MITOCW | Lec 17 | MIT 2.830J Control of Manufacturing Processes, S08

DUANE BONING: OK. So today's lecture, in some sense is actually out of sequence by one from Thursday, but based on when I was able to get Dan Frey to be able to come in and give his lecture. This is a lecture today on variance estimation, which is kind of a topic that probably ought to come after Thursday's, because Thursday's is still in the area of design of experiments, response surface modeling, and optimization. So it's kind of a nice wrap-up to the last three or four lectures that we've been doing. So we'll come back to that on Thursday.

But what I want to do today is this idea of nested variance components. And the readings for this material are not in either of the textbooks. There's a separate chapter that is on the website under Readings. And it's chapter 3 from this book by David Drain, *Statistical Methods for Industrial Process Control.*

I think he's still there, I'm not sure. He's a statistician for Intel. And actually, although the title doesn't indicate it, it's all about semiconductor manufacturing and statistical methods for that.

But this idea of variance components is actually, it does come up, and I've never actually seen it covered in any of the standard statistics texts. So I think it's a very powerful and important idea, and that's why I wanted to talk about it today. In addition to this book as a reading on the website, we also have spreadsheets. There's a spreadsheet with three different worksheets in it that are the two main examples out of the Drain book. He just pulls up the data, and then uses some statistics package to pull out things like the ANOVA table.

I actually show you the spreadsheet with all the intermediate calculations for both analysis of variance and this estimation of variance components in it. So you'll find that very useful. You may find it really useful as a template for some of the work on the problem set as well.

OK, so what I want to do is first off, refresh ourselves a little bit on analysis of variance. So in some sense, this is actually a little bit of a refresher, first on ANOVA, and then deepening of our understanding, hopefully, of ANOVA. So indirectly, it's a little bit of review for the quiz.

But then I want to talk a little bit about a different assumption of the model, or the underlying problem that's being looked at, that leads away from standard ANOVA, towards this idea of nested variance. And briefly, what nested invariance is, is structures in which you may have more than one source of random variation at work. And a classic example in semiconductor manufacturing would be this spatial hierarchy of structures like chips, you have 50, 100, 1000 chips on each wafer. And you may have chip-to-chip variance, for example. And then you may have a certain number of wafers in a lot or a boat of wafers that all are typically processed together.

So I might have 24 wafers in each lot. And I may have wafer-to-wafer variance. Similarly, I may run multiple lots over time in one line. So I have lot-to-lot variation. And very often, you're gathering lots and lots of data, much data.

And you're trying to say, OK, how much of it is die-to-die variation? How much of it is wafer-to-wafer variation? How much lot-to-lot? You want to separate out or decompose the total variation that you're seeing into each of these different components.

The standard ANOVA isn't really set up to do that. And that's what I'm going to talk about today, is try to show you what ANOVA assumes, which is looking for a fixed effect between different design points, or different replicates. And then this notion of nested structures, where you have things within other things within other things, or groups within other groups within other groups. And so that's the key plan.

So what I want to do is again, refresh us a little bit on standard ANOVA, then be fairly explicit about what the model is that underlies these nested variance structures. And then we'll actually go back and work through how we extend the ideas of ANOVA to actually be able to estimate these different variances.

And then finally, there are some implications of these variance structures for how you would design experiments. So for example, if you want to get the best estimate possible for that die-to-die variation, should you make more die measurements? Or should you have more wafer measurements? How would you allocate a total budget of measurements across that nested variance structure? So that's the plan.

And the key idea here, again, is there is a big difference, an important difference that I want to cover between the standard ANOVA and these nested variance structures. So what is it in standard ANOVA we were doing? So there was a very basic question in the standard ANOVA, which is, we're basically asking, if I were just having a single source of random variation at work, and we make, actually, the assumption that we're sampling from a normal distribution, when I look from one group to another, do I see evidence that something besides just random sampling from that single normal distribution is at work? Is there a fixed effect that is large enough that I think it's more than just a random chance that I got an observed difference between the means of a couple of different groups?

And the basic approach mechanically to do that was first off, we needed some estimate of the variance of just the natural variation, the pure replication error. And what we basically did is treated each of our treatment groups, looked at each one, and said, OK, there's a local mean of that treatment group. What's the replication variance around that local mean? I wasn't changing now the parameters. But for each of the ones with the design fixed, I just looked at replication.

So that was one estimate, that's our natural variation. Second, we looked at the group-to-group deviations. And we said, either this is due to a fixed effect, a systematic fixed effect that's different between those two groups, or it's just random chance because of sampling.

And then we looked at the ratio of those two variances, and said, by random chance alone, what would I expect based on the size of the samples, the number of data points going into each of those? Look at the F ratio for the statistics associated with that ratio of variance to see if it is likely to have occurred by chance alone. All sounds familiar? You could do that in your sleep? OK.

So here pictorially is what was going on with the ANOVA example. And in fact, I think this was the ANOVA example. I think that we did the actual data we used when we first introduced ANOVA a few weeks ago. It's the simplest possible ANOVA setup you could have with four data points. And what's beautiful about four data points is we could actually do the whole calculation by hand.

So here was the situation. I had two different groups, group 1, group 2, the data points 3 and 5, 7 and 9. And again, the basic idea is, we need to estimate number 1, the within group variation. And here we have two different local means with two different estimates of variance. We pool those to get a pooled estimate of natural variation.

