consider normal behavior.

And again, I want to be able to statistically say, what's the likelihood that some of these points or a new point belong to the natural variation or not? So we're going to typically use analytic probability density functions to do that. I think we've already talked about most of these things. I've been emphasizing, by the way, continuous values, these width or geometry parameters, but there's also a set of different statistical distribution than the normal distribution associated especially with the occurrence-- random occurrence of point defects.

And a little later in the term, we'll actually do some yield modeling dealing with things like Poisson distributions and in these sorts of discrete distributions that are more appropriate to discrete events or discrete occurrences. We're going to be, in the first part here, dealing pretty much with continuous parameters.

So this is just laying out, again, a little bit. We've talked about some of these. I want to build up just very briefly, get you going on some of the basic definitions. And then Chapter 4 in Spanos really articulates these much more. So we have our probability density function and cumulative density function, and there is that relationship between the two.

And then we can ask about certain moments or characteristics of these continuous probability density functions. These should be familiar to. We can ask, what is the mean or the expected value of one of these distributions? And here I'm using t as a stand-in for time or run number.

And if I have a probability density function associated with the random variable, and even time, then there we have a definition of the expected value of x as a function of time. Now, I'm kind of being pedantic here and giving this definition, because we also saw trends-- long-term trends in our data. And we could ask what the mean is as a function of time.

Again, we're in a situation where I'm trying to squeeze out those drifts. So usually we're in a situation where we want to avoid non-stationary drifting processes or analyses where some of these functions-- the probability density function and its parameters, like the mean-- drift over time. We typically want to be having stationary processes where the mean and other moments, like variance, are independent of time.

So in the stationary case, which is what we're mostly going to be dealing with, now I've removed the time dependence on the underlying statistical model. Let's say, from time to time, it should be stable. And then we have a mean, this mu of x-- the true, or theoretical true mean-- that we would calculate with the same formula. But again, the idea of the stationary process-- important characteristic there is that the mean is a constant, and not wandering or drifting.

Similarly, in the case when it's a stationary process, the variance, defined here as the second moment-- the expectation of the distance squared of your data from the mean-- [INAUDIBLE] expectation of that-- is the variance. That's just the definition. And in the stationary case, that is also a constant of the process-- has a nice alternative little description, if you work through the mathematics, where it can also be calculated as the expectation of the square of your data minus the mean squared.

By the way, you can ask the same questions-- variance, mean-- for any continuous distribution. They're perfectly good moments. In some cases, they completely characterize the distribution. In other cases, there may need to be additional moments. And an example here is we could ask, what is the mean of a uniform distribution-- uniform, where the probability density is equal across an entire range?

And so one could calculate those as well. Everybody is seeing the normal distribution. Again, we can actually have an explicit description for what the PDF is. We'll often be talking about the unit standard normal or unit normal distribution. The full distribution here actually has the mean in it, and it has the variance in it. So for your particular process, those may be very different for whatever collection of data.

In order to be able to talk generically about properties of the normal distribution, and to use tabulated values and so on, this normalization, where we subtract off the mean to get a 0 mean description, and then we divide by the standard deviation, gives us the unit normal distribution. And that's exactly always this distribution in terms of z units.

Now we've made it independent to the mean, independent of the variance. And the exact values, the probability densities for any values of z are well-known and tabulated. So we've made it 0 mean and made it unit norm-- unit variance. There is also a cumulative distribution here. I'm describing this as p of z, so now we can ask the cumulative distribution function for that unit normal, and plot that.

For the normal PDF, what's nice is this is the whole distribution, and all we need are those two parameters, mean and variance, and it completely characterizes the process. And then there's some statistical operations that we can do for when we-- that are really convenient when we have normal distributions.

For example, if we have multiple effects, we can look, and the sum of multiple normal random variables still has a normal distribution. So I'm just giving you a little bit of a peek here. You can do some reading, which is probably more effective than me just flashing by these things. There are important operations, for example, where, if we have an output y that is a sum of multiple random variables, there are things like the mean of the overall y is simply the sum of the mean of each of these individual random variables.

And if there's a scaling in the mean, that simply scales. There are also operations when these-- if these are independent random variables, the variances add. And you've got to be careful any constant. If you run that through the definition of variance, those are also squared. So you'll see the basic description of some of these mathematical operations and these distributions in May and Spanos.

And you'll also start to get a feel for using these distributions to ask some very standard questions, like what is the probability that you would observe a value in some range? So I think, with that, we'll close. If you want to get started on problem set 2, there's some problems that-- this is just basically giving you some familiarity with manipulating things with Gaussian distributions and other statistics. And then we'll start building up more detail next Thursday.