The following content is provided under a Creative Commons license. Your support will help MIT OpenCourseWare continue to offer high quality educational resources for free. To make a donation or to view additional materials from hundreds of MIT courses, visit MIT OpenCourseWare at ocw.mit.edu.

KAREN WILLCOX: So first up, bootstrapping, which is really aimed at this question how do we get estimates of the standard errors and our estimators that don't have known distribution? So remember we were talking on Wednesday. I don't even know what day it is today. Monday. We were talking on Monday about how you can run the Monte Carlo simulation and then we know, for example, the mean estimate and the limit then goes to infinity by the central limit theorem.

Then the distribution of the estimator itself, estimate for the main is normal. And the mean of that distribution is the actual mean that we're trying to estimate. It's unbiased. And the standard deviation of that distribution, remember, was the standard deviation of the quantity we're putting in an output divided by square root of n.

But we also said that, for example, the estimate of the variance-- remember Alex told you three different ways to estimate variance-- that those estimators don't have known distributions. They're not necessarily following a normal distribution or any other known distribution.

So bootstrapping is a trick, and actually, as I was putting together today's lecture, I [AUDIO OUT] tricks. Statisticians tend to have lots of tricks up their sleeves-- just going to pull out the section on the notes because I think you should have read this-- as to how bootstrapping works. And there's a lot of words here, but let's just come down to the [INAUDIBLE].

So what do we do in bootstrapping is perform our initial Monte Carlo sample, just like we've been talking about. So take the inputs, draw in realizations from the input distributions, run them each through the model, produce the samples, the yi's, and compute the desired estimator So this theta here might be a mean estimate or think of it as a variance instrument, one we can't necessarily easily do the confidence interval analysis on.

Then what bootstrapping says is, well, now you've got these N samples. And remember each one of these samples is being run through your finite element model, run through your CAD model. Let's now re sample those values. But N samples of y, N samples of the blade middle

hot side temperatures sitting in the bin, let's just re sample from those ones.

And why is that efficient? Because we don't have to run the model anymore. We don't have to do the finite element analysis. And we're just going to re sample from those with equal probability. So probability 1 over N int, we do it with replacement, so you could pick the same sample more than once.

So they're pretty clear. We had N samples. Then we're going to sample N time from those, but we're going to get a different set of N because, I mean, maybe we'll pick each one once, but more likely we're going to pick that one a couple of times and this one not at all. So we'll have a different sampling, we could compute the estimator, and we could keep doing that N minus 1 times. In the end, we're going to end up with N values to the estimator. And from there you could try to determine competence intervals.

So it's kind of a trick because it's kind of cheating. And it's not the same thing as running a new Monte Carlo, but hopefully, you can see that this is something you can do with almost no more computational cost, and it actually turns out that if you do some analysis, this is a reasonable thing to do and it take a lot of statisticians to analyze it.

So when people talk about bootstrapping in this context, this is what they're talking about, re sampling from your samples, putting the information together in different ways, and then using that to get a sense of the distribution and say that they're [? untestable. ?] Is that clear?

So [AUDIO OUT] internet, so I didn't get time to go through it on Monday. But I mostly want us to spin today talking about different ways to reduce variance. And again, the really important idea that should be in your heads is the quality of the estimators that come out depends on how many samples in the Monte Carlo. And if you can afford millions of samples, you can probably get really good estimates for the mean and maybe OK estimates of the variance.

But if, for example, you're interested in tail probabilities, so think about an aircraft design, a new flight critical system. Maybe they has to be a 10 to the minus 9 failure probability that we have to satisfy. So if we want to estimate that by Monte Carlo simulation, how many samples do we need? A lot. How many is a lot? At least how many? At least a billion.

If a probability is 10 to the minus 9 and we had a billion samples, then on average, you would expect one sample to be that failure. So that's not even going to be enough because you might get 9, you might get 1, you might get 2. So that means you're going to estimate this

probability differently.

So billions, tens of billions, hundreds of billions of samples to estimate what are called [? variates, ?] but even just in general, even just means, if you're talking about running a CSC code or a finite element simulation that takes minutes to run, we usually can't even afford hundreds of thousands of simulations. So the variance reduction method, trying to increase the accuracy of our Monte Carlo estimates for a given number of iterations.

And I think of them, their kind of tricks, and they're really clever tricks. I'll try to show you how, at least, important sampling works. And you can think of it in two ways. Here, if you've got a good number of samples, I'm willing to run a million samples, how much accuracy can I get in my estimators? Or you could think if I want to achieve a given level of accuracy in my estimator, how many samples do I need, and can we reduce the number of samples?

So there are a lot of different ways that people do variance reduction-- important sampling, what's called antithetic sampling, control variates, stratified sampling. And I've decided I was going to talk about control variates and importance sampling, but I think we'll just talk about importance sampling today because it's already quite a lot to think about.

So in general importance sampling, is a general technique for estimating properties of one distribution while only having samples generated from another different distribution. So keep this in mind as we go through, because it's a key idea. We're going to have samples generated from one distribution, and we're going to figure out how to manipulate those samples so that they can estimate things about a different distribution for us. And it seems kind of like a wacky thing to do, but what you'll see is that if you pick this distribution carefully, we can say things about the statistics of the other distribution in a more efficient way with the Monte Carlo estimates.

