DUANE BONING: So what I want to do today is pick up a little bit from last time, round our discussion of process optimization with a little bit of a different perspective today. Last time we talked about response surface methods and using those in optimization. What I want to do today is especially focus on-- geez. This arrow is not letting me draw. I want to focus on process robustness, and that's a little bit of a different perspective.

So far, most of the optimization we talked about was looking at getting a model of the mean response and then using that to drive the process to an optimum target, if you will. So we want to come back a little bit to ideas of things like CPK, quality loss function, and so on, that start to fold in additional requirements or additional goals in the optimization. Coming out of this is little bit more of an explicit consideration of variation modeling, not simply the mean response, but also thinking about this assumption that we've made that the variance is the same everywhere across the process and what to do when that is not the case, because in fact, it often is not the case.

So looking at the family of things that we might seek to optimize, there are really multiple things. Right? Some of which are easy to make in a quantitative fashion assessment of, models of, and then optimize against and some that are a little bit more qualitative or additional factors. So certainly things like quality is the main effect that we've been looking at.

That is to say, we're looking at the response of one or more outputs as a function of one or more control factors or inputs and then seeking to optimize the overall yield, overall output of the process. You want to be on target but you also want to be as immune as possible to variance. And we talked earlier in the term, not so much last time on Tuesday, but earlier in the term about CPK and quality loss functions.

But also, of course, another key driver in all manufacturing is cost, and there are costs associated with yield, of course. You're only getting revenue for things that are good. But there's also other cost factors, for example, picking of different points in the operating space may imply different things about cost of running of the process. And in fact, one of the important factors of cost is associated with rate.

Very often your process factors or what operating point you pick for your process has a direct parts per hour or other kind of throughput or rate implication, right? In fact, you'll very often have this trade off where you might expect if you run things faster, your quality might be a little bit reduced. Whereas if you run things at a reduced rate, you may be able to boost up yield or quality, but there's a trade off then in terms of how much production you can get out or how rapidly you can get that out.

In many of these cases, if you can actually fold all of those down into a direct dollar trade off then you can start to formulate an overall cost function or objective function, sort of an overall J function in our terminology from last time, that might actually trade off some of those things. But often, it's actually quite difficult to get really detailed information and build models of all of these factors. Now an additional factor is flexibility.

That's a little bit harder to get quantitative about. What do you think we mean by flexibility? I mean, we could find the process optimum, lock in our equipment, lock down the control knobs, and say that's where we're going to operate for the next four years. That would probably not be very flexible. But give me some thoughts on what kind of flexibility you might want in a process?

AUDIENCE: Change over time, like the process is easy to modify and change.

DUANE BONING: Excellent. Yeah, so the idea there was change over time. So a process. So some of those might be goals actually in the equipment design or the line design. Not necessarily picking an operating point, but definitely flexibility is the ability to rapidly do changeovers.

By the way, that often folds into some of these other factors, right? Can affect cost, quality, rate, change over time, how rapidly you do that, can have an impact on throughput. Changeover sometimes requires additional steps to warm up or prove in the new changeover. So furthermore if your equipment requires a little bit of adjustment after a changeover, it can have a quality impact. That's a good one. I like that. Other kinds of flexibility?

AUDIENCE: You could [INAUDIBLE] and try to make it more insensitive to the exact operating point to try to [INAUDIBLE] development, rather, something that gives more room for the process to [INAUDIBLE] on either side [INAUDIBLE].

DUANE BONING: Yeah, so what Neil is just saying is you might want to process that is perhaps less sensitive at an optimum. So this would be less sensitive so that as your inputs vary, your outputs would vary less. I think that's an important characteristic. That's a notion of robustness. I'm not sure I would associate that with flexibility, though.

I think it's a slightly different idea. I mean, certainly-- I'm trying to think how to think of that as-- it may be that what you would like is a process where this same characteristic might be true of multiple operating points so that one could rapidly, rapidly tune in the process, rapidly change over the process. Yeah.

AUDIENCE: Different rates. Output rates.

DUANE BONING: Different output rates. So what would the benefit of that be?

AUDIENCE: [INAUDIBLE] demand. We don't want to overproduce.

DUANE BONING: Interesting. Sure. So rate flexibility might be another characteristic. So I think we could identify multiple of these flexibility ideas. I think these are really important characteristics, especially in overall product design as well as process design.

We haven't really talked that much about it. We're really focused mostly, in this subject, on the manufacturing process control and the quality issues. But I did want to have that in there just as a reminder that there's a lot of these other kinds of factors that are important in a manufacturing setting. But let's focus a little bit, look at minimizing, well, quality loss and thinking a little bit more about other cost factors.

We've been saying that one of the important goals is we want the-- I'm not sure I would use x bar here. Let's call that y bar. If y is my output, I want my y bar output to at least hit or come as close as possible to some target. So again, if we assume that there is, in fact, some spread in the manufacturing process, I can never completely squeeze out all the variance.

We have our typical CPK-like picture where we've got some target, we've got spec limits, upper and lower, and again, we want to maximize yield. So one important characteristic of that is getting on to target. It turns out if you have a fairly rich process-- rich in the sense that you've got multiple inputs that you can use to affect the process-- maybe you also have multiple outputs. But for the moment imagine that I've got one output of concern, but I've got more than one x control factor or input factor.

