HUGH MCMANUS: Welcome to the Variability Simulation. This should look kind of familiar. This is a much simpler simulation than before. It's not going to take all day, but it's going to illustrate one point very well, which is what's the impact of variability on process performance? That's our learning objective and we're going to look at it through the poker chip simulation.

We're going to look at a computer simulation because one of the problems with variability is sometimes to understand a variable system, you need a lot of data, a lot of cases. And we're also going to explore some simple relationships that help us, including that cue-theory equation that Earll showed, to help us understand variable processes. This is a piece of a rather old data on an engineering process with high variability.

The traditional process shown here a the high cycle time and high variability was actually the process that was value stream mapped in the module we showed, I believe it was two rays ago, that had that very messy value stream map with all the arrows going everywhere. It was a very unpredictable process. It ended up having very high variability.

Now, the problem with that, given that that process was embedded in a larger process, which was to produce an airplane, it was just the drawing release process. The problem wasn't even so much the fact that it took a long time, although that wasn't great. The problem was that the variability was almost as high as the actual time, which means it could take almost no time, it could take double the amount of time. And how easy is it to do organized work in an environment where even the allegedly simple processes are so unpredictable? It's very bad.

Now, they managed to actually do a lot about that by applying Lean, but you can see how disruptive that high variability is going to be. The accounts payable module also reinforced that, that the high variability was very destructive to that process. On top of that, you're going to have quality problems with a variable process.

That's highly likely. They may be variable because of mistakes, or it may be that the variable process makes it easier to make mistakes. And then we're also going to talk, both in Six Sigma and today, about the process capability being affected by variability.

So we're going to do, I said this already, a dice game. We're going to do a computer sim and we're going to look at some simple relationships. We're also going to sort of foreshadow some stuff we're going to do in the Quality and Six Sigma.

Not on these bullets-- I'm talking fast right now because I'm going to try to carve out a little bit of time at the end to discuss some of the things you can do about variability because I think that's particularly important to our health care audience. To some extent, our engineers, but particularly our health care people. One of the big issues with applying any kind of process improvement in a very high variability environment is just getting your mind wrapped around what does an average mean, what does even a standard deviation mean in a process where the variability has weird distributions and odd special cases and outliers? We'll try to carve out a little bit of time to actually talk about some of the practical things you can do to get variability under control.

Here's a perfect system. What you are sitting in front of is a perfect, balanced, single-piece flow system, or near single. It's a very small-batch system. It has tasks arranged in order in, actually, a U-shaped shaped cell. It's not a pull cell, it's more of a push cell, but still, it's a U-shaped cell.

Every task has an inventory in front of it, a little buffer inventory to make sure that it has the things it needs to do to do its work. Every task is perfectly balanced, so there's no bottlenecks in the system. The only thing imperfect about the system is its variability.

And this half of the room, your mats actually say mailroom and PFR check and, well, you get the picture. We're looking at the accounts payable process here. You guys on the health care side actually have familiar-looking mats from yesterday because you had a process that had variability issues yesterday.

The point here, both sides of the room are identical. And the point we're actually trying to make here is an awful lot of processes look like this. Even if you get them perfect, if they have variability in them, they're going to behave in ways that are actually even worse than one might expect.

So let's find out how this system behaves. This is the game. It's a very simple game.

What we're going to ask you to do is every day-- we're going to call the shifts day-- we are going to process chips. We are going to process one six-sided die worth of chips, so anywhere from one to six chips. There's our variability.

The way we are going to process them is we are going to pick up chips from our inbox. If you are the customer or patient person, you have a big pile of chips, so you have an infinite quantity of chips. Just pick up chips from there.

Everybody else has a finite number of chips that they can pick up, the ones in their inbox. So you will be rolling a dice, picking up that number of chips from your inbox. If you don't have enough, OK, too bad. If you roll a 6 and you only have three, which is what everybody has now, then you just pick up the three.