And then we looked at the between group, the deviation between the means of the two groups. And we used that to estimate a group-to-group variance. And then we looked at the ratio between those two. And we saw it in this example. I can't remember, I think it was fairly marginal with this few data points, whether it was actually, in fact, statistically significant, whether this mean effect was real or not, because so few data points, it's really quite a large spread in the variances that you might actually come up with by chance.

OK, so what I want to do is go back and mathematically identify what the implied models were with the standard ANOVA. This is kind of pedantic here, because I think we do understand this. But I'm going to extend this in a minute and contrast it with the nested variance. So I want to be very, very clear.

Our null hypothesis in ANOVA, what we think is happening if there's no fixed effects, is basically every data point simply is being drawn from the same distribution. And it has some fixed mean, and then it's got some random variation. And each sample is being pulled from the same normal distribution with just one underlying variance at work. And recall that variance might be measurement variance together with process variance, but it's all identical.

So that's what's happening if there is no fixed effect. The alternative hypothesis that we're looking at, and we're trying to see is there evidence of, says basically that there is a fixed effect going on. Now I'm going to introduce a little bit of notation, and I actually spent some time trying to do the best I could to make this notation consistent throughout all of the slides today. We'll see if I succeeded in that, because there was a lot of changes to notation. I'm going to use the i subscript here in this case to indicate what later we'll see is the outermost level of variance.

So in this simple picture, now as we go through, i indicates here, which subgroup I'm in. So in our simple data, I had two subgroups. I had group 1 and group 2 with two data points.

So the i subscript here could be either 1 or 2, indicating group 1 or 2. And so the point is, there may be a fixed offset, either t1 or t2 from the grand mean, as a fixed effect associated with being a member of that group. And now the j is a subscript with this funky little notation down here, that indicates this is data point j, the jth replicate within, read those parens right there, within subgroup i.

So I had two replicates within each subgroup i. And the simple point here is that our alternative hypothesis is, there is this fixed offset, t1 or t2, at work, in addition to the random variation. And that's what we're trying to look at the estimate for.

But this is standard ANOVA. And so there is still this assumption that there really is only one random variation at work. There's only one random source of variation, and then there's this fixed effect on top of that that's repeatable. It's systematic.

Now, what I'm also going to do, this is a slide I added. I'm going to show a simple ANOVA table. This in fact, we showed you earlier a few weeks ago for that very simple set of data. And then I'll go back to the previous slide, just to be very careful on some of the subscripts and notation that I'm using.

But this is essentially the same data that I was talking about before. We have two groups. And within each group, we had two replicates.

And this part up here is the scratch worksheet that I might use to generate our ANOVA table down at the bottom. These are the sorts of calculations that either your statistics package would do, or you would do by hand. And it includes things like calculating the group average for each of those two groups. Remember the average of 3 and 5 is 4, and the average of 7 and 9 is 8.

And then some additional squared deviation calculations that go into ultimately the sum of squared deviations calculation that falls into the sum of squares column in your table. After which we divide by the degrees of freedom to get these mean square estimates. So I've got these little notations, s sub d, s sub g, s sub e, and then sum of squared deviations, sum of squared g, sum of squared e. And that's what I've tried to detail out here what these definitions are, back here in slide 7.

So this is still intermediate calculations, standard ANOVA. This is still all review. But some of the calculations, some of the terms that go into that, with a little bit of this funky new notation.

So again, recall what we need to do is look at OK, what is-- sometimes down here. Our first thing is our estimate of what just pure replicate error is. We're just looking for pure replicate error. And that's basically looking and saying, I've got my local mean per group i, group 1 or group 2. And then I've got the data within that. That's just the deviation of that individual point from its local mean.

I can take the square of that deviation. And then I take the sum of that over all of my data, sum of squared deviations, as my ss sub e. That's my sum of squared deviations in all of my data from the local mean.

That's not quite my estimate of variance. If I now take sse and divide it by the degrees of freedom, that gives me my mean square error, which is my estimate of the underlying variance. Right?

Now the other things that we do, we said the second piece here is we also look at the fixed effect. So here, we're just looking at the deviations of the group mean from the grand mean. And so here's the local mean, and here's the grand mean. That has the double bar over it, if you can quite see the notation there.

And so for that, I've also got a sum of squared errors here in ssg. This is the total sum of squared deviations, one for each data point. So each data point shares the same group mean.

So I'll have multiple entries in the table. Because if I have multiple replicates, they all share the same group mean. But I'm basically saying, OK, all of those contribute to group mean deviations in the total deviations in my data.

And then the last thing, I'm going to put a 0 by this. Because this last thing is sort of the total deviations in your data from the grand mean. This is just your raw sum of squared deviations.

So I just calculate my grand mean, and I treat all of my data equally, whether it's within a group or not. And that's just throwing all of my data into one big bucket, calculating a grand mean, and saying, there's a total, total amount of sum of squared deviations in that data. And what we want to do is actually take that total amount of deviations, and now divvy it up, and say, how much of it is going to be because of group to group variation, and how much of it is pure replication data? That's where we're going to want to get to.

But in standard ANOVA, what we simply do is remember that ssd is equal to the sum of the ssg and the sum of the sse. It was in fact, sort of a crude attempt at divvying up the total squared deviations. But what we'll see is that it's actually not the right way to get an estimate of the underlying variances.

And so now in this table, that's all that was going on. I just applied all of these formulas to get the scratchpad to work, and then put that into the ANOVA table. And in this case, we saw the total sum of squared deviations from the grand mean.