Conceptually do you think there's going to be more than one solution to or just one solution to a combination of those input factors to hit the particular output? If I only had one input and one output, seems like most likely I'm only going to have one setting for that input that is going to correspond to a maximum on that output. But if I've got multiple ones and they have some influence, you got an interesting situation here, right?

You end up with multiple solutions, if you will, to the maximum output problem. That is to say you may have contours in the multidimensional space, an x1, x2, y space that all achieve the same output. Turns out that's wonderful. It means, well, yes, you can match to target. But now that second input or the additional inputs give you a little bit of freedom to try to achieve other goals as well.

Now that's assuming that sort of the match to target is the paramount goal, and getting close to that is pretty close to a paramount goal because of the direct effect on yield and quality loss, but it wouldn't necessarily have to be so. But here I'm sort of assuming you've got to get to target first, and then after that it's an opportunity to choose your operating point to do other things. To do things like minimize the cycle time to get as fast a rate as possible or to start to fold in other qualitative or quantitative cost drivers.

A couple of simple examples might be at different operating points you might actually use more input materials or, say, manufacturing gases in a plasma reactor. At different operating points, you might be more or less efficient in the use of energy or other material resources. Another factor might be things like at different operating points, it has different side effects on the equipment, like wear of a machining tool. OK?

So these are places where you can start to add those in. And that actually means that you really would like to have multivariable models in the sense of some output y that in fact, does let you trade off two or more inputs. And this is just a visualization of what I was just describing earlier. Here's a very simple linear model, two inputs with an interaction. And now let's say that the target value for y is 5.

Well, that plane intersects with my response surface along some line, and you can sort of see that dashed line there. That indicates any combination of the x1 and x2 factors that can achieve that output. And so now, in fact, to find and decide what x1 and x2 I'm going to use, I actually need to think about some other criteria. Right? I guess if it doesn't matter, you can pick any of them and you might randomly pick.

I guess randomly picking is about the only thing you can do where you're not, I think, intuitively applying some additional filter or some additional criteria to the x1 and x2. So again, some of these might be picking x1 and x2 based on things like cost or rate. What are some other things that you might use to decide on which point along this space you would actually like to pick for an operating point? Yeah?

AUDIENCE: Minimum variability.

DUANE BONING: Minimum variability. Good. And we will talk about that for sure. Excellent.

AUDIENCE: If you were having [INAUDIBLE] x1 but not x2, then you will use x1 to control most of the [INAUDIBLE].

DUANE BONING: Interesting. Yeah, so that's getting at the notion of flexibility, so the idea was if I had some other product that also was using x1 and x2, there might be value in keeping one of those variables constant or using it to optimize all of them and then focusing in and letting one variable be more associated with the flexibility, if you will, the thing that you use to tune in across different products or different processes. I like that. Yep?

AUDIENCE: Physical constraints. Your machine might not be able to go that fast.

DUANE BONING: Physical constraints, absolutely. So depends on whether these are rate or other things in terms of whether the machine can go that fast, but certainly you may have explored a space in these coded variables in x1 and x2 that cover this broad range, but perhaps your equipment really doesn't want to run out here at one of the extreme low points. So that might actually drive you away from these extreme points or maybe it drives you towards one of those extreme points.

In general, I think-- let's see, I believe one of the case studies-- it's either case study two or three. Dave Hardt will come in and talk about run by run control, a little bit of use of some of these kinds of models for feedback control. And in general, an observation that comes out of that is you might want to avoid operating at one of the boundaries of your process, right?

If you were, in fact, picking one of these points that's at the edge of the space that you used to build your model, well, you're not really sure what happens if you adjust x to a little bit beyond that space. You might hope that the process continues to be well behaved and extrapolate out into that region. That's a little bit scarier to do.

But in general, if you pick a point that's closer to somewhere in the interior, almost maybe even the center of your space, that gives you the maximum freedom in an active control setting to make small modifications to your x1, x2 factors to drive the process back onto target. So I would be inclined to also try to sort of keep my factors in a place where I know I can keep moving them to control the process.

AUDIENCE: [INAUDIBLE], doesn't it make sense that if possible you actually need to test further out because you could have more optimal conditions if you could, say, push the boundary?

DUANE BONING: Yeah. So the question-- yeah, yeah. The question is kind of a general one about response surface models and optimal points. Is if you are on some boundary-- well, first off if I have linear models, very often you get to the constraints being a boundary especially if I fold in something, some additional factor. And so you could always ask the question, should I extend my experimental space and explore a little bit further?

And the answer is if your variables naturally can go beyond your initial guess of the range of exploration, absolutely. And as we saw last time, for example, in incremental or iterative optimization, you may have actually already-- I guess I'm sort of assuming you've already explored the space to cover the optimum region as much as you can.

That's a good observation as well. Good. Any others? Any other ideas? OK. Good. Good. I did want to hit on some of these.

Now we'll come back to this minimum variance because that is a great idea, right? So far here this picture is still looking at, even in this picture, it's still looking at just output and matching that output to the target, that shaded plane is hitting the target. And a bigger goal [INAUDIBLE], even just focused on quality is overall yield and minimizing either the effective variation or minimizing the variation.

So if we go back to our quality and variation equation, my output can vary due to a couple of different factors. We said alpha are sort of noise factors, and these are the actual process settings if you will or control factors. And so we can see that they'll have some sensitivity of the process to each of these kinds of factors.

If I tweak my x factor or u factor a little bit, my y is going to change some amount and normally I can control or I have power over those process settings, those options, but sometimes I don't. Sometimes I can dial it in, but I don't always get exactly what I dialed in. So there might be differences in sensitivity of the output where little deviations in x, especially with a non-linear or polynomial kind of model, might actually imply different amounts of delta y at different operating points.