And then everybody at the same time, to avoid confusion, passes those chips, the number of chips that they can pass, to the next person in line. And then at the end of each day, we will write down how many chips we passed. The only tricky part about that is that both the archive here and the discharge there, a certain number of chips every day will go into that mat, and the customers or the patients have to record that number.

That's kind of the overall throughput of the system. So that's a slight trickiness at the end. That's what we're going to do.

So for example, here's day one. Everybody has three chips. If this person rolls a 3, they pick up all three chips, send them along. This person rolls a 2, they can only pick up two chips. This person rolls a 5, they can only pick up three chips because that's all they have.

No waiting for this person to send the two along, so then you have five-- no, that's cheating. The chips only get passed one station per day. This person rolls a 1, et cetera. This person rolls a 6, they only had three, so they pass three.

And then this is what it looks like at the end of that day. Three have been passed along, two out, so the inventory went up. Three came out, two went in, so that went down a little bit. Three went in, only one left, so that's got more.

Let's see, one went in and three came out, so that person only has one. The overall throughput of the system was three. We're going to record how many did you do, how many did you pass along? So either your die roll or how many you had, whatever number is smaller. How many did you pass along?

And then at the end of the round, what is your work in progress? So everyone starts with three, so that number is already filled in. At the end of each day, you're going to record your number.

Yeah, so here's an example. Our person there passed one job along and some work built up, so at the end of that day, they had five, et cetera. The customer or patient worksheet is a little bit more complicated because they're sending the work out. That's basically always going to be the die roll because you have lots and lots of chips.

You're going to record the number of jobs that actually came through the system, and here, the U-shaped cell is going to help us. So the number of jobs that are added-- not in, but added-- to the finished pile at each round is recorded here. And then you're also going to keep track of the total WIP on the whole table, the total work in progress in the whole table, which starts at 12.

Now, you could just count every time, but that would be really slow, so instead, we're going to use a conservation law. Earll's an old aerodynamicist, so they love conservation laws, right? A former, a retired aerodynamicist. A former aerodynamicist.

They love conservation laws. This is the conservation of chips laws. If you know that you sent out six and three got finished, then three must have gone into the system. And they didn't fall on the floor or get eaten or anything, so this number must have gone up by three. So there's a little equation there that will help you figure out the new work in progress.

So everybody think they can do that much math? And I think that's actually written on your sheet, so you should be able to do that. OK, and there's an example. If three go in and three come out, it stays the same. If two go in and one comes out, then that extra chip must be sitting on the table, so it goes up by one. And we can keep doing that ad infinitum.

OK, what do we think should happen?

AUDIENCE: Maybe we should just ask if there are any questions.

HUGH MCMANUS: Any questions about the process?

AUDIENCE: So I [INAUDIBLE] and I roll the die, and I get like, 5. So I pass on three because I only have three chips.

HUGH MCMANUS: Because you only have the three, that's right.

AUDIENCE: So under patients completed, I'll have three there?

HUGH MCMANUS: You put three, yeah.

AUDIENCE: Then what will now be the work in progress?

HUGH MCMANUS: The work in progress is whatever the person next to you gives to you. Because everybody's going to pass them at the same time. We're going to be very formal about that. We're going to have a standard process where we're going to ask everybody to pass them all at the same time. So then yeah, you'll just have to look and see--

AUDIENCE: It's actually what you have left by the end of the round.

HUGH MCMANUS: That's right. That's correct.

AUDIENCE: If you rolled a 3 and had four chips here, you would move three, and you'd take one left plus whatever new stuff--

HUGH MCMANUS: Plus whatever payment. So it's whatever's sitting on that mat at the end of the round is your work in progress. So the average of a six-sided die is 3 and 1/2, so what do we think this system should be able to do over-- we're going to do 20 rounds. What do we think it should be able to do over 20 rounds? Is it going to process 3 and 1/2?

Are these die rolls going to even out? Maybe. Everyone's thinking.

AUDIENCE: 60?