We have an estimate there of the mean square. We had the group to group sum of square and the replicate sum of squares. We could use that to form an f test. And we see that it's only 89% confidence that in fact you would observe that big of a ratio by chance alone.

So if I had a cutoff of 95% confidence or 90% confidence with that data, I would have said, sorry, I don't have strong enough evidence to say that there's actually a group effect. There's not a fixed effect here in this data. OK?

So what's wrong with that? That's all great. That's all great. There's nothing wrong with that.

But if what you're really trying to do is deal with a different situation, a nested variance situation, that simple ANOVA can actually lead you astray. So now, we're actually shifting to a different model. But it's a model that I think comes up quite a bit, which is, instead of saying there was a fixed effect, so every time I had group 1, I would have a deviation t sub 1 for that group.

What if every time instead, I had pulled another sample that I just lumped together as group 1? Maybe it's a new wafer. It in fact, has a wafer random variation, coming from a different source than the replicate data within that wafer.

So in other words, I'm moving away from a fixed effect group to group, to a random effect group to group. There's a different random model, a g sub i here, which is not the fixed effect. I guess we called that a t sub i in the previous model. But in fact, this g sub i is now also drawn from a different random variation, or a different normal distribution, in this case. I have a very different model. I have a very different model.

Now, what I'm really interested in is dealing with a situation where there's two different variances at work. I still have replication variance, or the within group variance. So imagine again, I'm pulling now two different wafers. g1 is now wafer 1, g2 is now wafer 2. And within that, I measure two different chips.

Now what I'm interested in knowing is what the chip-to-chip variance is, and what the wafer-to-wafer variance is, as if it were random. So that's the key difference. And what we want to do is still be able to decide, is there something significant going on? But really, what I want to do is estimate what those variances are.

So here's the same picture. This is the same data. But now, I have a different kind of picture.

I still need to estimate number one, which is the within group variances. But now instead of having a fixed effect, what I want to do-- let me call that 2 star. I want to attribute the group-to-group deviations that I observe as indications of or samples of group-to-group variances.

So is this setup clear? OK. So our real goal here is to estimate these two variances. And by the way, point estimates are not going to be good enough.

We actually want confidence intervals on these estimates. So that's going to get a little bit tricky. But what we would like to actually do is decompose the total variance we observe, estimate what the total variance is, and decompose it into these two different sources.

So, here's our first attempt. I'll call this the naive attempt. Why don't we just reuse all those calculations we already had for ANOVA? Aren't they really telling us the same thing?

Think of it. Back in ANOVA, we were already saying, OK, under the null hypothesis, we had replication variance, and we had a variance due to group-to-group variation.

AUDIENCE: [LAUGHS]

DUANE BONING: Can't I just use those directly as my estimates of the two variances? I had a ratio of those two. Why aren't those two just great estimates of those two variances?

So our naive attempt, our first attempt, is let's just use what we already did. Number one, we had an estimate of pure replication error. Think of that as within group variance, right? Let's just reuse that. That's our estimate of within group variance.

Number two, we had this between group thing. We had a group mean square, deviation of the group means. Why not simply estimate the group-to-group variance as the mean square of that group deviations? So some squared group-to-group deviations divided by the degrees of freedom.

And in terms of total variance, if I assume, again, there are two different variances at work, there is within group and group-to-group, and they're independent, then my total variance should just be the sum of those. Why don't I simply use that sum of squared deviation divided by the degrees of freedom? It's just a total mean squared deviation. Why not use that as my estimate, just my total variance? Throw all my data into one big bucket, calculate total variance. Isn't that my best estimate of the total variance in the system? Seems natural, doesn't it?

In fact, we can calculate the grand mean, I can calculate the grand variance. Isn't that a great estimate of total variance in the system under this model? And well, the good news is, at least something still sticks. The within group variance estimate is still good.

But we would be wrong, and I'll show you why, in this naive approach, with our estimate of between group variance. And bizarrely, we're even wrong on our estimate of total variance in the system, just using all of our data, undifferentiated. Yeah.

AUDIENCE: Are they supposed to be [INAUDIBLE] really the sum of the other two [INAUDIBLE]?

DUANE BONING: So the question here, for folks in Singapore, was, shouldn't the last one be equal to the sum of this one, a plus b? Shouldn't that equal the total? And the answer is no. If you actually look back at the ANOVA table, let me do that. If you look back at the ANOVA table, let's look at it down here at the bottom.

The sum of squares is the one that is conserved. So your total squared deviations, you break apart into within group or group-to-group. But by the time you divide it by degrees of freedom, the mean square or the variances do not add. And that's weird.

And it turns out, in our second problem, the nested variance structure, they should add. We're saying, there is two different sources of variance at work. There's within group invariance, there's group-to-group variance, they're independent.

So my total variance actually should be the sum of the two. And we're going to use that knowledge, actually, to help us with estimating things. So in fact, your insight is right, not for standard ANOVA, but it is right for the nested variance structure.

Now that I've broken down your confidence in using ANOVA, hopefully, and we can rebuild what it is we really want to do. I think we've already talked about this a little bit. Again, I think that these nested variance structures actually arise a lot in any kind of process where there's any kind of batch processing at work, either in time or in space. So it's easiest, in my mind, maybe because I'm just mostly familiar with semiconductor manufacturing, to really see that.