So that's an example where our assumption about-- let me just get some of you down here-- equal variance everywhere in the process space might actually not apply. But also, we have these noise factors. We have these variation sources, uncontrollable factors, things that we can't completely eliminate, and we might want to find an operating point where we minimize the sensitivity to those noise factors. OK?

So one way of thinking about this as we look back at just a sensitivity kind of analysis, where can I look at either using my response surface model, which usually has this in it, and minimize some sensitivity metric on that or start to try to deal a little bit with these unmodeled noise factors. One approach here we're going to talk about is, what if I want to include those in the model? I actually want to understand how variance may change as a function of operating point.

This is going to be a little bit overcoming the assumptions that we made. Now how can I get minimum variation? I mean, so one approach here is think about it from the sensitivity point of view. The other is think about it from a CPK perspective. You know? And we talked about needing to center the process as one part of that, but the other part would be the variance part.

And again, here if I'm doing anything other than mean centering, I have to know how the variance actually depends on anything. Our assumptions all along in most of what we've done so far, ANOVA and other kinds of tools that we've looked at, have sort of assumed that the variance was constant. We want to back off and relax that as well.

And another perspective that we'll look for a little bit here is to look back at the quality loss function, which can fold in together. Remember the quality loss function that says, as I deviate my output from my target I incur more and more loss? It actually has a penalty, not just it's good as long as it's above the lower spec limit and below the upper spec limit, but as I deviate from the overall target I have continuing incremental worse behavior in the product. So that's sort of a generalization a little bit of the focus away from just the target to also the variance.

So what this is saying is for many of these approaches, whether you look at it from a CPK or a quality loss function, you need to know more about the variance or you need to know more about the influence of the noise factors in order to really get to process robustness. With what we've done so far already, you could do a little bit of robustness just with the response surface models just in terms of reaction to modelled input, but that's a pretty weak notion of robustness. That's only looking at some of the variances.

Well, let's explore the CPK one. This is actually all just reminder. We've said CPK is a nice measure that looks at two effects, right? So the CP has an effect from the deviation from the mean of your process where the true mean is and how far away it is from your spec limits, so that's sort of a centering or a target point, and then it also has a notion of folding in the spread in your process, right?

So one could, in fact, formulate something like CP or CPK as your cost function, your penalty function, your J, and try to build a model-- either use your model or build a separate model-- and minimize some cost. Well, in this case, maximize the CPK because a bigger CPK, we said, corresponds to yield. In fact, we'd use different formulas. You could actually relate a CPK to a fraction of nonconforming parts given different values.

Another approach would actually be to extend that almost directly and say, what I'd like to do is actually use J directly or a model of CPK. Build a response surface model, if you will, of CPK, and that would be a way to fold in both mean information and process variance information. What are some gotchas in trying to do this? What are some difficulties?

We said one of our purposes here, it might be in fact, pick your process operating point to maximize CPK. Why not just directly use that? Use this function right here. There's my function, I want to maximize K. Let me broaden the question. What are some good things or some bad things about this? Yeah?

AUDIENCE: [INAUDIBLE].

DUANE BONING: OK. I'm sorry?

AUDIENCE: [INAUDIBLE] is going to depend on y.

DUANE BONING: OK, so the comment there-- by the way, you need to try to speak loudly. I'm trying to repeat the questions or the answers as much as I can, but I think it's a little hard for people to hear. So I'm sorry, now I lost the question-- or the answer.

AUDIENCE: One of the things might be that the sigma might depend on mu or [INAUDIBLE] may depend on [INAUDIBLE].

DUANE BONING: So one of the complexities you said is this might actually be a function of the process conditions. Yes. In fact I would even almost say, this approach here where I built a quality overall CPK and if I were able to actually evaluate CPK at each of my process operating points, that might actually be a way to incorporate or accommodate a change in variance or a change in standard deviation.

Whereas in some sense, if I build a response surface model up front with the assumption of a fixed variance, it might then be hard to actually use that fixed model with just J as a cost function. So these are two slightly different ideas, and I'm glad you mentioned that because-- I mean, this idea, the first idea up here is basically saying, I've already built my response model, maybe for y. I might, might, talk about a response model for s, and then I just use the model or use CPK as my cost that I'm minimizing.

The second idea, in contrast down here, is build a new response variable and actually build a model directly of CPK as a function of my input conditions. And then I optimize on that. They're slightly different ideas because in that case, in some sense this case is, perhaps, assuming we're mostly having a fixed s. This one, as I said, can perhaps more easily accommodate a process varying standard deviation.

There's one other potential challenge with the model down below. Well, either approach, really, with these two functions. What's something that's a little nasty about these functions? They have things like a min function in them. A minimum. And minimum between two variables can be very non-linear, so it may not be that smooth of a function. There may be a point at which you actually start to break more to the left side or to the right side and have not a continuously varying parameter.

So it can be a little bit of a challenge from a modeling point of view as well as in optimization. It's kind of a nasty function. OK? And in fact, that nonlinear nasty behavior with its possible discontinuity is-- let's see, I want to-- ope, I guess I don't have it in here.