HUGH MCMANUS: 60, so that would be 3ish. So it's not going to be perfect, but it's going to be close to evened out.

AUDIENCE: So we don't process more than 3. [INAUDIBLE]

HUGH MCMANUS: Yeah, in the first round, you're only going to get 3. More will come in as the customer person rolls their dice. The initiating person.

[INTERPOSING VOICES]

You think less? Less, OK.

[INTERPOSING VOICES]

Less than?

AUDIENCE: 35.

HUGH MCMANUS: Less than 35. OK, you're thinking a lot less, OK. Yeah.

[INTERPOSING VOICES]

OK. So yeah, people are struggling with this. Intuitively, this is hard. It's hard to figure out what this thing should do. I've already sort of biased the conversation by intimating that the system's going to behave badly.

But one's intuition, unfortunately, about random systems is that they should even out. The way the human brain works is badly biased towards just figuring random things will even out. There's all sorts of interesting gambling fallacies that are almost hardwired into people's brains.

The "law of averages"-- now, there is no law of averages. It tends that there is-- we'll see. I won't go into that any deeper than that because what we're going to do, given that we can't really figure it out a priori, and if we try to do statistics on it, it actually gets really complicated really fast, we're going to do it, all right? So let's go ahead and do our process.

OK, what I'm going to ask everybody to do the first couple of rounds is follow my cue. I'm going to dictate the process.

AUDIENCE: And we're going to try to check and make sure you're doing it.

HUGH MCMANUS: Right. And our facilitators will check and make sure you're doing it right. The first thing we do is pick up the dice. Everybody holding their dice? OK, roll it.

OK, pick up that many chips, or three if you only got three. And now everybody together pass, passing to the right, yeah. This is a properly done-- actually, more or less by accident, properly done U-shaped cell. We're passing to the right, unless you're left-handed, in which case it's not so proper.

OK, everybody done that? Now write down what you did, how many did you pass? OK, the customer person has a slightly harder job. Everybody figuring that out? Yeah, OK. So the only communication necessary should be the number of chips finished, right?

All right, day two. Everybody roll the dice. Pick up that number of dice or whatever you have, and pass it to the right. Write down your results. The amount that was finished needs to be communicated, no other talking please.

And finally, day 10. Not finally, but day 10. Roll dice, pick up chips, pass them to the right, and write everything down. Now we're going to pause and ask how this process is doing.

Who has only one chip? A couple of people have only one chip. Who has eight ships or more? Eight or more. 10 or more? Looks like we have a winner. How many?

AUDIENCE: 14.

HUGH MCMANUS: 14 chips.

AUDIENCE: Just yesterday.

HUGH MCMANUS: That's right. Yeah, 6's are good today. So yeah, that's the one difference. So what should we do about this process? We've got at least two people who are keeping their in-bin clear. They're sitting back, definitely they have extra capacity.

Yeah, we've got at least one person who just cannot keep up and probably has to go, right?

[INTERPOSING VOICES]

So does that mean there's a bottleneck there? Must be, right? No, of course not, she has the same dice as everybody else. This is a transparent-- this process is behaving badly and is actually throwing out bad clues about itself.

That looks like a bottleneck. It isn't, it's just bad luck. Some people look like they're doing really well. No, they're just lucky.

And let's see if the trends continue. Day 11, roll the dice, pick up the chips, pass them to the right, write things down.

And 20. Pick them up, hand them along, write everything down.

OK, so how did that go? Let's see, who has 14 or better? See if we broke our old record. Several people. What have we got?

AUDIENCE: I got 15.

HUGH MCMANUS: 15. What do we got here?

AUDIENCE: 13.

HUGH MCMANUS: 13, oh not quite. So 15, we have a new winner. How is our problem employee now?

AUDIENCE: A little bit better, only at eight.

HUGH MCMANUS: Eight, OK. Down to eight, so magically got better. How about our people that had one before? Anybody have one now? We still got some people with not so many.