So let me get the screen up. And so we're at number 3, important sampling. Not a very good piece of chalk. And so let's think about a random variable, x, that has pdf f sub x, and it has mean that we'll call mu sub x, which is the expectation of x. And I'm going to put a little x on the subscript on the expectation to just denote that we're taking an expectation under the density f sub x.

So what I mean by that? If you want to think about an expectation as an integral, and remember Alex talked about this on Monday, the expectation of x mu of x is the integral of what? x times it's pdf dx. And so again, you can see where we're going with this. Remember

we talked about having a different distribution. We want to be clear that when we're talking about the expectation of the random variable, x, we're talking about taking the expectation with respect to it or under its pdf f sub x. Is the notation clear? Yes?

So we have this random variable, x. We may be interested in estimating its mean, and we know that this is the definition of the mean. It's an integral of x weighted by the pdf of x. I didn't put limits here, but this is an integral over this 1/4 of x, which would be minus infinity to infinity if x was a normal interval. If it was uniform, it would be an integral over the ab range, what this 1/4 of x would be.

So now what we're going to do is we're going to choose a random variable that we'll call z. And we're going to ask z to be non-negative, and we're going to choose it subject to the constraints that the expectation of z taken under the pdf-- still the pdf here-- of fx is equal to 1. And why are we writing this? Well, what's the 6th notation? It's the integral of z f sub x of x dx. That's the expectation of the random variable, v, under the pdf fx.

So if this is equal to 1, maybe you can see that all we've done is really define another pdf that we can write as f sub z. Yep. Because it's going to satisfy the properties of a pdf. It's going to integrate to 1. Is that OK?

AUDIENCE: [INAUDIBLE].

KAREN WILLCOX: z is going to be just another random variable. Yeah, so you can maybe just think of z as being defined on the same support if you want to, or I think this thing is allowed to be 0 in some places, but it's not allowed to be 0 everywhere. There are some conditions related to where z can be 0. Is that OK for the lecturer to ask you a question?

AUDIENCE: [INAUDIBLE].

KAREN WILLCOX: Yeah. So I think the easiest thing is to think about x as having infinite support, then integrals away from minus infinity [INAUDIBLE].

So we haven't done anything yet. We've just manipulated things. So why are we doing this? So if we did this, we'd have to think about the mean of x, mu x, the thing here, which is, again, the expectation of x with respect to under its own pdf. And we said that this thing was integral of f times its pdf, weighted by pdf.

And so now what can we do? We can divide by z and multiply by z. Just multiply by 1. And now

we recognize that z f of x is just what we're now calling f of z, where this is x over z f of z dx. So all we did was multiply and divide by z, and then say that we're going to group the z times the pdf's of x together and [INAUDIBLE] f of z. And we can do that because [? opposite ?] constraint is that same as integrating to 1.

So what is this guy here? This guy here is the expectation of something else with respect to the pdf of z, the density of z. What is it the expectation of? x over z? So this is the expectation of x over z, but now taken with respect to the pdf fd.

All right. So does this kind of feel like we're just moving chairs around? Yeah. That's why I said they sort of seem a little bit like tricks. Is anybody uncomfortable with anything that we've done? It's OK? We've put a 1 over d in here, and we've changed the pdf, the weighting and the integral to accommodate to make it equal to same. Yeah, Kevin?

AUDIENCE: [INAUDIBLE].

KAREN WILLCOX: The constraint we have here is that this is the expectation of [? d1 ?] of f of x is 1. Well, this is going to be what leads us to find the pdf's. Yeah.

AUDIENCE: [INAUDIBLE].

KAREN WILLCOX: So why do we put [INAUDIBLE] underneath and then bring it about? So let's look at what we have. We have now two different ways to compute the mean of x. That's what we're after. We have [AUDIO OUT] what we normally think about, which is thinking about this mean as the expectation of x under the pdf. And now we have derived this alternate expression, which is the expectation of x over z where the expectation is taken now with the pdf f of z, the z, and the weighting.

So this is what we've been talking about. How do we estimate the mean this way? Well, we sample x by drawing samples from the distribution f sub x. We define the pdf of x input. We draw samples from it. We estimate the mean to be the sample mean. And the variance of the corresponding estimator to the estimator variance is what?

What was on the numerator? Not mu x.

AUDIENCE: [INAUDIBLE].

KAREN WILLCOX: Yeah, it's [INAUDIBLE] squared. So we'll call it the variance of x. And again, I'm going to put

the subjects here just to denote. So that's what we saw before. Actually, we sampled inputs and we propagated them through to the outputs, but basically, we're just sampling from this density, and given the sample mean and our estimator and the estimator of variance with the [INAUDIBLE].

So what we've done here is defined a parallel pattern analogous [? path. ?] So what we do here? We're going to sample x over z, and we're going to sample it under now the distribution or the density xz. And again, we take the samples, we take the sample mean. That would be our estimate. And the variance of that estimator, the estimator variance, is now going to be what? Variance of x over z under this divided by [? N. ?]

And now maybe this is where you see why these tricks work because putting the d in the denominator and bringing it out here didn't change what we're estimating. These two expressions are equal. I like to think about it is that because of the way that variance behaves non-linearly, pulling the z under and putting it out here in front does not mean that these two things are going to be equal. And maybe you see there. You could drive me crazy when I was learning statistics is that things like means would behave in kind of this linear way that made thins, but the variance never seemed to be doing the sensible thing. So basically we're going to exploit that.