It's kind of one of the main drivers for the whole notion of the quality loss function, that E of L, that quality loss function with a quadratic dependence on deviation from the mean is a nice way of folding in together in a more continuously varying function, both deviation from target and effects of variance. So we'll come back to that, but in fact, that's one of the main reasons why the quality loss function is used and has an advantage is it has a smoother behavior, a behavior that drives towards an optimum that pulls in or folds in both of those effects and doesn't have these discontinuities or not quite as much of the nastiness of these min functions. OK?

OK, so I guess this discussion and everything we've been leading up to is really starting to say, OK, look. In order to be able to fold in both deviations from target but also variation, we need better ways to model the variance. In particular, the variance that might not be constant at your entire process base but may vary as a function or depend on a operating point. OK?

Now remember, in some cases we've been careful to remind ourselves that there were assumptions on constant variance in our approaches. But sometimes we haven't even been explicitly mentioning that, but I want to highlight that some of the methods we've been using and developed actually implicitly continued that assumption that sigma squared or the variance was constant throughout the space.

So in ANOVA, the simple ANOVA that we did, in order to decide if an effect is real we were always comparing a fixed offset to natural variance in the process and we were always pulling the estimate of variance across all operating points equally. OK? So we were enforcing an assumption that there was only one natural underlying process variance at work. And then we were trying to say, are there variances associated with fixed offsets that are larger than the replicate noise?

So that was in there and we were careful in talking about that. But another place was actually in regression model fitting. When we formed a least squares regression problem and then solved it, that actually equally weights all of our data and all of our deviations from the model prediction equally and essentially, is also a equal variance assumption. Think of it this way. If I had, in fact-- wait. Let me draw it.

If I in fact had different operating points or locations where the variance was very different and I'm trying to fit a line through this, but over here I have a huge spread. It may in fact be the case that I have less confidence or a wider spread here in picking a point that corresponds to where that line is going to go than in other regions of this space.

Now they can still appeal to best estimates, but in terms of confidence in the use of each of your data points, there are alternative regression approaches, weighted regression approaches that might say in regions where I have less variance-- where I have more confidence, I know those data points are more solid-- I might actually give them more influence in the weighting in the regression than other points that I'm not so sure about. And so weighted regression approaches are available, and probably the most common weighted regression is to apply a weight that's inversely proportional to the variance associated with each of the operating regions in your space. OK?

So just wanted to highlight that there are approaches that go beyond this assumption. We've been assuming that in both our regression analysis and our ANOVA analysis that we have one variant. I think I alluded to a book earlier this semester called *Analysis of Messy Data.* It's a wonderful book.

Or actually, there's a series of books. There's either two or three of them. But one of the examples of messy data is when the variance is not constant and there's one of the volumes in that series is focused, I believe, on analysis of variance when you don't have constant variance throughout your space. So there are places you can look when things get messy. OK.

OK, so what do we do? We've been assuming that but in reality, I think we've made the point. The process variation may in fact vary or depend on our operating point. I've already talked about the impact in our variance equation. If I had imperfect control on a control factor, I may have different sensitivity-- that dy, du-- especially with a non-linear model can vary. But it's also just simply the case that the effect of these noise factors may be such that I have different variants at different operating points.

So what are some approaches? All I've said so far is variance may depend on your process settings. How are you going to study that and understand it? Well, let's do design of experiments but also include an explicit model for variance or standard deviation.

So all we're doing here is simply saying treat variants as another process output. And so it is possible to include another response variable, either s squared or s or some variance model and do our typical DOE here on our input factors, but now get enough data that I can actually build a model of variance at each of my operating points. Can I possibly do that without replication?

No. I have to have replicates at each of my design points, each of my DOE points. I have to gather multiple replicates in order to have an estimate of the noise at that replicate point. So fundamentally, you have to do more than a non-replicated, full factorial DOE like we were talking about that just explores the operating space. You've got to get more data if you want a model variance. You always have to have much more data.

So you have to pick some number of replicates at each of your design points. And now this goes back to the tools we talked about earlier in the term in terms of deciding what sample size you need in order to be able to have different confidence intervals on your estimate of variance or standard deviation. Right? So depending on how close, how accurate you want to be on your estimate of a variance, that will influence how many replicates you need. Yeah?

AUDIENCE: To get this at each level, you do [INAUDIBLE].

DUANE BONING: Right. Right. Absolutely. So the comment or question was, to do this you actually need replicates at all of the points. Yeah. So in some sense, the DOE that we were talking about, this would be, say, full factorial, and then if I picked my center point we talked about doing something where I had replicates only at the center points. And that's wonderful if you're continuing to assume that there's only one variance and it applies everywhere.

In fact, that's kind of the assumption we made, even without necessarily having good evidence for it. But clearly here if I want a model of both the average output but the variance at each point, I mean replicates at each point. And you would normally want to do that where you have the same number of replicates at each point.

It's possible you could do it without a balanced experiment where you had different numbers of replicates. Then it gets really messy because then your confidence interval for your estimate of variance at each of your points is different. You'd like to have at least the same confidence interval or the same number of replicates. Now there is some danger in some of these kinds of models. I wouldn't quite say danger, I would say some approximation going on.

So for example, one thing that makes me a little bit uncomfortable is actually building a model directly of standard deviation, because estimating directly standard deviation is difficult. It's got bias in it. Whereas building a model, perhaps, for example, of variance, you can use the data and get an unbiased estimate of variance. Whereas you're susceptible to slight biases if you go in and directly model standard deviation.