I think we have one repeat winner, do we not? OK. So there we go.

The system is, in fact, behaving fairly badly. We even had an unintended lesson in failure of standard process, right? When we got off standard process, we also had some high variability in the performance of our process.

There's some calculations you can do they're self-explanatory. We're going to give you a few minutes to do that. Only a few. If you don't quite get them finished, that's probably OK.

The one thing we would really like to see is the customer just getting the average throughput. That first average for the customer is an important number because we want to see what our performance looks like. Oh that's a new one. OK, very good.

So you got 2.7 average throughput on m1. What's your average throughput, 2.6? OK. Looks like people are finishing up the numbers. 2.6, average 3.05. 2.3 and got another average? Still working on it, OK.

What about utilizations? What was our average utilization at our first medical table?

AUDIENCE: 0.77

HUGH MCMANUS: 0.77, OK.

AUDIENCE: 0.74.

HUGH MCMANUS: Pardon?

AUDIENCE: 0.74.

HUGH MCMANUS: 0.74, OK.

AUDIENCE: 0.87.

HUGH MCMANUS: 0.87, OK. We got a good one there.

AUDIENCE: 0.66.

HUGH MCMANUS: 0.66 2.8, and what's the utilization?

AUDIENCE: 0.8.

HUGH MCMANUS: 0.8, OK. So interesting thing here. And we can look at some of the other details. Another thing, I just asked you to look at your personal sheet there. Is there any strong trend in the inventories?

Inventories were definitely going up from 3 maybe. Some of them were going down. What's the overall trend in inventory? Anyone see a trend?

AUDIENCE: It went up and stabilized.

HUGH MCMANUS: Went up and stabilized?

AUDIENCE: Give or take.

HUGH MCMANUS: Give or take. Went up and stabilized, give or take.

AUDIENCE: Ours just kept going up.

HUGH MCMANUS: Just kept going up. Went up? OK. That was even with three people with only one?

AUDIENCE: The average overall went up.

HUGH MCMANUS: The average overall still went up, that's interesting. That's interesting. And what was the personal productivity, the utilization factor, of, say, you, the one who never had any inventory? What was your personal productivity?

AUDIENCE: 0.74.

HUGH MCMANUS: 0.74. And what about the people that had lots of inventory?

AUDIENCE: 0.68

HUGH MCMANUS: 0.68.

AUDIENCE: Utilization?

HUGH MCMANUS: Yeah.

AUDIENCE: 0.88.

HUGH MCMANUS: 0.88, OK. So this person's working very efficiently but has lots of inventory. This person is working not very efficiently, but has lots of inventory. This person doesn't have very much-- this system is behaving very oddly. It's behaving in ways that are difficult to understand. So why?

AUDIENCE: We're not always passing the exact number that we roll.

HUGH MCMANUS: We're not always passing the exact number we roll.

AUDIENCE: In a lower average process [INAUDIBLE].

HUGH MCMANUS: OK, so there's some unused capacity, right? You roll a 6, you only have one. But the inventory's going up. Is that problem going to go away?

AUDIENCE: Eventually.

HUGH MCMANUS: Eventually. Good answer. I mean, we got at the root of it. It was also sort of a system problem, right? Which is OK, it's behaving badly because of that, which makes it behave worse because your inventory is unpredictable. So even if there's lots of inventory, the chance that there would be a low inventory and you'd waste a 6 is still there. It doesn't go away even as the inventory comes up.

We don't know what's going to happen and things depend on each other. The amount that you have in inventory depends on the person next to you who's rolling dice. Even if you're doing great, if they roll a whole string of 1's, you've got nothing to work with.

So even if you're rolling all 6's or precisely on 3 and 1/2, you're still not being able to perform because of the way the system is hooked together. And this is really the key-- every time anybody rolls a 6 and doesn't have enough chips, that capacity is wasted and it never comes back. You can never get it back.