So again, we can do this manipulation with this other variable, z, and not change the estimate, still be getting after the thing we want. But if we choose z in the right way, and we'll talk about how to do that, we could make this smaller, which means, again, for a given number of samples, if this is smaller than this, then our estimates are going to be tighter. Or to achieve a given confidence in the estimator, we could take less samples here than doing it this way. Yeah.

So We'll talk about sort of the general ideas, and then I'll show you exactly how you might come up with a z for probabilities. But the general ideas are first of all, that some values of x in the Monte Carlo simulation are going to be more important for estimating the parameter. The mean is not a very good example, but you can imagine if we're talking about tail probabilities, we want to get more samples around that failure probability.

What goes on over here we don't really necessarily care about because it's like doesn't fail, it doesn't fail, it doesn't fail. So the idea is that some values of x are going to be more important, and what we want to do when we choose the z here is to somehow emphasize those more

important values. And so really what it's going to come down to is how do we choose z again so that this estimated variance is going to be less than this guy. And we will talk about that in flow probabilities.

So any questions? I didn't ask you at the beginning of class, did Professor [? Nguyen ?] talk about importance sampling in 1609? No? Just as well.

All right. So let's talk about-- this is 3.2, importance sampling for probability estimation. And hopefully, maybe it will make this a little bit more concrete for you. So what did we already see? We saw that the probability of A, where A is some event that we're interested in is estimated by a Monte Carlo, or can be estimated by a Monte Carlo, with an estimator that we'll call P hat A.

So let's think about the case where, again, we're interested in the probability of failures, so A is the event where, say y is greater than y current. So y is some output of interest. It's the temperature in [? blades, ?] and we're interested in looking at when that temperature exceeds some critical value. So this shows up often, and if we are looking at a probability of failure.

So we're going to define an indicator function. Usually when you're working with probabilities and if you're told to find an indicator function, you guys should have seen those in 1609, yes? In 6041. So the indicator function, i, is the function of Ai. So this is going to be corresponding to the i-th sample in our Monte Carlo. So we're drawing a sample, where running it through our [? planet ?] element code, and we're looking to see whether event A happens.

And indicator function says if event A happens, then takes a value of 1, if not, take a value of 0. So this is going to be if the corresponding sample of y, the yi exceeds y current, and if not, then we can set indicator function to be 0. So indicator function is set to 0. 1 result. Failed. Didn't fail. A happened. A didn't happen.

AUDIENCE: [INAUDIBLE].

KAREN WILLCOX: i to the range. i is either 0 or 1.

AUDIENCE: [INAUDIBLE].

KAREN WILLCOX: Oh, little i. Sorry. Yeah, that's right. Little i from 1 up to n. Sorry. So if we just think about these things in this way, then the reason we do this is because then our estimator P hat of A-- what was he estimator for the probability? What was it? Yep. So the number of samples where A

occurred just divided by the [INAUDIBLE] number.

So what does that look like? What does that look like in terms of the indicator function? I'll put the denominator in there. What's the numerator? Yeah. Just the sum of the iA's. So it's 1 over N. The sum equals 1 to capital N of iA's. In other words, it's the sample mean of the indicator function. And so we can think of the estimator this way, and we've already seen this. We saw that the expected value of this estimator for the probability was equal to what? Probability of A, unbiased. This would be a big piece.

And do you remember the expressions to the variance of this? Does anyone remember it? Here, it's P of A [? deserves ?] 1 minus P of A. And what's always in the denominator? N. Right. Standard error of the estimator is 1/3 of that thing.

All right. So that's what we saw before. So now what we're going to do is think about how importance sampling can help us come up with an estimator, for P of A that's got smaller variance. Do you remember when we did that analysis?

We also did an analysis where if P is really small and you want to have a tolerance that's relative to the size of P, you end up needing billions, millions of samples. So we're going to see whether importance sampling or see how importance sampling can help us. Again, I'm going to erase this, but I find it helpful to think about these two paths. We could sample, take lots of samples and we're talking about the mean here, use our regular estimator. Or we could think about introducing another random variable with another pdf and changing what we sample from and then weighting the integral to account for the fact that we sampled from a different pdf and coming out with an estimate of the same thing but changing the variance.

So let's see how that would work out. So we'll go straight to a pdf. So we'll introduce another pdf, and it's going to be called w. And it's a function of x, so it's defined on the same support as x, whatever x is. What are we doing? Yeah, we have [INAUDIBLE].

AUDIENCE: [INAUDIBLE]

KAREN WILLCOX: Ah, fourth.

AUDIENCE: [? Don ?] Ulrich just wanted to say hello.

KAREN WILLCOX: Hi, Don.

AUDIENCE: How are you doing? You've got chalk all over your face.

KAREN WILLCOX: I have. You guys know who this is, right? [? Don ?] Ulrich, Satellite astronaut.

AUDIENCE: I'm one of the lucky guys who took a fall in space.

KAREN WILLCOX: They're excited in learning all about computational methods. And today we're talking about estimating low failure probabilities, so things like 10 to the minus 9, things that can go wrong. Do you know anything about that?

AUDIENCE: That's 10 to the minus 9. Well, you don't need to worry about that.