But for the most part, people pretty much ignore those. And are basically-- I think assuming that those kinds of biases get washed out in the noise because I don't have so many replicates. They tend not to be all that careful with it, but I would just put a little side note here. We should be a little bit careful about the influence of bias in these factors. And in fact, I think in one of the case studies we'll talk about approaches for modeling variance and some of the biases that can creep in.

But the main latch here, the main lever for being able to get some power over the problem is simply doing a DOE that includes explicit modeling of the variance so that one can build up some kind of a variance or a standard deviation response surface. And then you can use that in combination with your output to form a cost function that seeks to minimize in some way both of those factors. So for example, now we can go back.

We can go back to that surface that we had. That x1, x2 surface where I found the line or some function that minimized the error in my output from the target. But then I can explore over that within my x1 and x2 constrained space and look for the place where I've got the minimum variance. Right? So if I took back-- if I go back here to slide five-- I just heard a bell. Is everything still working there? You still have the slides in Singapore?

AUDIENCE: No, we don't We don't have the slides, it was just disconnected.

DUANE BONING: Yeah, I just--

AUDIENCE: You might want to share the desktop again.

DUANE BONING: OK. Yeah, it looked like it dropped the call. So while that's calling, so the basic point was if I have an additional response surface for variance or standard deviation, the simplest use of that is simply to still do my optimization on the target and then just look up or use this to disambiguate my xy parameters.

AUDIENCE: When you have [INAUDIBLE] experiment that we're doing, [INAUDIBLE]?

DUANE BONING: Yeah. So the question was, let's say you only did four replicates at each point. Well, you ask the question, how good is your estimate of variance at that point? And you know how to calculate that, right? And your point is with only four replicates, it's kind of iffy about how good you're-- you're only going to have resolving power to be able to see really big differences in variance.

If it's small differences in variance with that number of replicates, you're not going to really be able to disambiguate. So in essence with that small number-- oh, thank you-- that small number of replicates, you're really just looking for gross variance differences. Which if they exist, you definitely want to use.

Now, if you're looking for really trying to get a good process that has the ability to take advantage of small variance differences-- because over many, many thousands or millions of parts, even small variances can really make a difference. Then you'd want to run more replicates in your design of experiments. OK, are we back at Singapore now? You have the slides there?

AUDIENCE: No we don't.

AUDIENCE: They had it and then they dropped it.

AUDIENCE: Yeah, we can't see the slides.

DUANE BONING: OK. So this one, it thinks we're in a call. Should I unshare? What should I--

AUDIENCE: No, I think it's coming back.

DUANE BONING: Oh, it's coming back?

AUDIENCE: [INAUDIBLE]. Yeah, yeah. It's back.

DUANE BONING: OK, great. Great. Great. Good. Great, so here again, the first and easiest idea for combining models of both mean with a y and variance is use the variance to simply pick which of the x1, x2 combinations such that you're on target, give you minimum variance. Now you can also do that directly and form an overall optimization, and this is just working through that.

This is the direct approach where if I have a response surface for my mean and a response surface for my variance, I have a constraint here first off that I need an optimum point that satisfies my output hitting a target. And then I can simply solve together, pick some x that satisfies that, and plug that into the s surface. Continuing on, substitute that in into the s surface so that you basically get some function where I need to find the x sub 2 that minimizes my variance.

And then I can do a direct analytic variance minimization to solve for x2 and once I have x2, I can solve back for x1. So all I'm saying there is you can just algebraically use your two response surfaces. If they are simple and linear then it's nice and easy to drive that towards where you have the best y hat that's on target that also is minimum variance.

Another approach, once I have models for both-- again, if I have a model for the mean response and a model for variance-- is think about other cost functions that combine both of these effects. So this is a bit more general because it allows us to look at, say, the quality loss function and have terms that penalize us for deviation from-- I don't-- well, penalizes for deviations from the target. I'm just looking at this and saying, I don't know why we-- let's just stick with y here everywhere. I don't know why we switched to x.

So when I have deviations from my optimum point from the mean, that's going to cost me something and then I also have a term associated with the variance. And if both of those are functions of my input parameters, now I'm finding again the best input parameter that minimizes the overall cost function. Now there's a subtle difference here between-- or more power between one of these approaches.

This approach versus the direct solution approach in the previous slide. The direct solution on the previous slide basically said, I have two outputs and two inputs and I'm trying to pull those two things into one overall optimum. Yes?

AUDIENCE: Professor?

DUANE BONING: Yes, question?

AUDIENCE: I think we cannot see-- I think we can see a slide but then we cannot see any drawing, and actually we are now stuck at slide five.

DUANE BONING: Oh, OK. So that doesn't sound like it's changing there.

AUDIENCE: We don't really know which slide--

DUANE BONING: Yeah, I'm on slide 14. Yeah, I'm sorry?

AUDIENCE: What we have right now is slide 11.

DUANE BONING: Yeah, so I'm not sure if the sharing is working right. Are there just long delays? Should I unshare and reshare?

AUDIENCE: Yeah, I would do that first and then [INAUDIBLE].

DUANE BONING: Oh, OK. Well, let me reshare here. Oh, we killed the coffee. OK, do you guys see slide 14 there in Singapore yet?

AUDIENCE: Yeah, we can see it right now.

DUANE BONING: OK. Great. So let's see if this works. We'll try continuing here. The point I was going to make is if I go back to slide 13-- see if you got 13 now. Did 13 come up?

AUDIENCE: We can't see on the screen, but we can see behind your projectors.

DUANE BONING: Oh, OK. So that's what we're using in the-- all right.