But I think it's much more generally true anytime I have some grouping of stuff, of parts, material, whatever, that I want to look at stuff within that group, and then group-to-group, and maybe then there's a hierarchy where there's a larger group of groups. And I want to estimate variances within that. That kind of hierarchical or nested structure, I think comes up all the time. And an important point that I haven't really explicitly said, but would like to mention here, is that the reason we're interested in estimating these variances differently, is that very often, physically in the process, the source of the variation at each of those levels is actually different. It's a different orthogonal, different kind of variation source.

So for example, in the wafer case, the source of variation chip-to-chip may have to do with non-uniformity within the tool, maybe spatial or other kinds of random, in many cases even spatial systematic variations. But the point is, there's a different kind of set of physics governing how well-matched each of those chips are within one wafer. It's a different source of physics than every new wafer I put into that single wafer tool, from one time to the next. That may also have deviations due to how well I can control some of the parameters of that process in time from one run to the next.

So it seems very natural that indeed, the variance, the underlying variation in those two cases, are different. They are orthogonal. They are in some sense, additive in that sense. They are uncorrelated. And we'll use that assumption.

But there's that underlying assumption in here. If, in fact, it's just arbitrary grouping where there's only one source of variation at work, and it's just random pulling of things to form, in fact, different groups, then that's more like the standard ANOVA. But when there's a nested structure, items within items within other items, usually there's a different source of physics that were causing the variation at each level.

OK, so again, our goal is to estimate each of these sources of variation, both point estimates and confidence intervals. Here's some examples. I think we've already talked about this. The within wafer versus say, the wafer-to-wafer, or run-to-run variability.

OK, so let me build up a little bit of the explicit model for this case, especially with multiple layers of nesting. We'll start without nesting. Then we'll do one level of nesting. And then working up to the second example, we'll do a three-level, two levels of nesting. But we basically will have points within wafers within lots.

So in fact, you can keep extending that. You could have lots within products, and products within fabs, and who knows? You could get further detailed. But here's the simplest model. And this really is the pure variance case without any nesting at all.

We basically have our individual measurements, there's an overall mean. And then there's some random variation occurring. And here I'm indicating that I'm taking multiple measurements, multiple samples, multiple replicates with this m sub i. And the point is, it's simply a 0 mean normal distribution with some variance.

And as we've said, many of our assumptions still hold. And in fact, we often make these assumptions and use them, and we have to be careful in some of the spatial situations I've talked about, because they don't actually hold. In some sense, what we're assuming in this case is, I'm taking multiple measurements on the same wafer.

But I'm assuming randomness in those measurements. I'm assuming each of those individual measurements, or each of those individual replicates is IIND, identically and independently distributed from a normal distribution. That's 0 mean with all sharing the same variance.

And just as a precursor, we'll come back to this, actually, in one of the case studies. If I have within wafer variation, chip-to-chip variation, and I have a systematic variance, center to edge, maybe it's always thinner in the middle of the wafer and thicker on the edge of the wafer, that's a systematic effect that this model is not really good at capturing. So you've got to be careful. This is actually a fairly strong assumption of random sampling within that particular scenario.

OK, but this is the simple case. Now let's look at this variance structure. And here I'm using exactly the notation out of Drain's book. So when you read Drain's chapter, this should look familiar to you.

In this case, now I've got an overall mean. And again, we've got a wafer effect. It's not a fixed effect, it's a random effect. Meaning every time I pull a new wafer out of my samples, all of the data points on it will share the same offset.

It's an offset of w sub i. But w sub i itself, or wafer sub 1, or wafer sub 2, is itself drawn from a 0 mean normal distribution with variance sigma squared sub w. So the amount of wafer offset for that particular wafer is randomly sampled. And then within that, I make multiple measurements, I make j measurements within wafer sub i.

And each of those individual measurements is itself also randomly distributed. OK? So that's just repeating what our situation here was. But now, instead of using that generic model, I'm really trying to illustrate it with wafers and measurements within wafers.

So if I take this basic formula right here, and I basically ask, OK, now if I want to do variance calculations, if I do a variance on this, what is the variance I'm going to observe in my total data? Probably should have replicated that equation right up here. We had xij is equal to mu plus w sub i plus m j of i. So if I threw all of my data into one big bucket, xij, and I simply calculated the variance across all of my data, what should not just my data, but in fact, if I had the total population, infinite numbers of measurement, what would the variance of that be?

And we're saying that the w sub i and the measurements within that are independent. So that these variances add, there's no correlation between them. The variance of a constant is 0. We have the wafer-to-wafer variance, so that's my sigma square w. And I have my measurement variance.

And so my total variance, my total true variance, is simply the sum of those two other independent variances. And so that's your earlier intuition that Nalish was talking about. So the individual variances are assumed to be independent. And again, that was not true in that naive attempt.

So how do we do it? How do I use the data to actually get good estimates of these independent variances? And here's the key idea. We've got sampling at work. We've got sampling at work.

Essentially within the replicates, I've got additional variances going on, because of measurement or replication variance. And that is contaminating or adding some noise to my estimate of the group-to-group variance. So what we basically need to do is unwrap that, recognize that I've got multiple samples around that, and pull out that random variance away from our best estimate of the group-to-group variance. We need to account for the fact that I've got also sampling noise going on when I'm trying to estimate the group-to-group variance.