KAREN WILLCOX: I'll be at the dinner tonight. So I'll pick up a beer. Thanks. Where is the talk?

AUDIENCE: Marlar Lounge. Second floor. 4:00.

KAREN WILLCOX: Or you could come to office hours.

AUDIENCE: And make sure you learn this stuff because it's important.

KAREN WILLCOX: I don't if you guys know, but I made it to the finals at the astronaut selection last year, but didn't get picked so, Don was, yeah.

All right. So we just introduced a bit-- I'd much rather teach that stuff. It's more exciting than going into space, right?

AUDIENCE: [INAUDIBLE].

KAREN WILLCOX: So we've introduced to pdf w of x, and I want you to think is this as being like an alternative pdf for x. And x is just a generic at this point. This thing is called the biasing entity. And you can see what we're going to do is we're going to choose w of x so that this event, A, occurs more frequently.

So the idea is we don't want to have to wait around for a billion samples for the one event to occur. We're going to choose this alternate pdf w, in such a way that this event, A, occurs more frequently. So then what is our estimate to the probability look like? Probability of A is maybe you can see it directly from this expression. We could write it as being the expectation of the indicator function. And it's the expectation taken under f of x.

So it's the expectation of the indicator function with expectation as with respect to f of x. Maybe

up here I should define that then f of x is but the pdf f of x. And what is this? This is nothing but the integral of the indicator function weighted by the pdf and then [? graded ?] up for the support of x.

So we're going to do the same trick that we did before. We're going to multiply and divide by this w. So we've got our f of x. We're going divide by w. We're going to multiply by w. And maybe now you can see that what we have now, we have an expectation of [? sine. ?] We can think about lumping this thing together now taken with respect to the pdf w.

So this is just the integral of something times the pdf integrated over the support. So we have to write it up here. So the last line in that development-- the left-hand side is P of A, which we haven't discretized with Monte Carlo yet. We're still doing things exactly-- is expectation with respect or under this pdf w of-- now, the indicator function with this waiting, f of x over w of x.

So what does this mean? This means we could draw samples from w instead of from x. So draw samples from the pdf w. But if you do that, in order to get the right estimate for the probability, you have to weight the samples by the f over w to make up for the fact that you drew from a different distribution. Does that make sense?

So weight by f of x over w of x, I'd say to counter the sort of my informal way of thinking about it. And if you do that, then the expectation works out so that you are still getting the probability you're after.

So then the last step is just to write down what our Monte Carlo estimator would be-- so then our Monte Carlo importance sampling estimator for this probability of A is what? So let's call it P hat. We'll put an IS down here to indicate that's the importance sampling estimator. So help me write it. What does it look like?

Take this risk part. That's the easy part. One over N-- is it Libby or Casey? Oh, it's [INAUDIBLE]. All right. Now, one of you guys have to do the harder part. 1 over N what? There is a regular Monte Carlo estimator. Summation from I equals 1 to N of what? iA times f of xi over w of xi.

And it doesn't show up in the formula, but it's important to know that this is all when the xi's are drawn from the pdf w. That's that. We never noted that before because we only ever had one pdf floating around, but now that we've got two, so here's the estimator. If we drew the samples from f of x, which is the regular way of doing it, that would be the estimator, but if we

draw the samples from the pdf of w, then this is our estimator, and basically we just have to re weigh it. Is that clear? So can you see-- yeah.

AUDIENCE: [INAUDIBLE].

KAREN WILLCOX: Here? Yeah. We divided by w, and we multiplied it by w.

AUDIENCE: [INAUDIBLE].

KAREN WILLCOX: Well, yeah. w is a pdf. Maybe I could have written it like if w might have been-- yeah. Sorry. w is a pdf. Yeah.

AUDIENCE: [INAUDIBLE].

KAREN WILLCOX: This one is strange, actually. We approached it here by starting off by saying w is a pdf. So w itself is a pdf. When I introduce a general importance sampling, we started off with a random variable and said that's why we put that constraint on so that we could get to a pdf. But here we're just say let's introduce a secondary pdf or an accelerated pdf.

So w is a pdf, which means it has to satisfy the property of a pdf. Greg, did you have some--

AUDIENCE: [INAUDIBLE].

KAREN WILLCOX: This guy?

AUDIENCE: [INAUDIBLE].

KAREN WILLCOX: That's right.

AUDIENCE: [INAUDIBLE].

KAREN WILLCOX: Yeah. So this would be given to you. So one thing that's just a little bit confusing and maybe I haven't been entirely, so I'm going to put it over here, is that when we think about Monte Carlo, I'm going to use this completely different variable so that I don't mix up with w's and f's. The model has say, an input that might be d and an output that might be q. And we talk about drawing from some distribution of d. Say d is normal propagating it through and getting the corresponding whatever q looks like.

So here when I talk about estimating the probability of A by drawing from x-- x is really q. Because this is the pdf that defines if A is out in here, we want to draw from this distribution. I

mean, if I gave you just this distribution, q, and I asked you to estimate a probability, you would randomly sample from it and then you would count the number on the tail.

So there is kind of an extra step, which is that we're really drawing from the d and pushing it through the model to characterize q. Is that clear enough?

AUDIENCE: The f point, is that all the samples, or is that just the [? sales ?] proportion?