AUDIENCE: Yeah, it's all right.

DUANE BONING: OK, we'll just use the projector then. The point I was making here on slide 12 and 13 is in some sense I've got two outputs and two inputs. I can directly solve the x1 and x2 to achieve both a y on target and then a minimum s. If I had three x's or four input factors, this approach-- well, I might have additional power in terms of picking points.

But this approach actually can-- I guess it can work as well if my variance actually is different. It has a unique minimum. But the point I wanted to make here is that with minimizing of the quality loss function, it's a little bit more natural in the output and can combine or led to explore over a different mixture of input factors a little bit more naturally. OK?

So it's a little bit more general than assuming immediately I've just got an x1 and an x2 and I'm directly solving for the outputs individually. What it's letting me do is essentially combine them into a single cost function, and it also gives another little bit of flexibility that is an important difference between this and the previous approach. In the previous approach, we said the constraint is your output is exactly on target. Is that true here?

No. When we're minimizing this combined penalty function with two terms in it it's actually a little bit more flexible in another important way, which is to say it's allowing you to maybe pick a point that's off target slightly because it's a lower variance point and overall you'll win in terms of yield or quality loss. So it's a little bit more flexible in letting you make that trade off between variance and being exactly on target. It doesn't presuppose that your mean on target is absolutely your best point.

So I like the quality loss formalism a little bit if, in fact, I have a variance that also depends on what operating point I'm at, what x factors and therefore also what y point I'm at. It lets me combine those and trade those off in a natural way. Now by the way, the classic quality loss function, they also use a constant factor between the variance and being off target.

A generalization of that is you can actually use different weighting factors in those two cases. You may actually have, I don't know, a customer that says, I really want my mean to be-- maybe your customer actually has some reason to wait or be more concerned about variance. He likes a tighter variance, but actually has-- as long as a mean is somewhere in my spec region, I'm fine.

It's kind of an odd combination, but there might be reasons where you might weight variance or being near target a little bit differently, and so that's one nice thing about this minimizing of the expected loss. You can actually weight those. OK, so what are some challenges here when the variance is varying?

One challenge with minimizing or variance [INAUDIBLE] that we've already talked about is you have to model it. And what I want to explore a little bit are some of the response surface model or approaches for actually building up models of these dependencies. Are you guys back? I just heard another bell. I'm just going to keep going.

AUDIENCE: No, we're not but it's OK. We can see the projector.

DUANE BONING: OK, we'll continue with that. I'll ignore the little bells and the birds chirping and whatnot. Now as we start to look at ways to build the model of the process we can return to the question of what is it that causes non-constant variance. And part of the effect is that our noise factors may have different influence at different parts of the operating space.

So one approach for building the model and trying to get places that are robust, that is to say insensitive to variation, is actually explicitly play with some of these noise factors, additional factors in our process that we're not using directly to control the process and try to explore and model what effect those noise factors have directly on the process. OK? And so that gets to, essentially, an approach where we can do an extension to our design of experiments where we explicitly look not just at, say, our main control factors. So these might be our A and B control factors.

But at each of our design points, what we might do is not just do pure replicates but actually take some of these other factors, things we can kind of change in the process but in normal operation we would not want to be. Maybe it's a machine setting that has a big cost associated with changing it or a big time lag in it. It's a noise factor or some other factor. What we might want to do instead of just running pure replicates is actually look at these additional noise factors and explicitly, in the design of experiments, tweak them, play with them, bury them in a small region around that particular operating point. And what that does is back to our variation equation.

That's explicitly saying, I know what my alpha factors are, these noise factors, I'm going to actually modify them a little bit and build a model that gives me information about sensitivity to those factors. So this is referred to as inner and outer factors. What we've got if I do the same thing at each of these corner points, I've got the inner design and then I've also got an outer design where I play with these additional factors at each of the corners.

This, by the way, gets very close to approaches often referred to as Taguchi approaches-- after one of the big spokesmen for explicitly thinking about dealing with noise and modeling of noise and optimizing processes to be robust to noise-- where you've got this inner and outer factor mixed design. So let's see how this works. Here's a very simple example. We've got two main factors in our inner array.

So if I do just a full factorial in that, we've got my four combinations. Right? And then here, instead of running pure replicates, what we're doing is an outer array where we vary the levels in another little full factorial around each of those corner points. And so now you can imagine that that's giving us an ability to model the mean response, again, at that corner point, but also a notion of variance.

If I just treat those as pure replicates and build a model or an estimate of variance, that gives me kind of a funny, slightly different way of thinking about the variant. It's actually a perturbed variant, but I'm thinking of that as in my normal operation those might be noise factors that I'm not explicitly controlling and might randomly be picked at one of those settings and so I'm going to treat them, for the purpose of my process optimization, for the purpose of robustness, as overall noise. Right?

So in essence, it's lumping together these other factors in order to explicitly and proactively explore the space in those factors and get an estimate of variance at each of our operating points, at each of our corner points. OK? So then you can build a model. You could build a model both of the average output and of the variance.

Another part of the methodology popularized by Taguchi is to say, I don't want two outputs that I have to somehow then optimize separately. I want to fold these things in together into one combined cost function that I then optimize. And the particular cost function that he refers to or uses is a signal to noise ratio, where what you're trying to do is get on target or maximize some output, but also then minimize variance.