And that's where the "law of averages" fails. In mechanical engineering, we call it a ratchet effect. It can go one way, but it can't go the other. We can lose capacity, we never get it back. So that's why we're seeing these lower utilizations even though we did 20 rounds, which you'd think would be enough time for things to even out.

OK, now we're going to run a computer simulation. The first thing that any person when confronted with tangled data like this should think is we need more data. And in fact, we do. Although as we will see in this particular case, even a very large amount of data does not help us.

We're going to run a computer simulation, which does this little exercise for 20, but then also 200 and something days. Looks like this. What's shown here is days and this is the inventory level at the different stations. The notations here are of the AP case, but again, it's the same simulation for the medical case-- the four stations, the inventory at the four stations, versus time.

There's a couple of things down here. This is our processing capability. This is the WIP that's on the table, the total WIP. This is the cycle time. This is how long it takes the average chip to move through the system.

And this is just a little luck thing to just see if maybe some of this variation is just due to luck. Maybe our average die roll is better or worse, but we're rolling an awful lot of die, so we do expect the actual average of the die rolls to kind of even out even if the system itself doesn't.

OK, so first thing we can notice about the system is OK, we're running it 200 times. Is it settling down? It's not. Just waiting for it to settle down is not going to work.

And let's see, did that do anything? Oh, there we go. Select that.

Boink, I just ran it again. Came out different. If I run it again, different. Even the gross behavior of the system is not predictable. It doesn't like, get bigger and then kind of even off. If I ran it a billion times and took the average, I'd actually get sort of a logarithmic curve where it would go up and steadily flatten out, although never actually reach an equilibrium.

But the behavior even over a large number of rounds of any given instance of this system is extremely unpredictable. So perfect system except for variability. The variability is making the system behave very badly.

OK, what can we do about this? The other nice thing about a computer simulation is we can change it relatively easy. We could, for example, lower the customer input variability.

That's a favorite of bad processes. People say, it's our customer's fault. There's too much customer input variability. Like all patients show up Friday night, so it's not our fault.

Let's lower the customer input variability some. Didn't help very much. Let's lower it a lot. Didn't help very much. Why?

That appeared to help a little bit. Yeah, that's a little better. That's with a lot less customer input variability. But basically, both count. The process variability and the customer input variability count a lot.

Let's do this, though. Let's lower the customer demand just a little bit. Oh, it looks about the same as lowering the variability a lot. Let's lower the customer demand 20%. Doesn't look great, but what's changed?

It looks like there's a lot less variability in the process. And if we lower the customer variability by 30%, there's still variability, but the system can handle it. Essentially we're destressing the system, running it at a little bit less than 100% capacity, and it behaves a lot better.

And this was brought up in the AP system where they were running at 100% capacity. It's the same in this system. Basically, if we try to stress it, we only get about 70%. If we say, you know, it's only going to run at 70% anyway, let's just give it 70% of the load. It actually behaves pretty well.

This is a nonintuitive effect unless you're one of those workers. What happens if you have to work all the time and something goes wrong? Chaos, right? Crap, I have to stay up all night, but I can't, I'm too tired, so things just go to hell.

What happens if you have six hours of work in an eight-hour day and something goes wrong? You can fix it. From an individual worker's point of view, this is very easy to understand. The system effect seems kind of mysterious, but actually, if you think about what's happening on the ground, if the system is not fully stressed, it can handle variability. And you don't have to back off on the system all that much to handle quite a lot of variability.

Yeah, I understand that we're tight on time. We're coming to a close here fortunately. So that's kind of magic.

The other thing we can do is attack variability, and if we attack variability heroically-- let's reduce variability 30% across the board-- it helps some. Yeah, it's better, but it's still behaving kind of badly. If we reduce variability 70% across the board, this is truly heroic variability reduction.

That didn't work well. Oh, there we go. Notice when it was all but one, it was still no good. We have now reduced variability heroically across the board, we've got about the same level of effect as reducing the-- so apparently, according to the simulation, what we need is either heroic variability reduction or a modest decrease in the load on the system.