KAREN WILLCOX: For 10 points? Oh, the N? That's going to be all the samples. Yeah.

AUDIENCE: [INAUDIBLE].

KAREN WILLCOX: Yeah. So let me try to write a summary and draw some pictures to help you, because I don't want to mix up two things. But what we would we do here is we would sample randomly, and more samples are going to end up in here than in here. And every time we sample here, we generate a corresponding sample of this.

So when you think about the importance sampling, just don't really think about this part of it. Think about it as we're sampling from this. But it turns out we're sampling from this by something from this and going forward. But we're still sampling from this.

So when we do that and we put samples, most of them are going to end up over here, and then one in a billion times, we get one out here on average. Yeah. So what we're saying is we could define a different distribution. We're jumping here a little bit, but let's just conceptually say that distribution looks like this. And instead of sampling from this guy-- which by the way we do by sampling here and pushing forward, but that's not really relevant-- this sample from this guy so that now when I sample from here, instead of 1 in a billion, maybe 1 in 10 or maybe half of them actually fall in the regions I mentioned for them.

Yep. So now I'm going to have N samples from this guy. But now when I compute the estimate of the probability of A, I can't just take the fraction that are here divided by the total numbers. Clearly that would be the wrong estimate. But if I take each sample that [? folds ?] in here, the 1's, and multiply them by the ratio of this guy to this guy, that's going to recalibrate that to be the right estimate.

And you can kind of see it graphically, because what happens is you get a sample here. The f of x is basically 0, fairly small. This guy is big. That ratio is going to be really small. So even though it's the 1, it gets weighted by a 10 to the minus 9 or whatever it is, or 10 to the minus

So you can see how it works, and it all works out. Draw from this guy, get more sample than here, but then each sample doesn't get a 1. It gets a 1 times the ratio of this to this. And by doing that, we'll be weighting, [INAUDIBLE] actually estimating the right probability. But then the next question, that's kind of the next thing we're going to talk about is, how do you figure out what a good biasing distribution is? How do you come up with the distribution?

And it would be easier if we didn't have this part of the problem, but because we are generating the samples from the inputs and pushing them through the model, we don't necessarily know how to bias the distribution here so that we get a good biasing distribution here. And that's like saying do you know what combination of inputs makes your aircraft much vulnerable to failure. And sometimes you do, and sometimes you don't.

So again, that's kind of the separate part. Part of it is a little bit different to describe. Yeah.

AUDIENCE: So you could take like a Gaussian variance with a [? unit ?] and plot that over there?

KAREN WILLCOX: Yeah. You could.

AUDIENCE: Would that be like [? an ?] example?

KAREN WILLCOX: Yeah. So really simple things that people do is that they, and we'll look, they scale this thing to shift more mass. Well, what I drew here is a little bit extreme. They scale it to shift more mass, and they basically make this thing have better tails. And sometimes they actually just shift it. We'll take whatever the distribution is and just shift it and scale it. And then there are more sophisticated things you can do, but, in fact, this is somewhat of an open question is how do you, [INAUDIBLE] define a good biasing distribution to make your sampling really efficient?

AUDIENCE: I would think that it would [INAUDIBLE].

KAREN WILLCOX: That's right. And so usually in the cases where it's easy to come up with a biasing distribution, you kind of know what conditions including failure anywhere. And so it's sort of obvious way to look. The really difficult problems are when there's tens, hundreds, thousands of them that are through input here. We don't even know which combinations of input to the ones that make you most vulnerable to failure. And the only way to find out would be to sample them all, which we already said we don't want to do.

So aircraft design is a good one, also people who simulate earthquakes and all these kinds of things, I mean, where events of different weather stuff. It's not even clear sometimes that you can simulate and get good estimates for some of these events.

AUDIENCE: [INAUDIBLE].

KAREN WILLCOX: Yeah. In some places, it's really hard for us to characterize what q is out here. The financial markets are another one. I don't know if you guys realize *The Black Swan* book, so it's all about [? fact ?] tails and suit and models and don't account for things that are like way out here, that are really unlikely, but when they happen, they have a really big impact.

So if we knew what was going on here, we would already know the probabilities. And so the trick is that we don't know what's going on here, but we want to learn about what's going on here by finding better ways to sample. But to find better ways to sample, we need to know what we're looking for. So then there are things, I mean, this still always happens, and this is what keeps us busy in research is that's where you're using activity. You learn a little bit and sample.

So let's just quickly summarize the steps for importance sampling. So we're going to define a w of x and however we come up, we're going to draw samples of x from that pdf. So I'll just note here that this is a pdf. If you prefer to call it f sub w, that's fine. And again, the idea is we want to get more samples in the region of interest.

And then we're going to estimate this probability, but we're going to do it using this thing here that we define, the P hat, IS, where we have to do the weighting to account for the fact that we drew from the wrong distribution. So importance weights if x divided by the w, that's the sample point.

And one thing I should say actually is if you don't choose a good w, you can actually make it worse. You could make your Monte Carlo convergence worse. So actually, I think there's one final result to write down is what are the properties of this P hat IS? They are that the expected value of our importance sampling estimator for the probability of A. What do you think it is?

Hopefully, it's P of A. It is, indeed, P of A. And that's actually, I mean, we didn't do anything over here except move things around. We had equality all the way through. So that's key. We dumbed this down with the expected value. In other words, it's unbiased. But then the key is that-- you could put a w in here to be clear-- that the variance of the estimator, which again

controls basically how good it is for a given number of samples, has got this more complicated expression.