And he actually has multiple different factors, but one can imagine building a signal to noise ratio that looks at the ratio of output mean to variant. You take the log so that it becomes more of a linear model and form a signal to noise. And so here, for example in the picture, is an example of an overall integrated cost function, a signal to noise ratio function where you might want the maximum signal to noise-- larger output, smaller variance-- and this would drive you towards an optimum point in your space based on where you're going to achieve that notion of best. OK?

So there's a couple of ideas on this slide. One here is simply this notion of how you do the design with the inner and outer array and then the other is this idea of a natural way of combining both variants and target different than the expected loss function, but kind of related to it are the signal to noise ratios. And if you look in the literature, there's different ones depending on what kind of objectives you have qualitatively. For example, if you really want to be close to the nominal and be insensitive to the noise, then that cost function that we just described on the previous slide is a pretty good one.

There's other places where maybe your main driver is really trying to get as-- do I have this right? I may have this-- no, I think I have this right because of the negative there. You might have a slightly different signal to noise ratio if you're more driven towards you want a large output, a large y, or another one if you want a small y, you want to minimize y. So you can still formulate these things in slightly different signal to noise metrics.

So let's explore this inner and outer array design approach a little bit more. This is often referred to as a crossed ray method. One question that comes up is, gee, looks like an awful lot of experiments. So we do want to understand how many experiments there are. Are there ways to minimize that or reduce that?

And overall, it's clear that I've got something that requires both control factor tests in the inner array, but also noise factor tests associated with additional variations right around each of the points. OK? And it can actually be even worse than just sort of what I've drawn here, because if our goal is really to get to an optimum point that hits the target and really is an optimum point in a target, you may even need something more than just a full factorial on your x1 and x2 parameters.

Think back to what we talked about on Tuesday. We said if you're trying to drive the process to an optimum y, you might have locally linear models. But if you really have an optimum y-- not just at some target, but some overall optimum-- what you need to get to is some natural maximum or minimum point, which is a place where you've got something like, say, quadratic curvature.

So ultimately your model for y, if you really are trying to get to a true optimum it may require some notion of a quadratic model, at least in terms of your output as a function of your input. So one approach was it might actually require, for example, a three level model. And in that case, things get bad really fast, right?

If I really needed a true three level model on each of my x1, x2 inputs, now I go as 3 to the K sub C where K sub C is the number of control factors, K sub N is the number of noise factors. So that could grow really, really fast. Fortunately we know I don't necessarily have to do a full three level factorial. There are other approaches like central composite that can build me a quadratic model in the output without needing so much replication.

I don't have to have necessarily full three levels on all of my factors because I don't really have that much interaction going on. But the point here is if I'm really trying to explore the space for building up my estimate of my output, I might well need a quadratic model in that. Do you think it gets that bad for the noise factors? Do I need a quadratic kind of model for my noise factors?

In other words, why not? If I look at this little corner point, why don't I need a quadratic? Should I include a center point there? Should I include multiple exploration around that as well? Do I need a quadratic model of variance at each corner? What are we using this exploration in this corner array to do?

We're just using it to exercise these noise factors and build up an aggregate into one estimate of variance at that point. We are rarely actually trying to build a local quadratic model. I'm just using it-- think of it almost as structured replicates at the corner. I don't really need that kind of resolution.

So you typically will not see things like doing a central composite design at each corner, something like that. Something very simple like a factorial, either full or even fractional factorial. This is a beautiful place for fractional factorial in those corner points.

If I had eight noise factors, I do not really need to do the eighth exploration at each of my corner points because I don't really care if it's a third or fourth or sixth order interaction between noise factors, what I would like to do-- I don't care about confounding. Confounding is great, no problem. I want to lump them all together and just get one estimate, in fact, of overall noise variance there. So very reduced models-- whoops. I didn't want to do that. Very reduced models. Heavily fractional factorial in those corner points are very effective for this purpose.

Another approach that we can use here is we could actually go beyond that notion of just lumping it all together just trying to get an estimate for variance at that point and actually treat my control factors and my noise factors almost on equal basis. So the purposes of our design, why distinguish them? I might actually want a functional dependence on the noise factor, maybe because that allows me to sort of more naturally fold it all in together into one big experiment.

So in that case, you can easily imagine building a single integrated model that has linear interaction, maybe even quadratic dependencies on my control factors. But also that has terms in it that look like essentially different parts of the problem that we might be interested in. One of those would be terms like this.

I've got a gamma times z1. Well, what's that doing? That's telling me essentially the sensitivity. That's Dy to Dz1. That's one of these alpha noise factors. I directly have a measure there of sensitivity to noise, that noise factor at that point. So that might be a very useful thing to know and actually model and distinguish.

A second one are terms that look like this. Here's an x1 and z1. What's this telling me? Well, this is telling me there's an interaction between my x variable and that noise factor. So I know, depending on where I pick my x variable, I've got a different sensitivity to noise at that point.

So that's kind of breaking that part of the overall response out and letting you explicitly say, yes, now I understand that interaction. I understand how my noise interacts with that particular factor and I can use that to help guide my optimization and selection of input points. So there is some value in going ahead and doing an integrated effect and noise response surface approach where you can build up an integrated model together in here.

And then you can use the same techniques that we talked about last time, about aliasing and worrying about which interaction factors and how high your order interaction really needs so you can go ahead and get reduced models and fold some of those together. But it essentially allows us to explicitly think about both control factor responses, but interaction with the noise factor. Yeah?

AUDIENCE: Do we have a control on triggering the level of noise?