So what happens is, if I calculate the wafer average observed variance just from my data, So if I observe w bar, I observe wafer average 1, wafer average 2, wafer average 3, and I look at the variances of these, the observed variance in the wafer averages actually has the true wafer-to-wafer variance in it. But it's also got this sampling noise attached to it.

And so what we want to do is, to get to the true variance of just the wafer-to-wafer variance, I need to subtract this off of the observed. So if this, now again, is my observed wafer-to-wafer average variance, I subtract off the sampling noise, my best estimate because of sampling. And that gives me my best estimate of the actual wafer-to-wafer variance.

Now, I added a slide here that actually shows a derivation of this. But I hope, before I go into the derivation, hopefully there's good intuition here. This is just the sampling contamination of noise, right? And for example, if I had a million replicate measurements, so if m were a million, I would be averaging out all of those small measurements, pure replication deviations. There are 0 mean.

So on average, this term gets smaller and smaller with more and more sampling of the wafer. And what I would observe in that case, from my wafer-to-wafer average variance, is really, really close to the true one. This is only a problem when m is small. And I've got lots of replicate noise or measurement noise contaminating my estimate of the wafer-to-wafer variation. So hopefully there's some degree of intuition here that makes sense.

If I actually now go in and do the calculation, first off, what I said is we have three or four or five different wafers, and I have some number of measurements within each wafer. And I'm simply calculating for each of those wafers what the observed average is. Up here is the formula for the wafer average. And then down here, I'm just applying various mathematics to that formula.

And so what's going on here is, we can see, we've got our wafer average for wafer sub i, is simply 1 over the measurements for all of the j replicates, 1 through m. We've got m replicates of that within that wafer sub i. Which I've just expanded out, I just plugged in x sub i in here.

Now if I expand that summation out, I've got m replicates of mu. So that's my m mu. I've got m replicates, all with the same w sub i. So that's that right there.

And then I've got my remaining part of my sum for my individual replicates. And just multiplying that out, my overall wafer average for that wafer sub i is mu plus the shared offset because of the sample for that particular wafer, and then all my measurement noise. And now I can apply my variance to the observed, this is the observed wafer-to-wafer average, is the variance of the mean, which again, is 0, because that's just a constant. This is the true variance, sigma w squared.

And then here, I've got m replicates, all with the same variance. So I've got a constant, which gives me, when I do my variance math, my 1 constant squared, my 1 over m squared. And then the sum gives me m equal variances. So that's where we get back to our basically 1 over m sampling variance. So that's the derivation going on, if you actually want to see the whole detail.

But again, all that that's trying to say is, if I actually look, group-to-group in the observed means, it's got two things inside of it. It's got the true wafer variance. But it's also got noise.

And the noise comes from the underlying measurement noise or the underlying replicate noise, reduced by the factor of number of replicates I have. My typical 1 over n reduction in variance. But that's what's contaminating my noise.

So once I have that, there's one quick observation to make here. Before we actually use that to get back to our estimate of what the true variance is, is to go back to the earlier point we made. We said that the total variance, the true total variance should be the independent sum of these two sources of other variance. But the same sampling contamination also occurs not just for what our observed wafer-to-wafer variance is, but if I were, in fact, to actually calculate my grand variance in just the data that I observed, I took all of my data, threw it into one pool, calculated a mean, and then did the sum of squared deviations of all of my data from the grand mean, divided it by the number of data points I had minus 1, I used one degree of freedom to calculate the grand mean.

So I'm just estimating the total variance in all of my data. That's the sigma squared t observed. The observed total variance is different, is not equal to the actual true, total variance in my system at work. In fact, the observed variance will always be smaller than the actual observed variance.

And the reason is, if I look at the total observed variance, oops. I think this is an error here. I think this should be a d. That should be ss sub d. My total sum of squared deviations divided by m minus 1. I've also expanded it out into where that data is coming from.

And the point is, I've got my wafer variances, I've got my measurement variances, and I've got some number of replications of each of those. And if I expand that out, essentially what I've got is multiple samples at work. When I make my observed calculation of total variance, that are factors that are multiplying times the true underlying independent sources of variance, but with factors that are always smaller than 1. In other words, I've got m kinds of replications going on with sampling from multiple samples from replication.

And I always get that 1 over and reduction in variance. I've got that same thing happening, both within the measurement reduction, and within the wafer reduction. So the simple point here is, this is why that naive attempt to just use total variance and use that as my estimate of the true independent sum of variances at work, why that doesn't apply, why that doesn't work.

So, now we've got a strategy here. What we're going to do is number one, estimate within group variances. That's still OK.

Number two, we're going to see the observed group-to-group variance, but then account for sampling, subtract off the sigma squared m over m, from that, to get our best estimate of true group-to-group variance. Now I have group-to-group variance, I have within group variance, I can add those to get my best estimate of total variance. That's the strategy.

So let's go back to our really simple nested variance example, and use that strategy not the naive approach, but replacing the naive approach and see how the numbers come out. So again, here this was step one of our strategy. But within group variance, that's still the same. The observed group-to-group variance is simply this mean, which was 4, and this mean, which was 8 from the grand mean, which was 6.

So each of those group-to-group deviations is 2. And that's squared, I've got 4 plus 4 is 8. So that's my observed group-to-group variance. And that's actually the same number that we had calculated using regular ANOVA. And then we did the ratio between those two in order to decide if there was something significant going on.

But now the point is, that was our contaminated estimate of group-to-group. That's our observed. It may be hard to see there, that's the g to g bar. That's our group-to-group observed averages.