So the 1 over N is still out there. And then there is an expectation of i of A, f of x over w of x, and then there is a minus t of A squared. I think I got all the pieces. And that actually goes back to what I was just saying. It's not actually obvious whether this is smaller or bigger than what we had before with the regular Monte Carlo estimator, but of course, it all comes down to the choice of w. I mean, the idea of choose a w to make this small so that you get away with fewer sampling.

That's kind of neat trick, no? Multiplying and dividing by things, no?

AUDIENCE: [INAUDIBLE].

KAREN WILLCOX: It's what?

AUDIENCE: [INAUDIBLE].

KAREN WILLCOX: Yeah. It used to always drive me crazy when I was an undergrad.

AUDIENCE: [INAUDIBLE]

KAREN WILLCOX: So next just briefly how to pick, and I'm not going to give, sorry, [? Trey, ?] I'm not going to give you any definitive answers on this, but how to pick the biasing distributions. So what do you do?

So we'll just look at two yeah?

AUDIENCE: [INAUDIBLE].

KAREN WILLCOX: Oh, here? Yeah. Just like all of the variances of the estimators have like sigma y or P of A or whatever it is.

AUDIENCE: [INAUDIBLE].

KAREN WILLCOX: Yeah. Remember like the original. Even the variance to the mean estimate is sigma y, sigma squared over N, and you don't know sigma. So you can plug in the estimates, but then you're introducing additional error. Yeah. To actually compute them, you have to use an estimate for P of A. Yeah.

AUDIENCE: [INAUDIBLE].

KAREN WILLCOX: Yep.

AUDIENCE: [INAUDIBLE].

KAREN WILLCOX: Yeah. So this is the theoretical expression for the thing. How you actually compute this is a whole other question, but we've seen that with every single one of the variances of our estimator. We had sigma squared over N. We didn't know what this was. On the original one, we had t1 minus t over N. We didn't know what this was. So all of them have got the theoretical result and how you compute them.

And you could compute the estimates. You can do bootstrapping, and in this case you would be able to estimate everything. Yes.

So one simple approach is just to scale. So we're talking about picking w. And thank you. These are homeworks 7. Thank you. And it's scaled in such a way that we shift probability into the region of interest, shift probability into the event regions. Again, the idea being that we want to get more samples.

So how would that work? In that case, w of x could be some 1 over a times the original pdf. But now the function of x over a, where a is some constant that's greater than 1. So I have to draw these for them to make sense to me. So it's saying pick w to be the scaled version of f of x where we scale the argument and then also scale the result.

So if I can just sort of sketch emotionally what that might look like, I know it helps me to think about what it's doing. So let's say x and this is the f of x, the original pdf. So maybe it's Gaussian. And just to make it easier to think about, let's just set it at 0. It doesn't have to be. And it's high, which is going to be something. We'll call it q.

Then what is w? Does anyone want to have a go at [? throwing ?] w? So the height is q over a. OK, good. Am I going to choose a greater than 1? So it's going to be lower.

AUDIENCE: [INAUDIBLE].

KAREN WILLCOX: Good. You're much better at visualizing this than I am. I had to mentally do a few points because I'm not very good. It's not a very good fit Gaussian.

All right. So I had to actually sort of mentally thinking it through. I'm like at this point, we're at 1 and say a were 2, then this guy at 1 is going to get the mass from 1/2. Then it's going to scale it down by two. But because this thing is falling off quickly, it's still going to end up being set or relative to what was in the middle.

So it's exactly a scaling that's going to shift probability mass out. So this is what I was saying when I meant [? heads ?] or tails. Any problems you see with this if we're trying to estimate a probability of failure, say the probability that x is greater than some critical value? What's that?

AUDIENCE: [INAUDIBLE].

KAREN WILLCOX: Yeah, it might not still be. Anyway it's definitely going to generate more samples over here, but it's also going to generate more samples over here. So we've made this tail fatter, but we've also made this tail fatter. So

AUDIENCE: [INAUDIBLE]?

KAREN WILLCOX: Can you do one if f of x is uniform? I think about what w would be. But basically w can be anything you want.

AUDIENCE: [INAUDIBLE].

KAREN WILLCOX: Yeah, that's just one way to do it. But you've got to make sure that the support of w is it at least as wide [INAUDIBLE] actually going to be dividing by 0. So if this guy was uniform, I mean--

AUDIENCE: [INAUDIBLE].

KAREN WILLCOX: That's right. But you couldn't have this be uniform and then say, I'm going to take my w to be this guy. Well, what's generating from here would you-- you would be OK because you never generate points here. Yeah.

AUDIENCE: [INAUDIBLE].

KAREN WILLCOX: You could but then they would just get multiplied by 0. Then you'd be wasting [? holds. ?] Yes, I guess there's no reason why you couldn't do this. Yeah. So that's one simple thing to do is just to move mass out, maybe [? theta ?] tails.

Another simple thing you could do is translation, which just means that w of x would be f of x

minus c, where maybe c is greater than 0 if we're thinking about a right tail. And again, maybe this wouldn't make sense if you had a finite support like a uniform distribution because then you would be putting w, which you didn't care about. But what would that look like notionally? If that's x and that's f of x, we get a first term normal [INAUDIBLE] 0, then the w of x is just going to be shifted over by an amount, c. So again, all it's doing is putting more probability mass in this case in the right tail that we care about.