And this is actually a very powerful practical lesson and it comes from that same equation that Earll was showing. If we're mathematically inclined here, what happens when our utilization is very high on this term? Just put a 1 in there and see what happens mathematically. It goes to infinity, right? If the system is completely loaded and there's any variability at all, it will just continue to get worse and worse in its behavior if it's 100% loaded.

The variability was actually squared, so if we can reduce variability, that's good, but this effect is actually even more powerful. And if we plot the two effects, here's a plot of what happens if we control variability for various utilizations. If we have a heavily loaded system-- that's this one-- and we reduce variability, once we've driven it down really low-- this is the cue time or inventory, they're basically the same-- it will start to behave. But if there's noticeable variation on a heavily loaded system, it's going to behave badly.

If it's not so heavily loaded, it almost doesn't matter. This is a lightly loaded system and the variability doesn't matter, this is a moderately loaded system and reducing variability is good. And that's not really linear, but basically, sort of proportionally, if we reduce variability, we get better behavior. But in a very heavily loaded system, we really have to drive it to 0.

This is Six Sigma. If we have a system that we want to load really heavily, we want to make billions of ICs, then we have to drive the variability very close to 0 to get the system to behave. This is the converse thought. If we ease off on utilization for various variabilities, what we see is that blow-up effect, that divided by 0 as we get close to 1.

And it's almost independent of the variability level. At any level, if we back off some from 100% utilization, the behavior actually gets a lot better. Question?

AUDIENCE: I'm just wondering how organizational management deals with this fact if they know that they need to operate significantly under their capacity--

HUGH MCMANUS: That's a tough one, right?

AUDIENCE: --in order to control their system versus having a perfect system without variation.

HUGH MCMANUS: That's right. That's right. Managers like Six Sigma because if your system is perfect, you can load it at 100%. They want to hear that. That's a message they want to hear.

In systems like health care or engineering, getting the variability very, very low is very, very hard. And this is a very rich field. In the interest of time, I'm going to jump to a couple of useful tips.

One is that one of the purposes of Lean, and this was actually illustrated very nicely by the AP case study, is to free up those resources. Not in order to lay people off or necessarily save money even, but if you can take 10% or 20% of your non-value-added time out of your day, what does that get you? That gets you that headroom that allows you to deal with variability.

And it's one of the sort of root causes of why Lean seems to have a magic effect of increasing quality and throughput time when that's not even what you're aiming at. If you just take waste out of a system, suddenly everybody in the system has a little bit of extra time to deal with variation, to deal with things going wrong. So that's sort of the hint, number one.

The other thing is that in terms of variability reduction, although that's an art, there's no one answer. There are things that help a lot. Reducing something like health care variability to 0 is obviously not even possible, but one of the things about actually both health care and engineering-type processes is that the distributions aren't really normal. It tends to be like, a bunch of standard cases and then some wacko ones.

Well, what happens to the variability in terms of those equations if you just take the wacko ones off the table? Gets a lot better, right? It gets a lot better. So essentially, special case handling for the outliers.

In the simulation game, we originally designed the health care simulation to be part of a three-day seminar where you'd play that last round at the end, like after this lesson, and there's something in there called a patient advocate. Anybody even look at that card? What did it do?

AUDIENCE: It helped those patients that would take the most amount of time and helped shepherd them through the system.

HUGH MCMANUS: Right, exactly. In terms of the simulation, it took the 6's off, right? And if you looked at the distributions, they were highly non-standard. The 6's were horrible.

And so that's actually a real example of taking the really difficult cases out of the standard flow of the system, and that makes the standard flow work a lot better. Of course, you have to devote some resources, some special resources, to the difficult people. But there's ways you can kind of knock the top off of high variabilities that can help a lot on that side.

On the capacity side, the real key is to use the benefits of any kind of Lean to essentially plow them back into making people's lives easier so they have a little bit of extra time to absorb the variability.