And so now we want to account for the contamination, subtract off the sampling effect to get to our best estimate of the true group-to-group variance. And it's what we observed, the 8. But now, using sampling to subtract the true variance. So this is our sigma squared m divided by the number of measurements I had in each of those groups, the sampling effect, which was m equals 2, in this case.

So I'm subtracting that component off. So I'm peeling that part out. And I get a best estimate now of the true group-to-group variance is 7. So now I've got, within group is 2, group-to-group is 7. My total is now the sum of those two, or 9.

So that's different than what we had seen before. Our observed total variance, I can't remember what it was. But it was smaller than that.

OK, that's pretty much the core of the idea. Let's just do a couple of examples. And these are examples out of Drain's book. This is the one starting on page 196.

So it's a little bit more data than our four data points. But the basic idea is still there. And this example is looking at the resistivity variation across multiple wafers. So he has 6 different wafers, 1, 2, 3, 4, 5, and 6. And in each case, he's making three replicate measurements.

So we have to be careful again. We're assuming that he's randomly sampling within the wafer to get replicate measurements of the within wafer variation. And then we're looking at wafer-to-wafer. Qualitatively, before we start going and applying all this machinery, what is this data basically telling you, qualitatively? And then we'll see if, in fact, the calculations come out with something that looks consistent to that.

First off, do you think the within wafer variation is bigger, or is the wafer-to-wafer variation bigger? There's a total amount of deviation in this data. But where's the main source of this? Percentage-wise, what do you think the lion's share is?

AUDIENCE: You have the wafer-to-wafer variance.

DUANE BONING: Yeah, it's pretty clear. There's nice clustering within each wafer. There is some spread, we have to think about that.

But it looks like there's perhaps bigger wafer-to-wafer deviations than there are within wafer deviations. So I'd kind of be looking to see if I'm decomposing these two sources of variance. I am expecting the wafer-to-wafer variance to be a little bit larger. But I'm not completely sure, because they're spread within both. So I'd like to do the right thing and get good estimates of these two things.

So Drain then goes through, and does an ANOVA. By the way, you can still do the typical ANOVA, because in fact, a lot of the intermediate calculations you reuse in doing the estimate of variance. And you still want to ask the question, do I have evidence that the wafer-to-wafer variance is bigger than the within wafer variance? Is there group-to-group deviation going on? So the ANOVA table is still valid for asking that question.

And that's what he does here. He's got total deviations from the grand mean. Remember he had 6 wafers, 3 measurements each, 18 total measurements. So the sum of squares divided by 17 is an estimate of observed total variation.

And then he calculates the wafer-to-wafer sum of squares. He has the residual, because he's got three replicates at each case. He can form the f over that and look at the statistics associated with that.

And that ratio, 20 times as much mean square from wafer-to-wafer compared to within wafer, basically saying, that's very significant. There is definitely a wafer-to-wafer effect. That's not just the same as sampling coming from within wafer. Highly significant.

So this is standard ANOVA. And then he very nicely plops out the variance decomposition. The variance components.

And notice what he's done here. Part of it is based on directly the observed results coming directly from the ANOVA. In fact, if I were to take the total sum of squares, divide it by its degree of freedoms, that is my mean square. And that's my sigma squared t observed.

My wafer mean square, that's sigma squared w bar. The observed wafer-to-wafer variance, which was the sum of squares divided by its degree of freedom. And then this is my sigma squared measurement. That's my random estimates.

Notice again, these do not sum. [LAUGHS] This is the naive calculations. You do not want to use those for your variance component estimates.

In order to get to that, you start with number one, random variation, the replicate error. That's a good estimate. But then you unwrap to get your best estimate of sigma squared wafer-to-wafer, by subtracting off and then counting for the sampling effects. And then you sum those together to get sigma t squared.

And what this has done out here in the percent is simply now assigned a percentage of wafer-to-wafer versus within wafer variance. And so in this case, about 87% of the observed variance is because of wafer-to-wafer. And only about 13% is within wafer variance.

Now, how did he actually go and do this calculation? Well, it's those formulas that I gave you. Or he says, run SAC PROC NESTED in SAS. That's all he gives you. OK?

So what I tried to do is in that spreadsheet that I posted on the website, is actually go in and do the calculations that you need in order to get to those variance components, for this example. And it's basically just applying all these formulas and concepts that we've already been talking about.

The tricky piece is actually appropriately counting for sampling effects. How many samples go into the denominator to subtract off the sampling effect, the sigma squared over m? It's pretty easy in the two-level case. But now when I have measurements within wafers within lots, I've got two levels of sampling going on.

And I got to know what factors to use. It's not just sigma squared over m. We'll see in a moment, it's sigma squared of something divided by m times the number of wafers.

And my spreadsheets try to help keep careful track of that. In the one-level case, it's pretty easy. Let me get to two levels in just a second.

But the other thing that essentially Drain's book just pulls out of the air is the interval estimates, again based on SAC PROC NESTED. And there's no help at all in terms of where you come up with interval estimates. So basically, my best recommendation is simply use our concept of chi-squared distributions with the appropriate estimate of the number of degrees of freedom or the number of data points going into the estimate of that variance piece. And my spreadsheet also shows some of that for you.

By the way, if you actually do that, it turns out that the book claims for the total variance and the wafer variance and the error variances, that those are 95% confidence intervals. But I think those are actually 90% confidence intervals if you do the calculation with the chi-squared formula down here.