But thinking of biasing distribution, the simple problems, simple things probably work for really difficult problems with lots of input in very rare events, differently, still an open question. Any other questions about importance sampling. We sometimes talk about distributing analysis? Does it all makes sense that's on this screen? Yes.

AUDIENCE: [INAUDIBLE].

KAREN WILLCOX: Actually control variance are my favorites. I've mentioned some of the other variance reduction methods. Control variance is another way to do variance reduction that I think is also a kind of neat. So if you're interested in knowing other variance reduction methods, I could give you some stuff to read.

So let's now just in the last 15 minutes talk a little bit about sensitivity analysis in the context of Monte Carlos. So the question is now how do we use our Monte Carlo results to understand which uncertain inputs are contributing the most to output variability. So I want you to put this picture back in your mind, this guy here, which is that we've got uncertain inputs, and maybe there's a lot of them-- d1, d2, d3, however many.

And maybe there's one output of interest, or maybe there is lots. But we've run these Monte Carlos, we put pdf's in all these guys. We draw samples. We run through. We generate some kind of output histogram on q that looks like however it looks.

But now we want to ask the question, well, which one of these is really contributing to the variability, or which combinations of these are contributing to the variability? Which combinations are causing this big fat tail out here, which corresponds to [INAUDIBLE] that don't meet our design criteria?

So this is really now kind of the opposite question. You've done a forward propagation of uncertainty, and you've got the uncertainty in q, and now you want to figure out where did it come from. So forward propagation of uncertainty, sensitivity analysis is kind of like the

reverse question. And it's important for two reasons, one is that it helps us understand where we should focus our uncertainty reduction efforts.

But I think in many ways when you look say like an aircraft design and development program at a place like Boeing or wherever, so much of the process is about reducing uncertainty. It's about running experiments or tests or running models and trying to figure out exactly what the performance is of the thing that you're designing and making decisions as well.

So understanding where this big uncertainty can help you decide what kind of experiments to do, whether it will improve your models, if you're thinking about uncertainty in finished products, it can help guide you one where you should tighten tolerances in the new manufacturing process. It might tell you that you've got uncertainty because you don't know certain conditions. So maybe it says you need a sensor in your engine or on board your aircraft or wherever.

Sometimes this is called factor prioritization, figuring out what the priority order of inputs. Effect is just a term that is often used in the statistics community. Which one should be priorities for uncertainty reductions.

So it's kind of half a story, but the other half of story is that it's also really important to understand where the uncertainties are not important. That they could be distributions on d1, but this doesn't really matter because this thing is not sensitive to d1. And that's important because then you shouldn't waste your time worrying about d1, and you shouldn't waste your time arguing with other groups about whether d1 should be 1 or 1.1 or 1.2.

So it's important to understand where things are actually not important. And this is sometimes called factor fixing because this is where you can understand where it's not important to consider uncertainty, and you can just pick a deterministic value of an input and go forward. And this is really important, actually in the policy setting.

As I mentioned before, you can work with FAA. People argue about all kinds of things when they make policies. It's nice to be able to show that some of the things that you argue about actually don't matter for the uncertainty in the policy decision. Less things to argue about.

All right. So how can we do sensitivity analysis? There is something called VABO, vary-all-but-one [INAUDIBLE]. Think about doing. So here are the steps. First of all run your Monte Caro with all the inputs varying. Think about d1, d2, d3, up to however many we have of them.

We've got pdf's for all of them. We sample all of them. We generate the corresponding-- it's called q on that order over there. Then let's go and take input case, so it's like the first one, d1, and let's fix it to deterministic value so it's no longer got a pdf, it's no longer able to vary, and rerun the Monte Carlo with all the other ones-- all the other, however many there are d minus 1 inputs varying.

So now we have the results of two Monte Carlo simulations, one with a fixed, one was varying, and one where it wasn't. And you could compare the statistic to the output. You could look at the variance from run one, and the variance of run two. And you could then attribute the difference in the variance to that fixed factor. And then he would repeat it with 1, 3, 4, 5, 6. It varies at 6 2, 6 3, 6 4. So vary all but one.

I mean, maybe it's a useful thing to do, but of course, there's question. The first question you'd say, well, what value should we fix factor k? If we fix it to this value and ran the Monte Carlo, would we get a different result than if we fixed it a different value? And the answer for a complicated system is probably yes.

And then the other question maybe you should be wondering is what about possible interactions. If I fix input 1 at a value and analyze things, but what if I fixed input 1 and input 2 and there were interactions between them, and I never explored them. Are there interactions that might be important that I would be missing?

So to address those limitations is something that's called global sensitivity analysis. So let's get this off. And I think when you think about global sensitivity analysis, it's useful to have this vary all but month's color in your mind because it's conceptually a little bit the same. So we're on 4. And 4a was the vary-all-but-one.

So this is 4b, global sensitivity analysis. So I'm going to draw a picture here. I'm going to use different notation, but let me not try to change otherwise, I will end up mixing something up. So we've got a model. So then I'll call the inputs x1, x2, down to-- we're going to have [INAUDIBLE], so d random inputs.