DUANE BONING: Yeah. So the big assumption in here is, in fact, many of these noise factors in the design of experiments, while you're doing your upfront process optimization, that you actually can vary them.

AUDIENCE: That's a noise. [INAUDIBLE].

DUANE BONING: Yeah, it's kind of a funny term. Yeah, I'm glad you raised that question. The question is we've got these noise factors and the term calling the noise factors-- one strong interpretation of noise is it's something I have no control over. This is a weaker notion. This is basically saying there are things like in the normal operation of the equipment I can't vary them or in the normal operation they will be naturally varied over and I can't set and specify and keep it at that noise factor at one of these.

So here's what I'm thinking as a good example. Temperature, room temperature, say. So you know in normal operation throughout the day, the temperature from the morning to the afternoon might be varying in your factory and you've only got a limit. You can't really completely control that. Maybe you're operating in a place where it has a big range and you don't have any control over it.

So what you would want to do is understand that while you're doing your process design. Treat it as a noise factor, know it's impact, but then in actual operation, I can't pick my temperature. It's going to just vary, but you would like your process to be as robust as possible to that factor.

So that's the kind of noise factor here is things that you can either actively control during the design of experiments or naturally sample from during the design of experiments. Other things would be operator. So that's a very good question. I'm glad you asked that. OK.

Once you've got a noise response kind of surface, now what you can also do is build outputs. Not only as a function of-- mean responses as a function of your input where you would essentially ignore the noise factors, but also if you just simply take the variance of this relationship that gives you a model for the variance of the response when under the assumption that x are constant.

So in other words, what we do is we can build two of these, one under the assumption such that my noise factors are at the mean. So I'm basically just lumping all of those into some overall noise from my noise factors. And then in my variance case, I just look at the terms that relate to the interactions assuming some x is relatively constant.

So we ignore these terms or lump all these terms into just the overall gamma and then include some of the interaction terms with my z. So essentially what this lets us do if we built an overall response surface is separate out and explore both the mean response and also variance as we try to pick an x1 and our x2 to minimize the sensitivity to those noise factors. Yes, there was a question? Singapore?

AUDIENCE: Prof, do we need to check the r squared before calculating the mean and then the variance. Whatever the r squared is, it's not very good. It's [? small. ?]

DUANE BONING: Yeah. Yeah. So this sort of goes without saying. I was just assuming that you use all of the normal response surface modeling methodology. You have to check significance of each of these terms. You would only include terms if they are significant. So it's more than just checking r squared, you really are running ANOVA.

Does that noise factor have an effect or not? If not, you just lump it into an overall aspect. So all of the typical model construction you would still want to do. So anyway, the point here is now if you explicitly do a response surface model, now you can build models of the variance.

This is just an explosion or expansion of the variance surface in general you will observe because of this squaring in the variance expression. It typically is quadratic in your x1 and x2, which is nice because then if you're trying to minimize that variance if it's quadratic in x1 and x2, it does tend to drive towards a nice minimum point.

I have one quick example and then we'll be done. This is alluding back to our robust bending where I've got a punch and I'm trying to get a bend in a piece of sheet metal, and we might have a couple of control factors that again, we can use to optimize the process. Like how deep I push the punch and what the width of the width of the initial punch dye is, but here's an example of a couple additional noise factors.

It may be we know that some of the incoming material will have a range of yield points and thicknesses. Again, I'm not going to be able to necessarily select those. I want to process it, it's going to be robust to those. But in a design of experiments, I can certainly sample from different incoming material types that explore that.

So I could then do a design of experiments where I vary both yield point and thickness of the input. I can build up a model of z, I can build up a model of the angle, I can go in and do estimation of both my main effect on my control factors, on interaction terms. I could then estimate the size of those coefficients, go in and make sure that they are significant, and overall get down to say something that looks like a reduced model where I might be looking and saying, aw, OK. That little beta 1, 2, 3 is insignificant. I'm not going to include an x1, x2, x3 factor, for example, in the model.

Here's the example of the reduced order model. In this case, the mean surface is nice and linear with an the interaction term. So we could go and do optimization purely on the mean surface trying to hit some target angle, but I can also look at the variant surface. And in this case, the variant surface has-- it is a quadratic, it's just a weakly quadratic.

But you can see a very, very strong dependence in the variance on this x2 factor. I think that was the thickness factor. Very big differences in variance. And so then I could combine those two in different ways to form an optimization and try to get to a good robust operating point and that might be a Taguchi signal to noise or it might be an expected loss function.

In this case here, I've plotted out what the expected loss surface is as a function of x1 and x2, and if you're seeking to minimize that you can start to see a region there where you would want to be in terms of the x1, x2 to have a good on target or close to target and a low variance process. OK.

I'm going to skip over this slide. It's basically just talking about a summary of doing the combined response surface model with a little bit of an assessment of maybe the number of different design points that might be required in those cases. But what I wanted to do here is just conclude and give you the high level perspective here is we did lots of machinery on response surface modeling talking mostly about mean output, but that same machinery can also be used with some extra ideas of these inner and outer arrays for also modeling variance when it varies as a function of operating point. And then we can combine those together to try to get processes that are robust to not only selection of operating point, but also to these kinds of noise factors.

So with that, I'll conclude and we'll see you all on Tuesday. I think on Tuesday we'll talk about nested variances, so it's going to be a little bit different. We're going to pull back from the design of experiments and talk a little bit more about interesting variance structures often arising in places like semiconductor manufacturing and elsewhere. So we'll see you on Tuesday.