So I'm actually not sure exactly what's going into his tables. I get slightly different answers. But I think the best conservative estimate, which may have a slight amount of extra overcounting of variance, but it's a slightly larger confidence interval. That is to say it's conservative, you're not fooling yourself, and thinking things are significant when they're not, is simply use the chi-squared distribution. And so that's what my best recommendation for the interval estimates are.

OK, so we're pretty comfortable with two levels? Let's do three levels. Great fun. It's the same idea. But now I have, not only measurements within wafers. I have wafers within lots. So I may have a random lot-to-lot effect.

So I pull 24 wafers, I do lots of processing. I pull another 24 wafers, I do some processing. There is a lot average that may be different from another lot average, from different from another lot average, because there's a lot-to-lot variance at work, in addition now. OK? So this is two levels of nesting, or a three-level variance structure.

Now what happens with the observed lot-to-lot variance? It's the same idea. But now we've got multiple levels of sampling going on.

You may not be able to see it, but over that l right here, there is a bar. Again, this is sigma squared l bar. The observed lot-to-lot variation.

And what it's got in it is the true lot-to-lot variance. But it's also got wafer-to-wafer variance noise added onto it. And then on top of that, it's also got replication within wafer variance noise added onto it.

Now, the good thing is, I've got multiple wafers. Say I've got 24 wafers in each lot. So the effect of the 24 wafer-to-wafer variance noise gets reduced by my factor of w of equal to 24. And similarly, also, within each of those wafers, I may have 10 measurements.

And that noise is multiplicatively averaged out. I've got a factor of the number of measurements, say it's 10, and number of wafers in a lot, would say it was 24 wafers in each lot. That's an awful lot of data of measurement noise that has lots of chances with a big denominator here to average out. So that factor starts to get smaller fairly rapidly. The lowest levels start to become smaller.

But the basic strategy is going to be the same. A phrase I've heard, or maybe Drain uses, is, what we want to do is estimate the individual variance components, three levels, but think of it as peeling the onion from the inside out. I'm confident at the innermost level of the variance of measurements.

Once I have that, I have the observed wafer-to-wafer variance, and I could subtract the sampling out. So now I have the next inner level of the onion, a good estimate for that. Using that, I can subtract out that sampling from the outermost level of observed variances. So it's the same strategy, we work from the inside out to get estimates of the outer levels of variance. Yeah.

AUDIENCE: This one gives a different answer if you do two wafers and three measurements or three wafers and two measurements.

DUANE BONING: Absolutely.

AUDIENCE: It's saying there's a better way of doing things, like [INAUDIBLE] wafer with fewer measurements.

DUANE BONING: Absolutely. So what Nalish is saying here is, you can imagine you will get different answers if I had two measurements on each of 3 wafers, 6 total measurements, for example. Or if I took those same 6 measurements and I did 3 measurements on only 2 wafers. Those denominators are different. And your precision of your estimates will be different.

Both your point estimates may be slightly different. But also your confidence intervals will be different, because in essence, the amount of noise is going differently in the two cases, and the number of samples is going. So that's jumping ahead about two more slides, but it's exactly the point that this does have an important implication on sampling. How you construct your sampling plan, how you allocate your total measurement budget and total replication budget, depending on which variance maybe you want to estimate most accurately.

Let me just qualitatively show you the results here for the three-level example in Drain. This is building on the two-level example, so we're still looking at sheet resistance. We now have 3 wafers within each lot. So this is lot 1, lot 2, all the way up to 11 lots. And then within each lot, the two little triangles here, we're taking two measurements within each wafer.

Now qualitatively, what do you think is the biggest source of invariance in this data? Is it within wafer, wafer-to-wafer, or lot-to-lot?

AUDIENCE: Wafer-to-wafer.

DUANE BONING: Yeah. So I hear a vote, and I kind of concur with it, wafer-to-wafer looks pretty big. So for example, here's wafer-to-wafer, another wafer-to-wafer. Within wafer, it's pretty nicely clustered.

So I don't expect a big within-wafer variance. Lot-to-lot's a little harder to see, because I have to average these 3 wafers. But it looks like there is some lot-to-lot variations.

But it looks a little bit smaller. So it looks to me like sigma squared wafer is bigger than sigma squared lot, which is bigger than sigma squared measurement. So let's see if that comes out of our variance components.

I've given you also this giant spreadsheet table with all of that data. And again, the estimates. There's again, the standard ANOVA, and then the splitting out into the variance components.

In the standard ANOVA, you can actually ask the question, is there statistical evidence for wafer-to-wafer variation? So it's basically, that ratio right there is 62. And it's highly unlikely that that's by chance alone.

He also does, is there evidence for lot-to-lot variance? And in the standard ANOVA, it looks pretty weak. Given the amount of wafer-to-wafer variance, the ANOVA table is saying, what you may be observing for your lot-to-lot deviations, is because there's big wafer-to-wafer variation. It may not be a significant lot-to-lot variance.

So that's interesting. Let's come back to that in a second. Oops. Now we can go in, this is just the ANOVA mean squares. But then you do this unwrapping of the variance, accounting for sampling.

And what he observes is, the pure replication variance is pretty small. The wafer invariance is pretty large. There's a small remaining lot-to-lot point estimate of variance. And then we have our total variance.

So if I decompose that, it looks like about 89% is wafer-to-wafer, 3% is within wafer, very small within wafer. And here at this point estimate is 8% of the variance, that's my best guess. 8% of the variance is coming from separate lot-to-lot variance. That's point estimates.