Now, let's just think about a single random output, so then just one random output. So global sensitivity analysis defines [AUDIO OUT] called sensitivity indices. And there are two different kinds of sensitivity indices. So the first one is what's called a main effect. Did I talk about main effects in 1609, 6041? No. Maybe in 62x, anybody seen effects?

So if you've seen effects sensitivity index for input i, usually the set of [? sessions ?] of u say people would use it with factor here, but let's just call it an input. So this thing is [? the limitation ?] si. Let's write the formula and talk about what it means. It's the variance of Y minus the expected value of the variance of Y given xi all divided by the variance, Y.

So the variance of Y minus the expected value of the variance of Y given xi divided by the variance of Y. Does anyone want to have a go at explaining what this is?

AUDIENCE: [INAUDIBLE].

KAREN WILLCOX: Yes. So if this happened to be 0, then overall sensitivity index would be what? 1. And it would mean that all of the variance in Y was explained by xi, because when we fix xi the variance [INAUDIBLE]. What about if the variance of this thing were the variance of Y? Doesn't have any effect. Yes.

Why is there an expectation here? So we go with [AUDIO OUT] So we're saying Y give xi.

AUDIENCE: [INAUDIBLE].

KAREN WILLCOX: Yeah. If we picked any particular value which to effect xi, as I said given xie to some x star, then it would just be a number. But because this thing is a random variable, we don't know where to fix it. So this is addressing this question of we don't know where to fix it. So at least the way I think of it conceptually is that we're going to fix it at all possible places and then [INAUDIBLE] over there.

And so it's the expected reduction in the variance is relative. It's normalized. So it's the expected relative reduction of the variance if we learned everything about the factor xi. So it's giving us a measure of how much to does xi contribute to the variance in Y, but it's a measure where we're taking expectation over the possible values that xi could take on. So that's where the word global come from. [INAUDIBLE].

So let me just write it in words. So expected relative reduction in output variance-- and then output variance is variance of Y-- it's a true value of xi is [? learned. ?] And then another thing that's important to see is that this is a measure of the effects of varying xi alone. And if we wrote this in a slightly different way, we could write it, we can see that would be averaged over variations in other inputs.

We'll see you Monday when we talk about [? planet ?] experiments. It's another way of

computing of effects that [? finds ?] out more with the second interpretation. Let me say it again. Measure, this is what we talked about. Measure of the effect of varying xi alone averaged over the variations of other inputs.

And that's where the term main effect comes in. So if we were to change xi, if we were to vary xi, how does it change the output averaged over everything else in the [? x3, ?] x4 varying? Is that clear? So there are two different ways you can think about that.

AUDIENCE: [INAUDIBLE].

KAREN WILLCOX: So xi can at most be 1 because the most we can get here is 0 if i is going to be between 0 and 1. And the idea is that you could compute these then you would rank them. And if you were looking for where you focus your efforts, which variable you try to control in the manufacturing process, before we go after the one with the biggest sensitivity index that you occupy.

So actually turns out we can also compute higher order interaction indices. You can also compute effects for the interactions. And I won't go into details, but instead of now just being si, there could be an sij, which would be two variables, two inputs interacting or three variables interacting or four, all the way up to all of them interacting.

And what the match shows, I mean, we're not going to talk about it, is this idea that you can take the variance of Y, think of it as this pi, and you can decompose this variance of Y. So let's think about-- it's easier for me to draw fY depends on x1 and x2, so they're just two inputs. There would be a part of the variance that might be due to x1 acting alone, there might be a part of the variance that's due to x2 acting alone, and there might be a part of the variance of the interaction between x1 and x2.

And if we computed the sensitivity indices, f1, f2, and f12, they would all come up to 1. Or the variance due to x1, the variance due to x2, and the variance x12 would come out to be the variance of Y. And you can show there is something called a high dimensional model representation that shows how all of this comes out. So it's kind of a nice result that you can attribute.

It actually turns out that there's something else also called a total effect sensitivity index that you can also compute that tells you how much does x1 contribute, not just by itself but with all of the interactions as well. So the total effect sensitivity index, what include x1 and x12, x1. And the total effect x2 would include x2 and x12. And then all the other ones, if you had more

variables then I would add them all up, and it turns out to be a fairly easy way to compute that as well.

Last questions? Great.

AUDIENCE: [INAUDIBLE]?

KAREN WILLCOX: Yeah, that's a good question. So I actually haven't told you how to compute these things that all. These are the expressions for them. It turns out there are a couple different ways. I think the most common way to compute these things is a method called the [? Sogl ?] method. [? Sogl, ?] I think, was a Russian statistician, which involves Monte Carlo simulations. And you basically end up doing one Monte Carlo simulation and then you do a second Monte Carlo simulation where you free some of the variables and redraw other ones. And it's done in kind of a clever way so that you get to this.

So then the question of how [? converge ?] of the sensitivity indices comes to be a question of how converged are the variance estimates. And we've talked a lot about mean estimate and how they converge. It actually turns out that to get variance to converge, you usually have to take more Monte Carlo simulations. Most of the time when we use these things, we do as many samples as we can afford, and that ends up being the [INAUDIBLE].

So that's a good question. And it also depends on what you want. Do you actually care whether you get the sensitivity index to four decimal places, or do you just care about saying number two is the biggest?