There's something nagging me here about this ANOVA observation. That lot-to-lot variance wasn't significant. What if we looked at the confidence intervals? This now looks at the interval estimates.

And what we see here is, here's our point estimate for the replication. And it's got a range from 1 to about 3. Not too bad of a range for that. My wafer variance, I had about 56 as my point estimate. And that, based on the numbers of samples and everything that I've got, might range from 33 to 113.

And here's the interesting thing. If we do that point estimate for our lot variance, but actually look at the chi-squared or it's again, this weird, slightly something different than chi-squared but very close to it. What you get in this case is a negative estimate for the lower limit of lot variance.

Woohoo. When a confidence interval intersects 0, that tells you it might be 0. So in fact, we would set the lower bound to 0. If I still needed my best point estimate, I would still stick with the 5. But this is basically telling me, consistent with the ANOVA, that I don't have more than 95% confidence that there's a non-zero lot-to-lot variance at work.

And in fact, if I wanted to, I might go back and say, I'm going to set and assume in a new model that the lot-to-lot variance is 0. And I'm going to attribute that, lump that together with the wafer-to-wafer, and build just a two-level nested invariance where I don't include that as a separate variance source. Since it wasn't significant, you might not want to include that in your model. Is there a question?

AUDIENCE: Yes. So for this analysis, what's the degree of freedom that you're going to use in the chi-square?

DUANE BONING: Yeah. It's a little bit tricky, but it's in the spreadsheet. Basically, what you do is if you look at the denominator when you take the sampling effect, so it might be m times w, or m times w minus 1, because I have a grand mean. When I'm doing that for the lot-to-lot variance, that would be my degree of freedom. So the degree of freedom that you use, the n minus 1 in the chi-squared, changes depending on which variance you're estimating, which variance interval.

And I actually have both the definitions with the variable names in the spreadsheet, and then what the numbers are for this data. So that is tricky. But it basically, is anytime you're estimating a variance, you have a sum of squared deviations. And then you take the mean square. What's going down in the denominator in the mean square estimate, that's what you're using for the degree of freedom.

AUDIENCE: Are they still the same as ANOVA analysis?

DUANE BONING: Not quite. Where they're really based is based on these things. So you'll see that in the spreadsheet.

So the last point I wanted to make has to do with, we've already talked about this a little bit, how you allocate the measurement budget. And the simple observation is, when you're out in outer level, if I'm trying to estimate the lot-to-lot variance, what most strongly affects that? And it's basically the data in the outermost level, because the variance component of the innermost level gets averaged away fairly quickly. So the contamination of measurement variance can be reduced if you pick your sampling plan easily.

By the way, you might still actually care about sigma squared in the observed mean-to-mean averages. We said that it's not the best estimate of the true, say, wafer-to-wafer average. But if I were doing SPC, statistical process control charting, based on, I observed some sampling plan and I observe and I plot on my chart, an observed wafer average, that may be what I want to control on.

And so I actually might still want to use that, the sigma x bar, for setting of my control limits, because that's the data that I'm charting. So don't throw away our old idea of keeping track of the observed wafer average. Just recognize that it's actually got a mix of a couple of variance components inside of it.

OK, and then the last point here was simply the point that Nalish already made, is, if you're really looking for an outer level variance estimate, what you want to do is push more data to the lower levels, in order to reduce these things. So for example, if I allocated almost all of my data just to m, that reduces this factor a lot, but not this factor very much. So to get the biggest multiplicative bang for the buck, what you want to do is push it just barely outside of the factor that you're trying to estimate. So if I'm looking at lot-to-lot variance, I need at least multiple w's to get rid of the wafer-to-wafer effect.

And then I get the multiplicative effect also with the m. That already is multiplying up fairly rapidly. But if I want to suppress this factor, I need at least some number of wafer replicates. On the other hand, if I think that variance is very small, and this variance is very large, I might allocate more to the m factor. So this can influence your strategy for how you pick your sampling plans when you've got nested structures.

OK, so to summarize, we have been looking here at nested variance structures with this weird grouping within one group within another group within another group. First off, you should be able to recognize when you've got nested variance structures. Second, hopefully now you've got at least a feel for how you would estimate those separate variance components. And then there is a little bit of implications on design plans that hopefully you're alert to.

So you will have a chance to play around with this at least a little bit on the problem set, if you haven't started that already. Do look at the spreadsheet. I think that will be a big help to you on that.

So with that, we'll end. And I'll stick around for a minute, because it sounds like there's a question in the Singapore end.

AUDIENCE: Yeah, I just have a question. Should I ask now?

DUANE BONING: Yeah. But you guys should feel free to go if you want here.

AUDIENCE: How do you tell whether it's a fixed effect or if it's a nested variance?

DUANE BONING: Oh, good question. How do you tell if it's a fixed effect or nested variance? That's a model assumption.

So I think the basic idea is, if it's a fixed effect, and I think I'm changing my group-to-group by, say, a design, if wafer number 2 in the lot always has a delta of some size as opposed to being randomly sampled, that might be a systematic fixed effect. But just raw data, I don't know. You actually have to look at the setup of the situation to know whether each one is treated as a wafer replicate, or if I'm doing something different to each wafer intentionally. That would be a fixed effect.

OK? All right? So Thursday, we'll see you on Thursday with Dan Frey as a guest lecturer.